INFuture2009

The Future of
Information Sciences

# Digital Resources and Knowledge Sharing

Department of Information Sciences
Faculty of Humanities and Social Sciences
University of Zagreb

# THE FUTURE OF INFORMATION SCIENCES

2nd International Conference "The Future of Information Sciences: INFuture2009 – Digital Resources and Knowledge Sharing"
Zagreb, 4-6 November 2009

*Organizers*
Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
IBM Croatia



*Technical editor*
Hrvoje Stančić

*Cover design by*
Hrvoje Stančić

*Publisher*
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb

*Printed by*
Vjesnik

*Impression*
200 copies

All published papers were reviewed by international board of reviewers.

A CIP catalogue record for this book is available from the National and University Library in Zagreb under 719803.

ISBN 978-953-175-355-5

# THE FUTURE OF INFORMATION SCIENCES

## INFUTURE2009

# DIGITAL RESOURCES AND KNOWLEDGE SHARING

*Edited by*

Hrvoje Stančić, Sanja Seljan, David Bawden,
Jadranka Lasić-Lazić, Aida Slavić

Zagreb, November 2009

# CONTENTS

# Preface

The second conference *The Future of Information Sciences* – **INFuture 2009: Digital Resources and Knowledge Sharing** aims to bring together researchers, professionals, businessmen and project managers from the broad field of information sciences and related professions. The objective of the conference is to provide a platform for discussing theoretical and practical issues in the field of information sciences.

INFuture conferences explore the role of information sciences and related sciences through technological and educational perspective, research studies, organizational, cultural, communication and business aspects evolved form technological development, market needs, European policies and strategies, educational and research activities and current situation in Croatia.

The conference "Digital Resources and Knowledge Sharing" is the second in the series of INFuture conferences, this time held in cooperation of the Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb and IBM Croatia. In the book, more than 70 papers are presented, elaborating, through multidisciplinary approaches, relevant standards and business applications, on the specific domain of information technology in the context of digitisation and heritage preservation.

The book is divided in eight chapters: (1) Invited papers, (2) Digitization and preservation, (3) Using of open-source solutions in cultural heritage, (4) Language technologies, (5) Using information resources in research, education and presentation, (6) Virtual environment in education, (7) E-services, e-government and business applications, and (8) Knowledge management.

Through digitization, heritage preservation in the European context, through preservation of the Croatian national identity, use of open-source solutions in research and education, and through language technologies and multimedia content, the scientific and cultural cooperation at international level has been enabled.

We believe that this type of scientific and practical work, together with intellectual and material capital, gives a significant contribution for memorising Croatian cultural heritage in the European context and creates a platform for future cooperation through networking of researchers and professionals in the field of information sciences.

*Editorial Board*

# INVITED PAPERS

# Sharing Knowledge and Information: Three Views of the Future

David Bawden
City University London
Northampton Square
London EC1V 0HB
dbawden@soi.city.ac.uk

**Summary**

*This paper considers the ways in which library/information services can contribute to the sharing of knowledge over the next two decades. It builds on the imaginative thoughts of several commentators who have considered future scenarios, while including the implications of recent events. These latter include the current world economic problems, which are already affecting library/information services, sometimes in surprising ways. They also include the result of the steady move towards a largely digital information world. Within this, we should note particularly the introduction of 'cloud' computing, which offers a way to share knowledge and information, as well as music, movies and so on, 'on demand', avoiding the idea of 'ownership' or 'collection'. Three possible views of the future are presented. The first is a continuation of the current situation, with library/information services continuing, and perhaps growing in importance, in something very similar to their current form. The second is a change to the current situation, with some forms of library/information service diminishing, or even disappearing, and others expanding, and changing their nature considerably. The third, and most radical, sees the disappearance of most current forms of library/information service, and their replacement by a very different 'information landscape'. The likelihood of these views prevailing, and their consequences for library / information specialists, and for the sharing of knowledge generally, are discussed.*

## Introduction

We will first briefly examine some current trends and issues, and consider their possible significance for library / information services. Then we will look at some possible future scenarios, based on extrapolating these trends.

## Current trends and issues

It is well known that the library/information world has faced a variety of challenges over the past decade, associated with changes in the technical, social, business and economic environment. The information landscape has undergone

remarkable changes in consequence; most particularly as a result of the move towards a predominantly digital and networked provision of information. All sectors of library/information services – including national, public, academic, industrial/commercial and governmental services – have changed greatly in response, though in different ways according to their situation.In the past two years, these forces for change have intensified, and their effects – actual and potential – on the library/information world have been seen more clearly. While one might point to a very wide range of such forces and influences, it is possible to identify a few particularly significant examples, in order to examine their effects.

## The economic situation

The economic downturn worldwide, and the likelihood that it will last for some years, with enduring effects thereafter, cannot fail to affect all the library/information sectors. We may assume that many of these effects, at least initially, will be negative. In the UK, we have seen a local municipality attempting to close the majority of its public library branches in response to economic problems; this provoking appeals to central government to over-rule the action. We have similarly seen an immediate decline in information-related jobs in the commercial sector. No doubt such problems will continue, and will be observed worldwide. On the more positive side, in the UK at least, the economic situation has prompted a ‘back to basics’ mentality. There is some evidence that this includes a re-found enthusiasm for libraries, for books, for borrowing rather than buying, and for sharing and collaborating. Not least, there is a renewed emphasis on the sharing of knowledge and expertise. All of these things could, paradoxically perhaps, result in more, rather than less, support for libraries and other ‘information institutions’.

## Digital world

We have now reached a situation where, in most situations in Europe and North America at least, digital information is the norm. Printed materials are either by-products of digital originals, or are ripe for digitisation, as in the massive book digitisation projects being supported by Google, Microsoft and others. Academic and professional journals are now invariably wholly or mainly digital in form, as are almost all business information resources. E-books are only slowly gaining in popularity, but new developments in e-paper and e-readers imply that a tipping point to wide availability of e-books, and perhaps also e-newspapers, may come soon. This has led many commentators to make assumption that all libraries and other information collections are on a track towards a necessarily all-digital future. However, one of the aspects of the economic downturn noted above, and consequent feelings of uncertainty about the current and future state of society in the developed world, has been a renewed enthusiasm for the ‘real’. This has included an enthusiasm for collections of real things, including real

books. This offers an interesting possible corrective to the idea that the move to an all-digital world is inevitable and universally welcomed.

## Web 2 and the Amazoogle

The success of internet-based information systems, from search engines to online retailers, has led to such systems providing a 'ideal type', by which all information provision is to be judged. Google has had such an impact in this respect that 'to Google' is now a respectable English verb, meaning to search for information. Similarly the Amazon book retailer has provided a model for digital interactions which has proved more acceptable that the kind of interface offered by libraries and information services. Wikipedia, for al its known deficiencies, has become the model for reference information, for most people. The rapid penetration into everyday life of social networking sites such as Facebook, and other Web 2 facilities such as Twitter, has been another factor which has altered the way in which most people – including students and professional workers - expect to receive information. Library and information services have been slow to respond to these developments; not surprisingly, in view of their rapidity. Reponses have generally taken one of two forms. Attempts have been made to point out the limitations of these new models, and the dangers of relying on them; this has generally been ineffective, pitted against the great advantages of speed, simplicity and 'image' possessed by Web 2. Alternatively, libraries and other information providers have attempted to join the new world, by providing 'Google-like' interfaces, and by linking to social networking sites. While this approach offers advantages, there is a danger that the value of sophisticated access to structured collections – one of the main advantages offered in 'library-like' environments – may be dissipated.

## The new generations

Studies of, and anecdotal experience with, the younger generations who have grown up used to ubiquitous digital information give a mixed viewpoint. There seems little doubt that Gen Y, the Google generation, the 'digital natives', or however we might want to label them, have different expectations of information provision, and different information practices, to their predecessors. But studies do not always agree with received opinion. They show, for example, that younger people, while certainly confident with technology, lack an understanding of its nature, and even more lack understanding of information resources, and lack the skills to interpret and us information well. Their fluency with digital information is more apparent than real. And while it is often assumed that the new generations have no interest in libraries, and no desire to use them, studies show that they read more than their predecessors, and appreciate advice from 'information experts'. The idea that there is a novel capability for multitasking, and absorbing and using information from many sources, is countered by concerns about 'attention deficit syndrome' and 'continuous partial atten-

tion'. Solid information on our users of the future is largely lacking. It seems unwise to plan the future based on assumptions. It is this concern that has led the British Library to support the 'Google Generation' study mentioned in the further reading, and in 2009 to sponsor a study of the preceding 'Generation Y'.

## The cloud

One of the most interesting new trends is the movement towards 'cloud computing', whereby information resources of all kinds will not be held in a collection, owned by a person or an institution, but will be stored on a networked, with wireless access from any place. In 2009, this has been illustrated strongly by public enthusiasm for the Spotify software, which enables music and videos to be downloaded on request from the cloud. This, if widely accepted, could mean an end to personal collections of CDs, DVDs, etc. It may not be too far-fetched to imagine that books, newspaper items and journal articles could be delivered in the same way. This would call into question the whole need for the 'collection', which has been the basis for libraries and information services since their inception in the ancient world. There are of course many issues here, and whether it would be acceptable for all purposes, is a debatable point.

## Trend summary

This brief overview of some currently prominent issues gives us a rather mixed picture. We see some general trends, but often with 'counter-movements' to suggest that the future is not entirely clear.

## Futurology and scenarios

There is a long tradition of 'futurology' in the library and information sciences, with a particular interest in outlining possible scenarios of what the information world will be like in future years. Shuman's collections of scenarios for 'libraries of the future' have been especially influential, and the writings of Sapp and Pennevaria are also interesting. We will now look at three possible, very general, futures for the library / information world, based on, and accepting some of the contradictions of, the trends noted above.

## Scenario 1: business as usual

Libraries, and other information providers, have a generally good 'brand image', much goodwill, and an established place in many settings. One possible future would therefore be largely a continuation of the current situation, with library/information services continuing, and perhaps growing in importance, in something very similar to their current form. This seems unlikely, especially in view of the increasing emphasis of digital networked information. On the other hand, we might set against this the observed continuing, and to a degree renewed, enthusiasm for reading, for collections, for 'real' things in general, and for a more shared and community-minded approach. If these trends continue,

and in particular if there is a back-lash against immersion in a virtual environment, then we might expect to see the traditional 'collection and place' form of library and information service regaining something of its former status. This would almost certainly be true in some sectors more than others, and in all cases will involve a considerable extension into digital provision. But we should not assume, as it often is assumed, that radical change is inevitable.

## Scenario 2: changing landscapes

This scenario involves a change to the current situation, affecting the library/information sectors in various ways. Under the influence of the negative (at least for traditional library / information provision) trends noted above, we might expect to see some forms of library/information service diminishing, or even disappearing. Others, conversely, might be expected to expand, and to change their nature considerably. We can already see some evidence of the sort of change which may be anticipated from developments in university libraries in Europe and North America. Many are moving from a role purely as a rather passive information provider, to a more active involvement in teaching, and to provision of study spaces extending into a 'social space' dimension. Similar changes are to be seen in the public library sector, with UK libraries morphing into 'Ideas Stores' and 'Discovery Centres'. Throughout Europe, there is a trend to involve the public library service more closely in the wider social and cultural environment. In a commercial environment, this is seen as an increased emphasis on knowledge management, though perhaps not using that term. Diminishment can be seen in the UK public library examples noted above; falls in usage can be seen in many European countries, leading to fears for the long-term survival of such services. Reduction in provision can also be seen in commercial and industrial information services, as the economic downturn exacerbates longer-term trends. Conversely, information services in law and in healthcare are, if anything, expanding in the UK, albeit accompanied by many changes in their structure and functions. This illustrates the way in which the various sectors are responding differently to the changing environment. Perhaps the one constant in these factors is the extent to which the successful services are managing to refocus the balance between physical and digital presence. This, along with a good awareness of user needs, and an ability to 'play to the strengths' of the 'library brand', seems to still be the requisite for success.

## Scenario 3: into the clouds

This scenario is the most radical. It assumes that the 'cloud' computing model has become pervasive, that it delivers texts and images in the same way as it is currently beginning to deliver music and video, and that such delivery has become accepted as the norm by most people. We might also expect that much more of this material would be free to use, in an extension of current trends with entertainment material, of the open access movement for academic resources.

This would mark the end of the 'collection' – whether of books, of music, or of photographs – as a feature of the lives of those who replied on cloud delivery. Similar, even one's own documents would not be stored on any local device, but would be uploaded to the cloud, to be retrieved when needed. This would involve a fairly major change in attitudes, as well as information practices, and it is not clear to what extent this would be welcomed. Issues of preservation, both technical and cultural, would become a particularly important issue. The impact of library / information services would be profound. Their collections – hitherto their defining feature – would now take the form of criteria for selection of material from the cloud. This would take the process begun by the move to the 'digital library', with collections being defined by selection criteria rather than by ownership or by physical location, to a much greater limit. The library might cease to exist as an institution, and might take the form simply of advisory services, assisting patrons with selection of cloud material, and subsequent use. Lacking a physical location, and managed collection, library / information services would, if they were to survive at all, have to become more deeply embedded in the life and work of their patrons. Conversely however, we might find that a reaction might set in to the cloud environment. There might be a demand for a return to the 'traditional' values of managed and organised collections, and for a revival of the 'information place' for study, reflection, and social interaction focused on cultural issues. Libraries and information centres would be seem to be well placed to meet that need.

## Conclusions

Current trends show a number of paradoxical and contradictory developments. Predicting the future based on them cannot therefore be straightforward. The history of prediction and futurology in the library / information sciences has shown both the failures of imagination and of nerve identified by Arthur C Clarke. For the most part, the most striking developments, from the home computer, to the web, to the cloud, were not clearly anticipated before their arrival. Yet, despite many grim predictions to the contrary, libraries and information services have survived so far. They have done so by a mix of holding on to their traditional purposes and roles, and of reinventing themselves to suit their new environment. In the inherently unpredictable times to come, a mixture of these strategies will be needed. So, also, will be research and reflection on developments, and on the ways in which information providers can be respond. The sharing of knowledge is likely to be more important than ever in coming years, and information specialists have a great contribution to make.

## References

Atkinson, R., Contingency and Contradiction: the place(s) of the library at the dawn of the new millennium, Journal of the American Society for Information Science and Technology, 2001, 52(1), 3-11

Bawden, D., The nature of prediction and the information future: Arthur C Clarke's Odyssey vision, Aslib Proceedings, 1997, 49(3), 57-60

Pennavaria, K., Representations of books and libraries in depictions of the future, Libraries and Culture, 2002, 37(3), 229-248

Rowlands I., et.al., The Google generation: the information behaviour of the researcher of the future, *Aslib Proceedings*, 2008, 60(4), 290-310

Sapp, G., A brief history of the future of libraries, Lanham MD: Scarecrow Press, 2002

Shuman, B. A., Beyond the Library of the Future: more alternative futures for the public library, Englewood CO: Libraries Unlimited, 1997

Shuman, B. A., The Library of the Future: alternative scenarios for the information profession, Englewood CO: Libraries Unlimited, 1989

Williams, P., et.al., The 'Google Generation' – myths and realities about young people's information behaviour, in *Digital Consumers: reshaping the information profession*, edited by D Nicholas and I Rowlands, London: Facet Publishing, 2008, pages 159-192

## Links

http://soi.city.ac.uk/~dbawden

# DigitalPreservationEurope: A Way Forward in the Long Term Curation of Digital Materials

Chiara Cirinnà
Fondazione Rinascimento Digitale
Via Bufalini 6, Florence, Italy
cirinna@rinascimento-digitale.it

Maurizio Lunghi
Fondazione Rinascimento Digitale
Via Bufalini 6, Florence, Italy
lunghi@rinascimento-digitale.it

**Summary**

*We are shifting from an industrial economy to an economy in which knowledge is an input of increasing significance and this has been triggered by rapid progress in the evolving Age of Information, where repositories of digital information and the tools to mine, analyse and re-purpose them represent a society's intellectual capital. Effective and affordable digital preservation strategies and systems will transform archives into valuable assets.*
*The European funded coordination action DigitalPreservationEurope (DPE) has reached its conclusion in March 2009 and several results in the field of digital preservation have been accomplished.*
*The primary objective of the coordination action was to concentrate and share efforts towards the common purpose of assuring effective preservation of digital materials. To this end, the three main projects on digital preservation, CASPAR, Planets, DPE (co-funded under the European Commission Information Society Technology (IST) Sixth Framework Programme) have worked together to create the window on their synergistic activities*
*Before a repository is created, PLATTER, the Planning Tool for Trusted Electronic Repositories, provides a basis for a digital repository to plan the development of its goals, objectives and performance targets over the course of its lifetime in a manner which will contribute to the repository establishing trusted status amongst its stakeholders.*
*After the repository has been created, the digital curation can be considered as a risk-management activity. Based on practical research and developed jointly by the DCC (Digital Curation Centre) and DPE, the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) provides a methodology for self-assessment through a metric to enable an auditor to establish the organisa-*

*tional context and goals of a repository and then to assess how it is achieving these in terms of risk.*
*Digital preservation solutions are a little less distant then before.*

**Key words:** digital preservation, trusted digital repositories, risk-assessment

## Introduction

Digital preservation is a set of activities required to make sure digital objects can be located, rendered, used and understood in the future. This can include managing the object names and locations, updating the storage media, documenting the content and tracking hardware and software changes to make sure objects can still be opened and understood.
But what does 'long-term' mean in the context of digital preservation? DPE agrees with the CCSDS[1], that states “a period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future”.
That assumed, what do we need to preserve? Various aspects of the digital objects may be needed to be preserved. The lowest level of preservation requirements includes preservation of the bit stream, this does not however ensure understandability, readability or usefulness of the digital object. The biggest risk in terms of understandability is that the meaning (and even the names) associated with values in a dataset, although known to the data producers, is not available to the users; without this the data is essentially useless. Another aspect is that, even for users within the same sub-discipline, terminology drifts and meaning is lost; users in different (sub)disciplines will require even more help with the semantics of the data.
A more complex approach may strive to preserve also the meaning so that it remains readable and understandable. Such an approach requires the preservation of additional information (representation information, technical metadata etc.)
Even more ambitious preservation approaches try to preserve understandable content in such a way that the provenance and source of the digital object also remains clear. Thus the users can have trust that the object is authentic, accurate, and complete.

## Why should we care about digital preservation?

Digital objects are much more 'fragile' than traditional analogue documents such as books or other hard copy mediums. Digital objects are fragile because they

[1] Consultative Committee for Space Data Systems

require various layers of technological mediation before they can be heard, seen or understood by people. Digital objects are also much more venerable to physical damage. One scratch on CD-ROM containing 100 e-books can make the content inaccessible, whereas to damage 100 hard copy books by one scratching move is - fortunately - impossible. A flash memory stick can drop into glass of water or get magnetised, portable hard drive or laptop can slip from your hands and get irreparably damaged in a second.
Digital objects require pro-active intervention to remain accessible. While you can put a book on a shelf and return to it in upwards of 100 years and still open it and see the content as it was intended by the author/publisher, the same approach of benign neglect to a digital object is almost a guarantee that it will be inaccessible in the future.
Alternatively the software or file format can become obsolete for a number of reasons. For example software upgrades may not support legacy files; the format take up is low and the industry does not produce compatible software; software which supports the format may be bought by a competitor and withdrawn form the market place. Without the intervention of digital preservation techniques the information contained will no longer be accessible.
The following paragraphs describe some of the main results achieved by the DPE[2] project.

## A successful coalition

In order to reach bigger results in terms of disseminating knowledge, information and practice among a wide community, the DPE project with the two main projects on digital preservation issues, CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) and Planets (Preservation and Long-term Access through NETworked Services) have signed up an agreement to enable synergy and concerted action. The first results of this joint cooperation was the development and delivery of a collaborative web platform shared by the projects to serve as a common entry point to digital preservation and curation projects[3], provided with common services, a calendar of events, information board, resources. Moreover they have collaborated on the development of training and educational events and programmes in Europe and supported the dissemination of publications and the mechanisms to ensure their visibility (e.g. by automatic means such as OAI-PMH).
Several common events have been organised, like the joint conference in Lisbon in 2007 and in Nice in 2008.

[2] http://www.digitalpreservationeurope.eu

[3] http://www.wepreserve.eu

## PLATTER, a tool for achieving trust in repository planning

DPE has recognized that a critical step for creation of a repository is to early plan the development of its goals, objectives and performance targets. PLATTER, the Planning Tool for Trusted Electronic Repositories, is not in itself an audit or certification tool but is rather designed to complement existing audit and certification tools by providing a framework which will allow new repositories to incorporate the goal of achieving trust into their planning from an early stage. A repository planned using PLATTER will find itself in a strong position when it subsequently comes to apply one of the existing auditing tools.

When we deal with a “trusted repository”, we can state that a repository is “Trusted” if it can demonstrate its capacity to fulfil its specified functions, and if those specified functions satisfy an agreed set of minimal criteria which all Trusted Repositories are assumed to require.

Among all the different available approaches, a suitable compromise would be to allow repositories to identify their own goals within a broadly accepted framework of basic requirements relevant to all trusted repositories. Precisely such a framework is represented by the Ten Core Principles of Trust Repository Design which have been developed by the Center for Research Libraries (CRL), the Digital Curation Centre (DCC), DigitalPresevationEurope (DPE), and the German Network of Expertise in Digital long-term preservation (nestor) in the field of audit and certification at a meeting hosted by CRL in Chicago in January 2007. The principles state that a repository:

1. Commits to continuing maintenance of digital objects for identified community/communities.
2. Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfil its commitment.
3. Acquires and maintains requisite contractual and legal rights and fulfils responsibilities.
4. Has an effective and efficient policy framework.
5. Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.
6. Maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.
7. Creates and maintains requisite metadata about actions taken on digital objects during preservation as well as the relevant production, access support, and usage process contexts before preservation.
8. Fulfils requisite dissemination requirements.
9. Has a strategic program for preservation planning and action.
10. Has technical infrastructure adequate to continuing maintenance and security of its digital objects.

What remains open, and what PLATTER is designed to address, is how these principles can be incorporated into the design and planning of a repository so that it is “trust-ready” from the start.

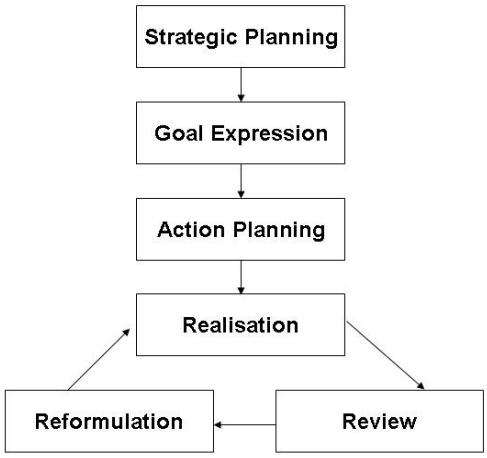
**Repository classification**

Since no "one-size-fits-all" approach can hope to apply to all types of repository, it is vital for a repository planner to be able to classify their repository in order to be able to compare its policies and practices with other similar repositories. The first stage of the PLATTER analysis is a taxonomic classification which will enable a repository to be compared with other similar repositories. Many possible schemes for such a classification could be developed, and in PLATTER we have chosen to characterise a repository along a number of independent axes grouped into four major descriptive classes:

- Purpose and Function: the purpose of this group of taxonomic axes is to determine the general functional type of the repository. The requirements of a national library may be quite different from those of an institutional or subject-based repository, a scientific data repository, or a national archive.
- Scale: in this group we consider the various factors which together define the overall scale of the repository, whether expressed in human, technical, or financial terms.
- Operation: this group of axes is primarily concerned with how material enters into the repository, the kind of material stored, and the extent to which that material may be accessed by end users.
- Implementation: this group of axes deals with the choices made in the implementation of the repository system.

The PLATTER planning cycle describes a semi-formalised set of steps intended to facilitate the processes of definition and expression of organisational objectives, and implementation and evaluation of the measures intended to meet them.

Table 1-The PLATTER Planning Cycle



The process is a cyclical one, and individual sections conform in many respects to parts of the DRAMBORA risk-analysis process. The following sections seek to describe in some more detail each stage of the PLATTER cycle, outlining their implicit parts and how they interrelate.

Strategic planning. Strategic planning is an invaluable means for maintaining a sufficiently broad and forward-facing organisational perspective, even when individuals are focusing on much more immediate and specific aspects of business activity. One, or more of three fundamental questions are posed during the process of strategic planning:

1. What do we do?
2. Who do we do it for?
3. How can we excel?

Responses to these three questions organisations will encapsulate the repository's mandate (or reference a non-self imposed, e.g., legislative mandate), detail the identities and broad expectations of primary stakeholders and describe in general, but tangible terms, the circumstances and performance levels that will represent success.

Definition of goals. Defined objectives must take into account the expectations and requirements of each major stakeholder. From the perspective of digital repositories, this may include management, funders, information creators, owners and depositors, and end users interested in accessing preserved content. Each must be related, either explicitly or implicitly with more fundamental strategic objectives, most immediately with the organisation's mission statement.

Undertake planning. The planning stage is bridge-building; between the determination of what must be achieved, and the tangible realisation of such achievements. Related experiences in comparable environments can be considered and where appropriate absorbed into the planning cycle. Action planning is only really feasible by adopting a global perspective of organisational constraints and influences. Many will have been formalised in the initial strategic planning stages, most notably those focused on establishing current situational awareness, and perceptions of any emerging contextual influences. Legislation, policy originating from parent organisations, stakeholder expectations and resource availability will all contribute to the success or otherwise of planned actions and must be given adequate consideration.

Deliver, review and reformulate implementation. In an iterative cycle, that may extend beyond the planning and development of the repository into full production phases, these three interrelated activities are fundamental to the ongoing improvement and developing maturity of the repository. An agile approach to all three will benefit the organisation and the pursuit of its objectives; no period of implementation should become too prolonged prior to the initial phases of review and reformulation.

The PLATTER process is centred around a group of Strategic Objective Plans (SOPs) through which a repository specifies its current objectives, targets, or key performance indicators in those areas which have been identified as central to the process of establishing trust.
In the future, PLATTER can and should be used as the basis for an electronic tool in which repositories will be able to compare their targets with those adopted by other similar (suitably anonymised) repositories. The intention is that the SOPs should be living documents which evolve with the repository, and PLATTER therefore defines a planning cycle through which the SOPs can develop symbiotically with the repository organisation.
The PLATTER tool is concerned exclusively with management of the objectives and targets of repository. It is not itself a tool for establishing trust and is not intended to compete with other initiatives in that area.
PLATTER is designed to complement DRAMBORA and a repository planned using PLATTER will be strongly placed to use DRAMBORA as a self-assessment tool.

## Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)

Most current digital repositories and most databases and collections used to help curate scientific data do not have specific mandates for long term preservation, nor do they have necessary long-term budgets. Instead they are mandated to support access and re-use in the near-term future. Long term preservation may be one of their aims, or at least hopes and wishes, but it is not (yet) a responsibility.
The DRAMBORA toolkit aims to complement other repository and certification work by addressing the full range of repositories, whether they aim for long term preservation or not. The toolkit is intended to facilitate internal audit by providing repository administrators with a means to assess their capabilities, identify their weaknesses, and recognise their strengths.
Within this toolkit the authentic and understandable digital object is positioned at the centre of a risk-based approach to audit; digital curation is characterised as a process of transforming controllable and uncontrollable uncertainties into a framework of manageable risks, classified according to a repository's activities, assets and regulatory context. The audit tool will encourage repository staff to identify and classify the risks posed at every stage of their activities, to assess the probability of their occurring, to appreciate their potential impact if they should arise.
Throughout a series of interactive stages, auditors are expected to develop a comprehensive image of their organisational objectives, the regulatory context within which they operate and the activities that must consequently be undertaken. Any risk management exercise will include these stages:

- Identifying the context where risks have to be managed

- Identifying risks
- Assessing and evaluating risks
- Defining measures to address and manage risks

The self-audit progresses through six stages:

Stage 1: Identify organisational context. In Stage 1, auditors document the mandate and derive both the goals and objectives of the repository. The ultimate purpose of this stage is to define the scope of the repository work, verifying internal awareness of the organisational framework, and at the same time ensuring that appropriate supporting documentation exists. Within this stage auditors must describe the overall purpose of the repository, in order to determine the characteristics that will undergo risk analysis and subsequent assessment. In particular auditors must identify the repository’s mandate which will be described in an organisational mission statement, then they will identify, within the mandate, each organisational goal and objective relevant to the repository.

Stage 2: Document policy and regulatory framework. This Stage gives auditors the opportunity to provide or refer that the repository:

- operates appropriately with respect to relevant regulatory frameworks;
- has an efficient and effective policy framework;
- is aware of the societal, ethical, juridical, and governance frameworks;
- is aware of the legal, contractual and regulatory requirements to which the repository is subject.

At this stage auditors need to determine what to look for, to collect information from documentary sources, and to compile a list of documents regulating the work of the repository.

Stage 3: Identify activities, assets and their owners. The purpose of Stage 3 is to develop a conceptual model of what the repository does and how it does it, by examining its activities and work processes, key assets and technology, and the staff involved. This Stage requires auditors to split the broad-level mission and goals of the repository into more specific activities or work processes that the repository carries out in order to achieve its aims.

Stage 4: Identify risks. The aim of this stage is to derive from organisational activities and assets a comprehensive selection of pertinent risks faced by the repository. Some risks can be derived from examining the mandate and objectives, regulatory environment and the model of the repository’s work (activities, assets, staffing, technology solutions).The principal outcome is the definition of an organisational ‘worry radius’, detailing the parameters within which risk management must be undertaken.

Stage 5: Assess risks. The aim of this stage is to characterise the risks and risk relationships derived within the previous stage, and to assess the severity of each. Each risk must be enriched with a number of additional attributes; among the most significant are values describing the probability and potential impact of

each, which cumulatively offer a quantitative insight into the overall riskiness of the repository's business activities.
Stage 6: Manage risks. A fundamental imperative with respect to this work is that risks must be managed appropriately. Once a risk has been assessed, a business decision must be made to determine how the risk is to be approached. This should consider the risk's potential impact, its frequency, its owners and its stakeholders. Risk mitigation strategies and tasks should be assigned, with accompanying deadlines for achieving predefined targets. There are several strategies that an organisation can pursue to deal with the negative impact of identified risks. In this Stage, auditors are asked to:

- choose a risk management strategy;
- describe the risk mitigation measure;
- assign responsibility for the risk mitigation activities;
- set target dates and/or results for the risk mitigation activities.

A principal outcome from the successful completion of this stage is a risk register with risk management features included. The risk management exercise cannot and should not stop with the creation of the risk register. Ongoing review and monitoring is essential to ensure that the risk management plan remains relevant. It is therefore necessary to repeat the risk management cycle regularity and review the target outcomes when their deadlines are reached.
Risk management is the final Stage and the end-result of this self-audit. The previous five Stages have created a comprehensive body of information that ultimately informs the risk treatment and management process.
This toolkit was developed as a collaboration between the Joint Information Systems Committee and Core eScience funded Digital Curation Centre (DCC) in the United Kingdom and the European Commission co-funded initiative DigitalPreservationEurope (DPE). These two initiatives will continue to work together to test and refine the toolkit, to manage the online tool, which is available at http://www.repositoryaudit.eu, and to foster its widest possible take up within Europe and broader international contexts.

## Conclusions

DPE has recognised that some of the benefits of digital preservation can be identified as the following:

- Legal. National legal frameworks often require organisations to provide adequate records of business processes, communications and many other types of data for many years after their creation.
- Accountability & protection from litigation. Recent legal cases have shown the importance of being able to search and recover archived emails quickly and in a legally admissible manner.
- Protecting the long term view. Access to digital data is critical to ensure business continuity and to support decision making with a long term

view. For research in particular preserving data may be crucial for identifying long-term trends.

- Protecting investment. The valuable intellectual assets of organisations are increasingly in digital form. This data represents both intellectual property and a considerable investment of time, effort and money. It would therefore be foolish not to protect and preserve these assets adequately.
- Reuse. Repositories of digital information and the tools to mine, analyse and re-purpose them represent a society's intellectual capital. Effective and affordable digital preservation solutions are essential to transfer digital data into valuable assets for business.

When PLATTER has been used for planning of objectives, a repository will be in a very strong position to carry out an effective DRAMBORA analysis because all its current objectives will be thoroughly documented. The combination of PLATTER and DRAMBORA therefore represents a powerful tool in the development of Trust.

## References

CCSDS (Consultative Committee for Space Data Systems). Reference Model for an Open Archival Information System (OAIS). Blue Book, Issue 1. Washington, DC (US), CCSDS Secretariat, January 2002. Technical report. CCSDS 650.0-B-1. Recommendation for Space Data System Standards. http://public.ccsds.org/publications/archive/650x0b1.pdf.

ERPANET. Workshop on audit and certification in digital preservation. 2004. http://www.erpanet.org/events/antwerpen/index.php.

JISC. Managing risk: a model business preservation strategy for corporate digital assets. 2005.

McHugh, Andrew; Ross, Seamus; Ruusaleep, Raivo; Hofman, Hans. The digital repository audit method based on risk assessment (DRAMBORA), 2007. ISBN: 978-1-906242-00-8. http://www.repositoryaudit.eu.

National Council on Archives. Your data at risk. Why you should be worried about preserving electronic records. 2005.

nestor Working Group on Trusted Repositories Certification. The catalogue of criteria for trusted digital repositories. Version 1. nestor Studies n. 8. 2006.

RLG/NARA Task Force. An audit checklist for the certification of trusted repositories. 2005. http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf.

RLG/OCLC Task Force. Trusted digital repositories: attributes and responsibilities. 2002. http://www.rlg.org/legacy/longterm/repositories.pdf.

Ross, Seamus; McHugh, Andrew. Audit and certification of digital repositories: creating a mandate for the digital curation centre (DCC). // *RLG DigiNews*. Issue index: October 2005. http://www.rlg.org/en/page.php?Page_ID=20793#article1.

Ross, Seamus; McHugh, Andrew. The role of evidence in establishing trust in repositories. *D-Lib Magazine*. July/August, vol.2, nos 7/8. http://www.dlib.org/dlib/july06/ross/07ross.html.

World Bank. Assessment of organisational capacity to manage records: a top level checklist. 2004.

# Knowledge Architecture: A Vision for the 21st Century

R. J. Bater
KnowPlexity Ltd.
43 Ashley Down Road
Horfield
Bristol
United Kingdom
bbater@knowplexity.com

**Summary**

*The realization that information and knowledge are key resources in the 21st century knowledge economy, and that they need to be managed, came late to many organizations. Few were well-equipped to face the challenge. For many, the result has been a 'BandAid' solution, where inadequate commercial off-the-shelf software has been applied to an ill-defined set of requirements at no small cost.*

*Two key problems facing organizations in this situation are identified. Firstly, the great majority of applications offered as 'knowledge management solutions' are actually information management solutions and fail to deal not only with the problems of managing email, instant messaging and Web 2.0, but also to engage with the essentially human behavioural and cultural issues. Secondly, understanding seems to be lacking among both vendors and practitioners regarding the complexities which can arise when trying to 'manage' this human feature called 'knowledge'. Leading thinkers in the field are drawing upon the unlikely combination of complexity theory and anthropology to forge a new paradigm of knowledge and information creation and transfer. Yet, this work is not widely known.*

*It is proposed that the relatively new discipline of Knowledge Architecture promises a heterogeneous, holistic framework within which information management can evolve to meet the challenges presented, by recognizing that organizations are complex adaptive systems and by the application of tools and techniques such as sensemaking and social network analysis.*

**Key words:** knowledge sharing, information management, knowledge management, knowledge architecture

# Introduction

## IM Awakening

A report by IDC in March 2008 (IDC/EMC, 2008) estimated that the digital universe in 2007 comprised 281 exabytes (1 exabyte = $10^{18}$ bytes), and that by 2011, the digital universe will be 10 times the size it was in 2006. Most people look at figures like this and see a threat. It portends challenges, certainly. But information professionals should also see opportunities.

If information production is growing so fast, then information must be important. So, information matters! And that happens to be the title of a report published by the UK government last November (Knowledge Council, 2008). In the Information Matters report, Sir Gus O'Donnell, the Head of the UK Civil Service and Cabinet Secretary writes in his introduction:

"Good information management needs to be partnered with good knowledge management. If it isn't, the value of information as an asset is undermined, and cost-effective, efficient service delivery is compromised."

Yet, only the year before, surveys revealed that:

- on average, staff spend 9.6 hours per week searching for information (IDC/EMC, 2008)
- at any given time, 3% – 5% of files can't be found. Recreating them costs on average €100 per document (Information Week Survey, 2008)

There is obviously a mismatch here between the aspirations of Information Matters and the reality of the surveys. If information professionals want to rectify that mismatch, that presents them with a considerable challenge. It's a challenge they welcome, of course, but to meet it on its own terms it will take time, resources, and a better understanding of how knowledge and information *together* generate value.

## BandAid Response

We long ago responded to the problems of managing money by devising complex systems for managing it. Information has not fared so well, and the response has proved to be piecemeal and blinkered, with each function attending to its own needs. When computerized information systems began arriving in large numbers in the 1960s, every information problem was tackled independently. In the absence of any overall scheme, the result was a collection of ad hoc, single-purpose tools, each serving their own organizational silo:

- a financial accounting system
- a customer database
- an employee database
- a payroll application
- management information systems

Today, we call that 'Information Scatter', that is, an organization's information is scattered among multiple, incompatible, non-communicating information

systems and data stores. A whole new generation of middleware has evolved in recent years to address this problem, like IBM's WebSphere application integration software, or Schemalogic's suite of applications to cross-map and manage metadata across multiple, siloed information systems. Such responses can be viewed as an attempt to derive order *a posteriori* out of a chaos that need not have existed in the first place. The result is complete bewilderment when users need to find specific information.

### Holistic Health

The problem is, knowledge needs information and information needs knowledge, a truism that seems to have been largely ignored by the recent histories of both information management and knowledge management. Without information to feed it, we can have no knowledge. And without knowledge, information is useless.

The 17th century English philosopher and scientist Francis Bacon is often quoted as saying "Knowledge is power." But this is almost a complete inversion of what he actually said. The quotation in full is:

"But mere knowledge is not power; it is only possibility. Action is power; and its highest manifestation is when it is directed by knowledge." (Bacon, 1597)

So, knowledge is the power to take effective action.

## Problems in Knowledge Architecture

### Information + Knowledge: A New Kind of Problem

A knowledge architecture is, if you like, an expression of the 'personality' of an organization - what it exists to do and for whom, how it goes about doing what it does, who does what, and what information and knowledge are used - and produced - along the way. Building a knowledge architecture which will work well for an organization, is therefore a bit like being a psychoanalyst. You have to delve deeply into motivations, the rationale behind patterns of behaviour, the influence of legacies from the past, whilst being prepared to come across all sorts of disorders and complexes lurking beneath the surface, just waiting to be discovered.

What's really needed is a framework where the jig-saw of business processes, information systems, information resources, and the knowledge in people's heads, can be put together to make a meaningful picture. That's the challenge of Knowledge Architecture. There are problems with information, and there are problems with knowledge.

### Information in the Wild

The problem of information scatter has already been mentioned. In reality, the information fragmentation problem is far worse. Much of the information we need to capture is wild. It doesn't reside in structured environments like data-

bases and libraries. It grows where and when it's needed – and often where it's not.

A recent working paper by UK National Archives CEO Natalie Ceeney estimated that 80% of an organization's information is of this type (Ceeney, 2009). Wild information resides in documents, emails, instant messages, blogs, wikis, del.icio.us, Technorati, Twitter – the list could be extended.

While some documents remain native to paper, most are now born-digital: Microsoft Office, Open Office, HTML, XML. We have EDRM systems for managing Office documents, but HTML and XML present a whole new set of problems. Email steadfastly refuses to limit its growth and to be manged, while instant messaging presents further problems of its own. Both almost certainly contain vital information. With Web 2.0 applications – blogs, wikis, URL collections like del.icio.us and Technorati – we must ask how we might identify what's important, how we might know when to capture something which is continually changing, and how to organize it? Some Web 2.0 commentators like Clay Shirky (Shirky, 2006) say that conventional information management techniques like classification and taxonomy are irrelevant anyway for Web-based resources.

**Taming Wild Information**

Even if we can capture what needs to be captured, we cannot assume that information from different sources can be easily aggregated without further cleansing, processing and mapping. Lou Rosenfeld, a founder of the North American Information Architecture movement, illustrates these problems as shown in Table 1 (after Rosenfeld, 2003).

Table 1: Metadata interoperability problems

| Source A | Source B | Source C |
|---|---|---|
| **Metadata are not interoperable** | | |
| **Create Date**: 071005<br>**Author**: William Jones<br>**Subject**: guidelines | **Date**: July 10, 2005<br>**Compiler**: Bill Jones<br>**Topic**: policies | **Created date**: 100705<br>**Creator**: Jones, W.<br>**Descriptor**: procedures |
| **Structural interoperability achieved through applying standards** | | |
| **Date**: 071005<br>**Creator:** William Jones<br>**Subject:** guidelines | **Date**: July 10, 2005<br>**Creator:** Bill Jones<br>**Subject:** policies | **Date**: 100705<br>**Creator:** Jones, W.<br>**Subject:** procedures |
| **Semantic interoperability achieved via controlled vocabulary** | | |
| **Date**: 2005-07-10<br>**Creator:** Jones, W.<br>**Subject:** policies | **Date**: 2005-07-10<br>**Creator:** Jones, W.<br>**Subject:** policies | **Date**: 2005-07-10<br>**Creator:** Jones, W.<br>**Subject:** policies |

Even internationally accepted classifications can pose problems. An EU institution had to deal with other economic institutions all over the globe. They used ISO 3166 (ISO 3166, 2006), which defines the international standard codes for representing countries, but it only supported some of their needs. For the rest, they had to improvise.
ISO 3166 is a geographical vocabulary, and the institution has to work with supranational, geoeconomic and other groupings as much as the geographical. For example:

Table 2: Geographical vocabulary problems

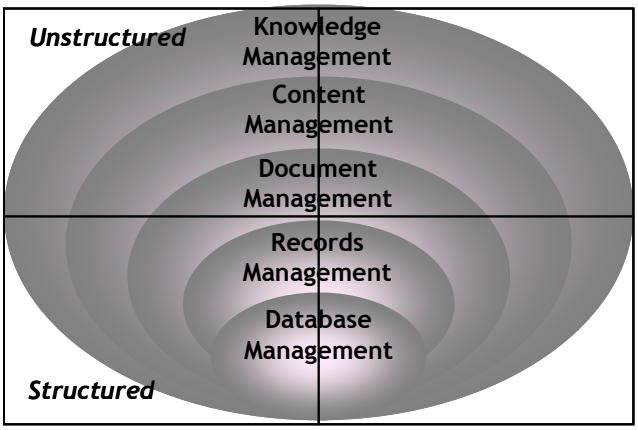
| **Supranational Groupings** | **Geoeconomic Groupings** |
|---|---|
| Near East | COMECON |
| Middle East | EFTA |
| Far East | OPEC |
| Europe | Eurozone |
| **Development Groupings** | **Legacy Geopolitical Groupings** |
| ACP (Africa, Caribbean & Pacific) | FYROM – Former Yugoslav Republic of Macedonia |
| ALA (Asia & Latin America) | |

So, for instance, you won't find supranational groupings in ISO 3166, like Near East or even Europe. And you won't find geoeconomic groupings like COMECON, EFTA or Eurozone either. Nor will you find common groupings used in development economics like those loose alliances of developing countries such as ACP and ALA. And of course, political groupings break up and reform. So the institution struggled to decide whether it should index documents under 'Macedonia' or 'FYROM'.
If we think these are substantial information management problems - and they are – how much more of a problem is it to try to manage something which is intangible, only partially controllable, highly personal and ever-changing? Well, the answer is 'we can't'. But we *can* encourage its growth and its flow, and. Web 2.0 applications offer new ways of doing that. They don't, however, lend themselves well to 'being managed'.
Yet, blogs & wikis only carry knowledge that gets written down. Alongside these 'visible channels' there are other, informal channels - like the telephone, conversations at the coffee machine or in the corridor - which carry a vast daily flow of unwritten knowledge which remains entirely untouched. We can't capture that knowledge directly and shouldn't try. But we can find out who talks to whom, and for what reason. In many cases, that will be enough.

**Accommodating Diverse Points of View**

Just as Architecture seeks to create spaces for defined purposes, so Knowledge Architecture must seek to create knowledge spaces for defined purposes. Knowledge and information are, of course, used for many different purposes in an organization. This paper proposes that these purposes can be clustered into five overlapping domains, each of which makes different demands of an architecture in terms of the attributes of information and knowledge they view as important.

Figure 1: Domains of use of knowledge and information




In Figure. 1, from bottom to top, the degree of structuring decreases The purpose of Database Management is to capture, store and make data retrievable. Database Management Systems are, of course, highly structured. The attributes which matter most are things like data type, cardinality and whether a column is a primary or foreign key.

Records Management has as its core purpose, the capture, secure storage and provision of controlled access to a subset of organizational information deemed to be records of the organization's activities. Much of its content will be derived from clusters of documents. The important attributes here are the organizational activity which produced the record, the applicable lifecycle (retention - preservation/destruction) as may be dictated by regulatory authorities, and appropriate access control. Structure is therefore essential.

Document Management, like Records Management, deals with information linked to defined organizational processes, but applies very different lifecycle parameters and access controls. Governance is driven more by internal needs for due diligence than by external regulators. Document Management focuses on the collaborative generation of discrete packages of unstructured information (documents) in the context of a defined organizational process. Controlled distribution, version control, approval workflows and sign-off procedures determine the important attributes and degree of structure.

Although not widely recognized as such, the main purpose of Content Management is publication, either to an internal or an external audience. Content Management people therefore need to maintain awareness of all organizational information which can and should be disclosed to defined audiences. Content Management is driven either by the requirements of applicable regulations, by the need to keep staff aware, or by marketing, promotional or public relations considerations.
Finally, the Knowledge Management domain - the least structured - is concerned with creating the conditions in which people can use and share the knowledge they have, combine it with knowledge shared by others, and fuel creativity by drawing on the information preserved in the organization's repositories.
Tools and techniques for the lower four domains are well developed but have now to be applied organization-wide, with all of the classification and metadata interoperability problems that implies.
But the tools and techniques for achieving the aims of knowledge management so far offered barely break the mould established by information management.

## A Vision for Knowledge Architecture

We need to look elsewhere for guidance on how Knowledge Architecture can accommodate knowledge alongside information and facilitate its 'management'; how it can be inclusive of all points-of-view on the organizational knowledge base and the diverse uses to which it is put. A number of innovative analytical frameworks - have emerged in recent years which may offer a solution to the impasse. Three, of particular significance are summarized here.

### The Cynefin Framework

When Dave Snowden was Director of the IBM Institute for Knowledge Management a few years ago, his research into the relevance of Complexity Theory to Knowledge Management produced what he called the Cynefin Framework (Cynefin Framework, 24 August 2009). Snowden has since developed this framework into a sophisticated methodology for what is known in management epistemology as 'sensemaking'.
The Cynefin Framework defines five domains in which we interact with experience, use information, and apply our knowledge:

- Simple
- Complicated
- Complex
- Chaotic
- Disorder

Cause-and-effect are collective truisms in the Simple domain, supporting logical decisions and best practice, but in the Complicated domain, cause-and-effect are not obvious and good practice is only revealed with effort or expertise. The

Complex domain affords no sense of predictable cause-and-effect, but it may be recognized in retrospect as the source of 'emergent practice'. The two remaining domains exhibit no discernible cause-and-effect relationships. Although 'novel practice' may reveal itself in the Chaotic, in Disorder, decisions will be made largely on intuition.
The Cynefin Framework helps us to understand how we need to define 'knowledge' according to the applicable domain, and to apply different approaches to 'managing' it. Sometimes, 'managing' means making no intervention at all. Instead, we must simply provide supporting systems, and 'Let it Be'.

**The I-Space**

In parallel with Snowden's work, Prof. Max Boisot, Professor of Strategic Management at the Birmingham Business School in the UK, proposed a model of information diffusion he calls the 'I-Space'. It is regarded as a seminal work in the realm of corporate anthropology (Boisot et al., 2007).
In the I-Space, the efficiency and effectiveness of transfer of information or knowledge is characterized along three key vectors, Concrete-Abstract, Uncodified-Codified, and Undiffused-Diffused. Concrete knowledge is purely experiential, unrefined, and contaminated with 'noise', but may be 'cleansed' and condensed to the abstract. Uncodified knowledge has little structure and requires effort to extract useful knowledge; codification makes it far more accessible.
Abstract, codified knowledge is most easily diffused and converted into organizational value, but abstraction and codification dilute semantics and may prove self-defeating in some cases. Differing degrees of uncertainty and risk therefore accompany certain types of knowledge, and this can have profound effects not only on how people use the different types, but also on the type of people motivated to use them.
The I-Space manifests itself in numerous ways, one which Boisot himself has studied being management culture. Boisot proposes four domains - POSSIBLE, PLAUSIBLE, PROBABLE and ACTUAL - which managers can inhabit in different proportions. As probability of outcome increases when one traverses from the POSSIBLE to the ACTUAL, so risk decreases. Different management styles crystallize, Boisot says, into two distinct types, the managerial and the entrepreneurial, each with its own mode of applying knowledge. It is difficult not to see the parallel between Snowden's four Cynefin domains and Boisot's.
The Snowden and Boisot frameworks can be seen as two similarly-structured sensemaking spaces addressing different, although related sets of referents. At their point of overlap, they share three interacting concepts: organizations, knowledge and people.If the Snowden and Boisot frameworks may be used to understand the processes by which knowledge acquires significance and the processes by which significant knowledge is made accessible, then an understanding of how such knowledge is acquired and exchanged would add a further dimension to any knowledge architecture.

The developing practice of Social Network Analysis (SNA) offers much in this respect. SNA uses a variety of methods to identify the nodes (people) active in a knowledge-sharing network, and to characterize their contributions in terms of various roles and types of 'ties'. For instance, some may act as gateways for knowledge flow between one cluster of nodes and another cluster; others may act as hubs, to which much of a cluster's knowledge gravitates.

## Conclusions

Knowledge Architecture then, has not only to integrate a multiplicity of forms and formats of information - documents, emails, instant messages, blogs, wikis - encodings like HTML, XML and numerous proprietary encodings from Microsoft, Open Office, and tens of others, but also has to cope with data representation standards, semi-structured and quasi-managed collections (blogs and wikis) and the special needs of specific points-of-view onto the organizational knowledge resource. Knowledge Architecture also has to engage with epistemology and account for what knowledge is, how we acquire and share it and why it is important.

Recognizing that organizations are complex adaptive systems, employing sensemaking techniques to understand how useful activity is sustained by knowledge and information, and identifying the social networks through which much of it flows, will become essential components of the knowledge architect's toolkit alongside the conventional tools of information management.

## References

Bacon 1597. Bacon, Sir Francis. Meditationes Sacræ. De Hæresibus. (1597)

Boisot et al., 2007. Boisot, Max H., MacMillan, Ian and Seok Han, Keyong. Explorations in Information Space: Knowledge, Agents, and Organizations, 2007. Oxford University Press

Ceeney, 2009. Ceeney, Natalie. Integrating Information Management into Business Processes: Project outcomes. The National Archives, undated. Available at: http://www.nationalarchives.gov.uk/electronicrecords/records-tools.htm

Cynefin Framework, 24 August 2009. Available at: http://en.wikipedia.org/wiki/Cynefin

IDC/EMC, 2008. The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide of Information Growth Through 2011. IDC/EMC, March 2008.

Information Week Survey, 2008.

Information Week, March 3, 2008.

ISO 3166, 2006. ISO 3166-1:2006. Codes for the representation of names of countries and their subdivisions - Part 1: Country codes. International Standards Organisation.

Knowledge Council, 2008. The Knowledge Council. Information matters: building government's capability in managing knowledge and information. Government Knowledge & Information Management Network, November 2008.

Rosenfeld, 2003. Available at: http://www.louisrosenfeld.com/home/bloug_archive/images/031019.pdf

Shirky, 2006. Ontology is Overrated: Categories, Links, and Tags. Available at: http://www.shirky.com/writings/ontology_overrated.html

# CLARIN: Where We Stand and Where We Need *Your* Input

Steven Krauwer
Utrecht institute of Linguistics UiL-OTS
Trans 10, 3512 JK Utrecht, Netherlands
s.krauwer@uu.nl

**Summary**

*In this paper we give a brief overview of what the CLARIN Research Infrastructure is, and where we stand in the process towards its construction. We present the current CLARIN position on a number of issues and we identify a number of challenges, highlighting the ones where we feel more input from our communities, both users and providers, is still needed.*

**Key words:** research infrastructures, language resources and tools, humanities, social sciences

## What is CLARIN

CLARIN is the short name for Common Language Resources and Technology Infrastructure. This is one of the proposed research infrastructures for Europe that have been selected by the ESFRI Roadmap process as a candidate for inclusion in the European research infrastructure landscape [6].

The objective is to create a European federation of digital archives containing language-based material (e.g. text and speech corpora, dictionaries, language descriptions, multimodal data, etc, etc) and tools. This federation should provide our target audience, which consists of scholars in the humanities and social sciences, easy access to data and tools, independent of location. In addition to this CLARIN will also provide access to language and speech technology tools through web services, so that –ideally- all these tools can operate on all the data types they were designed for, irrespective of location and origin of data and tools.

In the CLARIN philosophy all languages spoken or studied in the participating countries are equally important, irrespective of size or commercial potential. The CLARIN infrastructure should eventually cover all EU and associated countries. More information can be found on our website [1].

To illustrate what CLARIN wants to achieve we give some examples of what the researcher should be able to ask:

- give me digital copies of all contemporary documents that discuss the Great Plague of England (1348-1350)

- give me all negative remarks about Islam or about soccer in the 2008 proceedings of the European Parliament
- find TV interviews that involve German speakers with a Spanish accent
- summarize all articles in Le Figaro of August 2009 about Mr. Barroso – in Polish

Some of these examples may still sound futuristic, but others can already be realized on the basis of existing technology.

## Who are CLARIN

At this moment the CLARIN infrastructure as such doesn’t exist yet, but there is already a large and active community working on its design and construction. With financial support from the EC (grant EC-FP7-212230, under the FP7 Capacities programme) a consortium of 33 partners in 23 EU and associated countries is now working on the CLARIN Preparatory Phase Project, aimed at defining the future infrastructure and lining up the funding agencies and other stakeholders in the participating countries. In addition to that over 140 other institutions from 32 countries are actively involved in laying the foundations for the infrastructure (see [2] for a full list). The large majority of the participants are academic institutions or data repositories. Contributions from these participants to the project typically consist of data, technology and expertise.

In the Balkans region we have two participants in Croatia (University of Zagreb as a consortium partner [contact is Marko Tadic], and the Institute of Croatian Language and Linguistics as a member [contact is Damir Cavar]), and we have 5 participants in Bulgaria, 1 in Greece, 6 in Romania, 1 in Serbia, 2 in Slovenia, and 2 in Turkey.

## What is the time schedule

When I spoke at the INFuture conference in 2007 we had just been informed by the EC that our Preparatory Phase Project had been approved and that we could go ahead early 2008. On January 1st 2008 the preparatory phase started, and it will last until the end of 2010. The EC contribution to this phase is 4.1 M€. Governments from participating countries have contributed ca 14 M€ to this phase or to parallel, related projects. Contributions range from 50 K€ to 5 M€. If all goes well the next phase, the Construction Phase, will start on Jan 1st 2011. This phase will last until 2014 and will be used to build the infrastructure. Contrary to the construction of physical infrastructures such as space telescopes or particle accelerators we anticipate that the construction of CLARIN will be an evolutionary process, so that we expect to be able to deliver (initially limited) services already after the first year. We expect the cost of this phase to be ca 100 M€. Funding for this phase will have to come from the national governments in the participating countries. This looks like an awful lot of money but it has to be kept in mind that (i) this is not all ‘new’ money because many of the

existing operations by digital archives and language technology centers can be included in CLARIN at no or little extra cost, and that (ii) with 23 countries and a duration of 3 years the actual average cost per country per year will be less than 1.5 M€ (and even less per language). The Exploitation Phase, when the infrastructure will be fully up and running, is expected to start at the latest in 2015, but as we said before, we expect initial services to start much earlier than that. As the digital world is very dynamic we expect that even during the Exploitation Phase CLARIN will go through a continuous evolution process and take up new technologies as they emerge. At this moment we estimate that the total cost until 2018 will be 146 M€, but more precise estimations can be made towards the end of the current Preparatory Phase.
As part of our design activities we are now in the process of building a small experimental prototype, but this is not intended for end users. It will help us to check soundness and consistency of our design.
To what extent and by when users will be able to start using the services will mainly depend on what is going to happen in the various countries: every country is responsible for its own content and its own language. As some countries need more time for their internal preparations we expect some countries to start later than others.

## Main challenges

We see a number of challenges ahead of us. We will discuss a number of them, not just because we want to tell you about what we are doing but rather as an invitation to join our discussions aimed at taking away the main obstacles.

### Technical challenges

Interconnecting existing archives in such a way that the researcher who visits the archive will move seamlessly from one archive to the other and will be able to create virtual collections of documents without being bothered by differences in the ways different archives encode and describe their data is by no means an easy task. This type of task is not unique to CLARIN. In many countries libraries have interconnected their catalogues in similar ways and the customer doesn't see more than one catalogue. The technical problems that have to be solved in order to make this happen are far from trivial, but solvable. The real problem is the lack of standards in the way creators of digital data and owners of archives encode and describe their material.
What makes CLARIN unique is the service infrastructure that will allow researchers to use existing technology tools to retrieve, explore, analyse, exploit or transform their data. The challenge for CLARIN is not to design or build the tools. This happens in other programmes, aimed at technology development. The real challenge is to make tools work on data collections different from the ones they were designed for, and to ensure that they can be chained together to perform more complex operations. Here too the real problems are not techno-

logical in nature, but follow from the fact that every tool builder has his own ideas about the input on which it should operate and about the format of the output. The only possible way to address this problem is through better standards: if tool developers can agree on using a limited number of agreed upon standards the tools and the data can be combined just like Lego bricks. To get a quick overview of the CLARIN position on a number of technological (and linguistic) issues you can check our Short Guides [7].
Standards cannot be dictated or imposed: they can only be effective if they are based on best practice and widely supported by the user community. In CLARIN we have organized the discussions about standards in such a way that the whole community can participate, and through this paper we would like to strongly encourage the members of our community to join the discussion, if only to ensure that the standards that will eventually be supported by CLARIN are suitable for YOUR language and for YOUR research purposes. See our standardization action plan on [5] and join the discussions!

**Linguistic challenges**
CLARIN has the ambition to cover all languages relevant for the European research community, irrespective of market potential, status, or number of speakers. With the limited budget available, and the relatively small consortium (33 partners in 32 countries) there is no way we can ensure complete and adequate coverage for all languages. Broad consultation with the community at large outside the project consortium is necessary to make sure that the approach we adopt will fit all the languages that need to be covered.
We are also in the process of collecting as much information as we can about available language resources and technology and storing this information in our language resources technology registry, so that the existence of material (data collections and tools) can easily be discovered by researchers looking for facilities to support their research. The current registry is now accessible through our portal, the Virtual Language World [3].
Invitation: We invite and encourage the whole community to participate actively in this work, so that we can be sure that YOUR language, research interests and resources and tools will be covered by CLARIN.

**Take-up**
CLARIN’s target audiences are humanities and social sciences (HSS) scholars without technical background. Unfortunately in most of the HSS sub-disciplines there is very little tradition in making use of technological tools. There seems to be a wide gap between ‘the converted’ (as reflected by e.g. the annual Digital Humanities conferences organized by the ALLC [8], and the traditional pencil-and-paper scholars. Like in the case of standards the use of digital techniques can not and should not be imposed on people. It is our task to discover their potential needs and to make them aware of the possible benefits. This is mis-

sionary work that should be carried out on the work floor. Our invitation to you: talk to your analogue colleagues and show them the delights of going digital – and tell us about their needs.

**Legal and ethical issues**

Intellectual property rights (IPR) constitute a very hard problem, especially in the case of language-based material. We see three main issues here.

First of all we can observe that IPR legislation within Europe is far from uniform, which means that access to material –even if technically without any obstacles- may be constrained by the legislation of the country where data or researcher may be located. This is very hard to reconcile with the concept of the European Research Area, where there should be no obstacles for researchers and knowledge to move around. This legislation should be harmonized.

Secondly the CLARIN position is that data and other digital resources created with public funding (regional, national, or EU) should be completely free for other researchers to use (taking into account ethical considerations and a reasonable protection of the creator's interests).

The third problem is special for CLARIN: 're-purposed data'. This refers to data created for other purposes than research (e.g. novels made to entertain, newspapers and TV news made to inform, or recordings of telephone dialogues intended to communicate). There is a wealth of such material available, but very few creators of such data are willing to share it with the research community because they are afraid that giving it away might do damage to their interests. The CLARIN position is that legislation at the EU and/or national level should be adapted to ensure that such data can be used for research purposes and that at the same time the legitimate interests of the owners are sufficiently protected.

Within CLARIN we hope to be able to set up a light but secure licensing system, based on a small number of templates that should cover most of the cases, based on current best practice.

We urge all members of our community to communicate these messages to their legislative bodies at national and EU level and to participate in the modeling of the templates in order to ensure that these templates cover their needs.

**Business models**

Building and maintaining an infrastructure such as CLARIN costs money. Where should the money come from? The EU position is simple: research infrastructures are the responsibility of the national governments. The EU has contributed to the CLARIN Preparatory Phase, although even there the full responsibility to ensure that the design of the CLARIN infrastructure covers national needs lies with the national governments. If, e.g. Croatia or Slovenia or any other country do not participate actively (and at their own expenses) in the CLARIN design process no one will take care of the interests of their languages and their research communities. For the next phase, the so-called Construction

Phase no EC contribution is foreseen. It is not until the final phase, the Exploitation Phase, that a possible contribution from the EC (the running figure is up to 20% of the operational costs) is being discussed as an option for the 8th Framework Programme.
But apart from these global financing aspects there are other financial questions to be considered: who pays what to whom for what.
Expectations depend on your role in life (guess who says what):

- Everything should be available for free
- I want to be reimbursed for the extra effort to make my data and tools accessible through CLARIN
- I don’t want others to use my results to make a profit
- Funders should not pay for the creation of tools and data that can be bought on the market
- Funding infrastructures is primarily a national responsibility
- We fund you for now but we expect you to become self-sustaining in the future
- Creation of data and tools is the responsibility of the infrastructure

The current CLARIN position is the following:

- CLARIN is not a creator of digital data or technological tools: its role is to facilitate sharing and interconnecting existing and future resources and tools
- The creation and operation of the local (national) part of the infrastructure will be fully financed by the national authorities
- The overall coordination and management of the infrastructure are a joint responsibility (also financially) of the countries participating in the ERIC
- Standard use of the whole infrastructure for research purposes should be free to all research institutions in countries that have joined the ERIC
- For special services (through third parties) the extra cost may be charged to the user
- Research institutions from sites outside ERIC countries, or researchers with commercial objectives may be charged on a subscription or case by case basis

Given our commitment to the principle that results of public funding should be freely available to the research community we do not envisage facilities for e.g. university institutes to generate extra income through CLARIN by charging users for their services.

**Shape**

We see the future CLARIN infrastructure as a networked structure with one or more centers in most participating countries. These centers may be data centers or service centers (both with guaranteed 24/7 availability, guaranteed for a longer period of time), centers of expertise, and other centers (more loosely

connected to the infrastructure), as well as a small office to accommodate the organizational headquarters. We anticipate that all centers will be based on existing centers, which may have to extend their scope of activities or their service level. We do not anticipate any major investments in physical installations or network facilities.
In our documents we have specified the requirements for becoming a CLARIN center. If you are interested in hosting such a center we recommend that you first of all talk to your national CLARIN coordinator, and then read our documents about center types (see [4]).

**Governance**
The main challenge here is to find a legal form that allows 23 (or possible even more) countries to jointly build and operate an infrastructure distributed over all these countries, and to jointly fund and operate this infrastructure in a sustainable way (i.e. based on a firm long term commitment rather than on ad hoc grants for a couple of years).
The current CLARIN position is that we aim at the creation of an ERIC (European Research Infrastructure Consortium), which is a new type of international legal entity, created by the EC for the specific purpose of operating research infrastructures. Participants in such a consortium are governments (i.e. not research institutions), because only governments can make long-term commitments.
CLARIN will approach the relevant ministries in the participating countries with more details about this in the coming months.

## Sharing

What can be shared through CLARIN? The answer is simple: you can use CLAROIN to share anything that (i) might be relevant for our target user community, that (ii) satisfies certain quality criteria, and that (iii) you are legally allowed to share. This can apply to data (raw or enriched), tools, programs, expertise, etc.
Why would one want to share at all? For researchers there are a number of possible motives: idealism, hope for fame, hope that others will share with you or simply because your funder tells you to share. For the funder the obvious motive to insist on sharing is of course that it will ensure a better return on their (i.e. the tax payers') investment.
Why would one not want to share? One reason might be that it may involve an extra effort (adapting it to representation or interoperability standards, creating metadata, writing documentation). Another might be that it is possible that others do brilliant things with your material, which you had never thought of doing. Or others might criticize it because it isn't as good as it should have been.
The CLARIN position here is that the key to this issue is in the hand of the public funding bodies: they should make sharing through CLARIN (including

the extra efforts it might require) a contractual obligation in their funding contract.

How can you share through CLARIN? The best way to do it is to register and deposit your material at one of the CLARIN centers. These centers will be especially equipped to handle your material, keep it accessible in a sustainable way on a 24/7 basis, and handle all the licensing that should ensure that only authorized people can put their hands on it.

Alternatively you can make it available in a more traditional way (e.g. through your university’s web servers), but it should be noted that this situation is far from ideal as universities tend to be extremely unstable and unreliable in this respect: they change URLs whenever they have contracted a new company to create a new web presence, and there is no guarantee that you material will remain accessible after your project is over or when you leave for another job or for retirement. Your resources are definitely better off in the hands of specialized digital repositories.

## Concluding remarks

I have referred to a number of CLARIN documents (see links below), and you may wonder how you can get access to them. If your organization is a CLARIN member you can apply for an account on the member space of our website [9], which will give you access to all the documents produced and discussed in our project working groups. If they are not a member and if they qualify for membership you should ask them to join [2].

In this paper I have tried to give you a more or less up-to-date picture of where CLARIN stands, but it is important to stress that CLARIN doesn’t exist yet and is still full of challenges that need to be addressed and for which we urgently need your input, especially if you want to make sure that your own needs and requirements are taken into account. Be selfish and join CLARIN!

We urgently need better connections with the humanities and social sciences communities at large, because we don’t want CLARIN to be based on the needs and priorities of the language and speech technology community alone. Here too we need your help!

Finally, to avoid any misunderstandings or false expectations: CLARIN is not about content creation (which should be the responsibility of other national or EU programmes) but about sharing, and it aims at doing this by providing service-based access to what exists and to what will exist in the future.

## Links

[1] The CLARIN website. http://www.clarin.eu (16.10.2009)
[2] List of CLARIN members and how to join. http://www.clarin.eu/members (16.10.2009)
[3] CLARIN’s Virtual Language World portal. http://www.clarin.eu/vlw (16.10.2009)
[4] CLARIN centers. http://www.clarin.eu/files/wg2-1-centers-doc-v8.pdf (16.10.2009)

[5] Standardisation action plan.
http://www.clarin.eu/system/files/private/Standardisation%20action%20plan-v8.pdf
(16.10.2009)
[6] ESFRI Roadmap. http://cordis.europa.eu/esfri/roadmap.htm (16.10.2009)
[7] CLARIN Short Guides. http://www.clarin.eu/documents/short-guides (16.10.2009)
[8] ALLC (Digital Humanities Conferences). http://www.allc.org (16.10.2009)
[9] Join a Working Group. http://www.clarin.eu/join-a-working-group (16.10.2009)

# A Progress Report on Bitext Parsing

Alexander Fraser
Institute for Natural Language Processing, University of Stuttgart
E-mail: fraser@ims.uni-stuttgart.de

Renjing Wang
Institute for Natural Language Processing, University of Stuttgart
E-mail: wangrg@ims.uni-stuttgart.de

Hinrich Schütze
Institute for Natural Language Processing, University of Stuttgart
Azenbergstrasse 11, 70180 Stuttgart, Germany

**Summary**

*Recent work has shown that a reranking approach can be used to improve the syntactic parsing of a sentence given a translation of that sentence, an automatically generated parse of that translation, and a word alignment between them. Such approaches rely on reducing syntactic divergence as measured using overlapping feature functions capturing different types of divergence. These feature functions are combined in a log-linear model which is trained to maximize parsing accuracy.*
*We conduct our research in the framework of N-best parse reranking. However, we apply reranking to bitext and add only features based on syntactic projection from German to English. The system takes as input (i) English sentences with a list of automatically generated syntactic parses, (ii) a translation of the English sentences into German, (iii) an automatically generated parse of the German translation, and (iv) an automatically generated word alignment between the original sentences and the translations. The system is trained using the gold standard trees of 3718 sentences from the Penn English treebank that have been translated into German. We achieve an improvement in F1 on held out test data and this improvement is statistically significant.*

**Key words:** syntactic parsing, multilinguality, treebanks, machine translation, annotation projection

## Introduction

Recent work [1,2] has shown that a reranking approach can be used to improve the syntactic parsing of a sentence given a translation of that sentence, an automatically generated parse of that translation, and a word alignment between

them. Such approaches rely on reducing syntactic divergence as measured using overlapping feature functions capturing different types of divergence. These feature functions are combined in a log-linear model which is trained to maximize parsing accuracy.

In this work we extend the approach of Fraser, Wang and Schuetze [1]. We view this work as part of a research program aimed at finding alternative sources of supervision for syntactic parsing which can augment small and expensive to create syntactic treebanks. We analyze the gains in parsing accuracy obtained by this approach and provide examples, with a special focus on features which are estimated using the baseline parser on 1.4 million parallel sentences from the Europarl corpus (for which we do not have gold standard parses). This is similar to the self-training approach of McClosky, Charniak and Johnson [5]. We then augment the approach with two new feature functions which capture difficult prepositional phrase attachment phenomena, resulting in a further gain in performance as measured through cross-validation on sentences taken from the Penn Treebank. Finally, we discuss applying the system to the Europarl corpus and discuss possible improvements.

We conduct our research in the framework of N-best parse reranking (following Collins [6], but see also Riezler et. al. [7]). However, we apply reranking to bitext and add only features based on syntactic projection from German to English. The system takes as input:

1. English sentences with a list of automatically generated syntactic parses
2. A translation of the English sentences into German
3. An automatically generated parse of the German translation
4. An automatically generated word alignment between the original sentences and the translations

The system is trained using the gold standard trees of 3718 sentences from the Penn English treebank that have been translated into German. We achieve an improvement in F1 on held out test data (measured by using cross-validation) and this improvement is statistically significant.

## Bitext Parsing

As a motivating example consider the English sentence “He saw a baby and a woman who had gray hair”. Suppose that the baseline parser generates two parses, one where it is attached high (to both of the NPs), and one where “who had gray hair” is attached only to the woman. Suppose further, that the second parse is the correct parse in this context. How can we determine that the second parse should be favored? Since we are parsing bitext, we can observe the German translation which is “Er sah ein Baby und eine Frau, die graue Haare hatte” (glossed: “he saw a baby and a woman, who gray hair had”). The singular verb in the subordinate clause (“hatte”: “had”) indicates that the subordinate S must be attached low to “woman” (“Frau”) because the subject is singular. In cor-

rectly resolving the attachment ambiguity, we are using the human translator's disambiguation of the English syntax (performed while translating to German).
To accomplish this automatically, we follow Collins' approach [6] to discriminative reranking. The approach begins with a generative model which models the joint generation of a sentence and its parse tree and then reranks the 100-best hypothesized parses. Given a new sentence to parse, we first select the best N parse trees according to a generative model. Then we use new features to learn discriminatively how to rerank the parses in this N-best list. We use features derived using projections of the 1-best German parse onto the hypothesized English parse under consideration. Because our features are based on bilingual projection, they are complementary to the features used in previous parse reranking work.
In more detail, we take the 100 best English parses from the BitPar parser [8] and rerank them. We have a good chance of finding the optimal parse among the 100-best hypothesized parses. An automatically generated word alignment determines translational correspondence between German and English.
We use features which measure *syntactic divergence* between the German and English trees to try to rank the English trees which have less divergence higher.
Our test set is 3718 sentences from the English Penn treebank which were translated into German. We hold out these sentences, and train BitPar on the remaining Penn treebank training sentences. The average F1 parsing accuracy of BitPar on this test set is 87.89%, which is our baseline. The test set is very challenging, containing English sentences of up to 99 tokens.
We implement features based on projecting the German parse to each of the English 100-best parses in turn via the word alignment. All parses and the word alignment are generated automatically.
By performing cross-validation and measuring test performance within each fold, we compare our new system with the baseline on the 3718 sentence set. The overall test accuracy we reach is 88.59%, a statistically significant improvement over baseline of 0.70.
Given a word alignment of the bitext, the system performs the following steps for each English sentence to be parsed:

1. Run BitPar trained on English to generate 100-best parses for the English sentence
2. Run BitPar trained on German to generate the 1-best parse for the German sentence
3. Calculate feature function values which measure different kinds of syntactic divergence
4. Apply a model that combines the feature function values to score each of the 100-best parses
5. Pick the best parse according to the model

## Related Work

The most directly related work is that of Burkett and Klein [2], which is work that was published after [1] was submitted for publication. Similarly to our previous work, they used feature functions defined on triples of (English parse tree, Chinese parse tree, alignment) which were combined in a log-linear model. To train this model they used a small parallel treebank which contains gold standard trees for parallel sentences in Chinese and English, while we only required access to gold standard trees for the English side of our training corpus in order to improve English parse quality. They defined similar features to the coarse features defined in our previous work and trained a system which improves first the Chinese parse and then the English parse and iterates. In addition they try experiments allowing the alignment to vary, but these experiments are inconclusive. Our additional features go beyond the coarse syntactic divergence features in their work to address specific problems we observed through error analysis, and to incorporate self-training features. Two other interesting works in this area are those of Fossum and Knight [3]; and of Huang, Jiang and Liu [4]. They improve English prepositional phrase attachment using features from a Chinese sentence. However, unlike our approach, they do not require a Chinese syntactic parse as the word order in Chinese is sufficient to unambiguously determine the correct attachment point of the prepositional phrase in the English sentence without using a Chinese syntactic parse.

## Model

We define feature functions which measure syntactic divergence. We use a model combining feature functions in a linear fashion, a log-linear model, to choose the best English parse (see the first equation below). The feature functions **h** are functions on the hypothesized English parse **e**, the German parse **g**, and the word alignment **a**, and they assign a score (varying between 0 and infinity) that measures *syntactic divergence*.

The alignment of a sentence pair is a function that, for each English word, returns a set of German words that the English word is aligned with. Feature function values are calculated either by taking the negative log of a probability, or by using a heuristic function which scales in a similar fashion (for example, a probability of 1 is a feature value of 0, while a low probability is a feature value which is a large magnitude positive number. Note also that we define the value of log 0 to be –infinity for the purposes of this work, though in practice we do not work with probabilities of 0).

Given a vector of weights $\lambda$, the best English parse $\hat{e}$ can be found by solving the second equation below. The model is trained by finding the weight vector $\lambda$ which maximizes accuracy. This is done by reranking the output of the generative model for a set of sentences for which we have gold standard parses. Training is discussed in detail later in the paper.

$$p_\lambda(e|g,a) = \frac{exp(-\sum_i \lambda_i h_i(e,g,a))}{\sum_{e'} exp(-\sum_i \lambda_i h_i(e',g,a))}$$

$$\begin{aligned}\hat{e} &= \underset{e}{\operatorname{argmax}}\, p_\lambda(e|g,a) \\ &= \underset{e}{\operatorname{argmin}}\, exp(\sum_i \lambda_i h_i(e,g,a))\end{aligned}$$

## Training

Log-linear models are often trained using the Maximum Entropy criterion, but we train our model directly to maximize F1. We score F1 by comparing hypothesized parses for the discriminative training set with the gold standard. To try to find the optimal $\lambda$ vector, we perform direct accuracy maximization, meaning that we search for the $\lambda$ vector which directly optimizes F1 on the training set, using the algorithm of [10]. See [1] for further details.

## Feature Functions

We first briefly describe the feature functions we found useful in our previous work, and then introduce two new feature functions which we defined after an error analysis. The basic idea behind our feature functions is that any constituent in a sentence should play approximately the same syntactic role and have a similar span as the corresponding constituent in a translation. If there is an obvious disagreement, it is probably caused by wrong attachment or other syntactic mistakes in parsing. Sometimes in translation the syntactic role of a given semantic constituent changes; we assume that our model penalizes all hypothesized parses equally in this case.

For the initial experiments, we used a set of 34 probabilistic and heuristic feature functions, but we do not have space to briefly describe all 34 features.

**BitPar LogProb** (the only monolingual feature) is the negative log probability assigned by BitPar to the English parse. This feature is important, as it encodes the monolingually derived knowledge which is inherent in BitPar's model. The rest of the feature functions are bilingual and encode additional sources of knowledge derived from the parse of the German translation.

### Count Feature Functions

We now introduce feature functions which *count* projection constraint violations.

Feature **CrdBin** counts binary events involving the heads of coordinated phrases. If in the English parse we have a coordination where the English CC is aligned only with a German KON, and both have two siblings, then the value contributed to **CrdBin** is 1 (indicating a constraint violation) unless the head of

the English left conjunct is aligned with the head of the German left conjunct and likewise the right conjuncts are aligned.
Feature Q simply captures a mismatch between questions and statements. If an English sentence is parsed as a question but the parallel German sentence is not, or vice versa, the feature value is 1; otherwise the value is 0.

**Span Projection Feature Functions**
Span projection features calculate the percentage difference between a constituent's span and the span of its projection. Span size is measured in characters or words. To project a constituent in a parse, we use the word alignment to project all word positions covered by the constituent and then look for the smallest covering constituent in the parse of the parallel sentence.
**CrdPrj** is a feature that measures the divergence in the size of coordination constituents and their projections. If we have a constituent (XP1 CC XP2) in English that is projected to a German coordination, we expect the English and German left conjuncts to span a similar percentage of their respective sentences, as should the right conjuncts. The feature computes a character-based percentage difference.
**POSParentPrj** is based on computing the span difference between all the parent constituents of POS tags in a German parse and their respective coverage in the corresponding hypothesized parse. The feature value is the sum of all the differences. The projection direction is from German to English, and the feature computes a percentage difference which is character-based.
**AbovePOSPrj** is similar to **POSParentPrj**, but it is word-based and the projection direction is from English to German. Unlike **POSParentPrj** the feature value is calculated over all constituents above the POS level in the English tree.
Another span projection feature function is **DTNNPrj**, which projects English constituents of the form (NP(DT)(NN)). The feature computes a percentage difference which is word-based. It is designed to disprefer parses where constituents starting with “DT NN”, e.g., (NP (DT NN NN NN)), are incorrectly split into two NPs, e.g., (NP (DT NN)) and (NP (NN NN)). This feature fires in this case, and projects the (NP (DT NN)) into German. If the German projection is a surprisingly large number of words (as should be the case if the German also consists of a determiner followed by several nouns) then the penalty paid by this feature is large. This feature is important as (NP (DT NN)) is a very common construction.

**Probabilistic Feature Functions**
We use Europarl corpus of Koehn [9], from which we extract a parallel corpus of approximately 1.22 million sentence pairs, to estimate the probabilistic feature functions described in this section.
For the **PDepth** feature, we estimate English parse depth probability conditioned on German parse depth from Europarl by calculating a simple probability

distribution over the 1-best parse pairs for each parallel sentence. A very deep German parse is unlikely to correspond to a flat English parse and we can penalize such a parse using **PDepth**.

The feature **PTagEParentGPOSGParent** measures tagging inconsistency based on estimating the probability that for an English word at position i, the parent of its POS tag has a particular label. Consider (S(NP(NN fruit))(VP(V flies))) and (NP(NN fruit)(NNS flies)) with the translation (NP(NNS Fruchtfliegen)). Assume that "fruit" and "flies" are aligned with the German compound noun "Fruchtfliegen". In the incorrect English parse the parent of the POS of "fruit" is NP and the parent of the POS of "flies" is VP, while in the correct parse the parent of the POS of "fruit" is NP and the parent of the POS of "flies" is NP. In the German parse the compound noun is POS-tagged as an NNS and the parent is an NP. The probabilities considered for the two English parses are p(NP|NNS, NP) for "fruit" in both parses, p(VP|NNS, NP) for "flies" in the incorrect parse, and p(NP|NNS, NP) for "flies" in the correct parse. A German NNS in an NP has a higher probability of being aligned with a word in an English NP than with a word in an English VP, so the second parse will be preferred. As with the **PDepth** feature, we use relative frequency to estimate this feature.

Note that when an English word is aligned with two words, estimation is more complex. We heuristically give each English and German pair in the alignment unit one count. The value calculated by the feature function also works differently. If an English word is aligned with multiple German words, we use the geometric mean of the pairwise probabilities (i.e., each English word has the same overall weight regardless of whether it was aligned with one or with more German words).

**Other Features**

Our best system uses the nine features we have described in detail so far. In addition, we implemented 25 other features, which did not appear to improve performance, as we showed using a feature analysis in our previous work, see [1] for further details.

**New Feature Functions**

After conducting an error analysis of our system, we noticed that it had systematic failures in PP attachment which occurred higher in the tree than our previous feature functions had addressed. We define two new feature functions below, where we make use of the node numbering we introduced in [1] (briefly, all nodes in the tree including POS tags are assigned a unique integer; by convention **i** refers to a node in the English tree, and **j** refers to a node in the German tree). Recall also that higher values indicate penalized behaviour (these values scale like negative log probabilities).

The first feature **PPinNPPP c**hecks whether a PP inside of a NP or PP in German attaches to the same (projected) constituent in English.

```
For each German node j
  if j is PP and parent(i) is NP or PP
    let j' be the nearest sibling to the left of j that is a NN, NP or PP
    if j' is defined
      let English node i = project(j)
      let English node i' = project(j')
      value += 1 if i' is not a sibling of i
                    or i' not the nearest sibling to the left of i that is a NN, NP or PP
```

**EngPPinSVP** checks whether a PP inside of a S or VP in English attaches to the same (projected) constituent in German (note in the feature definition the attachment in the German can be to the left or to the right).

```
For each English node i
  if i is PP and parent(i) is S or VP
    let i' be nearest sibling to the left of i that is a POS(V*) or VP
    if i' is defined
      let German node j = project(i)
      let German node j' = project(i')
      value += 1 if j' is not POS(V*) or VP, or j' is not a sibling of j
```

## Experiments

We used the subset of the Wall Street Journal which consists of all sentences that have at least one prepositional phrase attachment ambiguity for our experiments. An example of such an ambiguity is (VP bring (NP attention) (PP to the problem)) vs. (VP bring ((NP attention) (PP to the problem))). The first 500 sentences of this set were translated from English to German by a graduate student and an additional 3218 sentences by a translation bureau. We withheld these 3718 English sentences (and an additional 1000 reserved sentences) when we trained BitPar on the Penn treebank.

### Parses

We use the BitPar parser [8] which is based on a bit-vector implementation of the Cocke-Younger-Kasami (CKY) algorithm. It computes a compact parse forest for all possible analyses. BitPar is particularly useful for N-best parsing as the N-best parses can be computed efficiently.

For the 3718 sentences in the translated set, we created 100-best English parses and 1-best German parses. The German parser was trained on the TIGER treebank. For the Europarl corpus, we created 1-best parses for both languages.

**Word Alignment**

We use a word alignment of the translated sentences from the Penn treebank, as well as a word alignment of the Europarl corpus. We align these two data sets together with data from the JRC Acquis to try to obtain better quality alignments. To generate word alignments, we used IBM Model 4 [13], as implemented in GIZA++ [11]. As is standard practice, we trained Model 4 with English as the source language, and then trained Model 4 with German as the source language, resulting in two Viterbi alignments. These were combined using the *Grow Diag Final And* symmetrization heuristic [12].

**Experimental Analysis**

In [1] we reached the best performance by performing a greedy feature selection. We started with a $\lambda$ vector that is zero for all features, and then ran the error minimization (without random generation of $\lambda$ vectors, which makes the algorithm deterministic). One feature at a time is added. This greedy algorithm produced a vector with many zero weights, with a good fit to the training set resulting in a 0.93 improvement on the training set. The resulting performance of 0.66 on the test set over the baseline was our best result.

When we repeated this process using all of the features used in our previous work together with our two new features, we obtained an improvement of only 0.80 on the training set. On the test set, the improvement was only 0.55 over the baseline. This shows that with the addition of the 2 new features we are having problems with *search errors* for the $\lambda$ vector which optimizes F1. We know that a vector exists which results in a better fit to the training data, it is simply the vector we had before, with the addition of two zeros added for the weights of our two new feature functions. Presumably, there might be another vector which assigns non-zero weights to our two new feature functions which will result in a further improvement.

We therefore went back to performing 5 trials per fold of our 7-fold cross-validation using the non-deterministic algorithm (these are combined by averaging). This algorithm differs in that it tries one thousand randomly determined $\lambda$ vectors at the beginning of each iteration in an attempt to escape local minima which cannot be escaped from using the one-dimensional search. Using this algorithm our previous result was an improvement of 0.82 on train, and 0.55 on test. With the two new features we obtained an improved fit on train of 0.93 and an improvement on test of 0.70. This shows that the new features are effective. However, due to the search errors with the greedy algorithm we are unable to effectively apply it, even though we found it superior previously.

## Conclusion

In this work we have introduced two new feature functions which improve over the system we presented in [1]. Although we had worse performance with the greedy feature selection algorithm we previously, performance with the non-deterministic algorithm improved by an additional 0.15% F1 resulting in a total improvement of 0.70% F1 over the baseline. We are currently preparing to apply this system to generate improved parses of the entire Europarl corpus, and we hope to obtain additional increases in parse quality through self-training [5] by retraining our baseline parser on the improved parses.

## Acknowledgments

This work was supported in part by Deutsche Forschungsgemeinschaft Grant SFB 732. We thank Helmut Schmid for his comments and for support of BitPar.

## References

[1] Fraser, Alexander; Wang, Renjing; Schütze, Hinrich. 2009. Rich bitext projection features for parse reranking. In EACL.

[2] Burkett, David; Klein, Dan. 2008. Two Languages are Better than One (for Syntactic Parsing). In EMNLP.

[3] Fossum, Victoria; Knight, Kevin. 2008. Using Bilingual Chinese-English Word Alignments to Resolve PP-attachment Ambiguity in English. In AMTA.

[4] Huang, Liang; Jiang, Wenbin; Liu, Qun. 2009. Bilingually-constrained (monolingual) shift-reduce parsing. In EMNLP.

[5] McClosky, David; Charniak, Eugene; Johnson, Mark. 2006. Effective self-training for parsing. In NAACL.

[6] Collins, Michael. 2000. Discriminative Reranking for Natural Language Parsing. In ICML.

[7] Riezler, Stefan; King, Tracy; Kaplan, Ron; Crouch, Dick; Maxwell John; Johnson, Mark. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In ACL.

[8] Schmid, Helmut. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In COLING.

[9] Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In MT Summit.

[10] Och, Franz. 2003. Minimum Error Rate Training in Statistical Machine Translation. In ACL.

[11] Och, Franz; Ney, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29:1, 19-51.

[12] Koehn, Philipp; Och, Franz; Marcu, Daniel. 2003. Statistical Phrase-Based Translation. In HLT-NAACL.

[13] Brown, Peter; Della Pietra, Stephen; Della Pietra, Vincent; Mercer, Robert. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19:2, 263-311.

# DIGITIZATION AND PRESERVATION

# Planning and Designing of Digital Archival Information Systems

Arian Rajh
Agency for Medicinal Products and Medical Devices
Ksaverska cesta 4, 10000 Zagreb, Croatia
arian.rajh@almp.hr

Hrvoje Stančić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
hrvoje.stancic@zg.t-com.hr

**Summary**

*Digital archival information systems can be planned and designed by following the most prominent records management methodologies like ISO 15489, related standards and DIRKS, or by following project management approach. Project management approach means planning and designing of digital archival information system as a cost and time-bound project and performing monitoring and evaluation activities in these perimeters. This approach can be done by following different project management methodologies like PRINCE2 or Project Cycle Management. The aim of this article is to explain these available options, to compare them and to make recommendations based on assessment of capabilities of an organisation in which the potential digital archival information system is to be planned and designed. Project cycle management is suitable methodology for organisations in which digital archive developers do not have sufficient administrative capabilities or enough staff with technical know-how to develop digital archive without outsourced technical assistance, but are allowed to apply for European financial aid, such as IPA or other funds in the future.*

**Key words:** digital archival information system, planning, records management, ISO standards, DIRKS, PRINCE2, PCM, IPA, EU funds

## 1. Introduction

The process of planning and designing of digital archival information systems, with preparatory planning activities, as well as indicated implementation and post-implementation activities, such as evaluation of the system, can be conducted either by following prominent and internationally standardised record-keeping methodologies or by following project approaches. The aim of this article is to show what the available planning and designing options are considering

digital archival information systems. The differences between these two major options, standardised records management methodologies and project management methodologies, will be explained by showing their particular phases and other characteristics, compared and analysed.

The first option for planning and designing of digital archival information system is to use ISO 15489:2001 and ISO/TR 15489:2001 standards (HRN ISO 15489-1, HRN ISO/TR 15489-2). ISO 15489:2001 provides internationally recognised records management methodology for planning and designing of recordkeeping systems. Designing part in the process of planning and designing of digital archival information systems is additionally described in supplemental ISO documents such are ISO 26122:2008(E), ISO 23081-1:2006 (HRN ISO 23081-1) and ISO ISO/TR 23081-2:2007 (HRN ISO 23081-2). ISO 15489 recordkeeping planning and development process is elaborated in detail in DIRKS manual. Together with the reference model for open archival information system (OAIS, ISO 14721:2003) these ISO standards and pertaining DIRKS manual can provide inclusive methodology for planning and designing of digital archives.

ISO 15489 brings out the framework for entire recordkeeping practice in organisations. Although ISO 15489 can not be reduced to planning method, one of the most pragmatic parts of that standard is eight-step methodology used to plan and develop recordkeeping systems. ISO 26122 standard provides further insight in one phase of the ISO 15489 recordkeeping systems planning methodology because it recommends and standardises work process analysis for records. ISO 23081 standard can also help the developer of a digital archival information systems. This standard gives framework for defining and modelling metadata scheme and that is usually one of the most sensitive concerns when designing a recordkeeping systems. DIRKS manual describe each step of the ISO 15489’s eight-step methodology in detail and it is recommended to refer to it in initial phases of establishing a digital archive.

The second option is to use project management approaches or methodologies in digital archives planning and designing stage. Project management is a domain that deals with projects’ planning, execution and management phases. Projects are resource and time-bounded activities aiming to deliver measurable and clearly assessable results. There are numerous specific approaches in the project management but this paper is focusing just on PRINCE2 and Project Cycle Management (PCM) methodologies because of their paradigmatic value. Due to simplicity and comprehensibility of these methodologies they are considered as exemplary.

Planning and developing of a digital archival information system should be done systematically and methodologically. That means that planning and development of digital archives should be done by choosing and following one methodology or several coherent methodologies. Often it is not easy to decide which methodology to choose. The decision on methodology which is to be used for

planning and designing of a digital archive should be based on appropriateness of a specific methodology to the organisational capacities and restrains in which future digital archive will be designed and implemented. Such a decision depends on the level of readiness of an organisation to build or purchase, develop and upgrade, test and evaluate a digital archival system. Understanding what most common methodologies could offer and understanding the level of organisational readiness for activities of establishing a digital archive is a good foundation for deciding on a particular methodology.
An example of planning and designing of a digital archival information system in the Croatian Agency for Medicinal Products and Medical Devices is given later in the text.

## 2. Planning and designing procedures based on ISO 15489, related standards and DIRKS methodology

ISO 15489 standard was prepared by the technical committee ISO/TC 46 Information and documentation. Its scope is records management, recordkeeping systems, recordkeeping policies, and synchronisation of records management with family of ISO quality management standards. ISO 15489 is linked with ISO documents for work process analysis and for records and metadata schemes development.
DIRKS manual was published by the National Archives of Australia in 2001, and it was revised in 2003. Although DIRKS manual in the first place represents expansion of the Australian recordkeeping standard AS 4390-1996 and application of the ISO 15489 standard to Australian archival practice, it can be useful to any archival expert planning to develop a digital archive.
ISO 15489 methods for developing recordkeeping systems consist of eight steps. These steps are not strictly consequent because some of them can be preformed in iteration. ISO methodology includes preliminary investigation, business process analysis, identification of recordkeeping requirements, assessment of existing system(s), identification of recordkeeping strategies, design of recordkeeping system, implementation of the system and post-implementation review and testing. Strategy components for designing recordkeeping system or digital archive under ISO 15489 include system designing, system development, education of the staff, conversion of records for ingest into system, development of recordkeeping policies, benchmarking of the system according to policies that were developed, and development of retention policy and appraisal of the records. DIRKS manual describes given methodology with more detail about primary and feedback paths between particular steps.
After conducting preliminary investigation on the needs and requirements for the new digital archival information system, experts should analyse business activities and use this output to identify specific recordkeeping requirements and to design the recordkeeping system. At the same time, feedback from identification of recordkeeping requirement should enhance business process analy-

sis while feedback from designing of recordkeeping system should improve the identification of recordkeeping requirements. Output of identification of recordkeeping requirements should lead to identification and selection of appropriate recordkeeping strategies. After certain achievements in the design of a recordkeeping system, experts should start developing it and use feedback from the development to improve designing process. When the digital archive is finally implemented, it is necessary to use feedback of post-implementation review to verify significance and quality of identified recordkeeping requirements, and to constantly improve designed and implemented digital archive. Post-implementation feedback is expected to be also used in the future preliminary investigations for upgrading of newly built system or for designing another one.

ISO/TR 26122 (ISO/TR 26122:2008(E) Information and documentation – Work process analysis for records) is expansion of the ISO 15489. Its scope is narrower and it focuses on identification of relations between business processes and their context, business processes and legal requirements, hierarchical decomposition of business processes and sequential dependence of their immanent transactions. This procedure of ISO 26122 is directly related to the analysis of business processes and indirectly it affects identification of recordkeeping requirements and designing of a recordkeeping system because business processes are linked with records and characteristics of records are affecting design of systems.

ISO 23081 (ISO 23081 Information and documentation – Metadata for records, Part 1 and Part 2) represent ISO 15489 expansion related to designing of metadata elements and procedures which should be included in design of the system. Metadata are considered evidence of relationship between system and business context entities. Entities could be: records, business processes, business mandates or business rules, policies and legislation, and agents that conduct business and handle records. Metadata groups are: identity metadata, description metadata, use metadata, relation metadata, and event plan and event history metadata. With description of required metadata and metadata groups, ISO 23081 can provide digital archive developers with a valuable insight from the archival point of view which can facilitate various legal and business compliances of their future digital archival systems.

## 3. Planning and designing procedures based on PRINCE2 and PCM methodologies

A digital archival information system could also be set up through a project approach. This means that establishing of a digital archive must include strictly planned fulfilment of assumptions and preliminary prerequisites, scheduled activities, measurable results and explicated objectives according to one of the project management methodologies. Setting up a digital archive as a project requires very different planning and designing approach in comparison to the record management and recordkeeping methodologies.

PRINCE2 stands for **PR**oject **IN C**ontrolled **E**nvironment and it is process-based methodology for managing projects. PRINCE methodology was firstly developed in the United Kingdom, and is used since 1989 by the Central Computer and Telecommunications Agency. PRINCE2 methodology dates from 1996 as the second major improvement of PRINCE methodology and is a trademark of Office for Government Commerce of HM Treasury. Last revision was issued in 2009.

PRINCE2 manages projects through several phases: (1) start up the project; (2) directing a project; (3) initiating a project; (4) management of stage boundaries – a critical phase for project managers who have to keep their mind on the project scope at all times; (5) controlling a stage; (6) managing project delivery – a process which ensures that all outputs of the project meet previously stated requirements; and (7) project closing. PRINCE2 methodology also includes (8) planning phase which is initiated after start up a project phase, but conducted iteratively and simultaneously with other project phases.[1]

The most important agents defined by PRINCE2 are Project board, Project manager and Team manager. A PRINCE2 project is organised according to customer-supplier relationship model in which customers are the future users of the planned digital archival information system. Various documents are produced in each of these project phases – business case, project initiation document, stage plan, team plans, quality logs, project report, lessons learned report, follow-on actions etc.

PCM (**P**roject **C**ycle **M**anagement methodology) is a different kind of project management approach. PCM has been used globally after it was adopted and popularised by the United States Agency for International Development (USAID) in the late 1960s. In 1993 European Commission started to use the same approach as an obligatory project planning methodology. PCM is also widespread due to its simplicity and transparency. Because of intelligibility of the logical framework diagram, PCM provides non-experts with valuable insight into project proposals to evaluate and approve them with less effort and in shorter time.[2]

PCM is divided in fewer phases than PRINCE2. It consists of five phases: programming, identification, formulation, implementation and evaluation phase. For internationally financed projects, in (1) the programming phase it is usually

[1] "PRINCE2 uses three levels of plans (...) Project Plan, Stage Plan, and Team Plan. In cases where a Stage Plan or Team Plan exceeds predetermined tolerances (time or money) then an Exception Plan can be produced to replace the plan that has exceeded its tolerances." Adam, Azad. Implementing Electronic Document and Record Management Systems. New York: Auerbach Publications, 2007, p. 136. For more detailed explanation of development and implementation of digital archival information systems see Azad, pp. 127-140.

[2] There are several available methodologies useful for making decisions on adoption of project proposals like AHP (Analytic Hierarchy Process). However, PCM remains equally useful for decision-makers and for experts with responsibility to propose a project.

necessary to prepare strategic documents on the highest or higher levels so they could be used as the basis for later assessments on project proposals. Such documents are national and regional strategic plans, multi-annual indicative planning documents for selected future period, national progress reports or similar documents. In (2) the identification phase competent authorities together with the key players in each sector must recognize problems in sectors and perceive potential ideas that are proposed by project planners as solutions to the recognised problems. In (3) the formulation phase experts involved in project planning and designing have the most important role. They have to prepare the project proposal and other project documentation which will be the basis for achieving decision on project's launch or termination (go – no go decision). If the project is approved, formulation stage is followed by (4) the implementation of project activities and (5) the evaluation phase.

## 4. IPA projects and an example of digital archive project

The European Union's financial aid is allocated through the pre-accession programmes, the community programmes and the structural and cohesion funds. Instrument for Pre-Accession Assistance (IPA) has consolidated all former pre-accession programmes (CARDS, PHARE, ISPA, SAPARD). Croatia is involved in several groups of community programmes like FP7 (7th Framework Programme), CIP (Competitiveness and Innovation Framework Programme), IDABC (Interchange of Data between Administrations), Progress, Fiscalis, Marco Polo II, Media 2007 etc. At the moment the best way to launch implementation projects in Croatia is through FP7, if projects have strong research and development features and potential beneficiary is willing to cooperate with partners from various EU countries. If projects are set up to be solutions of some serious administrative problems, especially in the public sector, the best way is to apply them for IPA financial aid. IPA project should be placed under one IPA component and it should be linked to an issue stated in one of the negotiation chapters in order to justify the project and acquire financial help.

IPA 2009 TAIB (Transition Assistance and Institution Building) project "Preparations for eCTD and Implementation of Digital Archival Information System" planned in the Croatian Agency for Medicinal Products and Medical Devices (ALMP) will be used as an example of planning and designing a digital archival information system. This project deals with implementing new Europe-based eCTD standard for electronic pharmaceutical documentation[3] and for that it is linked to strengthening the ability to assume the obligations of the EU membership and chapter of negotiations that deals with the free movement of goods.

3 eCTD – electronic Common Technical Document, standard for electronic resources for medicine' regulations (http://esubmission.emea.europa.eu/whatisesubmission.htm). These records consist of granulated folder structure (modules), PDF files and XML backbone files. eCTD standard facilitate communication between marketing authorisation holders that submit records and agencies that receive them, easily providing insight on last authorised version of registration documentation.

ALMP is a national competent authority with mandate to control and to approve pharmaceuticals for the Croatian market. Heads of Member States' agencies agreed in Reykjavik in 2005 that the European agencies will shift to electronic records for registration of medicines from 2009 onwards and this should be obligatory for Croatia after accession. Since eCTD records will be used, ALMP needs to develop a digital archival information system in order to receive the documentation from marketing authorisation holders in pharmaceutical industry and other European agencies, to process it, maintain it and preserve it as evidence of registration activity. The project also deals with digitisation and microfilming of the existing paper records with the aim to interlink them with eCTD records and thus maintain business critical resources in a unique digital archive.

The project was developed by PCM methodology. In identification phase ALMP stated the problem of potential inability of co-operation with the EU Member States' agencies after their shift to the eCTD resources. The problem was addressed to the Croatian Ministry of Health and Social Welfare, so it could be included in the analysis of problems in the health sector and financial help could be requested.

In the formulation phase the project fiche was drafted and submitted to The Central Office for Development Strategy and Coordination of EU Funds (CODEF)[4]. For this project the project fiche was linked with Stabilization and association agreement, Accession partnership, National programme for the integration of the Republic of Croatia into the European Union, Croatia progress report, European partnership, and Multi-annual indicative planning document in order to justify the project and financial help requested. *Overall objective* of the project is Croatian participation in the European medicines network based on sharing current common standards. *The purpose* of the project is implementation of current European-based digital resources for regulation of medicines.

The project consists of three major groups of activities: (1) related to business process analysis and redesign, (2) related to digitisation and microfilming of paper records, and (3) related to development and customisation of digital archival information system.

Business process analysis activities involve business process analysis, redesign of business processes and definition of workflows. Digitisation and microfilming activities involve preparation for digitisation and microfilming of approximately sixteen million pages, digitisation, microfilming, and disposition and weeding-out of paper according to the retention schedule. Development of digital archival information system activities consist of development and customisation of the document and records management software.

The project predicts two instruments and therefore the implementation of activities will be divided to two contracts. The first contract will be fee-based ser-

[4] CODEF (Croatian: SDURF) is the governmental authority with mandate to prepare development strategy related to the Croatian accession, and monitor strategy implementation.

vice contract and it will cover business process analysis and digitisation and microfilming activities while the second contract will be global price service contract and it will cover activities of development and customising document and records management software.
After the project fiche was drafted, controlled by CODEF and pre-approved by the Delegation of European Commission and other authorities with such mandate, the Agency started to work on tendering dossier for the project. Tendering dossier in this particular case includes Terms of Reference for both contracts. All activities are planned to be outsourced due to the lack of particular experts, and administrative and space capacities in the Agency[5].
At the moment of writing this paper the IPA project "Preparations for eCTD and Implementation of Digital Archival Information System" is in final part of formulation phase and tendering procedure is predicted to start after signing of Financial agreement for IPA 2009 cycle between the EU authorities and the Croatian authorities.

## 5. Comparison of methodologies and recommendations for digital archival information system planning and designing

ISO-based and project management based approaches described in this article can be compared by the number of their phases and by emphasis on particular exertion in each phase. Because of various number of phases that derive from different methodologies, for this comparison it was necessary to consider a wider common delimiter of phase boundaries and thus to establish division of methodologies into wider common phases. Delimiters of phase boundaries were defined according to the issues that need to be assessed during lifecycle of potential digital archive. Issues that need to be assessed in particular lifecycle moments are: stated problem, proposed solution, planned activities, and results of the implementation. In this sense the stated problem is firstly assessed by its relevancy, secondly by accordance of solution compared to the problem, thirdly by adequacy of planned activities for implementation, and finally by the quality of accomplished results. After defining delimiters it was easier to define common phases for all selected methodologies and to compare them. Methodologies could thus be divided to (1) initial phase, (2) planning phase, (3) designing and implementation phase and (4) final and post implementation phase. It is showed that in PCM's initial phase is more emphasised than in ISO 15489, DIRKS and PRINCE2 methodology while ISO 15489, related standards and DIRKS approach are mostly focused on the planning phase and PRINCE2 on the designing and implementation phase.

[5] The Agency collaborated in preparation of the tendering dossier with the Central Finance and Contracting Agency (CFCA, Croatian: SAFU). CFCA is the Croatian governmental implementation agency for decentralised EU fund implementation system and is in charge of preparing projects for tendering and implementation and for monitoring financially related aspects of projects.

Table 1: Methodologies compared by their phases with common phase delimiters

| ISO 15489 and DIRKS | PRINCE2 | PCM |
|---|---|---|
| Preliminary investigation | Starting up the project | Programming |
| Business process analysis | Planning | Identification |
| Identification of record-keeping requirements | Directing a project | Formulation |
| Assessment of existing system/systems | Initiating the project | Implementation |
| Identification of record-keeping strategies | Controlling the stage(s) | Evaluation |
| Design of recordkeeping system | Managing product delivery | |
| Implementation | Managing stage boundaries | |
| Post-implementation review and testing | Closing the project | |

Initial phases (assessment of problems)
Planning phases (assessment of solution)
Designing and implementation phases (assessment of activities)
Final and post-implementation phases (assessment of results)

Each methodology could be used for planning, designing, implementation and evaluation of digital archival information system, but selection of methodology must depend on an estimation of the most critical phase in setting up the particular digital archive and on current state of affairs in organisation in which the archive will be established. Estimation of the most critical phase means that designers of the digital archive must recognise potential risks in advance. PCM should be considered if approving and initiating the project are at stake. It is also possible to use more than one methodology, for example, to prepare proposal document using PCM and to implement digital archive using an approach which is more adequate for solving implementation related entanglements. Selection on the basis of type of current organisational environment in which digital archive will be set up refers to administrative, financial and other capacities of organisation, including human resources.

As the case of ALMP showed, the selection of financial aid (fund) and other instruments was made by estimating potential project scope, costs and technological characteristics and by examining the possibility of compliance with strategic and negotiating documents for Croatia, which is necessary to have for pre-accession help. Since PCM is mandatory methodology for IPA projects, and estimation showed that the Agency's project would be perfect IPA candidate, this methodology was used from the beginning of project planning. Even if the Agency decided to develop digital archive with its own means PCM would be used because of decision making process that have to deal with approving substantial financial assets for the project. However, in the implementation phase

team leaders that serve either under the first or under the second contract will be allowed to use different development methodologies if it will be shown that advantages of the proposed methodology prevails PCM advantages in further development of the digital archival information system.

It can be concluded that PCM emphasise initial phase and is most adequate for decision-making on necessity of particular projects. Therefore it should be primarily used to evaluate projects before implementation. PRINCE was developed for project managers as it is based on implementation of issues related on project activities. ISO 15489 standard and DIRKS manual have a wider scope and are not limited to the recordkeeping systems, such as digital archival information systems. They are primarily developed as reference methodologies that can be used for establishing whole records management environments by cautious planning of recordkeeping policies and systems.

## References

Adam, Azad. Implementing Electronic Document and Record Management Systems. New York: Auerbach Publications, 2007

Begičević, N.; Divjak, B.; Hunjak, T. Decision making on project selection in high education sector using the Analytic hierarchy process. // Proceedings of the ITI 2009. Lužar-Stiffler, V.; Jarec, I.; Bekić, Z. (ur.). Zagreb: SRCE, 2009 (547-552)

Dirks manual – Users Guide, Part I. National Archives of Australia, 2003. http://www.naa.gov.au/Images/dirks_part1_tcm2-935.pdf (July 2009)

Divjak, Blaženka (ur.) Projekti u znanosti i razvoju: Europski programi. Varaždin: Tiva tiskara, FOI, 2009.

Europski fondovi za hrvatske projekte. priručnik o financijskoj suradnji i programima koje u Hrvatskoj podupire Europska Unija. Zagreb: SDURF, 2009

HRN ISO 15489-1 Information and documentation – Records management – Part 1: General (ISO 15489-1:2001)

HRN ISO 15489-1 Information and documentation – Records management – Part 2: Guidelines (ISO/TR 15489-2:2001)

HRN ISO 23081-1 Information and documentation – Records management processes - Metadata for records – Part 1: Principles (ISO 23081-1:2006)

HRN ISO 23081-1 Information and documentation – Records management processes - Metadata for records – Part 2: Conceptual and implementation issues (ISO/TS 23081-2:2007)

ISO 14721:2003 Space data and information transfer systems – Open archival information system – Reference model

ISO/TR 26122:2008(E) Information and documentation – Work process analysis for records.

Komunikacija i vidljivost. Priručnik za vanjske aktivnosti EU. Zagreb: SDURF, 2008

Načini pružanja pomoći, sv.1. Smjernice za upravljanje projektnim ciklusom. Podrška učinkovitoj provedbi vanjske pomoći EK. Zagreb: SDURF, 2008

PIU Manual_version 4.0. April 2009. http://www.safu.hr/hr/Dokumenti (April 2009)

Practical guide to contract procedures for EC external actions. 2009. http://ec.europa.eu/europeaid/work/procedures/implementation/practical_guide/index_en.htm (May 2009)

Priručnik za komponentu I programa IPA: Pomoć u tranziciji i izgradnja institucija. Zagreb: SDURF, 2007

Strateški okvir za razvoj 2006.-2013. Zagreb: Vlada Republike Hrvatske, SDURF, 2006

# Electronic Records Management System Requirements

Marko Lukičić
Ericsson Nikola Tesla
Krapinska 45, Zagreb, Croatia
marko.lukicic@ericsson.com

Vlado Sruk
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, Zagreb, Croatia
vlado.sruk@fer.hr

**Summary**

*This paper describes objectives and achievements of Enterprise Document and Records Management (EDRM) System standards in Europe through a work on MoReq2 specification. MoReq2 is a comprehensive catalogue of generic requirements for an Enterprise Records Management (ERM) system. It builds on the original MoReq specification, which was published in 2001. Specification is intended for use in public and private sector organizations which wish to use ERM systems.*
*First, we introduce ERM systems by discussing record definition and ERM key characteristics. The short overview of most popular European standards for managing electronic records is given, as well as MoReq specification, its purpose, organization and content. Subsequently the final MoReq2 specification and its modules are presented.*

**Key words:** MoReq2, specification, Enterprise Records Management, standards

## Introduction

With the use of computers, Internet and World Wide Web, our ability to share, understand and generate digital information has increased over the last few years. Thus, its level of importance in everyday business is higher than ever before. However, this change is having a great impact on common understandings about information, communication and knowledge raising many crucial questions such as: What is reliable information? How do we communicate effectively? How do we develop and maintain knowledge in our archives? Before considering a business solution, these questions have to be answered in three different domains: technological, organizational and governmental.

Challenges in the technological domain have to cope with problems of manipulating and securing massive volumes of geographically distributed business critical and sensitive data. Organizational challenges include reform of business processes and ways of handling and storing paper documents in classical archives. To enable an archive to be aware of digital data and to take over retention, management, retrieval and disposal processes of electronic data, it is necessary to redefine user roles and its responsibilities too. Challenges in the governmental domain present the greatest obstacle in implementing an electronic archive. Its critical task is to ensure legality and trustworthiness of digital data in respect to legislative and regulative.

The key benefits of the electronic archives are well known. Data can be easily and centrally managed and secured. The digital form of documents eliminates costs of physical storage because it doesn’t require special necessities as large personnel, big rooms secured from fire, flood, freezing or high temperature that could cause the occurrence of fungi, etc., and it bypasses the barriers of distributed offices. However, the main reason of implementing a digital archive is to achieve pure electronic business backbone. A business information system based on electronic archive enables organizations to automate the whole process from data acquisition or generation, to its processing, classification and archival as a true and legal evidence of business activity[1].

The Enterprise Records Management (ERM) systems are designed as a result of this necessity. Its main purpose is to provide a backbone for building the digital archive capable of managing electronic and physical records. Main functions of ERM systems are: content creation and capture, storing content, content retrieval, short- and long-term preservation and content disposition and disposal.

Today, such systems have become necessary for forming the electronic writing office, also known as the paperless office[23]. The key role of ERM systems is to reduce the response times for information requests, eliminate paper redundancy and duplication, and finally remove paper from the records management cycle while maintaining legality and trustworthiness of digital data in respect to legislative and regulative.

[1] Cornu, Jean-Michel. DLM-Forum. Guidelines on best practices for using electronic information. Office for Official publications of the European Communities. ISBN:92-828-2285-0. Luxembourg. 1997.

[2] Baumann, Stephan; Malburg, Michael; Meyer Auf’m Hofe, Harald; Wenzel, Claudia. From paper to a corporate memory. KI-97 Workshop on KBS for Knowledge Management in Enterprises. Freiburg, Germany. 1997 Sep 9-12; p. 16.

[3] Volarevic, Marijo; Strasberger, Vito; Pacelat, Elvis. A Philosophy of the Electronic Document Management. Proc. of the 22nd International Conference on Information Technology Interfaces; 2000 Jun 13-16; Pula, Croatia. Zagreb: SRCE University Computing Centre, University of Zagreb; 2000. p. 141-146.

## Enterprise Records Management Systems

There is no general definition what ERM systems are. Nowadays, definitions are mainly short descriptions of products for managing electronic records defined by the enterprise content management system vendors. However, ERM functionalities are mainly dictated by requirements defined by the national archives. And these requirements are results of long evolution and tradition in managing and organizing physical records in archives. Thus, differences between particular ERM systems are mainly in algorithm realization (ex. security algorithms), technology support (ex. integration with other systems) and in additional functionalities (ex. extended object model). Still, there is one document that has become most referenced in recent years when considering ERM system definition: ISO 15489-1:2001 (ISO) standard[4].

### Record

The ISO 15489-1 standard defines a record as: „recorded information, in any format, that is created, received and maintains as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business“. There are two key points that must be noted in ISO 15489-1 definition of a record. First, definition is opened for all types of records (ex. digital records, paper records, physical objects, etc.). And second, record is an evidence of business action, transaction or any other activity (ex. contract).
By ISO definition, to be authoritative, records must be:

- Authentic (ex. to have been created or send by the purported person);
- Reliable (trusted contents which accurately reflect the documented activity);
- Have integrity (records must be complete and unaltered);
- Useable (records can be located, retrieved, presented and interpreted).

It is important to recognize similarities and differences between documents and records as shown in Table 1.

Table 1: Differences between documents and records.

| **Document** | **Record** |
|---|---|
| A “piece” of information you can handle or manage | A “piece” of information you can handle or manage |
| May be important, or not | Represents important evidence of decision or act |
| Under the management of its “owner” (usually author) | Under corporate management |
| Can be changed at will | Cannot be changed |
| Can be deleted at will | Cannot usually be deleted |

[4] ISO 15489: Information and documentation – Records management. Reference number: ISO 15489-1:2001(E). International Organization for Standardization. 2001.

**Records Management**

Besides records, ISO 15489-1 standard defines records management as: "the field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposal of records, including processes for capturing and maintaining evidence of an information about business activities and transactions in the form of records."

ERM systems are systems that ensure management of records (as defined in ISO records management definition), whether they are in electronic of physical format, while ensuring all records characteristics (as defined in ISO record definition). By the ISO standard, ERM systems should provide:

- Reliability of complete, organized, accessible and protected records;
- Protected integrity by authority control systems;
- Compliance with legislative, regulative and appropriate business requirements;
- Reflected comprehensive range of appropriate business activities;
- Systematic creation, preservation and management of records.

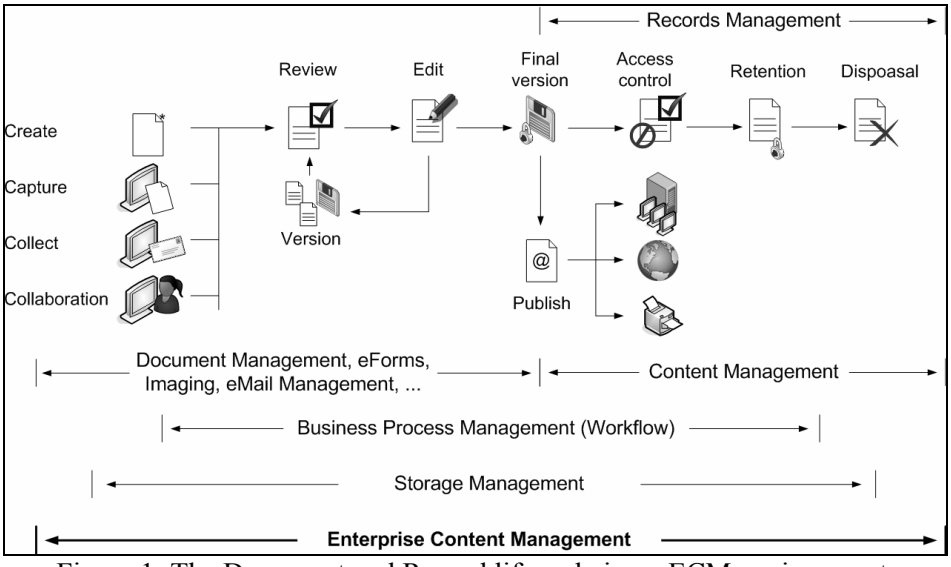



Figure 1: The Document and Record lifecycle in an ECM environment.

Nowadays, ERM systems are usually integrated with an Electronic Document Management (EDM) system to form an Enterprise Document and Records Management (EDRM) system. Synergy of those two systems combines document oriented collaboration functionalities of EDM systems with classification, compliance, preservation and disposition and disposal functionalities of ERM systems. Business Process Management (BPM) systems may be used as a bridge between a case file in ERM system and appropriate case file workflow.

Other technologies such as electronic forms (eForms) can be used as additional tools for user interface customization and records capture automation.
ERM (or EDRM) systems typically fall in Enterprise Content Management (ECM) system – system for managing all organizational information assets over their lifecycle, Figure 1.

## Records management standards

Although national archives indirectly shape functionalities of ERM systems, there are still pending functionality problems to be solved. The main problem is differences in practices between national archives of different countries. Despite the similarity of records management practices among European Union (EU) countries, there are still minor dissimilarities, which require individual implementation of specific functionality[5].
Moreover, almost every EU country has prescribed its own standard for records management. Because of non-existent records management standard at EU level, major ERM vendors mainly choose to certify its ERM systems for British PRO/TNA standard only. This retarded digitization of archives, and indirectly implementation of true electronic businesses, in smaller EU countries.
In continuation, we bring the short overview of most popular records management standards in EU countries.
**DOMEA (Germany)[6]**. DOMEA concept (Document Management and Electronic Archiving in Electronic Business, also known as "Paperless Office Concept") is the most important guideline for the implementation of electronic records in Germany. It consists of three main sections: Organization concept, Requirements catalogue and expansion modules. Although IT vendors are not obligated to certify its ERM systems against DOMEA, there are 170,000 approved DOEMA licenses in Germany, Austria and Switzerland.
**ELAK (Austria)[7]**. ELAK (Electronic Act) is a program of the Austrian Federal Government for a simplification and consolidation of the federal internal management of records. In addition to DOMEA, the ELAK concept describes requirements and functions of the ERM systems in more technical detail. Moreover, it provides examples what has to be considered in invitations for public tenders.

[5] PICTURE consortium. Integrating and Strengthening the European Research Area. ICT Research for Innovative Government. Project No 027717. PICTURE consortium. 2007.

[6] Bundesministerium des Innern. DOMEA Concept. Organisational Concept 2.0. Document Management and Electronic Archiving in Electronic Courses of Business. Koordinierungs- und Beratungsstelle der Bundesregierung fur Informationstechnik in der Bundesverwaltung, KBSt. ISSN 0179-7263. November 2005.

[7] Bundesministerium fur offentliche Leistung und Sport. ELAK-Konzept. Funktionsbeschreibung. November 2001. URL:http://www.digitales.oesterreich.gv.at/site/5286/default.aspx

**Gever (Switzerland)**[8]. The Swiss Gever is collection of five standards that introduces management of electronic records and paper based records administration abandonment. The five standards are: Business Administration, Methods and functions with regard on legal defaults, Business model GEVER Federation, Service catalogue of GEVER applications and GEVER metadata.
**Protocollo Informatico/CNIPA (Italy)**[9]. CNIPA (National Centre for Information Technologies in Public Administration) is the government organization responsible to give support to the Italian public administrations in creating information systems to further improve the quality of services and keep the administrative costs down. Protocollo Informatico is published by CNIPA that describes the electronic protocol as a framework of resources used by administrations for managing documents.
**ReMANO (Netherlands)**[10]. ReMANO is a catalogue of software specifications for ERM systems in Duch government bodies. It is published in 2004 by "Nederlands Instituut voor Archiefonderwijs en – onderzoek".
**NOARK (Norway)**[11]. NOARK-4 is functional requirements specification for ERM and case management systems used in all public authorities in Norway.
**PRO/TNA (United Kingdom)**[12]. The PRO/TNA document is developed by Public Records Office (The National Archive). Its main purpose is to provide a tool for benchmarking ability of government departments to support electronic records management. Although, PRO/TNA was the most comprehensive and popular standard in EU, it is replaced with MoReq specification.

## MoReq specification

Though there are a large number of records management national standards in EU, the absence of the EU-wide standard complicates an interoperable delivery of European electronic government services to public administrations, business and citizens. This was in confrontation with i2010 eGovernment Action Plan

[8] eCH, eGovernment-Standards. eCH-0037 Hilfsmittel GEVER Vorgaben Bund. Verein eCH. April 2005.

[9] Centro Nazionale per l'Informatica nella Pubblica Amministrazione, CNIPA. Protocollo informatico. CNIPA. URL: http://www.cnipa.gov.it/site/it-IT/Attivit%C3%A0/Protocollo_informatico/

[10] Nederlands Instituut voor Archiefonderwijs en – onderzoek. Softwarespecificaties voor Records Management Applicaties voor de Nederlandse Overheid (ReMANO). Archiefschool. Mart 2004.

[11] Riksarkivet - The National Archives of Norway. Norwegian recordkeeping system, Version 4 (NOARK). Functional description and specification of requirements. Riksarkivet. 2000.

[12] Public Record Office, United Kingdom. Requirements for Electronic Records Management Systems. The National Archives. 2002.

which one of main objectives is to accelerate the delivery of tangible benefits for citizens and business through eGovernment[13].

The need for a comprehensive specification of ERM system requirements for government authorities was first articulated by DLM Forum in 1996. DLM Forum (Donnees Lisibles par Machine) is a multi-disciplinary forum constituted by European Commission. Its main goal is to investigate, promote and implement - in close cooperation with Member States - possibilities for wider cooperation in the field of electronic archives both between the Member States and at EU level.

In 1999 DLM Forum issued the following action point: development of a reference model for managing electronic documents and records in public administration.

The work on specification began in 2000 and it was completed in 2001. MoReq, first became available electronically in 2001, was published by the European Commission as an INSAR (Information Summary on Archives publication) supplement in early 2002.

The specification contains functional and non-functional requirements for ERM systems. Functional requirements cover following topics: overview of ERM system functionalities, classification scheme, control and security, retention and disposal, capturing records, search, retrieval and rendering, administrative functionalities and other functionalities. Non-functional requirements, such as the metadata model, can vary enormously between environments. Thus, MoReq specification identifies and describes non-functional requirements only in outline.

The specification suggests that ERM system should be introduced not only to Administrators and Archivists (that is, staff responsible for records management), but to all general offices and operational staff that are involved in a creation, receiving and retrieving of the records. Therefore, MoReq specification embraces records management closely-related requirements such as document and case management. However, these requirements are described in less detail than functional requirements.

Because of its comprehensiveness, MoReq becomes accepted and used worldwide. However, some problems have been raised in the last few years. The main problem is that MoReq is not a formal standard but a guideline. Non-existence of the testing regime disables ERM system vendors to provide conceivable proof of MoReq compliance. Furthermore, there has been no advancement since 2001. Technology has moved on. And the lack of governance caused uncontrolled MoReq translations referenced in particular ERM systems. Something had to be done.

[13] European Communities. i2010 eGovernment Action Plan. Communication from the Commission, of 25 April 2006, i2010 eGovernment Action Plan - Accelerating eGovernment in Europe for the Benefit of All. 2006. http://europa.eu/scadplus/leg/en/lvb/l24226j.htm

**The MoReq2 project**

The revision of MoReq specification was proposed by Ian MacFarlane in "Plans for MoReq, a report on scoping of a MoReq 2" paper[14]. This document contains key conclusions of DLM Forum discussion about MoReq revision. In 2006, DLM Forum published "Scoping report for the development of the Model Requirements for the management of electronic records (MoReq2)". This document outlines details of changes in the old document. The overall aims for the MoReq2 development, as described in Scoping report, are to develop extended functional requirements within a European context, and to support a compliance scheme by:

- Strengthening from MoReq what have in the interim become key areas and covering important new areas of requirements with clarity;
- Ensuring that the functional requirements are testable and developing test materials to enable products to be tested for compliance with the requirements;
- Making the requirements modular to assist application in the various environments in which they will be used.

As stated in the report, to provide compatibility with earlier version, MoReq2 is to be an evolutionary update to the original MoReq, not a radically different product.

The MoReq2 project started in 2007, and the MoReq2 specification is formally published at the beginning of 2008.

**MoReq2 specification**

The MoReq2 specification is a collection of required and optional functional and non-functional requirements for the ERM systems. While required functionalities are mandatory for MoReq2 compliance, optional requirements correspond for desirable but not mandatory characteristics of the ERM systems.

Required and optional requirements are grouped in core module (mandatory module for MoReq2 compliance) and optional modules (ERM system providers may choose to additionally certify software for particular optional modules).

Core module contains requirements regarding classification scheme and file organization, controls and security, retention and disposition, capturing and declaring records, referencing, searching, retrieval and presentation, and ERM system administration.

Optional modules are: Management of Physical (Non-electronic) Files and Records, Disposition of Physical Records, Document Management and Collaborative Working, Workflow, Casework, Integration with Content Management Systems, Electronic Signatures, Encryption, Digital Rights Management,

[14] Ian MacFarlane. The Plans for MoReq (Model requirements for the management of electronic records): A Report on the Scoping of the MoReq2. *DLM Forum Conference*. Budapest. 2005. http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/12_macfarlane_06-10-05am.pdf.

Distributed Systems, Offline and Remote Working, Fax Integration and Security Categories[15].

The metadata requirements present another important part of MoReq2 specification. MoReq2 metadata, based on Dublin Core Metadata Element Set, includes indexing information and other data needed for effective records management, such as access restriction information. As is not possible to define all the metadata requirements for all possible kinds of ERMS implementation, MoReq2 suggests minimum requirements, which are intended as the starting point for customization and expansion. These minimum requirements are closely related to lists of specific metadata "elements" which the ERM systems must be able to capture and process.

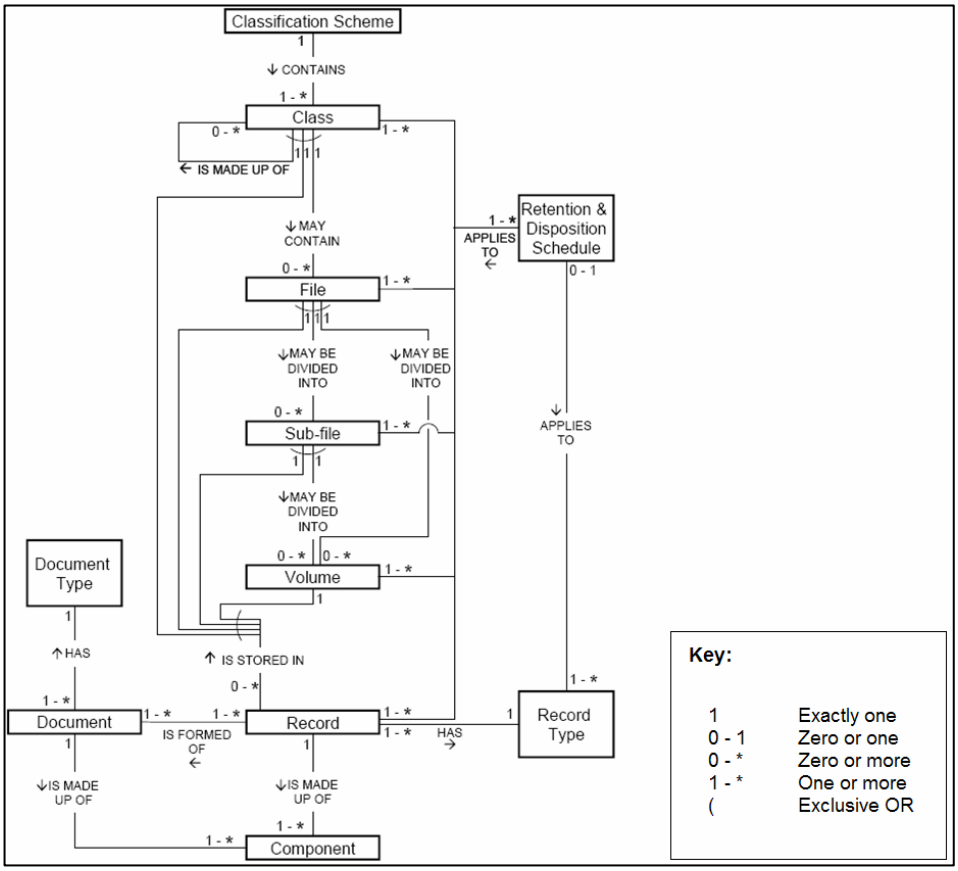



Figure 2: MoReq2 entity-relationship model.

[15] Cornwell Affiliates plc. Model Requirements for the management of electronic records. MoReq2 specification. Office for Official publications of the European Communities as INSAR supplement VIII. Bruxelles. Luxembourg. 2008.

Sub-files and components are new entities added to the MoReq object model, Figure 2. Sub-file is an intellectual subdivision of a file. It is often used in case management environments. Typical examples of sub-files are “invoices”, “assessments” and “correspondences”. Component is a bit stream that, alone or combined with other bit streams, makes up a record or document. Typical example of a component is a JPEG image of an HTML document.
Hybrid file, file that contains electronic and physical records, is omitted from object model. Therefore MoReq2 allow classes, files, sub-files and volumes to contain electronic records and physical records together, in any combination.
A special care was taken regarding specification localization. As each country may have need for specific requirements regarding managing electronic records, in MoReq2 structure is imbedded “Chapter zero”. This chapter can be used to represent specific needs of a particular country. The only restriction on expanding MoReq2 specification with this chapter is that content of the chapter should not contradict the content of the rest of MoReq2.

## Conclusion

With the increase of produced information and wide diversity of information formats, challenges for the management of electronic records have never been greater. To cope with implementations of new technologies and to achieve trustworthy of digital records in respect to legislative and regulative, particular countries started introducing specifications regarding managing of electronic records. However, multiplicity of different specifications makes interoperability of data between administrations of EU countries difficult.
MoReq2 represents a step forward in the process of unifying records management software standards and practices across Europe. It provides governments and corporations a single approach to managing their most important records. Thus, MoReq2 will significantly contribute to the accomplishment of greater interoperability between administrations, business and citizens, that is the achievement of the goals of the Europe Union’s i2010 eGovernment action plan.

## References

Baumann, Stephan; Malburg, Michael; Meyer Auf’m Hofe, Harald; Wenzel, Claudia. From paper to a corporate memory. KI-97 Workshop on KBS for Knowledge Management in Enterprises. Freiburg, Germany. 1997 Sep 9-12; p. 16.

Bundesministerium des Innern. DOMEA Concept. Organisational Concept 2.0. Document Management and Electronic Archiving in Electronic Courses of Business. Koordinierungs- und Beratungsstelle der Bundesregierung fur Informationstechnik in der Bundesverwaltung, KBSt. ISSN 0179-7263. November 2005.

Bundesministerium fur offentliche Leistung und Sport. ELAK-Konzept. Funktionsbeschreibung. November 2001. URL:http://www.digitales.oesterreich.gv.at/site/5286/default.aspx

Centro Nazionale per l'Informatica nella Pubblica Amministrazione, CNIPA. Protocollo informatico. CNIPA. URL: http://www.cnipa.gov.it/site/it-IT/Attivit%C3%A0/Protocollo_informatico/

Cornu, Jean-Michel. DLM-Forum. Guidelines on best practices for using electronic information. Office for Official publications of the European Communities. ISBN:92-828-2285-0. Luxembourg. 1997.

Cornwell Affiliates plc. Model Requirements for the management of electronic records. Office for Official publications of the European Communities as INSAR supplement VI. ISBN:92-894-1290-9. Bruxelles. Luxembourg. 2001.

Cornwell Affiliates plc. Model Requirements for the management of electronic records. MoReq2 specification. Office for Official publications of the European Communities as INSAR supplement VIII. Bruxelles. Luxembourg. 2008.

eCH, eGovernment-Standards. eCH-0037 Hilfsmittel GEVER Vorgaben Bund. Verein eCH. April 2005.

European Communities. i2010 eGovernment Action Plan. Communication from the Commission, of 25 April 2006, i2010 eGovernment Action Plan - Accelerating eGovernment in Europe for the Benefit of All. 2006. http://europa.eu/scadplus/leg/en/lvb/l24226j.htm

Ian MacFarlane. The Plans for MoReq (Model requirements for the management of electronic records): A Report on the Scoping of the MoReq2. DLM Forum Conference. Budapest. 2005. http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/12_macfarlane_06-10-05am.pdf.

ISO 15489: Information and documentation – Records management. Reference number: ISO 15489-1:2001(E). International Organization for Standardization. 2001.

Nederlands Instituut voor Archiefonderwijs en – onderzoek. Softwarespecificaties voor Records Management Applicaties voor de Nederlandse Overheid (ReMANO). Archiefschool. Mart 2004.

PICTURE consortium. Integrating and Strengthening the European Research Area. ICT Research for Innovative Government. Project No 027717. PICTURE consortium. 2007.

Public Record Office, United Kingdom. Requirements for Electronic Records Management Systems. The National Archives. 2002.

Riksarkivet - The National Archives of Norway. Norwegian recordkeeping system, Version 4 (NOARK). Functional description and specification of requirements. Riksarkivet. 2000.

Volarevic, Marijo; Strasberger, Vito; Pacelat, Elvis. A Philosophy of the Electronic Document Management. Proc. of the 22nd International Conference on Information Technology Interfaces; 2000 Jun 13-16; Pula, Croatia. Zagreb: SRCE University Computing Centre, University of Zagreb; 2000. p. 141-146.

# Preservation of Interactive Multimedia Systems with an Ontology based Approach

Kia Ng, Eleni Mikroyannidi, Bee Ong
ICSRiM – University of Leeds,
School of Computing & School of Music,
Leeds LS2 9JT, UK
kia@icsrim.org.uk, kia@kcng.org, kia@comp.leeds.ac.uk

David Giaretta
STFC, Rutherford Appleton Laboratory,
Oxfordshire OX11 0QX, UK

**Summary**

*Digital preservation aims to address changes that inevitably occur in hardware or software, in the designated community, i.e. the users of the preserved information. In order preserve digital information so that they are usable and understandable in the future, digital information has to be enriched with metadata, usually referred to as Representation Information, which can be used for the interpretation of information. Representation Information needs to be connected to the Knowledge Base of the designated community with appropriate terminologies for better interpretation and representations. Ontologies offer the means for organizing and representing the semantics of this knowledge base. The paper presents the CASPAR project[1] (supported under the EC IST Framework programme) which aims to build a pioneering framework to support the end-to-end preservation lifecycle of scientific, artistic and cultural information. This paper is focused on the contemporary arts testbed with a particular attention on interactive multimedia performances (IMP)[2]. The paper describes several different IMP systems and presents an archival system, which has been designed and implemented based on the CASPAR framework and components for preserving Interactive Multimedia Performances.*

**Key words:** digital preservation, interactive multimedia, ontology, performing arts, OAIS, gesture, motion

[1] http://www.casparpreserves.eu

[2] http://www.icsrim.org.uk/caspar/

## Introduction

Interactive Multimedia Performance (IMP) preservation is part of the Contemporary Arts testbed of the CASPAR project. IMP is chosen as part of the testbeds for its challenges due to the complexity and multiple dependencies and typically involves several difference categories of digital media data. Generally, an IMP involves one or more performers who interact with a computer based multimedia system making use of multimedia contents that may be prepared as well as generated in real-time including music, audio, video, animation, graphics, and many others.[3, 4]

The interactions between the performer(s) and the multimedia system[5, 6, 7] can be done in a wide range of different approaches, such as body motions (for example, see Music via Motion (MvM)[8, 9]), movements of traditional musical instruments or other interfaces, sounds generated by these instruments, tension of body muscle using bio-feedback,[10] heart beats, sensors systems, and many others. These "signals" from performers are captured and processed by multimedia systems. Depending on specific performances, the input can be mapped onto multimedia contents and/or as control parameters to generate live contents/feedback using a mapping strategy.

Traditional music notation as an abstract representation of a performance it is not sufficient to store all the information and data required to reconstruct the performance with all the specific details. In order to keep an IMP performance

[3] Ng, Kia (ed). Proceedings of the COST287-ConGAS 2nd International Symposium on Gesture Interface for Multimedia Systems (GIMS2006), 9-10 May 2006, Leeds, UK. http://www.i-maestro.org/documenti/view_documenti.php?doc_id=1052

[4] Ng, Kia; Nasi, Paolo (eds). Interactive Multimedia Music Technologies IGI Global, Information Science Reference, Library of Congress 2007023452, 2008.

[5] Young, D.; Nunn, P.; Vassiliev, A. Composing for Hyperbow: A Collaboration between MIT and the Royal Academy of Music. in *Proc. of the New Interfaces for Musical Expression International Conference (NIME)*. Paris, France. 2006

[6] Overholt, D. The Overtone Violin. In *Proc. of the International Conference on New Interfaces for Musical Expression*. Vancouver, BC, Canada. 2005.

[7] Lévy, Benjamin; Ng, Kia. Audio-driven Augmentations for the Cello, in Ng (ed), in *Proc. of the 4th i-Maestro Workshop on Technology-Enhanced Music Education*, co-located with the 8th International Conference on New Interfaces for Musical Expression (NIME 2008), Genova, Italy, ISBN: 978 0 85316 269 8, pp. 15-20, 4 June 2008

[8] MvM. Music via Motion, http://www.kcng.org/mvm or http://www.leeds.ac.uk/icsrim/mvm [last accessed 28/9/2009]

[9] Ng, Kia. Music via Motion: Transdomain Mapping of Motion and Sound for Interactive Performances, in *Proceedings of the IEEE*, vol. 92, 2004.

[10] Nagashima, Y. Bio-Sensing Systems and Bio-Feedback Systems for Interactive Media Arts. in *Proc. of the New Interfaces for Musical Expression International Conference (NIME-03)*. Montreal, Canada. 2003.

alive through time, not only its output, but also the whole production process to create the output needs to be preserved.

## Interactive Multimedia Performance (IMP) Systems

In this section we describe several different IMP systems and software with different types of interaction and different types of data while the following section explains how the CASPAR framework is used for their preservation.

### The 3D Augmented Mirror (AMIR) System

The 3D Augmented Mirror (AMIR)[11, 12] is an example IMP system which has been developed in the context of the i-Maestro project (http://www.i-maestro.org), [13] for the analysis of gesture and posture in string practice training. Similar to many other performing arts, string players (e.g. violinist, cellists) often use mirrors to observe themselves practicing to understand and improve awareness of their playing gesture and posture. More recently, video has also been used. However, this is generally not effective due to the inherent limitations of 2D perspective views of the media.

The i-Maestro 3D Augmented Mirror is designed to support the teaching and learning of bowing technique, by providing multimodal feedback based on real-time analysis of 3D motion capture data. Figure 1 shows a screenshot of the i-Maestro 3D Augmented Mirror interface which explore visualization and sonification (e.g. 3D bow motion pathway trajectories and patterns) to provide gesture and posture support. It uses many different types of data including 3D motion data (from a 12-camera motion capture system), pressure sensor, audio, video and balance.

The i-Maestro AMIR multimodal recording, which includes 3D motion data, audio, video and other optional sensor data (e.g. balance, etc) can be very useful to provide in-depth information beyond the classical audio visual recording many different purposes including technology-enhanced learning, and in this

[11] Ng, Kia; Weyde, Tillman; Larkin, Oliver ; Neubarth, Kerstin; Koerselman, Thijs; Ong, Bee. 3D Augmented Mirror: A Multimodal Interface for String Instrument Learning and Teaching with Gesture Support, in *Proc. of the 9th International Conference on Multimodal Interfaces*, Nagoya, Japan, pp. 339-345, ISBN: 978-1-59593-817-6, ACM, SIGCHI, DOI: http://doi.acm.org/10.1145/1322192.1322252, 2007

[12] Ng, Kia; Ong, Bee; Weyde, Tillman; Neubarth, Kerstin. Interactive Multimedia Technology-Enhanced Learning for Music with i-Maestro, in *Proc. of ED-MEDIA 2008 World Conference on Education Multimedia, Hypermedia & Telecommunications*, Vienna, Austria, 30 June – 4 July 2008.

[13] Ng, Kia (ed). Proceedings of the 4th i-Maestro Workshop on Technology-Enhanced Music Education, co-located with the 8th International Conference on New Interfaces for Musical Expression (NIME 2008), Genova, Italy, ISBN: 978 0 85316 269 8, 4 June 2008. http://www.i-maestro.org/workshop/

context for the preservation of playing gesture and style for detailed musicological analysis (now and in the future).

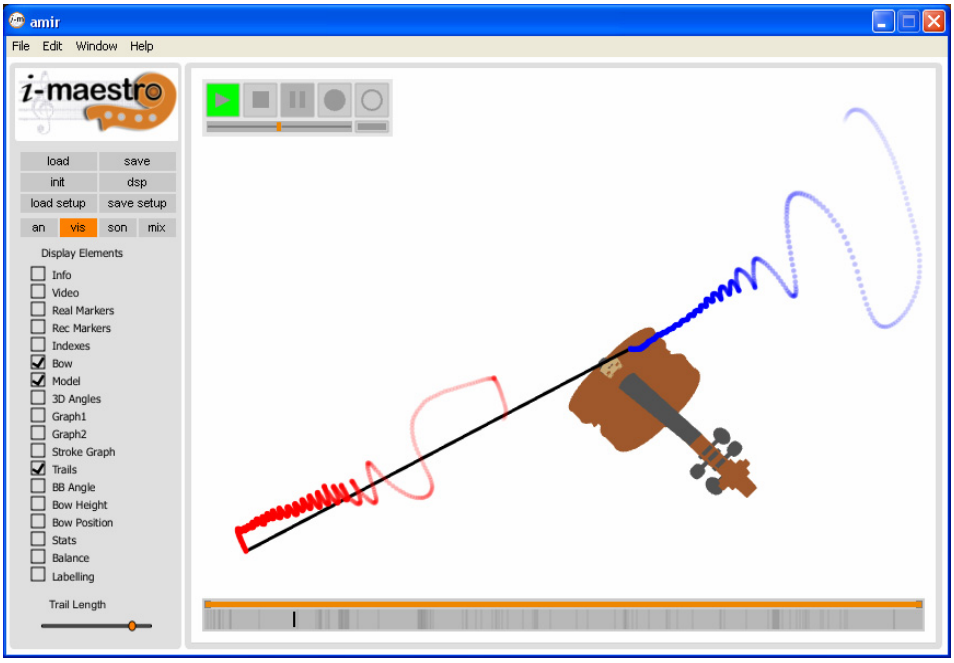



Figure 1: The i-Maestro 3D Augmented Mirror System

**ICSRiM Conducting Interface**

The ICSRiM Conducting System is another IMP system example. It has been developed for the tracking and analysis of a conductor’s hand movements.[14, 15] The system is aiming at supporting students learning and practicing conducting, and also provides a multimodal recording (and playback) interface to capture/measure detailed conducting gesture in 3D for the preservation of the performance.

A portable motion capture system composed by multiple Nintendo Wiimotes is used to capture the conductor’s gesture. The Nintendo Wiimote has several ad-

[14] Bradshaw, David; Ng, Kia. Tracking Conductors Hand Movements Using Multiple Wiimotes, in *Proc. of the International Conference on Automated Solutions for Cross Media Content and Multi-channel Distribution (AXMEDIS 2008)*, Florence, Italy, pp. 93-99, Digital Object Identifier 10.1109/AXMEDIS.2008.40, IEEE Computer Society Press, ISBN: 978-0-7695-3406-0. 4. 17-19 Nov. 2008.

[15] Bradshaw, David; Ng, Kia. Analyzing a Conductor’s Gestures with the Wiimote, in *Proc. of EVA London 2008: the International Conference of Electronic Visualisation and the Arts*, British Computer Society, 5 Southampton Street, London WC2E 7HA, UK, 22-24 July 2008.

vantages as it combines both optical and sensor based motion tracking capabilities, it is portable, affordable and easily attainable. The captured data are analyzed and presented to the user highlighting important factors and offer helpful and informative monitoring for raising self-awareness that can be used during a lesson or for self-practice. Figure 2 shows a screenshot of the Conducting System Interface with one of the four main visualization mode.

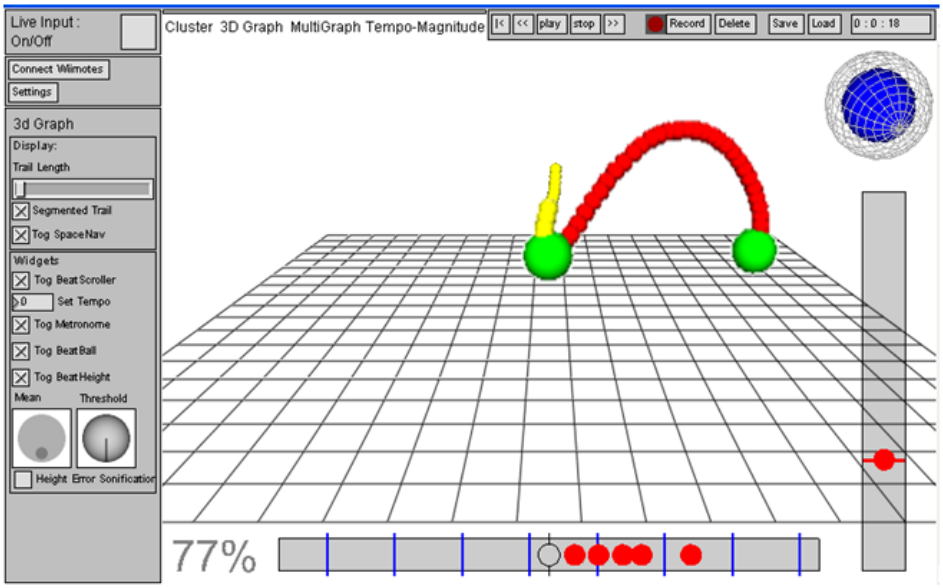



Figure 2: The ICSRiM Conducting interface showing a conducting gesture with 3D visualisation

## Preservation with Ontology models

The preservation of these IMP systems is of great importance in order to allow future re-performance, understanding and analysis. The multimodal recordings of these systems offer an additional level of detail for the preservation of musical gesture and performance (style, interpretation issues and others) that may be vital for the musicologist of the future.

Preserving an interactive multimedia performance is not easy. Preserving the single digital media object for a longer term is already a challenging issue. However, putting all the necessary digital objects together does not reconstruct the full system to allow a re-performance. For the preservation of IMP, we proposed to preserve the whole production process with all the digital objects involved together with their inter-relationships and additional information considering the reconstruction issues. It is a challenging issue since it is difficult to preserve the knowledge about the logical and temporal components, and all the

objects such as the captured 3D motion data, Max/MSP patches, configuration files, etc, in order to be properly connected for the reproduction of a performance.[16]

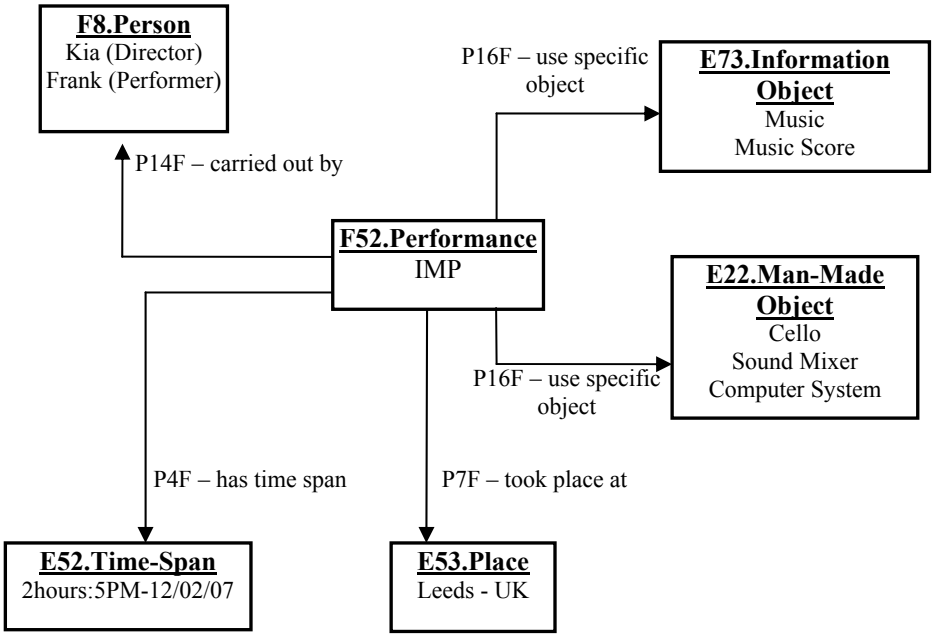



Figure 3: Modeling an IMP with the use of the CIDOC-CRM and FRBR ontologies

Due to these multiple dependencies, the preservation of an IMP requires robust representation and association of the digital resources. This can be performed using entities and properties defined for CIDOC-CRM and FRBRoo. The CIDOC Conceptual Reference Model (CRM) is being proposed as a standard ontology for enabling interoperability amongst digital archives. [17]

[16] Ng, Kia; Pham, Tran Vu; Ong, Bee; Mikroyannidis, Alexander; Giaretta, David. *Preservation of interactive multimedia performances*, International Journal of Metadata, Semantics and Ontologies 2008 - Vol. 3, No.3 pp. 183 – 196, DOI: 10.1504/IJMSO.2008.023567. 2009.

[17] Ng, Kia; Mikroyannidis, Alexander; Ong, Bee; Bonardi, Alain; Barthélemy, Jérôme; Ciavarella, Raffaele; Boutard, Guillaume. Ontology Management for Preservation of Interactive Multimedia Performances, in *Proc. of the International Computer Music Conference (ICMC)*, Belfast, 24-29 August 2008.

CIDOC-CRM defines a core set concepts for physical as well as temporal entities.[18, 19] CIDOC-CRM was originally designed for describing cultural heritage collections in museum archives. A harmonisation effort has also been carried out to align the Functional Requirements for Bibliographic Records (FRBR)[20] to CIDOC-CRM for describing artistic contents. The result is an object oriented version of FRBR, called FRBRoo.[21] The concepts and relations of the FRBRoo are directly mapped to CIDOC-CRM.
Figure 3 shows how the CIDOC-CRM and FRBR ontologies are used for the modelling of an IMP.

## ICSRiM IMP Archival System

The CASPAR project evaluated a set of preservation scenarios and strategies in order to validate its conceptual model and architectural solutions within the different testbed domains. In this case, our scenarios are related with the ingestion, retrieval and preservation of IMPs.
The ICSRiM IMP Archival System has been designed and developed with the CASPAR framework integrating a number of selected CASPAR components via web services. The system has been used to implement and validate the preservation scenarios.
The archival system is a web interface, which communicates with a Repository containing the IMPs and the necessary metadata for preserving the IMPs. The first step for preserving an IMP is to create its description based on the CIDOC-CRM and FRBRoo ontology. This information is generated in RDF/XML format with the use of the CASPAR Cyclops tool. The Cyclops tool[22] is used to capture appropriate Representation Information to enhance virtualisation and future re-use of the IMP. In particular, this web tool is integrated into the Archival System and it used in order to model various IMPs.
During ingestion, the IMP files and the metadata are uploaded and stored in the Repository with the use of the web-based IMP Archival System. For the retrieval of an IMP, queries are performed on the metadata and the related objects are returned to the user. Figure 4 shows the web interface of the ICSRiM IMP Archival system.

---

[18] Gill, T. Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model. First Monday, 9. 2004.

[19] Doerr, M. The CIDOC CRM - an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 24, 2003.

[20] FRBR. Functional Requirements for Bibliographic Records - Final Report. Frankfurt am Main, Germany, International Federation of Library Associations and Institutions (IFLA). 1997.

[21] Doerr, M.; Leboeuf, P. FRBRoo Introduction. http://cidoc.ics.forth.gr/frbr_inro.html (last accessed: 1/10/2009). 2006

[22] http://www.utc.fr/caspar/wiki/pmwiki.php?n=Main.Proto

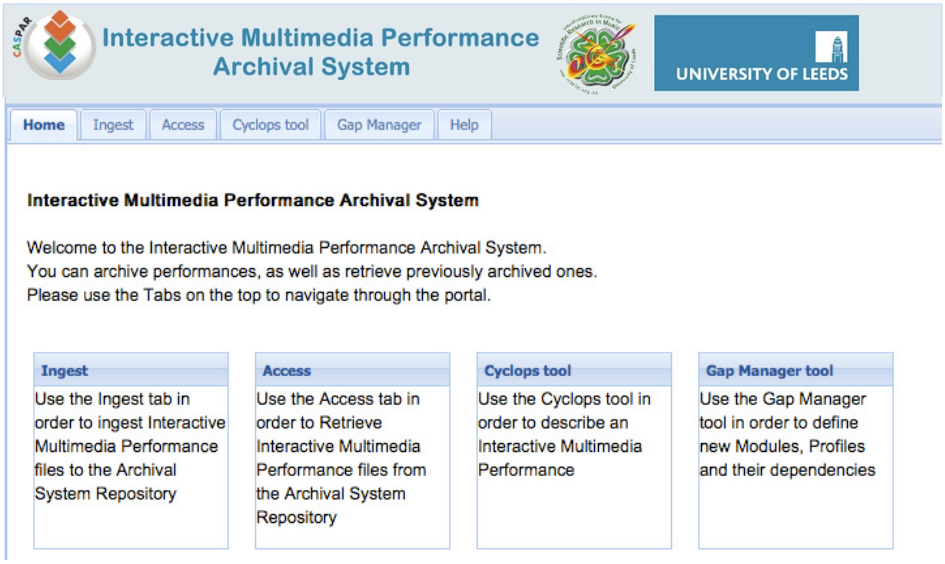



Figure 4: The Interface of the Web Archival System.

In case a change occurs in the dataset of an IMP, such as the release of a new version of the software, the user has the ability to update the Representation Information and the dataset of the IMP with the new modules (e.g. the version of new software). A future user will be able to understand which one is the latest version of a component and how these components can be reassembled for the reproduction of the Performance by retrieving the Representation Information of the IMP.

## Conclusion

This paper briefly introduces the usages and applications of interactive multimedia for contemporary performing arts as well as its usefulness for capturing/measuring multimedia and multimodal data that are able to better represent the playing gesture and/or interactions. With two example IMP systems, it discusses key requirements and complexities of the preservation considerations and presents a digital preservation framework based on ontologies for Interactive Multimedia Performances.

With the CASPAR framework, standard ontology models were adopted in order to define the relations between the individual components that are used for the re-performance. The paper also described the development and implementation of a web-based archival system using the CASPAR framework and components.

The ICSRiM IMP Archival System has been successfully validated by users who have created their own IMP systems using their own work for ingestion

and using ingested works from others (without any prior knowledge) to reconstruct a performance with only the instruction and information provided by the archival system.

## Acknowledgement

Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project CASPAR. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.
The research is also supported in part by the European Commission under Contract IST-026883 I-MAESTRO. The authors would like to acknowledge the EC IST FP6 for the partial funding of the I-MAESTRO project (http://www.i-maestro.org), and to express gratitude to all I-MAESTRO project partners and participants, for their interests, contributions and collaborations.

## References

Bradshaw, David; Ng, Kia. Tracking Conductors Hand Movements Using Multiple Wiimotes, in *Proceedings of the International Conference on Automated Solutions for Cross Media Content and Multi-channel Distribution (AXMEDIS 2008)*, Florence, Italy, pp. 93-99, Digital Object Identifier 10.1109/AXMEDIS.2008.40, IEEE Computer Society Press, ISBN: 978-0-7695-3406-0. 4. 17-19 Nov. 2008.

Bradshaw, David; Ng, Kia. Analyzing a Conductor's Gestures with the Wiimote, in *Proceedings of EVA London 2008: the International Conference of Electronic Visualisation and the Arts*, British Computer Society, 5 Southampton Street, London WC2E 7HA, UK, 22-24 July 2008.

Doerr, M. The CIDOC CRM - an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 24, 2003.

Lévy, Benjamin; Ng, Kia. Audio-driven Augmentations for the Cello, in Ng (ed), *Proceedings of the 4th i-Maestro Workshop on Technology-Enhanced Music Education*, co-located with the 8th International Conference on New Interfaces for Musical Expression (NIME 2008), Genova, Italy, ISBN: 978 0 85316 269 8, pp. 15-20, 4 June 2008

Doerr, M.; Leboeuf, P. FRBRoo Introduction. http://cidoc.ics.forth.gr/frbr_inro.html (last accessed: 1/10/2009). 2006

FRBR. Functional Requirements for Bibliographic Records - Final Report. Frankfurt am Main, Germany, International Federation of Library Associations and Institutions (IFLA). 1997.

Gill, T. Building semantic bridges between museums, libraries and archives: The CIDOC Conceptual Reference Model. First Monday, 9. 2004.

Mikroyannidis, Alexander; Ong, Bee; Ng, Kia; Giaretta, David. Ontology-Based Temporal Modelling of Provenance Information, in *Proceedings of the 14th IEEE Mediterranean Electrotechnical Conference (MELECON'2008)*, Ajaccio, France, pp. 176-181, 2008.

MvM. Music via Motion, http://www.kcng.org/mvm or http://www.leeds.ac.uk/icsrim/mvm [last accessed 28/9/2009]

Nagashima, Y. Bio-Sensing Systems and Bio-Feedback Systems for Interactive Media Arts. in *Proceedings of the New Interfaces for Musical Expression International Conference (NIME-03)*. Montreal, Canada. 2003.

Ng, Kia. Music via Motion: Transdomain Mapping of Motion and Sound for Interactive Performances, in *Proceedings of the IEEE*, vol. 92, 2004.

Ng, Kia (ed). Proceedings of the COST287-ConGAS 2nd International Symposium on Gesture Interface for Multimedia Systems (GIMS2006), 9-10 May 2006, Leeds, UK. http://www.i-maestro.org/documenti/view_documenti.php?doc_id=1052

Ng, Kia (ed). Proceedings of the 4th i-Maestro Workshop on Technology-Enhanced Music Education, co-located with the 8th International Conference on New Interfaces for Musical Expression (NIME 2008), Genova, Italy, ISBN: 978 0 85316 269 8, 4 June 2008. http://www.i-maestro.org/workshop/

Ng, Kia; Mikroyannidis, Alexander; Ong, Bee; Giaretta, David. Practicing Ontology Modelling for Preservation of Interactive Multimedia Performances, in *Proceedings of the International Conference on Automated Solutions for Cross Media Content and Multi-channel Distribution (AXMEDIS 2008)*, Florence, Italy, pp. 276-281, Digital Object Identifier 10.1109/AXMEDIS.2008.43, IEEE Computer Society Press, ISBN: 978-0-7695-3406-0. 17-19 Nov. 2008,

Ng, Kia; Mikroyannidis, Alexander; Ong, Bee; Bonardi, Alain; Barthélemy, Jérôme; Ciavarella, Raffaele; Boutard, Guillaume. Ontology Management for Preservation of Interactive Multimedia Performances, in *Proceedings of the International Computer Music Conference (ICMC)*, Belfast, 24-29 August 2008.

Ng, Kia; Nasi, Paolo (eds). Interactive Multimedia Music Technologies IGI Global, Information Science Reference, Library of Congress 2007023452, 2008.

Ng, Kia; Ong, Bee; Weyde, Tillman; Neubarth, Kerstin. Interactive Multimedia Technology-Enhanced Learning for Music with i-Maestro, in *Proceedings of ED-MEDIA 2008 World Conference on Education Multimedia, Hypermedia & Telecommunications*, Vienna, Austria, 30 June – 4 July 2008.

Ng, Kia; Pham, Tran Vu; Ong, Bee; Mikroyannidis, Alexander; Giaretta, David. *Preservation of interactive multimedia performances*, International Journal of Metadata, Semantics and Ontologies 2008 - Vol. 3, No.3 pp. 183 – 196, DOI: 10.1504/IJMSO.2008.023567. 2009.

Ng, Kia; Weyde, Tillman; Larkin, Oliver ; Neubarth, Kerstin; Koerselman, Thijs; Ong, Bee. 3D Augmented Mirror: A Multimodal Interface for String Instrument Learning and Teaching with Gesture Support, in *Proceedings of the 9th International Conference on Multimodal Interfaces*, Nagoya, Japan, pp. 339-345, ISBN: 978-1-59593-817-6, ACM, SIGCHI, DOI: http://doi.acm.org/10.1145/1322192.1322252, 2007

Overholt, D. The Overtone Violin. International Conference on New Interfaces for Musical Expression. Vancouver, BC, Canada. 2005.

Young, D.; Nunn, P.; Vassiliev, A. Composing for Hyperbow: A Collaboration between MIT and the Royal Academy of Music. in *Proceedings of the New Interfaces for Musical Expression International Conference (NIME)*. Paris, France. 2006.

# Legal Contexts of Digitization and Preservation of Written Heritage

Jasmina Smolčić
Veleučilište u Požegi
Vukovarska 17, 34000 Požega, Croatia
jsmolcic@vup.hr

Antonija Valešić
Veleučilište u Požegi
Vukovarska 17, 34000 Požega, Croatia
avalesic@vup.hr

**Summary**

*Purpose of this paper is comparing valid acts of Republic Croatia which regulate matters of preservation of written heritage in correlation to related valid contexts in European union.*
*In continuation of paper we will put an accent on an obligatory example(OP) and protection of author rights, also on advantages, disadvantages and main reasons for formation of digital contents.*
*In this paper we report facts of 27 countries, members of European union linked with digitization of written heritage, and we represent their influence on the further development of digital contents in Republic Croatia.*

**Key words:** juridical context, obligated example, protection of author rights, access to information, digitization of written heritage, digital contents in Republic Croatia, Europeana

## The legal framework of digitizing

Information society development and diffusion of new information technologies and communications have substantially influenced the policies of the member countries as well as their initiatives in relation to the preservation and evaluation of huge wealth of European cultural and scientific heritage. Government of the countries of the European Union ensure that appropriate legislation should guarantee obtaining the authority for storage, protection and handling of electronic materials as a cultural good, the attainment of compulsory copies of electronic materials and access, search and use of such materials as the foundation of a democratic society.
Croatia is not yet an EU member but has devoted considerable attention in the preparation of the group law on the protection of Croatian heritage in electronic

media as well as priorities by way of selection, storage, permanent storage and use. The Croatian legislation emphasize the important laws that provide legal framework for regulation of electronic material matter, and that the following laws: Law Library (NN 105/97), Law on archival matter and archives (NN105/ 97, 65/09), Law of museums (NN 142/98), Law on the protection and preservation of cultural goods (NN157/03), Law on Copyright and Related Rights (NN 167/03), Law on the Freedom of Information Act (NN 172/03).

Croatia in the drafting of the national program of digitizing the archive, library and museum collection is taken into account the principles and experience in the field of digitization of cultural heritage in the European Union. Accepted the Lund principles (Principles Lund, Sweden, 2001. G.) and Lund Action Plan (Action Plan Lund, Sweden, 2001)[1], which define the basic objectives that can be summarized in the basic recommendations: development of a mechanism for promoting good practice in order to harmonization and optimization; Dissemination of European scientific and cultural content, development of reference criteria for the actions of digitization; Encouraging quality and promote the availability of the contents of the citizens of Europe.

On the basic principles of the digitization of cultural heritage defined in Lund, Sweden, adopted and Parm Charter (Charter of Parma, 2003.) Used by the operational support of the European network of Minerva. Network Minerva (Minerva Network) is directed to establish a common European technology platforms and content, making recommendations and guidelines for the digitization, creation of data, ensuring long-term availability and protection of digital content. Year 2002. Minerva network is put forward his proposal the extension of the network to new countries before entering the EU and in Russia and Israel, which is concretrize in MinervaPlus project that began with the application in early 2004. year.

Group of national representatives appointed by the institutions responsible for culture at the level of EU member states, who met in Paris 19th November 2003. The agreement on a definition of the conclusions and principles Parm charters given in the following articles: Art. 1. Intelligent application of new technologies, art. 2. Availability, art. 3. Quality, Art. 4. Protection of intellectual property rights and respect for privacy, art. 5. Interoperability and standards, Art. 6. Inventories and multilingualism, art. 7. Comparative evaluation (assessment), art. 8. Cooperation at national, European and international level, Art. 9. Enlargement, Art. 10. Joint construction of the future direction of society.[2]

[1] http://www.minervaeurope.org/structure/nrg/documents/chapterparma031119final-hr.htm (2009 -06-19)

[2] http://www.minervaeurope.org/structure/nrg/documents/chapterparma031119final-hr.htm (2009 -06-19)

The fundamental law of the related laws governing the matter of protection of cultural heritage is the Law of Copyright and Related Rights (here in after the Act). These kinds of laws includes the legal system of the state to safeguard the authors since it is assumed that only the author of which is guaranteed compensation for the creation will be driven to continue creating. It is important to emphasize that the Constitution of the Republic of Croatia guarantees the protection of the moral and material rights deriving from scientific, cultural, artistic, intellectual and other creative work. Digitization of books, and a growing share of sales of electronic books in total turnover, all are actually setting up the problem of protection of copyright. It is clear that the full protection of intellectual property could also prevent the free access to copyright works, and do not want nor the authors nor the publishers nor the state itself. Copyright under the Act (Article 99) lasts for the whole of life and seventy years after his death, regardless of when the dealer lawfully published. Current programs include digitizing mostly old material, ie the one you should not pay the copyright holder.

Law, and exceptions from the protection for the otherwise protected works may be used without paying the usual fees, and exemptions usually allow for research, personal training, teaching. The artist's name and source must always be mentioned. It is interesting that the law does not mention the possibility that cultural institutions may duplicate the individual works for their needs, such as the protection of materials or other lending institution. Law on Copyright and Related Rights of the Republic of Slovenia provides for the free copying of up to three copies for the personal needs of the individual, but also for the internal use of public archives, public libraries, educational and scientific institutions if the duplication is performing from their own copies.

IFLA in 2000 – the adopted Statement of copyright in the digital environment in which stresses that the legislation on copyright affect almost all library services for users and determine the conditions under which the material can be accessed. Believes that libraries must maintain a balance between the interests of users and their rights to free access and the interests of the author to just compensation for their intellectual work. The basic attitude is IFLA's digital environment that is not so different from the analogue to justify enhanced protection to the author user. The European Union and of the coordinated law (Article 84) allows for limitations to the benefit of individual institutions. The law provides that public archives, public libraries, educational and scientific libraries, which their services can not charge its own copies of copyright work reproduced on any other basis to a maximum of one copy. This allows the digitization of analogue materials for the purpose of its protection, but not access to the material via the Internet or its use. Libraries, archives and museums before digitizing of copyright protected works should negotiate and conclude an agreement with the bearers of rights, or the institutions they represent, about the conditions of accessing digital material. When obtaining a license, there are sometimes difficul-

ties in determining the rights holders and the project states that a portion of material belongs to the acts for which one can not establish the bearer of rights. Happens if the project manager will need to enclose a given statement at the time when such digitized material to become available to the public.[3]
The whole article is the emphasis on the protection of digitization of text and cultural heritage, but all the protection there is no purpose if the written heritage is not available to the public. Croatia, which rests on the foundations of democracy in 2003. year passed the Law on the Freedom of Information Act (hereinafter the Law), whose aim is to facilitate and ensure the realization of the right of access to information, physical and legal persons through openness and public action by public authorities. The law expires in Article 35th paragraph to the right of access to information includes the right beneficiary to claim and obtain information as well as the obligation to the public authorities to allow access in requested information, or to publish information when it is no specific request for such disclosure represents the obligation of a law. According to the rules of Nomotehnic in this law there are exceptions to the right of access to information that are prescribed by Article 8 Law. Is an interesting paragraph 2 type specified in Article 6 determined that the right to deny access to information if there are grounds for suspicion that its publication would endanger the right of intellectual property, except in the case of express written consent of the author or owner.

## Digitization of written heritage

Digitizing is the process of recording, storing and processing the content using digital cameras, scanners and computers.[4]
In recent decades, new technologies and new forms of communication among people have changes in all social levels and the role and importance of libraries in society. Almost no library that does not affect the concept of digitizing or not digitalizes or thinking about how to digitalizes their holdings in order to protect and made available to a broader audience.
Digitization, by itself it seems tempting to all libraries and educational institutions, however, as the rest of it in himself has certain advantages and disadvantages.
What the digitization process, and I make very high quality are larger and faster availability of material, then you can build high-quality copies since the duplication is not lost on quality, quality time with not disappear because digital stuff does not usage. With the advantages of step with them and are certain disadvantages. Comparing with the paper, digital formats are short-lived and are not

[3] http://www.ffzg/.../HorvatPravniaspekti.html (2009-06-21)

[4] Digitization as a Method of Preservation? : Final report of a working group of the Deutsche Forschungsgemeinschaft / H. Weber. 1997. http://www.knaw.nl/ecpa/PUBL /weber.html

readable to the eye. Specifically, documents that are printed on paper without substance and guarded in the cold space with a little moisture can last several hundred years and the paper remains the eye readable without any additional equipment. The big disadvantage of this procedure is its high cost, which includes labor, rapid technology changes that require a switch to new formats and media, then the lack of standards for digital formats. But the biggest drawback is acceptance and long-term data storage.

This question is tackled by prof. Aparac-Jelušić stating that this problem undoubtedly depends on the national strategy digitize each country, and the established plan of raising the compulsory copies of various electronic materials.[5]

Such a step in the process of digitization of which the Professor spoke in 2001, launched in 2004 and had a task to create a Draft National Program for digitizing the Republic of Croatia.[6]

Document to that end seeks to provide a framework for shaping long-term policy digitize, planning and organization of national, institutional and cooperative projects of construction of digital collections in the institutions that want to protect and improve their access to their collections.

In the framework of our National Program for digitization set out some important reasons that encourage such a process. Primarily to the digitization performed to protect the source, increase the availability and ability to use written heritage for the creation of offers, or customer service or complete an existing fund.[7]

Digitization in order to protect the original increases greater than the possible damage during use, transmission, transportation or other proceedings. On the other hand, digitization in order to improve the availability becomes available remotely releasing the digital content via the Internet, regardless of where the user is located in the area reviewed material, it is available to her. The third reason is the creation of new offers. These may be the reason not only to offer new features but also new services. To complete the digitization of the fund allows its users the completion of those portions of the funds that are needed and states that users search for themselves. Digital copy will then be the only form in which the material exists in the institution.

[5] Aparac-Jelušić, Tatjana.Digitalna baština u nacionalnim programima zaštite baštine. URL: http://dzs.ffzg.hr/text/Digitalnabastina-aparac.htm (2009-06-16)

[6] Prijedlog nacionalnog programa digitalizacije arhivske, knjižne i muzejske građe. URL: http://www.daz.hr/bastna/nacProgramDigit(2).pdf. (2009-06-14)

[7] Prijedlog nacionalnog programa digitalizacije arhivske, knjižne i muzejske građe. URL: http://www.daz.hr/bastna/nacProgramDigit(2).pdf. (2009-06-14)

## Croatian projects under the projects EU Member States

Ministry of Culture of the Republic of Croatia, following the digitization of cultural heritage in the world, and especially in Europe, familiar with all the initiatives that are as basic guidelines adopted countries of the European Union and in accordance with the process of digitization of the cultural heritage recognized important objective in the achievement of cultural policy and the cultural development of our country.[8] Understanding to thereby increase the possibility to process the availability of valuable materials at the national and international level, the Ministry of Culture appointed union people who are in charge in order to achieve these goals. And in February 2006 played the plan.

Back two years made the survey in 27 EU countries that represent the statistics of digital cultural heritage in order to detect speed and cost, which marks the process of digitization.[9] From its data can be seen as the digital projects at the low level of financial resources (0.6%). What is actually digitalizes, regardless of where the institution for, whether it is an archives, museums, national and other types of libraries, is actually an old and rare stuff that keeps this process and also became available not only to users of those institutions but also other audiences that are interested in this type of cultural property.

According to statistics, most EU countries there is still no developed plan for the digitization, were even 2/3 (66%). The exceptions are Germany, Estonia, Lithuania, Slovenia and the Netherlands who have written and accurately formulated plans for the digitization of its cultural resources. Surprising is the fact that says that 2/3 of institutions within the surveyed countries, there is no on-line catalogues, so that then we should not be surprised that certain countries have developed plans for digitization.

However, due to high cost of proceedings, most countries are still waiting to digitize their material intended for it.

Croatia has also conducted a survey in order to achieve the real state of libraries and their opportunities in today's world where technology has advanced. The research results showed deflating figures that say that only 8% of the library has access catalogues via the Internet, while only 18% have general access to catalogues.[10] The data only show that we belong to 2/3 EU countries are still struggling to improve unsatisfactory infrastructure and establish a library of on-line catalogues.

But in a good way because we are trying to integrate the library and create a digital collection. This confirms the fact that we first partner outside the European Union countries called the European Partnership. The Republic of Croatia,

[8] Isto

[9] http://www.numeric.ws/uploaded_files/NUMERIC%20Newsletter%20Nov%202008%20Issue%204.pdf (2009-06-16)

[10] http://www.niska.hr/dokumenti/sadrzaj.html (2009-06-20)

only confirms its regional leader position in the process of digitization of cultural heritage, and culture in general. Specifically, the project of digitizing the archive, library and museum collection "Croatian cultural heritage", boosted by the Ministry of Culture, is recognized as an organization that has the desire, opportunity and ability to actively contribute to further development of Europeana. Europeana is a multimedia online library of internet users around the world provides access to more than 2 million books, maps, records, photographs, archival documents, pictures and movies from national libraries and other institutions in the culture from 27 EU countries (digital collections of Spain, France (Gallico) and other countries joined in the European digital library.[11]

Europeana opening a new way of using the European heritage, anyone who is interested in literature, arts, science, politics, history, architecture, music or film now has a free and fast access to the largest European collections using web portals in all the languages of the EU. Recently held conference in the Czech Republic on the subject of digital cultural heritage (26.5.2009), discussed how to extend the European project and European cultural heritage closer to the world and much more is available to witness the rich European cultural diversity. Within this we'll list our most important projects that will eventually be of great importance not only at national level as now but also at the international level when they set on the European.

These are digitized heritage NSK, Silvije Strahimir Kranjčević, Peter Preradović, Đuro Sudeta on the web, Naša sloga, Glas Podravine and the recently completed project of the HAZU.

It is also of great importance and the Society for the Advancement of literature on new media.[12] Represents a non-governmental organization founded with the aim of promoting literature in new media, primarily on the Internet and CD-ROM, as well as the promotion of literature among the users of new media and technology. It is planned to publish 20 of free electronic books. Specifically, potential users should be English, Internet users, then the Croatian diaspora which is contemporary Croatian books unavailable, people who lived in the area of former Yugoslavia, and understand the English language. Cultural exchange with these countries is very low intensity. But this project, the Croatian culture becomes available outside of its borders thus that some books are translated and the Slovenian, Czech, English and German. It is important to mention also that the potential users and people with special needs, especially poorly mobile and immobile persons and partially sighted people because in such a project allows you to access books from home, and today provides software to increase the letters. And finally the last group would be the world's academic community, its slavistic part.

[11] http://europeana.eu/links.php#2 (2009-06-20)

[12] http://www.donacije.info/seek_deatils.php (2009-06-22)

The world is such a project known as Google Book Search[13] originated from two sources publisher and library. As our project, which was probably prompted hereby, including finding books that are very difficult to find. The aim of the project is to create a searchable virtual catalogues of books in all languages so that users found the new books, and publishers find new customers. And finally we have mentioned a project the World Digital Library (World Digital Library) encouraged Congress Library and Google to include and unified all the libraries throughout the world. Initiator Billington said that this project should bring together old and unique materials kept in the U.S. and the Western repository with other beautiful cultures that lie across Europe, including in the more than one billion Chinese people in eastern Asia, India and the world of Islam.

## Conclusion

Digital materials is an essential phase that must reach all of Europe’s cultural institutions to protect and enhance a common cultural heritage, to improve education and tourism and all in order to contribute to the development of new digital content. It is necessary to emphasize that the Republic of Croatia on a series of laws and protection of digitization of cultural heritage is trying to be European and international level but for the conduct of the life laws of all the necessary commitment of all entities, both public bodies responsible for law enforcement and non-government organizations, NGOs and the overall public.

## References

Aparac-Jelušić, Tatjana. Digitalna baština u nacionalnim programima zaštite baštine. URL: http://dzs.ffzg.hr/text/Digitalnabastina-aparac.htm

Digitization as a Method of Preservation? : Final report of a working group of the Deutsche Forschungsgemeinschaft / H. Weber. 1997. http://www.knaw.nl/ecpa/PUBL /weber.html

Prijedlog nacionalnog programa digitalizacije arhivske, knjižne i muzejske građe. URL: http://www.daz.hr/bastna/nacProgramDigit(2).pdf

## Links

http://books.google.com/googlebooks/library.html
http://europeana.eu/links.php#2
http://www.donacije.info/seek_deatils.php
http://www.ffzg/.../HorvatPravniaspekti.html
http://www.minervaeurope.org/structure/nrg/documents/chapterparma031119final-hr.htm
http://www.niska.hr/dokumenti/sadrzaj.html
http://www.numeric.ws/uploaded_files/NUMERIC%20Newsletter%20Nov%202008%20Issue%204.pdf

[13] http://books.google.com/googlebooks/library.html(2009-06-22)

# Managing and Presenting Digital Content in the ARHiNET System

Vlatka Lemić
Hrvatski državni arhiv
Marulićev trg 21, 10 000 Zagreb
vlemic@arhiv.hr

Hrvoje Čabrajić
Avicena Software d.o.o.
Put Supavla 30, 21 00 Split
hrvoje.cabrajic@st.t-com.hr

**Summary**

*ARHiNET is a network information system for describing, processing and managing archival records created in 2006 by the Croatian State Archives and Avicena Software Company. It is a national archival system in Croatia, recognized by the Ministry of Culture as national project, as well as part of the e-Croatia program, the operational plan of the Government of the Republic of Croatia. Development of the archival information and institutions network is a long-term strategic archival service project and ARHiNET implementation enhanced the standardization of the archival institutions work, and enabled establishment of a unique system of processing and description of archival material, as well as data integration and exchange between the institutions that keep archival records. All archives in Croatia are included in the implementation of this unique archival information system that comprises all business processes in archival institutions, together with some other records holders under the state archives supervision. Currently, there are about 700 registered users from more than 150 institutions. Designing, realization, introduction, use, maintenance and development of such a complex program solution enclose permanent activities on system improvement, finding new functionalities and solutions, as well as upgrading of the present ones. During the three years of the system operating, more than 300 versions of program solutions have been developed and put in production, and experiences gained from work and user education led to the development of the version 2.0 that was released in February 2009. In this article authors present solutions and functionalities concerning managing, indexing and presentation of digital content developed and implemented within the ARHiNET program solution.*

**Key words**: archival information system, archives and Internet, digital content, digitization of archives, digital records management

## ARHiNET system

ARHiNET is a network information system for describing, processing and managing archival records created in 2006 by the Croatian State Archives (CSA) and Avicena Software Company. It is implemented as the national archival system in Croatia, with specific objectives:

- establishment of the unique system based on international standards,
- providing efficient and user oriented system of collecting, processing and presenting archival material,
- inclusion of all important elements of archival records management and management of business processes in the archival institutions into one comprehensive system,
- facilitating work of archival professionals,
- standardizing and assuring quality of services and products provided by archives,
- ensuring preservation and presentation of data by using information-communication technologies,
- introducing new technologies and technological solutions in the archival institutions.

ARHiNET is created on modular basis which enables design and implementation of particular modules as separate projects in a relatively short period of time and their continuous integration into the unique information system. Advantages of such a solution are the creation of an integrated base and a unique system of data protection with minimal costs.

ARHiNET system structure comprises of two parts: the open one is intended for external users who want to search databases and catalogues and use other offered services, and the protected part, intended for the employees in archives and other institutions, in which all professional-business processes that define processing and management of archival material are taking place. The program solution consists of several databases organized according to the logic of records type and user type/roles that define access to particular records:

**ARHiNET Public** – database containing records for access by external users. This database is read-only, that is, records are not added or changed, but only retrieved. Database is optimized for faster searching and records retrieval. Records in ARHiNET Public are daily automatically replicated from ARHiNET Master database according to the authorization criteria, i.e. only records available for searching are imported. Main purpose of defining ARHiNET Public database is database search optimization and acceleration, protection of access to the Master database and possibility of simultaneous work on the records so the units of data can be available to public while being edited.

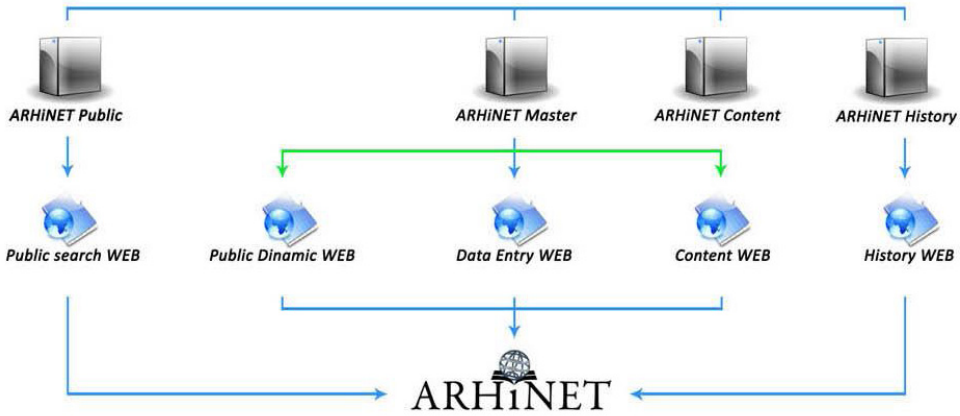
**ARHiNET Master** – database containing records that originate from descriptions of archival and current records, their creators and holders, special types of

archival material, archival documentation, lists and registers and all other data related to business processes management in archives (holding, processing, preservation and use).

**ARHiNET Content** – database containing records that are not directly related to archival records, but are used for administration and facilitating work within the ARHiNET system. It contains elements that can be found on external web pages, professional topics, archival forum and advices, help files, etc. The idea behind it is to allow users dynamic management of all modules and functionalities without programmer's additional help. It also enables system localization (translating user interface) and data import and export.

**ARHiNET History** – database containing all changes made in the units of data (editing, updating, deleting etc). Records life cycle management is defined with MoReq specification and implemented in ARHiNET with purpose of tracking the changes of every single unit recorded within databases, which is very important system functionality. Administrator has exclusive access to this part of the system, and every change of data is recorded in this database: who changed the record, when was the record changed and what data content has been changed.

Scheme 1: ARHiNET Structure




## Digital records inside the ARHiNET system

An important segment of the new information system inside the Internet environment was to provide accessibility of digital content and professional description of digital material. Administration, processing and presentation of digital content are defined into the separate module inside the ARHiNET, and its design was accompanied with adoption of firm rules of managing digital content in the archival institutions, both based on detailed analysis of current policies and practices.

Possibilities of new technologies extended and improved ways of protection and access to archival material, and because digitization of archival records considerably facilitates their availability, archives are faced with mass production of digital records. Besides large quantities of documents and technical problems, issues concerning digitization that archives are faced with mostly refer on selecting and preparing records for digitization, their organization and presentation, as well as availability. Such condition is a reflection of complex nature of archival material and differences in provenances, arrangement and types of records kept in archives.

The ARHiNET enables description of archival material of any type and content (textual, graphic, cartographic, audiovisual, electronic, objects, photographs etc.) according to the international standards for the archival description as well as other relevant specifications. Every record is described with set of general data elements, special data depending on type of records and related tables of additional data which are available in the form of special lists. They are defined in several basic, mutually linked data sets:

- fonds and collections,
- records creators
- records holders

that all together provide data integration and saving the content and context of all records.

Since digitalization for majority of Croatian record creators and holders imply mass digitization, while digital preservation is currently in the professional background, a first step inside the ARHiNET was to provide support for digitizing archival material, and second to implement procedures for preservation and accessibility of “born digital” records. Concerning considerable efforts and resources being invested in digitization, basic principle of ARHiNET program solution – integration of data and reducing of costs – was useful in relation to problems of facilitating large-scale digitization and its cost-effectiveness.

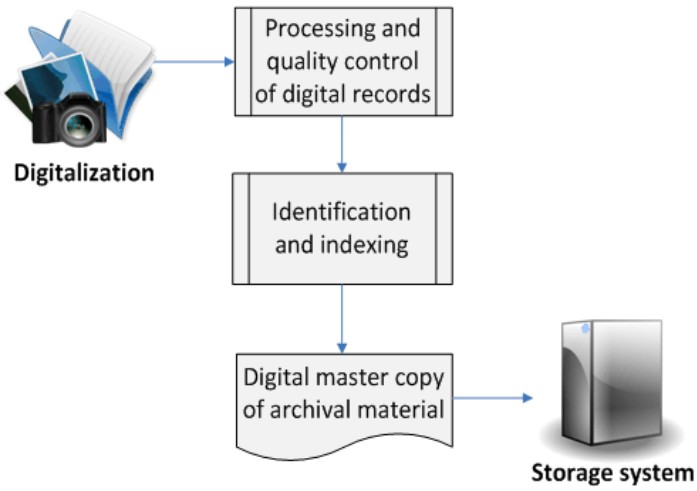
Several business processes were encompassed by digitization procedure in the ARHiNET framework:

- digitization of archival records,
- processing digitized records,
- saving master copies in the storage system,
- automatic creation of web copies in JPEG (or other) format,
- linking of archival units and digital records, and
- presentation of digital content within program solution.

With such a defined and adopted concept, the intact quality and protection of the master copy of digitized records is ensured. External users are granted an access and review of digitized content through web copies, with the possibility of their downloading and printing, while availability of master copies should be granted by the institutions which hold those records.

Basic operational unit for working with digital records inside the ARHiNET system is the digital master copy. Since activities of describing, indexing and managing user copies are depending on quality of master copies system of creation, indexing, and storage of master copies is defined by statutory procedure.

Scheme 2: Procedure of making digital master copies

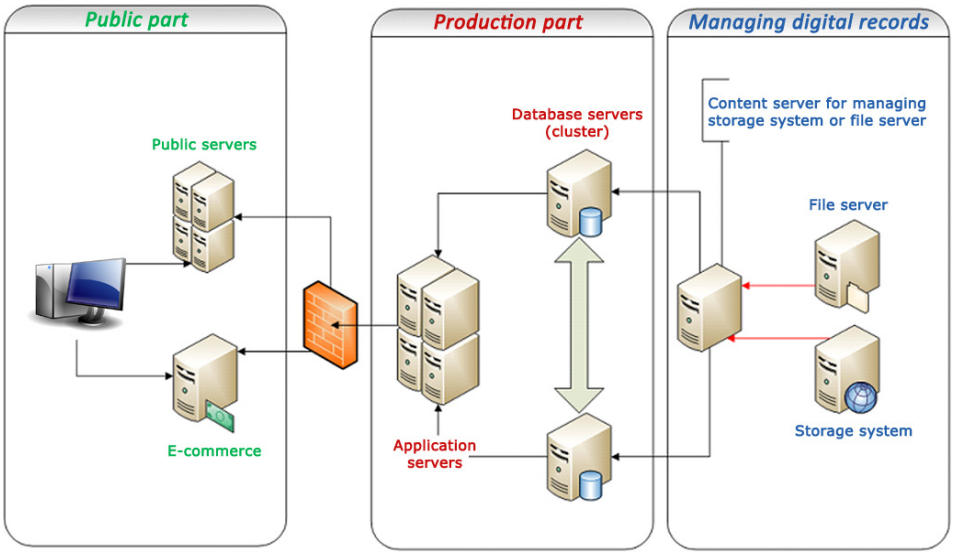



Term master copy, i.e. original digital reproduction of archival unit/record, in the framework of working inside the ARHiNET system is applied for digital copy of a single archival unit which is completely analogous with original record. Technical characteristics of master copy represent optimum of resolution and quality of digital record, depending of type and material of the original record.
Master copy represents digital material from which all user and other types of copies are made. Master copy is kept inside the storage system marked with a unique identifier, and cannot be subsequently altered after the processing, description, controlling and authorization have been finished. Access to master copies is allowed only to operators who are authorised for periodical quality control, migration and making copies.
Every archival and other institution which actively participates in national archival information system must adopt one's own set of rules for all these activities in correspondence to CSA standard procedures for ARHiNET operational work.
Managing digital content is functionally integrated on various levels with other modules inside the ARHiNET system:

- browsing and searching data,
- browsing and searching digital content,
- managing digital content on the storage system,

- integration and administration of ARHiNET databases and modules,
- managing archival material,
- managing records holders.

Scheme 3: ARHiNET system support




## Protection of digital records

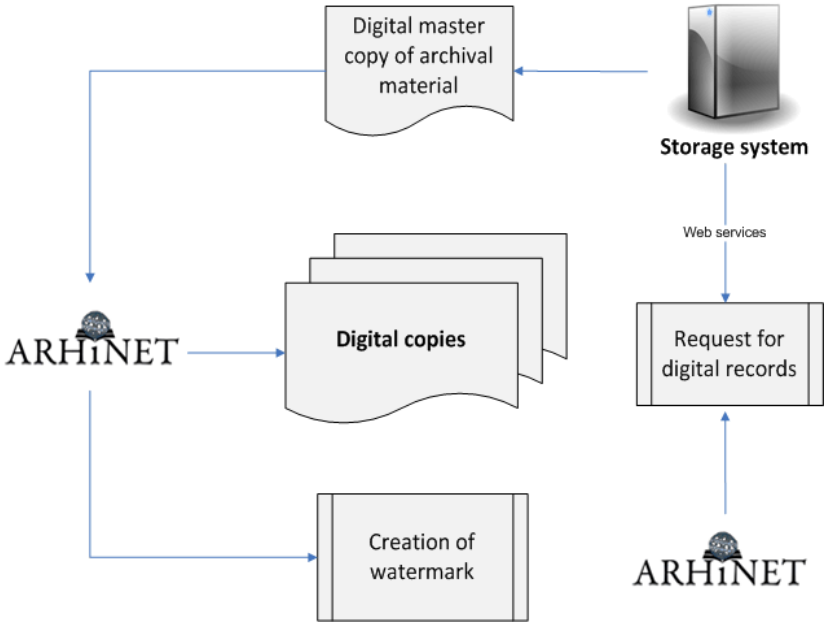
Protection of digital copies of archival material is a significant segment of ARHiNET system because using of digital records cannot be implemented without providing a system of supervision and protection of digital copies from unauthorised copying, multiplication and distribution.

By analyzing present solutions the ARHiNET project team decided to develop its own method for protection and management of digital copies in order to create unique and efficient system of controlling distribution of digital copies and to establish a mechanism of copyright protection for the material, of which originals are held in archives or other institutions/record holders. Functionalities of such a system are:

- providing implementation of visible watermark sign on every single record,
- providing implementation of invisible unique identifier for every single record,
- providing a system of control and tracking of eventual frauds of digital copies,
- recording all changes into database.

After the user has searched the ARHiNET data bases on archival material and has ordered digital copies ARHiNET automatically starts the procedure of digital master copy retrieval from storage system by using web service. After the receipt of virtual master copy, ARHiNET will automatically create digital copy with watermark consisted of name of the record holder and date of the creation.

Scheme 4: Procedure of creation of the watermark




After the creation of watermark, the next sequential part of automatic procession will add a unique identification mark on each copy of the digital record. Each digital copy is defined by pixel scale and each pixel is uniquely defined by its place and colour, while colours are defined by custom palette for colours. Having in mind these settings, it is understandable that in the cases of changes of a single pixel in relation to the master copy ($\pm$ 1) digital copy will be uniquely changed comparing to original. Regarding possible number of combinations (number of pixels × changing colour shades × possibilities of simultaneous changes of one or more pixels) it is done on unlimited number of combinations which allows unique identification and indexing of every single copy, without affected their quality.

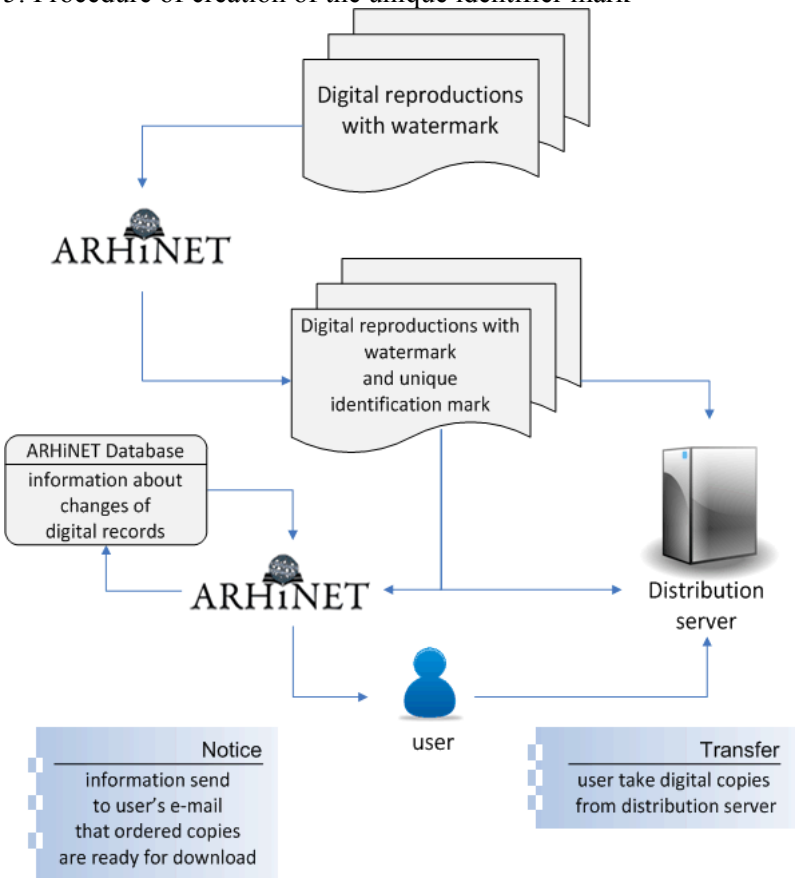
Adding a unique identification mark on each digital copy is recorded inside data base with folowing parametres:

- master copy ID,
- location of altered pixel,

- colour, and
- date of change.

Those data are linked with data about users and order forms of digital copies which all together represent base for documenting use of archival material and tracking changes.

Scheme 5: Procedure of creation of the unique identifier mark

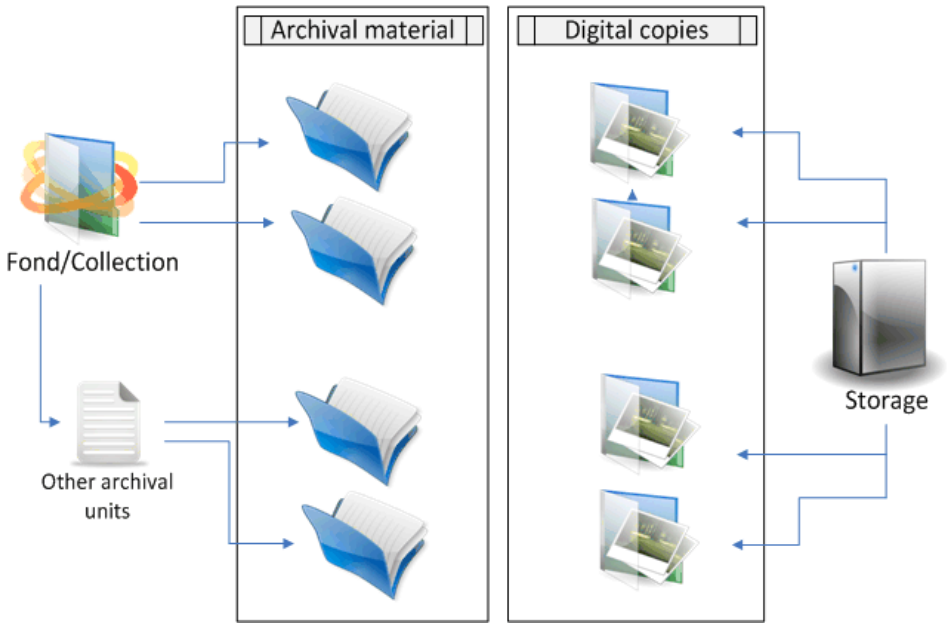



## Providing context for digital material

Data on archival material are inevitably changed in time because of accessions and transfers of new and old records (new fonds and collections are created or existing are supplemented), processing of records and other procedures in the process of archival arrangement. Descriptive elements, following general standards and rules for description of archival material, elaborated and used within ARHiNET are defined not just for archival units, but also for digitized records. Data and metadata made by processing digitized records are connected with data of archival unit's descriptions which enables accuracy in efficiency, as well

as, facilitate every day work. Once those relations are made, every change of data in unit of description will automatically be reflected on digitized records so there is no need for multiple editing of same change. This is realized by providing inside the system a list connecting ID of archival units in ARHiNET with ID of master copies stored on the storage system.

Scheme 6: Connection of archival unit's descriptions with digital copies




This connection requires implementation of content management system for managing storage system, and maintenance of digital master copies data base comprising of two basic tables. One lists elements important for managing digital copies:

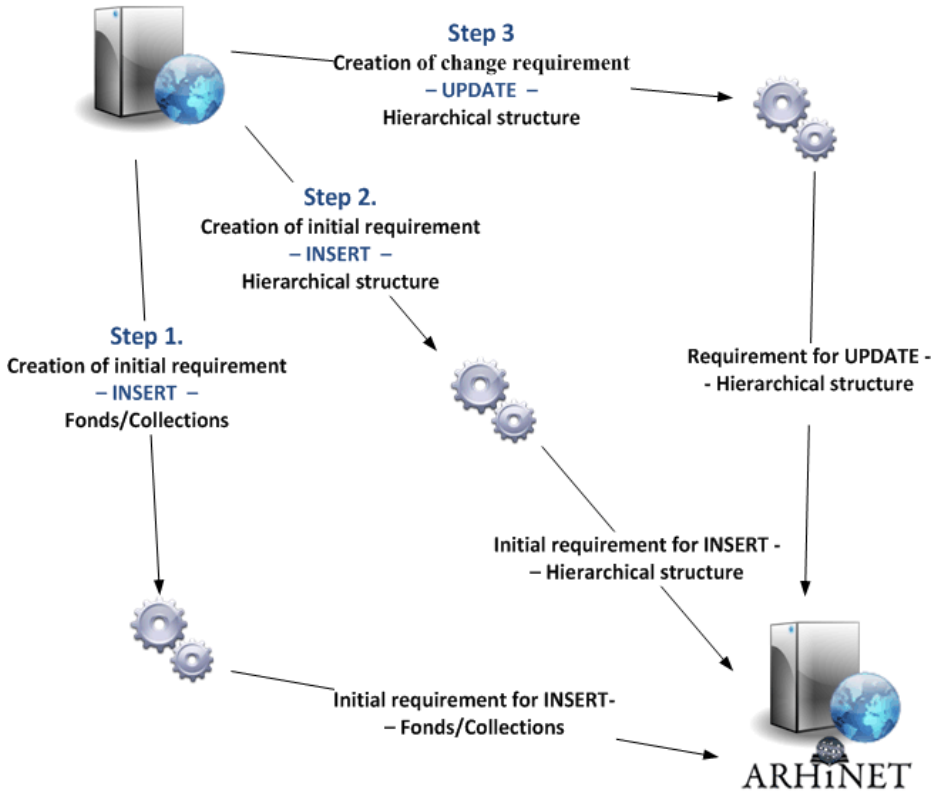
- master copy ID,
- path – position where master copy is placed on the storage system
- order – enables defining sequence of presentation of pictures inside archival unit
- accessibility – information on restriction of access and use ,
- ID of related structure,

while other administrates elements important for defining structure:

- holder and fond/collection structure,
- hierarchical structure of unit of description – level of description
- export/import of data in XML format.

Operational work of all mentioned elements insures ARHiNET and storage system communication through web service.

Scheme 7: XML scheme for connection of archival unit's descriptions with digital copies




## Digital preservation – future development

Production, dissemination and filing of documents in electronic form present some of the biggest problems for modern archives. Although ARHiNET system already supports description and integration of all type of archival material, current development is directed toward upgrading present functionalities with options of online access to digital documents and their search and retrieval. This will be realized through ARHiNET Central Data Poll Model (CDP) which is designed for digital preservation and access to digital data such as databases and multimedia records. It defines XML structure, datasets and files list which enables preservation of structure, content and context of digital record and their management and use in one unique system. Implementation of such system will provide integration of traditional and digital archives, as well as, bring archives closer to their major goal - to ensure authentic, reliable and preservable records, regardless of the form and physical medium they have been created and preserved on.

# Long-term Inactive Data Retention through Tape Storage Technology*

Ivan Vican
Metronet telekomunikacije d.d.
Ulica Grada Vukovara 269d, Zagreb, Croatia
ivan.vican1@zg.t-com.hr

Hrvoje Stančić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
hrvoje.stancic@zg.t-com.hr

**Summary**

*Increasingly the need to retain digital documents indefinitely for legal, administrative or historical purposes is simply leading to a "save everything forever" approach. The authors argue that due to the technological reasons it is much easier to preserve large amount of documents in the electronic than in the paper form. Thus the selection procedures tend to be less restrictive than they used to be. Nevertheless, for most organizations it would be impossible to sustain this data growth forever. Archives, libraries, museums, institutions holding cultural heritage, as well as other companies and firms, are implementing solutions for creating digital archives, digital libraries, digital repositories and other types of storage systems aiming at long-term preservation of digital materials. Most of the data held in such systems are inactive for a long time, i.e. only a small set of data is frequently retrieved. Therefore, due to the specific needs of every organization, the storage planning process and the technology that is going to be used for storage and long-term preservation requires individual approach. The focus of this paper is on the retention of the long-term inactive data through tape storage technology. The authors will discuss current state of the art tape storage capabilities, and their advantages and disadvantages as a long-term storage and preservation solution.*

**Key words:** long-term preservation, storage systems, tape storage, archive, library, museum**,** data, electronic material

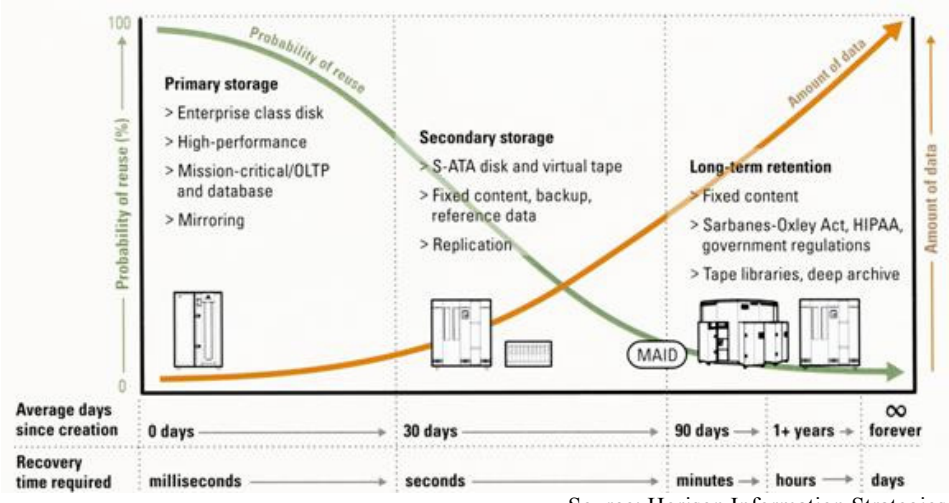
* The authors are solely responsible for the content of this paper. It does not represent the opinion of the institutions they work in, and those institutions are not responsible for any use that might be made of data appearing in the paper.

## Introduction

The need to retain expanding volume of digital data for regulatory, business, personal or heritage purposes is leading to a challenge on which technology data should be stored and archived. Capabilities of digital archives, libraries and repositories are multiple: faster and less complicated manipulation, concurrent access to resources, less space is required, access to high valued material… and they vary among the storage technologies. Other important variable that influences the choice of storage technology is frequency of reuse, i.e. how often data is needed.

Over the time, relevance of data is changing whereby its value is also changing. The implication is that most data will become inactive over the time and the need to access such data is declining. Data lifecycle is providing insight in data value fluctuation, which is measured by frequency in given time. This kind of approach to data lifecycle is specific to business enterprises. By expiration of retention period they will probably delete archived data (Graph 1). However, institutions such as archives, museums, libraries and institutions which focus lies on preservation of heritage need to retain data indefinitely. In other words data has a constant value. Specificity of each system implies individual approach when it comes to retention, preservation and archiving across different information systems and institutions. However, the need to archive and the storage technology that is being used are quite common.

Graph 1: Data reference over time




Source: Horison Information Strategies

As a possible answer to the rising needs of archiving digital data, tape storage technology is offered. Rapid pace of innovations in tape storage technology, especially during the last decade, is reviewing capabilities of long-term retention,

preservation and archiving through this technology. The intention of paper is to debate about advantages and disadvantages of tape storage technologies' capabilities when it comes to the long-term retention, preservation and archiving.

## Tape Storage Technology

First commercially available magnetic tape was introduced in 1952 with the capacity of 1.4 MB. Immediately after introduction, tape replaced punched cards to became the first real removable storage medium. From its very beginning, tape was connected with mainframe system in order to store bulk data. Modern usage of tape storage is mainly connected to backup and archiving systems. The first recording technology used in tape systems was linear recording technology which dominated until middle of 1980s. After that period helical record technology took primate. In the last ten years, both technologies are improved considerably. Leverage has turned to linear technology primarily because of the possibility of higher record density due to the development of the linear serpentine technology and faster data transfer rate.[1] Over the 50 years of development, the tape storage systems came out in numerous standards and formats. Prevailing standard nowadays is the linear serpentine recording technology on tape with half inch wide reel in single-hub cartridge. The half inch tape width is the most frequently used magnetic tape in history. The medium is produced with the metal particle technology and their variations like the advanced metal particle.

By the appearance of affordable disk and optical storage technologies, the magnetic tape storage was pushed down with the future not so clear. Over the last ten years, with the growing need for data storage space, tape storage is recognized as a medium that can bear with these growing challenges. This generated the explosion of tape storage technologies and formats. Among numerous formats and tape technologies, two are representing actual pinnacle in the development of magnetic tape storage: Enterprise-class tape and Linear Tape-Open formats.

### Tape Systems

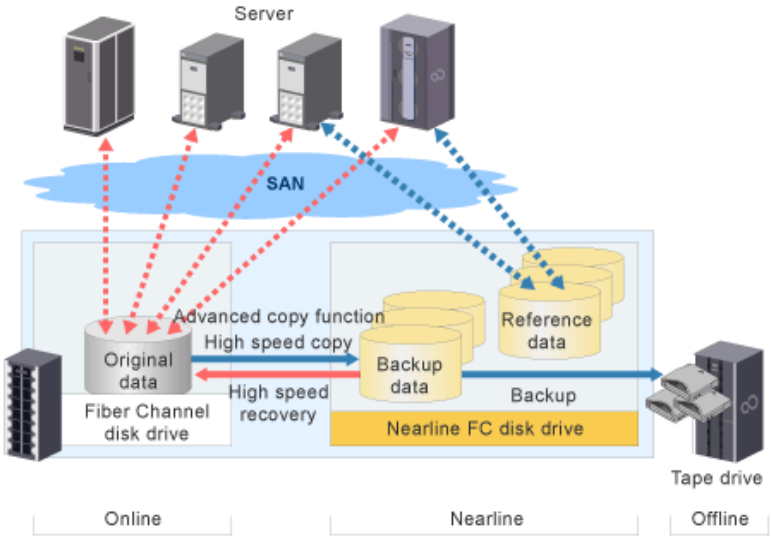
In order to utilize tape medium, proper devices are required. Such device goes by name tape system and it is divided in three types of systems: tape drive, autoloader and library. *Tape drive* represents a basic element of the system as it provides physical and logical structure for reading and writing processes.[2] It allows connection with other devices via SCSI, SAS and Fiber Channel network technologies.

[1] See: Haeusser, Babette; Kessel, Wolfgang; Silvestri, Mauro; Villalobos, Claudio; Zhu, Chen. IBM System Storage Tape Library Guide for Open Systems // IBM Redbook, Seventh Edition, 2008. http://www.redbooks.ibm.com/redbooks/pdfs/sg245946.pdf (last access: 18 August 2009).
[2] Ibid.

*Tape autoloader* consists of a tape drive and an automated tape cartridge exchange system with up to ten tape cartridges in the housing. With added automation feature, the autoloader becomes an autonomous tape drive which does not require constant human intervention in order to exchange tape cartridges.
*Tape library* is able to meet the most demanding archiving needs and because of that it is the most complicated tape system. Such systems have two or more tape drives, depending on the quantity of tape cartridges which can rise up to few thousands. The library layout permits simultaneous access to multiple tape cartridges. The exchange of cartridges is operated by a robotic mechanism and it takes only few seconds to exchange tapes.

**Tiered Storage: Position of Tape Storage**
In the traditional information system tape storage is classified as an offline (archival) tier, as opposed to the disk systems which are online (primary) or near-online (secondary) storage (Picture 1).[3] However, thanks to tape libraries, tape storage is increasingly seen as near-online tier while tape drive and tape autoloader are considered as offline tier.

Picture 1: Tape in a network storage environment




Source: Fujitsu Corporation

Hierarchy of storage classes is enabling consolidation, scalability and faster work of an information system. Storage classes are defined according to the re-

[3] See: Brooks, Charlotte; Byrne, Frank; Higuera, Leonardo; Krax, Carsten; Kuo, John. Redbook: IBM System Storage Solutions Handbook // IBM Redbook, Seventh Edition, 2006. http://www.redbooks.ibm.com/redbooks/pdfs/sg245250.pdf (last access: 20 August 2009).

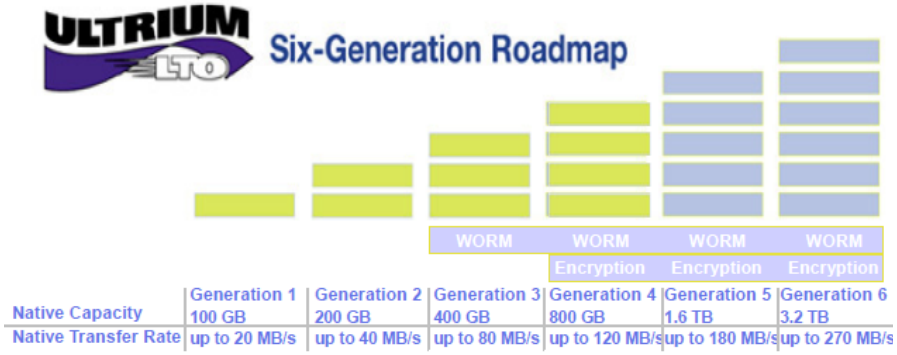
trieval speed, therefore depending to the storage technology. For example, if the data is being accessed on daily basis it will be stored at the primary disk storage tier. Predefined data policy, with the help of storage and archival software, are automating routing processes towards the designated storage device cutting down the load on network and servers. In addition, the storage area network (SAN) technology is enabling direct connection of storage devices with computer systems or with other storage devices. Thereby, it is possible to move data between tiers without the server intervention.

## Linear Tape-Open and Enterprise Tape Storage Technology

Also known as LTO, it was developed at the end of 1990s by LTO Consortium. Main goal was to define and to manufacture the first open format that will offer high-capacity, high performance of tape storage devices to midrange IT systems. The standard format of LTO is known as Ultrium.[4] From 2002 until today, LTO Ultrium is most commonly used tape ever.[5] The reason for that can be found in innovative technology and accessibility.

LTO Ultrium format has defined Six Generation roadmap for growth and scalability (Picture 2). The roadmap represents goals and there is no guarantee that these goals will be achieved. However, each available generation was released with doubled performance and capacity. The latest available generation is LTO-4 released in 2007 with native capacity of 800GB, 120MB/s of native data transfer rate and data encryption at device level. LTO-5 is coming out in 2009.

Picture 2: LTO Generations




Source: LTO Program

Data compression (DC) techniques are quite common in tape storage. LTO-DC is called Streaming Lossless Data Compression (SDLC) and it is able to pass

[4] Ibid.

[5] See: LTO Ultrium format reaches new heights with over 100 million cartridges shipped // LTO, 2008. http://lto.org/pdf/LTO%20100%20Million%20Cartridge%20Milestone.pdf (last access: 20 August 2009).

through already compressed data such as JPEG, MPEG and MP3.[6] LTO-DC algorithm is able to achieve 2:1 compression which gives LTO-4 1.6 TB of compressed capacity and 240MB/s of compressed data transfer rate (Table 1).

Table 1: LTO Tape drive specifications

| | Data transfer rate | Data transfer compressed | Native capacity | Compressed capacity | MTBF |
|---|---|---|---|---|---|
| LTO-4 | 120 MB/s | 240 MB/s | 800 GB | 1.6 TB | 250,000 hr |

Source: LTO Program

Write Once Read Many (WORM) capability was introduced in the third generation. WORM format is designed for long-term and temper-resistant data retention, which is most useful for legal regulations. This is achieved via Cartridge memory chip which holds information about specific cartridge, media in that cartridge and the data on that cartridge.[7]

Compatibility issues are common in tape generations. LTO is designed for backward compatibility for two generations according to the following rules: read/write compatible with one generation prior, read only compatible with two generation prior. For example, LTO-4 is able to perform read/write on LTO-3 and to read from LTO-2 generation. However, it is not possible for LTO-4 to expand the capacity of LTO-3.[8]

In the past, reliability was the weakest point of a tape storage technology. Tape suffered from incorrectly written data, jammed heads and short life period because of mechanical wear out. These drawbacks were solved with the following technical features: read after write verification, surface control guiding mechanism for less damage to tape, error detection/correction for data integrity, magneto-resistive head, large internal data buffer, automated cleaning system and speed matching towards host adapter.[9] Anyway, tapes should be checked once a year for medium deterioration. In case of possible data loss due to deterioration, data should be refreshed, i.e. moved to a new tape. LTO drives are automatically checking tape deterioration every time a tape is mounted.

Advancement in reliability has positively affected the availability and the predicted durability (Table 2) of tape medium. However, the tape wares off after repeated read/write operations which as an effect can have increase number of errors at tape recorded data. The LTO tape cartridge is made for 5,000 load/unload cycles.[10] With the appropriate handling and average usage of four times a week it can last approximately 30 years. This applies only to read op-

[6] See: IBM System Storage Tape Library Guide.
[7] Ibid.
[8] Ibid.
[9] Ibid.
[10] See: Sun StorageTek Linear Tape Open (LTO) Ultrium Data Cartridges // Sun Microsystems. http://www.sun.com/storage/tape_storage/tape_media/lto/specs. xml (last access: 22 July 2009).

erations. If a tape is rewritten in full once a month it will last for approximately 17 years.

Table 2: LTO-4 Tape cartridge reliability

| | Full file passes | Media durability | Archive life |
|---|---|---|---|
| LTO-4 | 260 | 5,000 load/unload cycles | Up to 30 years |

Source: LTO Program

**Enterprise Tape Storage Technology**

The start of a modern enterprise tape storage technology is dated in the first years of 1980s. The technology was primary developed for the needs of mainframe systems. [11] Today, they are still the most common tape technology attached to the mainframe systems with added interoperability towards open platforms as well. At first glance, the enterprise tape storage technology and LTO are quite similar. LTO has succeeded many technical features from the enterprise tape storage technology. WORM capability was first introduced in this technology. Differences can be found in generations of the same technical features. For example, larger data buffer and cartridge memory can be used. When it comes to mechanical components, tape drive and tape cartridge are more robust (Table 3) then LTO. The reason for that rests within the enterprise tape storage working environment.

Table 3: Enterprise class Tape cartridge reliability

| | Full file passes | Media durability | Archive life |
|---|---|---|---|
| T10000B | 360 | 15,000 load/unload cycles | Up to 30 years |
| TS1130 | 300* | 20,000 load/unload cycles | Up to 30 years |

*TS1120

Source: Sun Storage Tek, IBM Corporation

In the mainframe environment tape storage is used for transactional process with application such as LOB, OLTP, CRM and other high duty cycle applications. All this requires lots of starts and stops which puts tremendous physical stress at the tape drive and tape cartridge.

In order to achieve even faster backups and recovery processes, Virtual Tape Library (VTL) technology was developed. VTL is using disk array to emulate tape drives and tapes. Disk is a random access medium which results with higher performance rate. After some time data from virtual tapes that are spinning on disks will be migrated to the physical tapes. This is called disk-to-disk-to-tape (D2D2T). Enterprise tape technology is dominant in the VTL because it is able to sustain heavy duty cycles.

Proprietary IBM 3592 and Sun Storage Tek T10000A/B tape drives and medium are representing top of the peak in Enterprise tape storage technology.

[11] See: IBM System Storage Tape Library Guide.

Sun Storage Tek T10000B was the first available tape cartridge medium with the native capacity of 1TB. It was released in 2008 as a successor to the T10000A released in 2006. T10000A can be reformatted to T10000B capacity. The drive is not compatible with any previously released Sun/STK tape formats. T10000B tape cartridges are available in two formats: *sport cartridge*, with rapid access over less capacity, and *standard cartridge*. Both formats can feature WORM capability.[12]

Table 4: Enterprise class tape drive specification

| | Data transfer rate | Data transfer compressed | Native capacity | Compressed capacity | MTBF |
|---|---|---|---|---|---|
| T10000B | 120 MB/s | 360 MB/s | 1 TB | 2 TB | N/A |
| TS1130 | 160 MB/s | 350 MB/s | 1 TB | 2 TB | 290,000 hr |

Source: Sun Storage Tek, IBM Corporation

The IBM TS1130 represents third generation of 3592 tape technology. The first generation was introduced in 2003, while the third generation came out in 2008. TS1130 uses existing 3592 tapes and provides backwards compatibility, supporting read and write for 3592 generation 2 and read only for 3592 generation 1. Three formats of tape cartridges are available: *short-length* – providing rapid access, *standard* – providing high capacity and *extended*. Cartridges are available in WROM and rewritable format.[13]

## Conclusion and Recommendations

In general, tape storage technology is the most affordable storage technology today[14]. When it comes to archiving, both LTO-4 and enterprise tape systems are suitable. However, LTO-4 format is offering more than sufficient capacity and performances for archiving purposes at lower costs than the enterprise tape systems. In addition, LTO-4 is designed to work with the open system platforms while enterprise tape has remained primarily in the proprietary mainframe systems. Since a lot of information and storage systems in archives, museums and libraries are build using open system platforms, LTO-4 could be a more appropriate solution for such institutions.

It could be suggested to these institutions to hold dual tape systems. The primary system should consist of disk storage which complements tape storage. In that case the data is virtualized at disk storage while it is being retrieved from tape storage. The layout of system should support fast data access and retrieval which grants utilization of archive by users. There should also be the secondary

[12] See: Storage Tek T10000 Tape Drive, Operators Guide // Sun Microsystems Inc. Broomfield : Storage Technical Publications, 2009. http://dlc.sun.com/pdf/96174revEC/96174revEC.pdf (last access: 2 August 2009).

[13] See: IBM System Storage Tape Library Guide.

[14] In US $ per MB of storage.

system, which is called electronic vault and it is usually placed off site. Users should not have access to this archive. The main purpose of an electronic vault is the disaster recovery, archiving for future usage and migration to the new technologies. Only tape storage system, without disk storage, should be sufficient for the needs of an electronic vault.

The most applicable type of a large storage system for archives, libraries and museums is the tape library. Thanks to their modular design, the tape libraries can be easily reconfigured and upgraded to the new tape technologies. Entry level LTO-4 libraries are scalable up to native capacity of 20 TB, 40 TB and 2-4 tape drives. For example, a library which has 5 TB of data equals to approximately 500,000 books[15]. All this data could be stored at six LTO-4 tape cartridges without compression. If the plan is to digitalize 1TB of video content per year in the next 3 years, the library can be extended with additional three cartridges. They are also more reliable than autoloader because the system is set up that in case of one tape drive failure the other will take its place. At the same time inappropriate cartridge handling is minimized[16]. Multiple tape drives could also enable simultaneous write/read operations on multiple tapes. However, current LTO-4 technology will became obsolete in approximately six years. At that time tape drives and tapes inside the present libraries should be replaced with the new LTO generations of drives and tapes. This will be possible due the modular design of LTO libraries and the life of library can thus be extended to approximately ten years. The pricing of entry tape LTO-4 library with two tape drive and twenty tape cartridges is up to 15,000 €. This should be affordable for any institution planning serious digitization or already holding large amount of digital data on unstable media and thinking about migration.

We strongly suggest that archives, libraries, museums and other information institutions involved in the digitization of records and cultural heritage should consider recommended tape technology when building large storage systems. It could provide space and reliability for large collections thus adding to the positive perception and trust among the users and financial supporters while at the same time preparing the ground for future certification processes of the system and the applied storage and archiving procedures.

[15] Approximation: 10 MB per electronic document.

[16] The reported main reason for tape damage is its accidental dropping on the floor.

## References

Blair, Colin; Currie, Julie; Goodall, Eric; McElyea, Kevin; Miller, George; Poston, Ben. IBM Medical Archive Solution // IBM Redbook, First Edition, 2004. http://www.redbooks.ibm.com/redpapers/pdfs/redp9130.pdf (last access: 15 August 2009)

Brooks, Charlotte; Byrne, Frank; Higuera, Leonardo; Krax, Carsten; Kuo, John. IBM System Storage Solutions Handbook // IBM Redbook, Seventh Edition, 2006. http://www.redbooks.ibm.com/redbooks/pdfs/sg245250.pdf (last access: 20 August 2009)

Castets, Gustavo; McLure, Chris; Koutsoupias, Yotta. IBM TotalStorage Tape Selection and Differentiation Guide // IBM Redbook, Third Edition, 2004. http://www.redbooks.ibm.com/redbooks/pdfs/sg246946.pdf (last access: 25 July 2009)

Haeusser, Babette; Kessel, Wolfgang; Silvestri, Mauro; Villalobos, Claudio; Zhu, Chen. IBM System Storage Tape Library Guide for Open Systems // IBM Redbook, Seventh Edition, 2008. http://www.redbooks.ibm.com/redbooks/pdfs/sg245946.pdf (last access: 18 August 2009)

LTO Ultrium format reaches new heights with over 100 million cartridges shipped // LTO, 2008. http://lto.org/pdf/LTO%20100%20Million%20Cartridge%20Milestone.pdf (last access: 20 August 2009)

Reine, David; Kahn, Mike. Clipper Notes: Disk and Tape Square Off Again – Tape Remains King of the Hill with LTO-4. Wellesley : The Clipper Group Inc., 2008. http://www.dell.com/downloads/global/corporate/iar/Clipper_Tape_v_Disk_2008.pdf (last access: 17 August 2009)

Storage Tek T10000 Tape Drive, Operators Guide // Sun Microsystems Inc. Broomfield : Storage Technical Publications, 2009. http://dlc.sun.com/pdf/96174revEC/96174revEC.pdf (last access: 2 August 2009)

Sun StorageTek Linear Tape Open (LTO) Ultrium Data Cartridges // Sun Microsystems. http://www.sun.com/storage/tape_storage/tape_media/lto/specs.xml (last access: 22 July 2009)

# Trends in Preserving Scholarly Electronic Journals

Golnessa Galyani Moghaddam
Shahed University
Department of Library and Information Science, Shahed University, Persian Gulf Highway, Tehran, IRAN, Postal code: 3319118651
g_galyani@yahoo.com

Mostafa Moballeghi*
Karaj Islamic Azad University (KIAU)
Department of Industrial Management, Islamic Azad University (IAU) - Karaj Branch
P.O. Box 31485-313, Karaj, IRAN.
m_moballeghi@yahoo.com

**Summary**

*Scholarly electronic journals have become the largest and fastest growing segment of digital collections for most libraries. Many issues and concerns for managing electronic journals relate to preserving and providing continued access to them. The preserving of scholarly electronic journals is a complex issue with various aspects and is largely different from archiving of print-based scholarly journals. In this paper the author deals with issues concerning archiving of scholarly electronic journals. The purpose of this paper is to identify and discuss different issues related to preserving scholarly electronic journals. The following issues are discussed: differences between print and digital media, shift in the responsibility of archiving, copyright and intellectual property rights, cost of archiving, expertise, selection, redundancy, organizational issues, etc. Technical issues and challenges related to digital preservation include a lack of practical implementations of preservation standards and a lack of technical knowledge, in general, of what information is required to support the digital preservation process within organizations. Nevertheless, digital preservation has received considerably more prominence in recent years, gaining the attention of entities such as national libraries, national archives and other organizations.*

**Key words:** Digital Preservation, Electronic Archiving, Scholarly Electronic journals

* Mostafa Moballeghi is corresponding author.

## Introduction

The preserving of scholarly electronic journals is a new ground of research with various aspects. The purpose of preserving the electronic journals is to ensure that they remain accessible now and future. In this paper the author addresses some of the important issues surrounding preservation of digital resources especially scholarly electronic journals. The author deals with issues concerning archiving of scholarly electronic journals such as differences between print and digital media, shift in the responsibility of archiving, copyright and intellectual property rights, cost of archiving, expertise, selection, redundancy, organizational issues, etc. Following the discussion of challenges and issues by details, the author attempts to trace the trends in digital preservation of scholarly electronic journals.

## Background

The word *archiving* often refers to the process of storing physical objects, generally though not exclusively paper-based, in a physical location, such as a room or a building, to maintain that object's physical integrity and its intellectual context as could be represented by other objects within the archive.[1] Digital archiving has little to do with physical objects or physical storage and it is different from the traditional meaning of *archive*, even some experts prefer to use *digital preservation* instead of *digital archiving*. The vocabulary such as *digital preservation* are being used in handbook of Digital Preservation Coalition[2] to define all the activities employed to ensure continued access to digital resources which have retained properties of authenticity, integrity and functionality. This is a richer interpretation and means that there will need to be more thought and preparation given to what resources are stored, how they are maintained and subsequently accessed and by whom.

In a new definition by S. Rabinovici-Cohen and his colleagues digital preservation comprises two aspects: *bit preservation*, which is the ability to access the bits of the digital record, and *logical preservation*, which is the ability to use and understand the data in the future. In addition, logical preservation must support tracking the provenance of a record and ensuring its authenticity and integrity. Bit preservation issues are mostly well understood and are supported by existing products used to migrate data between different generations of storage technologies. In contrast, logical preservation is still mostly an unsolved problem.[3]

[1] Seadle, 2006.

[2] The Digital Preservation Coalition (DPC) was formed in July 2001 to raise awareness of the issues raised by the need to keep and to re-use for a decade or more digital assets and resources which institutions have created or purchased. Further information on the DPC is available from its website (http://www.dpconline.org/).

[3] Rabinovici-Cohen, et al, 2008.

With a view into technological context today, developments in information technology have obviously changed the traditional system of publishing, distributing, and even the use of scholarly journals. The initial communication for publishing a paper is so quick now, especially with the help of e-mail. An accepted manuscript can be accessed online before the date of publication. Even the patterns of use of scholarly journals are changing in the digital environment.[4] With the impact of information technology preservation of scholarly journals is more complicated than print-based journal and it has social, economics, legal, organizational and technical dimensions.

In January 2008, *Portico* and *Ithaka* invited 1,371 library directors of four-year colleges and universities in the United States to respond to a survey examining current perspectives on preservation of e-journals. A strong response has yielded interesting findings that we now share with the community in the hope they will spark useful discussion among library directors, funders, and administrators regarding strategic library priorities. The survey finds that a large majority of library directors across the spectrum strongly agree or agree that the potential loss of e-journals is unacceptable, and a significant majority believe their own institution has a responsibility to take action to prevent an intolerable loss of the scholarly record. Most larger libraries responding now support one or more e-journal preservation initiatives; however, the majority of respondents from smaller libraries have yet to support any preservation effort and secure permanent access to e-journals for their institutions. The survey shows that this majority is significantly uncertain about their options for e-journal preservation and how urgent is the need to act.[5]

## Differences between Print and Digital Media

Electronic journals have many advantages over print journals. Online access allows for easier searching and retrieval of a topic. Electronic journals can be accessed anywhere (given proper equipment and software), and they have the ability to link to other people and resources beyond locations or *place*. With the greater capability of the digital medium, however, the content of digital files may be lost to future scholars not just because the physical item deteriorates, but because the information cannot be extracted and interpreted correctly. A scholarly journal on the printed page can be viewed and read without any special equipment as long as one knows the language in which it is written. Digital scholarly journals, however, cannot be viewed without special equipment, such as a computer, an Internet connection, and the required software.

With the machine dependency, archiving of electronic journals is more complicated than archiving print journals. The life expectancy of digital media is an-

[4] Liu, 2005; Tenopir, 2005.

[5] Digital Preservation of E-Journals in 2008, 2008.

other issue of concern. The short lifecycle of digital media is a threat for digital archiving because digital media become obsolete much faster than print media. The format of the digital resources can be damaged or lost and may no longer be intact, retrievable, understandable, or displayable. The technology used to store the publication is likely to become obsolete even before that happens.[6] Therefore, continued access to archived resources is a big issue in digital archiving, while *access* was not a big issue to traditional archiving.

## Shift in the Responsibility of Archiving

Information technology has caused substantial changes in the traditional roles of libraries and publishers.[7] One of the major changes is a shift in responsibility of archiving from libraries to publishers in an electronic environment. Historically, archiving records and documents has long been the responsibility of librarians, and publishers largely shade away from this role. Several libraries hold many research journals in print from their first volumes. Few publishers have complete journal collections archived for posterity. In the electronic environment, publishers and producers of scholarly journals are practically undertaking the responsibility of archiving, however. Magie Jones (2003) has pointed out this issue as follow:

> "the transition from purchasing print journals, which the library then owned forever, to licensing access to e-journals for a defined period of time has major Implications for libraries and publishers. In terms of archiving responsibilities, there are no longer any clear-cut distinctions between who should be doing what. There is a lack of clarity regarding responsibilities and uncertainty about precisely what libraries are paying for when they license journals. This has meant that the transition from print to electronic has been more problematic than it might otherwise have been."[8]

With a historic view, the trends in responsibility of digital archiving have been as follow: (1) The people with long traditions of preserving physical artifacts (e.g. archivists, librarians, museum curators) increasingly recognized that it is their responsibility which is now digital. (2) The people with long traditions of managing computer-dependent data sets (e.g. scientific data center personal, technology managers) increasingly recognized that it is their responsibility. There is a debate over responsibility of digital archiving among all stakeholders, but at the case of electronic journal it seems that the publishers practically have to accept the responsibility of digital preservation; as in digital environment, electronic publications (particularly electronic journals) are not physically

[6] Steenbakkers, 2005.

[7] Steenbakkers, 2005.

[8] Jones, 2003.

owned by libraries. Although libraries traditionally owned the resources forever once they paid to publisher, now they license access from the publisher. In fact, licenses are an agreement for legal use of electronic resources not for ownership and publishers will remain the owner of electronic resources. Libraries as subscribers are therefore concerned that publishers do not consider the archiving and preservation of these works and include archiving and perpetual access to back issues in licensing of these works.

## Copyright and Intellectual Property Rights

Copyright and other intellectual property rights (IPR) are two important issues because of their substantial impact on digital preservation. We know that copyright law was originated long before there was a thought of the World Wide Web. Copyright seems to be established well for traditional archiving but not for electronic materials. The copyright and intellectual property rights issues for digital materials are much more complex and significant than for traditional media. If these issues are not addressed adequately, preservation will be curtailed. Both contents of digital resources and their associated software needs to be taken into consideration. It may be noted that the current archiving initiatives (such as JSTOR, Portico, E-Print Repositories, LOCKSS, OCLC Digital Archive, JISC, PubMed Central, Open Access Model, e-Depot, etc), have adopted many divergent approaches to preserving intellectual contents over time because of complexity of copyright law in digital environment.[9]

Copyrights issues have not got a quick solution in digital preservation, as copyright law allows only fair use and it can prohibit a successful preservation to some extents. Some experts suggest to put away copyright in digital preservation or make some changes in law, though it is not easy to do. They reason if current law does not allow copying for digital preservation, the most obvious solution is to change the law and if libraries want to preserve information, they need to be able to carry out the required activities.[10]

Although making changes in law or licensing practice is difficult, rights holders and libraries have to understand and cooperate with each other to progress. There are many stakeholders who may have an interest in archiving electronic journals. Mary Feeney describes in detail the stakeholders as authors, publishers, libraries, archive centers, distributors, networked information service providers, IT suppliers, legal depositories, consortia, universities, and research funders. Feeney also suggested that the relationship of the stakeholder to the digital material archiving needs to be taken into consideration.[11]

[9] Galyani M., 2008.

[10] Muir, 2004.

[11] Feeney, 1999.

## Cost of Archiving

The other important issue in digital preservation is cost of archiving. Digital preservation is essentially about preserving access over time and therefore the costs for all parts of the digital life cycle are relevant. Of course, digital access has many advantages over paper-based or microform access in terms of convenience and functionality, however, providing continued access is an important concern for digital librarians. Cost of digital preservation seems to be much higher than the cost of traditional preservation. Access to digital resource with the rapid technological changes is not easy and needs expert staff and considerable expenditure on technological needs.

Mary Feeney (1999) gives a thorough breakdown of cost considerations based on one of the studies commissioned by the Digital Archiving Working Group (DAWG). She pointed out:

> "One clear message that has emerged is that a great deal of money can be wasted if digitization projects are undertaken without due regard to long-term preservation. It is now relatively easy to produce digital versions of texts or images. However, if there is no plan in place for archiving the digital files, long-term preservation will be expensive, or may even result in the work having to be repeated".[12]

Calculation of costs in digital archiving is not easy, however, is a valuable and necessary task to establish a cost-effective and reliable business model. Costs for *maintaining* the digital copy also need to be considered from the beginning whether those materials are produced as a result of digitising analogue materials or they are *born digital*. It may be noted that other issues such as organizational mission and goals including the type and size of collections, the level of preservation committed to and the quantity and level of access required, and time frame proposed for action should take into consideration.

## Expertise

Digital preservation needs high skilled staff while in the traditional archiving the scenario was different. Montgomery and Hedstrom pointed out

"The need for digital preservation expertise is high: asked to rate staff as expert, intermediate, or novice, only 8 of the 54 institutions considered their staff at the expert level."[13]

It is obvious that the ability to employ and develop staff with appropriate skills is made more difficult by the speed of technological change and the range of skills needed. Continuous training and *learning by doing* are the methods that can be adopted while both methods have their own limitations. Libraries need to ensure their existing staff and members can develop, and continue to develop,

[12] Feeney, 1999.

[13] Hedstrom and Montgomery, 1998.

the range of competencies they need to manage the digital materials in their care.

## Selection

*Selection* is another important issue in electronic archiving. The huge quantity of information being produced digitally, its variable quality, and the resource constraints on those taking responsibility to preserve long-term access make selectivity inevitable for archiving. Traditionally, lack of selection for preservation may not necessarily mean that the item will be lost, but in the digital environment non-selection for preservation will almost certainly mean loss of the item. Although not all resources can or need to be preserved forever, some will not need to be preserved at all, others will need to be preserved only for a defined period of time, and a relatively small sub-set will need to be preserved indefinitely. The decision should be made as early as possible to help save resources for the most valuable digital assets.

In digital preservation where there are multiple versions of an item, decisions must be made in selecting which version is the best one for preservation, or whether more than one should be selected. The importance of selection has been acknowledged by many stakeholders, e. g. the National Library of Canada (NLC)'s guidelines state,

> "The main difficulty in extending legal deposit to network publishing is that legal deposit is a relatively indiscriminate acquisition mechanism that aims at comprehensiveness. In the network environment, any individual with access to the Internet can be a publisher, and the network publishing process does not always provide the initial screening and selection at the manuscript stage on which libraries have traditionally relied in the print environment. Selection policies are, therefore, needed to ensure the collection of publications of lasting cultural and research value."[14]

## Redundancy

In traditional archiving, some level of redundancy with multiple copies was inevitable in different repositories, but the story is different in the electronic environment. Some authors, such as Dale Flecker (2001), believed that there was large-scale redundancy in the storage of journals in the print era, as many different institutions collected the same titles. Theoretically, in a digital environment, a single institution can provide worldwide access and accept preservation responsibility, although there is a debate whether a level of redundancy should exist in the digital environment.[15] In order to avoid the danger of losing access over time, at least one copy of materials should be stored in two different re-

[14] NLC 1998.

[15] Flecker, 2001.

positories. Librarians should make clear who will undertake preservation responsibility and for what period of time. Making appropriate documentation for each level of preservation, selection process and responsibility can give some assurance to have successful preservation strategies.

## Organizational Issues

There are many organizational issues regarding digital preservation. Digital preservation requires new workflows, new skills and close co-operation across different professions ranging from traditional preservation management skills to computing science. The organizational structure to support this is not yet in place.[16] There is lack of clarity in roles and responsibilities between organizations and between different stakeholders. The Digital Preservation Coalition (DPC) carried out a UK-wide survey to assess the nation's preservation needs in 2006. One striking result of the survey is the common lack of clarity in responsibilities for digital preservation, which has been seen by a majority of the respondents as a barrier to digital preservation.[17]

We may be noted that although the situation in digital archiving has been improved since 2006, the organizational issues still need to be taken into consideration. Organizations need to understand digital preservation needs, expertise, technological infrastructures, costs and prepare proper strategies to ensure a successful digital preservations.

## Discussion and Conclusion

The preserving of scholarly electronic journals is a complex issue with various aspects and is largely different from archiving of print-based scholarly journals. With a broad view, preserving of scholarly journals has social, economics, legal, organizational and technical dimensions. The issue of differences between print and digital media, shift in the responsibility of archiving, copyright and intellectual property rights, cost of archiving, expertise, selection, redundancy, organizational issues are discussed and covered in this paper. Digital preservation seems to be a complex process and there are many unsolved organizational, managerial and technical issues that make digital preservation a challenging task for all stakeholders.

Technical issues and challenges related to digital preservation include a lack of practical implementations of preservation standards and a lack of technical knowledge, in general, of what information is required to support the digital preservation process within organizations. The challenges associated with digital preservation are not purely technical. In order for digital archives to be sustainable over time, the organizations responsible for the archives must have ap-

[16] Hockx-Yu, 2006.

[17] Waller and Sharpe, 2006.

propriate expertise, resources, and political/institutional mandate to carry out the work required. Given the cost and complexity of digital archives, as well the potential to exploit the rich sets of relationship across individual collections, coordination of work across social boundaries (institutional, regional, disciplinary, organizational and professional) is also important.

We may note there are some threats for long-term availability of electronic resources. Data mismanagement, technological dependency, media degradation and technological obsolescence have all threatened the long-term accessibility of resources stored in digital formats.

Nevertheless, digital preservation has received considerably more prominence in recent years, gaining the attention of entities such as national libraries, national archives and other organizations. It has to come to be recognized as a legitimate and essential area of research and development. Many stakeholders of scientific publishing have begun to consider importance of electronic archiving and take initial steps to meet their responsibility effectively. The new concerns of electronic archiving led to a series of meetings over the past few years among publishers, librarians, and technologists sponsored by a variety of organizations. In order to manage the archiving issues, different initiatives and projects (such as JSTOR, Portico, E-Print Repositories, LOCKSS, OCLC Digital Archive, JISC, PubMed Central, Open Access Model, e-Depot, etc) were created by various organizations and institutions.

Finally, digital preservation requires new workflows, new skills and close cooperation across different professions ranging from traditional preservation management skills to computing science. There is a need toward more awareness of digital preservation among all stakeholders. This field still is in its infancy.

## References

Feeney, M. (1999), "Towards a National Strategy for Archiving Digital Materials", *Alexandria,* Vol.11, No.2, pp. 107–122.

Flecker, D. (2001), "Preserving Scholarly E-Journals", *D-Lib Magazine* (September) Vol.7, No.9. Available at: http://www.dlib.org/dlib/september01/flecker/09flecker.html (accessed September 20, 2008).

Digital Preservation of E-Journals in 2008: Urgent Action Revisited Released (2008), Available at: http://digital-scholarship.org/digitalkoans/2008/06/06/（Accessed July 17, 2009).

Galyani Moghaddam, G. (2008), "Preserving Scientific Electronic Journals: A Study of Archiving Initiatives", *The Electronic Library*, Vol.26, No.1, pp.83-96.

Jones, M. (2003), "Archiving E-Journals Consultancy - Final Report Commissioned by the Joint Information Systems Committee (JISC)", October. Available at: http://www.jisc.ac.uk/uploaded_documents/ejournalsfinal.pdf.（Accessed August 18, 2008).

Hedstrom, M. and S. Montgomery (1998), "Digital Preservation Needs and Requirements in RLG Member Institutions" A study commissioned by the Research Libraries Group. December 1998. Available at: http://www.rlg.org/preserv/digpres.html（Accessed August 17, 2008).

Hockx-Yu, H. (2006), "Digital Preservation in the Context of Institutional Repositories", *Program: Electronic Library and Information Systems,* Vol. 40, No. 3: pp. 232-243.

Liu, Z. (2005), “Reading Behavior in the Digital Environment: Changes in Reading Behavior Over the Past Ten Years”, *Journal of Documentation*, Vol.61, No. 6, pp.700–712.

Muir, A. (2004), “Digital Preservation: Awareness, Responsibility and Rights”, *Journal of Information Science,* Vol.30, No.1, pp. 73-92.

NLC (1998), National Library of Canada, Electronic Collections Coordinating Group. *Networked Electronic Publications Policy and Guidelines*, October 1998. Available at: http://www.nlc-bnc.ca/pubs/irm/eneppg.htm（Accessed July 17, 2009).

Seadle, M. (2006), “A Social Model for Archiving Digital Serials: LOCKSS”, *Serials Review*, Vol.32, No.2, pp.73-77.

Steenbakkers, J.F. (2005), “Digital Archiving in the Twenty-First Century: Practice at the National Library of the Netherlands,” *Library Trends*, Summer.

Tenopir, C. (2005), “Inundated with Data”, September. Available at: http://www.libraryjournal.com (Accessed December 15, 2008).

Rabinovici-Cohen, S. et al. (2008) “Preservation DataStores: New storage paradigm for preservation environments” *International Business Machines Corporation* (IBM), J. RES. & DEV. Vol. 52 No. 4/5 JULY/SEPTEMBER.

Waller, M. and Sharpe, R. (2006), “Mind the Gap: Assessing Digital Preservation Needs in the UK, Digital Preservation Coalition, York”, available at: http://www.dpconline.org/docs/reports/uknamindthegap.pdf (accessed January 16, 2008).

Original scientific paper

# Monument as a Form of Collective Memory and Public Knowledge

Marija Kulišić
Department of Art and Restauration, University of Dubrovniku
Ćira Carića 4, 20000 Dubrovnik, Hrvatska
mkulisic@unidu.hr

Miroslav Tuđman
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lucića 3, 10000 Zagreb, Hrvatska
mtudman@ffzg.hr

**Summary**

*Monument is a term that occurs in Western cultures as a product of different social processes, and therefore it is not enough to research and document only its materiality, but also its function, which changes depending on the society itself. Likewise, public knowledge depends on the society in which it exists - it is constantly dynamic in terms of its structure and organization. Furthermore, the way the corpus of public knowledge is being formed is changing, just like the public space of contemporary Western societies in which the cultural monuments exist. This is so because public space is shaped by this corpus of public knowledge. The phenomenon that clearly defines the relation between the monument and the public knowledge is collective memory. The feelings of belonging and forming an identity are influenced by collective memory and at the same time, these are some of the main characteristics of both, monuments and public knowledge.*
*Social reality is created by public knowledge, but it is also mirrored in monuments. It is therefore necessary to analyse the relation between monuments and public knowledge, so that on the one hand we can better understand the logic of forming and organizing the corpus of public knowledge in public space, and on the other, clearly explain the active social role of monuments.*

**Keywords:** monument, collective memory, public knowledge

## Introduction

At the beginning of this article, it is important to explain the usage of certain terms, since as in many other fields, including the field of cultural monuments and collective memory, terminology is often translated from other languages,

and usage may vary. For that reason we shall try to eliminate some of possible doubts in this introduction, so that the following text is more comprehensible.
In this paper we do not wish to discuss the distinction between the terms "monument" and "historic monument", as it is in this case irrelevant. We shall use the term cultural monument[1] in the way that it is defined by I. Maroević, which includes both terms. Throughout this article, the term "cultural monument" is frequently replaced by the shorter term, "monument". These two terms may be considered as synonyms. It is similar with the terms social and collective memory. In many articles in Croatian, the term "collective memory" has been translated as "collective memory" *(kolektivno pamćenje),* and we shall use it as such in this paper. However, in Croatian, the term "social memory" *(društveno pamćenje* lately occurs with increasing frequency as more appropriate, and it seems closer to the meaning of cultural monument itself.
This paper is an attempt to clarify the source and relation between cultural monument and public knowledge in contemporary society. The phenomenon of collective memory provides the fitting theoretical framework for such discussion. It is a phenomenon that is being established through communication, it proves belonging to the group participating in identity construction and tends to crystallize itself in space and time through past reconstruction being a part of present and future[2].
However, in the paper we shall show only the basic relations between those two notions. Thus. issues of communication patterns and cultural memory patterns shall be left out. Recognizing such patterns in forms of public message such as cultural monuments, is a subject for further research.
Let us consider an assertion that cultural monuments are forms of collective memory and well organised sets of messages that format public knowledge in public space.[3] The verification is even indicated in original meaning of the word monument as any artefact erected by community of individuals to commemorate or to remind future generations of individuals, events, sacrifices, practices or beliefs, and therefore the monument has a direct influence on memory function[4]. F. Choay claims that the past that is invoked and called forth is not just any past: it is localized and selected to a critical end, to the degree that it is capable of directly contributing to the maintenance and preservation of the identity of an ethnic, religious, national, tribal, or familial community. The very es-

[1] Maroević, I. Uvod u muzeologiju. Zagreb: Zavod za informacijske studije, 1993; p. 139.

[2] Halbwachs M. On Collective Memory. / Lewis A. Coser. (ed). Chicago: The University of Chicago, 1992; pp. 41-120.

[3] Tuđman, M. Informacijsko ratište i informacijska znanosti, Zagreb, 2008., p. 93

[4] Choay, F. The invention of historic monument. Cambridge: Cambridge University Press, 2001; p. 6.

sence of the monument lies in its relationship between the present and the memory, in other words, in its anthropological function.[5]

## Notes on the etymology and history of monuments

The bond between cultural monuments and collective memory is easily perceivable if we analise one of the first interpretations of the meaning of monuments. In the time when this term appeared in Western Europe, famous French writer on architecture and esthetics, Quatremère de Quincy (1755-1849), defined a monument as a sign that evokes events, objects and individuals, and the word itself is applicable to many works of art, from smallest medals to largest edifices[6]. According to Quincy, the term "monument" expresses luxury and brilliance that is particularly suitable for public edifices which are designated primarily to serve peoples needs. He recognises instinctive compatibility between an edifice and its purpose, and for him art is just an outside attribute of monument that indicates its validity and purpose.[7] His reflection on monuments and their role in societies and cities probably came out of his earlier studies of ancient architecture. Quincy's other works include a comparative study of Egyptian and Greek architecture.[8]

In ancient Egypt, the main form of collective memory consists precisely of monuments - temples around which collective memory was organised and materialized. More about this is written by contemporary egyptologist J. Assman who, analyses, within his studies about cultural memory, analises how societies of ancient civilizations like Egypt, Israel and Greece relate to monuments, from written texts to great temples. Assmann believes that Egyptian temple presents builded memory and also a medium for state to manifest itself and the eternal order. For that reason, in Egypt, unlike in other ancient cultures, *monumental discourse* was established. The state disposes with temples and at the same time with media that make collective identity visible and at the same time ensures continued duration in collective memory, even after death. For an individual in Egyptian society the *monumental discourse* was the way for salvation that secures the place in eternity. Assmann states that *monumental discourse* is discourse of *merit* (k word *ma'at* that also means justice, truth and order), *eternity* and *political affiliation.*[9]

---

[5] Ibid. p. 7.

[6] Quatremère de Quincy, A. Restauriranje, Restaurirati, Restituiranje, Ruina, Ruine, Spomenik. // Anatomija povijesnog spomenika / Špikić, Marko. (ed.). Zagreb: Institut za povijest umjetnosti, 2006; p. 86.

[7] Ibid. p. 87.

[8] Špikić, M. Uvod. Kontemplacije i invektive. // Anatomija povijesnog spomenika / Špikić, Marko. (ed.). Zagreb: Institut za povijest umjetnosti, 2006; p. 15.

[9] Assman, J. Kulturno pamćenje. Zenica: Vrijeme, 2005; p. 198.

Egypt is one of few cultures that placed the accent on visual media as a main bearer of collective memory. Architecture and art, same as hieroglyphic writing like a form of art in ancient Egypt, served to shape sacred public space that ensured durability, attachment, truth and justice. In this text we shall not give forms of collective memory in Israel, Greece or elsewhere because there collective memory was primary organised around ancient texts and oral tradition. Still, example of Egypt is important for us to understand that form of materialised collective memory is not arbitrary, but together with written texts and oral traditions unexceptionally constructs collective memory as a phenomenon we recognize today. That becomes even more important in contemporary culture where visual media are becoming leading devices for communication of different kind of messages. Although, multimedia starts to be even more represented and this includes different forms of expression including audio, textual, tactile, dancing, preforming etc. Medium is any form with which we can transmit a message. This was apparently understandable for old Egyptians who built their temples to be sacred places where works of art will be made, hieroglyphic text written, where festivals and rituals will take place that will, together with temples itself, send over explicit messages of Egyptian culture and civilization through space and time, present, past and future.

English writer J.Ruskin (1819.-1900.) who is responsible for initiation and development of the conservation idea, in his famous book *The Seven Lamps of Architecture,* stresses memory as the sixth pillar of architecture, since according to him, that is the purpose of making buildings to be more lasting, more monumental and worth of memory, and therefore decoration on them are more vivid, metaphoric or imbued with historical meaning.[10] Ruskin believes that there are only two strong conquerors of human forgetting: Poetry and Architecture.[11] However, architecture includes poetry and it is more powerful in the process of memory, because "we have learned much more about Greece from the ruins of its sculptures than from its sweet poetry or military historians."[12]

## Monument as a form of communication object

Let us return to the present and consider a recent assumption that modern memory is, above all, archival, and according to P. Nora it relies entirely on the materiality of the trace, the immediacy of the recording, the visibility of the image. Our age has become obsessed with the archive, and it exists only through exte-

[10] Ruskin, J. Luč pamćenja. // Anatomija povijesnog spomenika / Špikić, Marko. (ed.). Zagreb: Institut za povijest umjetnosti, 2006; p. 292.

[11] Ruskin was studying ancient architecture and was fascinated by Babylon architects.

[12] Ruskin, J. Ibid. p. 291.

rior scaffolding and outward signs.[13] From the point of view of information science, we could say that our age has become obsessed with INDOC objects that resist entropy and forgetting, since their function is to memorise, that is, to endure and transmit and save given content or potencial message through time.[14] Therefore it is understandable that in contemporary age, cultural monument is defined as a document, which makes the monument the medium and the message at the same time.[15]

German and Comparative Literature professor A.Huyssen believes that in our days we can not discuss personal, generational, or public memory separately from the enormous influence of the new media as carriers of all forms of memory.[16] Huyssen notes that we are going through transformation of temporality, processes of time-space compression, brought on by complex intersection of technological change, mass media, and new patterns of consumption, work, and global mobility. Space and time are fundamental categories of human experience and perception, and our society, because of the informational and perceptional overload combined with a cultural acceleration, attempts to secure some continuity within time, to provide some extension of lived space within which we can breathe and move. According to Huyssen, cultural needs in a globalizing world can be reduced to slowing down rather then speeding up, expending the nature of public debate, trying to heal the wounds inflicted in the past, nurturing and expanding liveable space rather than destroying it for the sake of some future promise, securing "quality time".

Local memories are intimately linked to articulation of those needs, nevertheless they express the growing need to, as Huyssen calls it, spatial and temporal anchoring in a world of increasing flux.[17] And although Huyssen believes that monuments, like any other media, are a socially changeable category[18], we could assume that for a while cultural monuments (at least the existing ones) will still be perceived as that kind of anchor, just because they are fixed in space

[13] Nora, Pierre. Između Pamćenja i Historije. Problematika mjesta. // Kultura pamćenja i historija / Brkljačić, Maja; Prlender, Sandra. (ed.). Zagreb: Golden marketing-Tehnička knjiga, 2006; str. 30.

[14] Tuđman, M. Struktura kulturne informacije. Zavod za kulturu Hrvatske, Zagreb, 1983; p. 57.

[15] Ibid; p. 140.

[16] Huyssen, A. Present Pasts: Urban Palimpsets and the Politics of Memory. Stanford: Stanford University Press, 2003; p. 18.

[17] Ibid. p. 21-27.

[18] Huyssens postulate is that only if we historicize the category of monumentality itself can we step out of the double shadow of a kitsch monumentalism of the nineteenth century and the bellicose antimonumentalism of modernism and postmodernism alike. Only then can we ask the question about monumentality in potentially new ways, about which this paper is not about. Ibid. p. 40.

and time. We could say that monuments are references for spacial and temporal interpretation, because on the one hand they are lasting and they transmit messages through space, but also they are themself a message in the space, and on the other hand they evoke memory and remembering sending off messages through time taking over completely the role of media in which communication with users is actualized.

As we have mentioned before, transmission and memorizing messages through time is a task of any INDOC object - in this case a monument - which we define as communication object within a structure of communication process. There are different kinds of communication objects[19], but we classify monuments as spacious and plastic, that is, ambiental objects[20] which are according to their characteristics of communicational form lasting, unreplicative and analog. Communication objects within communication process are defined as messages,[21] so cultural monuments are in fact lasting, unreplicative and analog ambiental messages, or more precisely, forms of collective memory.[22] Memory lives and it is maintained in communication, since we only remember what we communicate and what, according to Halbwach, we can locate within the social frameworks of memory[23]. Seen that way, cultural monuments are not just admirable virtuous works of art and architecture, but they have an active social role in creating and communicating messages of public space and collective memory as well.

## Public knowledge and cultural monuments

Let us be reminded that knowledge is symbolic product which is defined by four functions: cognition, communication, information and memory. Different types of knowledge are historical categories that often disappear, change, or die together with the societies and circumstances in which they appear.[24] In the present time we distinguish open access knowledge and controlled knowledge, but in this paper we shall focus on the first category, the open access knowledge, which can be divided in two different types of knowledge: social and public. Social knowledge is defined as knowledge that includes tradition, historical and cultural heritage of all nations, but also civilisation inheritance that society col-

[19] More about communication objects see Tuđman, Ibid. pp. 56-68.

[20] Ibid. p. 58.

[21] Ibid. p. 57.

[22] Tuđman, M. Informacijsko ratište i informacijska znanosti, Zagreb, 2008; p. 94.

[23] Assman, J. Ibid. p. 43.

[24] Tuđman, M. Svijet znanja i sudbina knjige // Aleksandru Stipčeviću s poštovanjem. Zagreb: Zavod za informacijske studije, 2008; p. 181.

lects, stores and exchanges with other cultures and communities.[25] On the other hand, public knowledge is dominant knowledge in public information space, and it represents the dominant form of knowledge in Western cultures.

Public knowledge is a type of knowledge that has not existed at all times, but was institutionalised through history, mainly coinciding with the expansion of printing, but also appearance of books and their role. It is based on new attitude towards knowledge as a result of knowledge valorisation through communication process, after which the knowledge is accepted by scientific and social community by consensus[26]. It is important to note that in contemporary Western societies dominant knowledge is determined by public knowledge in public space. Space that is prevailed by public and mass media, and their primary function is to control and supervise public knowledge.[27] The person controling the public space also controls total outflow of knowledge, and is able to ensure the dominance of messages with which he interprets and reaches his personal goals.[28]

According to the type of knowledge, cultural monuments could be classified as a part of social knowledge, taking into consideration just its physical features and the time of monument duration and development, in other words, its structural and functional identity[29]. However, a monument cannot be separated from its public life[30]. Since monuments exist, that is to say, "live" in public space, we have to take into account their context that indicates spatial and social components which are both present in their chronological and social time.[31] Thus, cultural monuments are components of the public space organisation and therefore their coded messages, and thus the coded collective memory, are an integral part of public knowledge. Moreover, since there is a consensus concerning public knowledge, or to be more precise, majority approval for dominant messages of governing elite, coded collective memory in monuments is an integral part of social order.[32] As we have have mentioned before, the same object situation was present in ancient Egypt with *monumental discourse* that was not only a com-

[25] Ibid., p. 182.

[26] Tuđman, M. Obavijest i znanje. Zagreb: Zavod za informacijske studije, 1990; p. 108.

[27] Ibid.

[28] Ibid. p. 187.

[29] See Maroević, I. Ibid. pp. 134-135.

[30] Young, James: Tekstura sjećanja. // Kultura pamćenja i historija / Brkljačić, Maja; Prlender, Sandra. (ed.). Zagreb: Golden marketing-Tehnička knjiga, 2006; p. 213.

[31] Maroević, I. Ibid. p. 135.

[32] Tuđman, M. Informacijsko ratište i informacijska znanosti, Zagreb, 2008; p. 95.

munication medium, but medium in which the state manifestes itself and the eternal order[33].

Therefore we can say that in the present age, the fate of monuments as well as the fate of collective memory is in a way determined by dominant knowledge in public space, i.e. the public knowledge that is being controlled by public and mass media.

## Conclusion

The meaning and the social role of cultural monuments change depending on the context that is in our time defined by public and mass media. Nevertheless, monuments are not passive observers although they are fixed in space and time, but because of their ability to be communication objects, they are actually active participants in social events. Cultural monuments as communication objects are not a replica of the reality, but they constitute that reality[34], and not just any kind of reality.

Since monuments are a form of collective memory, they reconstruct the past in such a way that they are taking part in the present and the future. And just because of such features, it is very hard to imagine near future without cultural monuments. Even Ruskin himself believed that the value of monuments is in their ability to continuously testify about people, about the passage of time, with the purpose of linking together forgotten and future periods, and they almost build the identity of entire nations by adding their affections[35].

In the globalization society, the concept of cultural monument once again confirms its essence as its function to construct and maintain identity. What is more, the difference in the relationship that certain monuments have with the past, the memory and the knowledge condition the way they are protected and preserved[36].

[33] Assman, J.Ibid. p.198.

[34] Tuđman, M. Struktura kulturne informacije, Zavod za kulturu Hrvatske, Zagreb, 1983; p. 77.

[35] Ruskin, J. Luč pamćenja. // Anatomija povijesnog spomenika / Špikić, Marko. (ed.). Zagreb: Institut za povijest umjetnosti, 2006; p. 302.

[36] Choay, F. Ibid. p. 13.

## References

Assman, Jan. Kulturno pamćenje. Zenica: Vrijeme, 2005.

Choay, Françoise. The invention of historic monument. Cambridge: Cambridge University Press, 2001.

Halbwachs Maurice, On Collective Memory. / Lewis A. Coser. (ed). Chicago: The University of Chicago, 1992.

Huyssen, Andreas. Present Pasts: Urban Palimpsets and the Politics of Memory. Stanford: Stanford University Press, 2003.

Nora, Pierre. Između Pamćenja i Historije. Problematika mjesta. // Kultura pamćenja i historija /Brkljačić, Maja;Prlender, Sandra. (ed.). Zagreb: Golden marketing-Tehnička knjiga, 2006, pp. 21-43.

Maroević, Ivo. Uvod u muzeologiju. Zageb: Zavod za informacijske studije, 1993.

Quatremère de Quincy, Antoine-Chrysôstome. Restauriranje, Restaurirati, Restituiranje, Ruina, Ruine, Spomenik. // Anatomija povijesnog spomenika / Špikić, Marko. (ed.). Zagreb: Institut za povijest umjetnosti, 2006; pp. 71-88.

Ruskin, John. Luč pamćenja. // Anatomija povijesnog spomenika / Špikić, Marko. (ed.). Zagreb: Institut za povijest umjetnosti, 2006; pp. 287-314.

Špikić, Marko. Uvod. Kontemplacije i invektive. // Anatomija povijesnog spomenika / Špikić, Marko. (ed.). Zagreb: Institut za povijest umjetnosti, 2006; pp. 13-24.

Tuđman, Miroslav.Struktura kulturne informacije. Zavod za kulturu Hrvatske, Zagreb, 1983.

Tuđman, Miroslav. Obavijest i znanje. Zagreb: Zavod za informacijske studije, 1990.

Tuđman, Miroslav. Epistemologijski postav informacijske znanosti // Odabrana poglavlja iz organizacije znanja. Zagreb: Zavod za informacijske studije, 2004; pp. 102-111.

Tuđman, Miroslav. Svijet znanja i sudbina knjige // Aleksandru Stipčeviću s poštovanjem. Zagreb: Zavod za informacijske studije, 2008; pp. 167-219.

Tuđman, Miroslav. Informacijsko ratište i informacijska znanosti, Zagreb, 2008.

Young, James: Tekstura sjećanja. //Kultura pamćenja i historija /Brkljačić, Maja; Prlender, Sandra. (ed.). Zagreb: Golden marketing-Tehnička knjiga, 2006; pp. 197-216.

# Liberating Narratives – Museums and Web 2.0

Darko Babić
Faculty of Humanities and Social Sciences.
Deprtment of Information Sciences
Ivana Lučića 3, Zagreb, Croatia
dbabic@ffzg.hr

Željka Miklošević
Faculty of Humanities and Social Sciences,
Deprtment of Information Sciences
Ivana Lučića 3, Zagreb, Croatia
zmiklose@ffzg.hr

**Summary**

*The paper attempts to put into relation the social web environment and museums. In a retrospective view on the formation and dissemination of knowledge in museums, discernible are several stages also connected to the public access and the social role of these cultural institutions. The virtual environment is seen yet as another stage which with the Web 2.0 technologies creates possibilities for a redefined role of the museum at a more socially profound level that might be characterized as multivocal and collaborative.*

**Key words:** Museums, knowledge formation, Web 2.0, users

As it has been known throughout history, a revolution usually occurs in stages and is often not fully recognized until majority of people unintentionally accept the pattern of behaviour, thus creating a new world. Although the World Wide Web was born in 1989, it took next five to six years for the general public to start using it. This new tool for spreading information enabled contents to be available, though theoretically[1], to everybody. They were digitally published on web sites as predominantly static, not interactive and proprietary[2]. In short, those websites included "read-only" material and provided one-way flow of information. Further development of the web that happened in the next half decade brought about a new method, the one which emerged in the business sector

[1] The reality is that even today the luxury and advantage of the web technology is not available to everybody.

[2] O'Reilly, Tim. "What is Web 2.0." O'Reilly Media. September 30, 2005. http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (June 20, 2009)

and offered new possibilities of information exchange. For Tim O'Reilly and web entrepreneurs "the bursting of the dot-com bubble in the fall of 2001 marked a turning point for the web" since the only businesses that survived were those that used the technologies now popularly know as Web 2.0[3]. This new label seems to mark a specific advance in Web technology that differs from its predecessor, the so called, Web 1.0. However, it rather refers to a set of techniques for Web page design and execution, and represents a model defined by a change of mindset. The shift from Web 1.0 to Web 2.0 is a direct result of the change in the behaviour of those who use the World Wide Web. As opposed to Web 1.0, Web 2.0 is characterised by a new approach in content building in which numerous users simultaneously produce and consume information. The main attribute of these protagonists of the Web 2.0 era is that "they have embraced the power of the web to harness collective intelligence"[4]. In other words, users become contributors to various forms of content, whether by producing or distributing them, providing their comments or marking content with their own individual (associative) meaning. They are free to share and reuse, openly communicate and enforce decentralisation of authority. All they produce becomes incorporated into the structure of the web and available to all users to discover and become engaged. The web grows as a result of collective activities bringing improvement to services with a growing number of users.

"…participatory production of content, collaborative categorization of sites by freely chosen keywords…"[5] reflect an alternative direction in the formation and organization of knowledge in the virtual sphere where *expertise* is shifting from the position of *a few* to the position of *many*, as well as from *professionals* to *amateurs.*

In addition to the business sector, such developments strongly influence the cultural sector, more specifically museums, which had stepped into cyber space somewhat bashfully but soon realized the possibilities of "unlimited space for display and communication"[6]. However, apart from being of relevance, digital environment also presents a challenge to some of the basic museums tasks such as collecting, preserving, and communicating which has additionally been stressed by the emergence of Web 2.0. The museum, as traditionally authoritative institution, faces yet another reality to which it is called to adapt in a newly created social context, even a virtual one. With regard to its role in the creation and dissemination of knowledge this transformation follows a series of modifications that occurred in different periods and contexts in the course of its history.

[3] Ibid

[4] Ibid

[5] Ibid

[6] Müller, Klaus. Museums and Virtuality. // Curator. 1(2002), 45; p.23

## The Story of Museum

In today's world which is dominated by both mechanical and digital reproduction it would almost be impossible to rely solely on authenticity of materiality in attempts to communicate museum meanings. Still, museums have popularly been regarded as treasure houses of authentic objects. One reason for that might lie in the fact that the concept of collection and material culture studies have formed constituent elements of museum practice and created an "image which points to the central importance of objects, to the material world, in constructing narratives of cultural authority"[7].

Looking back into the history of not only the museum but also to forms of its etymological predecessors, objects were of central importance and their use, as in the case of Greek statuary, ranged from religious veneration (ancient Greece) to trophies of conquest and cultural veneration (ancient Rome) and profane aesthetic contemplation (Christianity). However, it was the Renaissance culture that brought back the interest in Aristotle's methods of study[8], the culture in which the Medici Palace emerged only to be later cited as "the identity of origin for European 'museums' and for European collecting practices"[9]. During the sixteenth century, collections, mostly owned by princes and scholars, became rather commonplace in Europe. Those *cabinets of curiosities,* as they have often been called, were organised differently but what they all shared was a "strange" juxtaposition of objects we would today call unsystematic. Their representational system reflected idiosyncrasy, the subjective worldview of the person who owned them. The underlying principle of collecting was founded on relativity, aesthetic impression, resemblance, emphasising the magical aspects of the world. Moreover, the prince/scholar stood in the centre of his collected objects dominating over them, and at the same time dominating the world. They constituted a unity in which various objects were placed in one space expressed the sameness (books were placed next to antiquities, natural things, instruments…). Clearly, we today would consider all those things as disparate elements which need to be separated exactly because of their difference. The *naturalia* would, therefore, find their appropriate place in an institution such a natural history museum, just as books would be placed in libraries.

Community, that is, the "museum audience" was in this particular case of little significance than it was going to be in subsequent periods since in the Renaissance world, the order of the macrocosm (the world) resembled that of the mi-

[7] Witcomb, Adrea. Re-Imagining the Museum – Beyond the Mausoleum. London: Routledge, 2003, p.102

[8] During his travels to the island of Lesbos, Aristotle began collecting, studying and classifying botanical specimen

[9] Hooper-Greenhill, Eilean. Museums and the Shaping of Knowledge. London: Routledge, 1995, p. 23

crocosm (man) and secrete knowledge gave the prince specific powers and advantages in assuming symbolic position in the ruling strata by presenting their collections to few relevant people. In The Birth of the Museum, T. Bennet speaks about the social role of the cabinets of curiosities as public representation of the princely person not to the masses but to an exclusive group of people, such as foreign emissaries and political opponents. Objects were safely stored in a hidden part of the house and kept away from the view of the masses.
The beginning of the seventeenth century brought a change in the relationship between objects in the scholarly approach to material culture and in the context of knowledge formation. What used to function as a unifying principle of analogous correspondence now translated into the practices of contrasting in order to discover identities and differences by way of measuring and ordering. "Order established elements, the simplest that could be found, and arranges differences according to the smallest possible degrees. Difference was defined by visual morphological features, rather than by the interpretation of hidden resemblances. The seeing of things was now privileged over the reading of things. To see was to know"…. The ordering of things by means of signs constituted a knowledge based upon identity and difference"[10].
This new epistemological paradigm induced changes in the practice of exhibiting. Thus, what used to be placed together in order to illustrate the variety and richness of the world and "tied" together according to hidden resemblances, was now classed into the same family on the basis of morphological features. What looked the same was placed together and it was important to finish the series. The spatial arrangement in the exhibiting spaces divided objects into "sections". In other words, the space became strictly defined and controlled and one in which "deliverers of knowledge" became scientists whose primary objective was to acquire knowledge through newly founded institutions - scientific societies. "Institutionalisation was seen as a more productive way of pursuing scientific enterprise"[11] and it called for the distribution of the knowledge to people, much the same as churches spread religion. Along with scientific societies that started emerging in Europe in the end of the seventeenth and the beginning of the eighteenth century science was to be facilitated by the establishment of museums (among other institutions such as libraries and botanical gardens, to name just a few). However, during the period of their articulations, museums were still exclusive institutions where access was socially restricted to the newly formed bourgeois public who was differentiated by the attendance to museums from the "rough" general public. It was not before the mid nineteenth century that museums as institutions opened to the whole population. They were born within a newly formed cultural and historical context and connected to new po-

[10] Hooper-Greenhill, Museums and the Shaping of Knowledge, p. 135

[11] Ibid, p.142

litical and social purposes. The museum as cultured place was transformed, as Bennett says, into a space of homogenisation at the same time setting the classes apart. "In its new openness the museum was envisaged to be an exemplary space in which the rough and raucous might learn to civilize themselves by modelling their conduct on the middle-class codes of behaviour to which museum attendance would expose them"[12]. On the other hand, it was a place that offered an opportunity to people to share what used to be kept and enjoyed by princes, kings and scholars. Thus it symbolically proclaimed liberty, democracy and triumph over tyranny. It was a programme of the government to "manage" population. Naturally, new display practices needed to be introduced in order to achieve this goal. A great amount of material that was coming to museums required care and administration. Private property became the property of the state which the state could filter, reorganize and transform. New narratives supporting democracy and egalitarianism were formed out of the royal and aristocratic ones. Physical objects became exponents of the "true history" based on structural relations among exhibited museum objects. "The selection of items that were to be displayed and the separation of these from the items that would be stored or otherwise disposed of, led to the development of new categories of inclusion/exclusion, and to new "curatorial" processes"[13]. Curatorial research and organisation of collections were the practices that gave the authority to the institution resulting in a divide between the subjects who produced knowledge in the hidden rooms of the museum and the subjects who consumed it in the public space. Museum objects played a central role in the formation of the grand narrative through their constructed relationship, facilitated by the academic principles and the space itself which reflected the scientific principles of the "order of things". The museum entered into space between History and various histories showing difference, development and progress, as opposed to the earlier, eighteenth-century arrangement by strictly visible features such as size and material.

A shift in the view on objects occurred in the modern museum where material things no longer represented themselves in their physical existence and historical development. Their connection based on taxonomy was expanded to form relation to human beings. Material things were composed as objects through their connection to histories, stories and people. An important change in respect to this new relationship of objects removed the object from its central position and placed an emphasis on narration. The social history approach introduced a practice in which object of little monetary value but important for the life of a certain community started to be collected together with the object that once demanded traditional connoisseurship. Ideas are now more important and curators

[12] Bennett, Tony. The Birth of the Museum. London: Routledge, 1997, p. 28

[13] Hooper-Greenhill, Museums and the Shaping of Knowledge, p. 179

choose objects which illustrate the story they form. Various sorts of information have received significance that emphasise not only the "representative" but functional and commonplace. New disciplines of social sciences such as sociology and psychology deployed in museum practice prompted another point of interest for the institution – the museum user/visitor. The public function of the museum which used to be disciplinary and instructional now offered new educational methods and place for entertainment.

This change of perspective and the emergence of "post-museum" which Hooper-Greenhill takes it should "play the role of partner, colleague, learner (itself), and service provider in order to remain visible as an institution"[14], was to a great degree supported by electronic technologies and mass media which created a modern public sphere. This sphere was in later stages reinforced with the virtual sphere on the Internet that has presented itself as a "non-hierarchical space of communication which encourages social interaction"[15]. In such an environment, the museum was bound to become responsive to the new social structures and to satisfy their quench for information. The focus on objects had yet to be changed. Nevertheless, objects have remained important but in as much as they "emit" information which can be communicated through different media. However, the break in the concept of *auratic* and authentic quality of the physical object occurred before the advent of the electronic age. Mechanical reproduction gave birth to the notion of multiplicity of the museum object and increased access to them[16]. In the similar way digital reproduction, (the Internet) obliterates the unique existence of objects in time and space thus in a way influencing changes in the issues of ownership and access. However, in addition to the displacement of both objects and their place, digital reproduction created possibilities for simultaneous processes, such as merging of information about dispersed museum items, or linking objects with distant sites, perhaps of their origin. What is most significant about it is that the architectural space of the museum can now be extended into cyberspace where the exclusiveness of the institution is to a great degree undermined, and which presents an environment that facilitates less restricted access to people as well as opportunities for sociocultural interaction that is just "a click away". This openness was to a certain degree concurrent with the increased role of the museum in the community and with the development of "user-centred philosophies for the creation and deliv-

[14] Hooper-Greenhill, Eilean. Museums and the Interpretation of Visual Culture. London: Routledge, 2004, p. ix

[15] Witcomb, Adrea. Re-Imagining the Museum – Beyond the Mausoleum. London: Routledge, 2003, p. 109

[16] Benjamin, Walter. The Work of Art in the Age of Mechanical Reproduction. // Illuminations – Essays and Reflections / Arendt, H. (ed). New York: Schocken Books, 2007

ery of networked information resources"[17]. Systematic documentation that became the fundamental principle of curatorial practice in the museum of the second half of the 20th century created easy ways of information search and retrieval on the web. The virtual environment that many museums took use of in digitally representing themselves and their collections furthers the insistence on ideas and story-telling as well as it presents advantage in the form of an unlimited space and new creative ways for display, communication and knowledge sharing as added features to the physical museum reality.

## Influence and Use of Web 2.0 in Museum

Museums found themselves engulfed by the new media in the mid-1990s when the need to face the potential of the new technology was recognized by ICOM in the 1995 policy statement that recommended museums to actively contribute to internet information with their own programmes and collections in order to more thoroughly play their role in society[18]. However, it is with the occurrence of Web 2.0 that more particular changes in the interaction between museums and their users could be encouraged. Museum audience have for the last two decades been in the focal interest of the institution and information about users has been obtained through surveys as part of a museum market research. The main purpose of the research is to "identify the users" (…) "determine their needs, characteristics, attitudes and behaviour"[19] since the visitors' book no longer provided sufficient information. The Web 2.0 applications, in this respect, offer even greater possibilities of acquaintanceship as well as interaction. Unlike the first stage of museums' extension in the virtual world where museum activities and events were broadcasted to the internet users and completely produced by the institutions themselves, at the second stage with Web 2.0, museums could proffer a democratic approach to their audiences, draw things out of them and call on their expertise. The Web 2.0 applications, namely blogs, wikis, social bookmarking or tagging and podcasts present this opportunity in the context of museums.

Blogs as a conversational mode present new means of creating wider community of museum users and outreach possibilities. With markedly participatory characteristics, blogs might be taken as a way of encouraging free comments from the public unsolicited by the museum provided that the institution takes them into consideration. Museums can use blogs in a more traditional way as a

[17] Trant, Jennifer. Social Classification and Folksonomy in Art Museums: early data from the steve.museum tagger prototype. http://dlist.sir.arizona.edu/1728/01/trant-asist-CR-steve-0611.pdf (July 27, 2009)

[18] Parry, Ross. Recoding the Museum – Digital Heritage and the Technologies of Change. London: Routledge, 2007, p.93

[19] Šola, Tomislav. Marketing u muzejima. Zagreb : Hrvatsko muzejsko društvo, 2001, p.148

sort of virtual visitors' book, or as a tool for promoting discussion on specific museum activities. An advanced way could include picking of the mental attitude of the community and using it as a feedback, taken integrally, and incorporating it in the managerial mindset.

Similarly to blogs, wikis help capture and collect community's knowledge making it accessible to everyone . The museum thus becomes a place for discussion and functions on the principle of peer collaboration and editing. It actively invites participation allowing the public to give their own knowledge about a certain item or topic related to the museum collections and objects. Therefore, it could be a powerful tool for creating stories based on collective memory of community/ies which strongly resembles O'Reilly's concept of harnessing collective intelligence. In addition to predominantly textual tools for sharing content, podcasting includes audio and video material to be created and broadcasted on the web. With this technology, a big section of distant audience can download the latest material automatically from the web. The use of podcasts in museums can enable people to explore a sample of the collection or enjoy virtual tours of the museum while they're on the move. Yet another advantage of this software is that it might encourage people to produce their own audio and video material in relation to museum collections, but also museums to create audio and video narratives based on people's contribution. Unlike the above mentioned applications which allow contribution in content creation, something that is more in line with museum interpretation/communication, tagging is an activity more related to documentation. Documentation has always been an essential part of museum practice based on taxonomy and standardization of data which gave the museum its authored voice. With tagging, this strictly professional approach to objects preceding interpretation might be moderated in a way as to allow users to create additional means of access to museum objects. This new sort of openness of the institution is important in promoting social engagement with its audiences. Yet, another way which is a sort of infiltration into the virtual community is social networking which helps in building a "relationship with an on-line community so other institutions, organisations, groups of interest use our [museum] data to create more complex and richer on-line experiences (and vice versa) "[20]

In that way the museum can act as a subject in the virtual sphere, thus adapting to different socially and culturally defined groups of social networks (thus being active in managing its own virtual identity), or as an object in which case it is a reference that people use in order to form a group around it (thus being passive).

[20] Methven, D.; Hart, T. Organisational Change for the On-line World – Steering the Good Ship Museum Victoria. // *Museums and the Web 2009*/ Trant, J.; Bearman D. (eds). Toronto: Archives & Museum Informatics, 2009. http://www.archimuse.com/mw2009/papers/methven/methven.html (June 14, 2009)

All the mentioned means of (virtual) public engagement with museum activities can serve to empower users to create knowledge and share it with others, but at the same time their involvement can encourage institutional advancements in the matters of knowledge creation and presentation, and, in effect, democratization. Social media can stimulate engagement of museum users and encourage them to develop a relationship and response to museums that could be meaningful for the institutions themselves. Seen as a platform built by the new media, museum can be a site where users could establish cultural dialogue between themselves and in such a way prompt a two-way relationship between museums and communities – the one in which the museum is formed by the communities and in which communities shares the values which are being formed by them in collaboration with the museum. On the basis of such an interactive platform, the museum becomes a site "for exploring the complex subjective relationship between individuals, communities, objects and power within the broader project of social transformation"[21].

## Instead of conclusion

Seen in retrospective, museums as social institutions have undergone a process of different degrees of openness and accessibility to the public – from the moment they opened their doors to restricted groups of people in end of the 18th century wider circles during the 19th century, to the 20th century and the present day when the institution became accessible to everyone. Each step of the way material culture, which has been the basis of museums in most cases, was intricately connected to the formation of knowledge and degrees of accessibility. The recent shift of focus from physical characteristics of objects to narratives is not to say that materiality ceased to be relevant and important in museums and that the virtual sphere with digital representation of objects and emphasis on ideas represent a break with the past practices. On the contrary, museums need to use both material and immaterial sources of knowledge but in a way as to invite a multiplicity of interpretations by allowing the community to step in. Thus, regarding the connection of material objects and the creation of knowledge in museums four main aspects could be discerned – those closely tied with ownership (objects exclusively owned by an individual who forms the knowledge on the collection), connoisseurship (objects analysed only by experts and presented to the public from a single and unquestioned cultural perspective), contextualisation (museum narratives formed by professionals based on objects in connection to people's experiences) and collective collaboration (narratives weaved together by professionals and community).

[21] Graham Janna; Yasin Shadya. Reframing Participation in the Museum: A Syncopated Discussion. // Museums after Modernism: strategies of Engagement / Pollock, G.; Zeman J. (ed). Blackwell Publishing Ltd, 2007, p. 167

Museums have always had one of the essential roles in the formation and dissemination of knowledge. The changes at the beginning of the 21st century create space for the museum to accept new possibilities of bringing into play entire collective memory, provided by an each individual’s contribution, for the benefit of humankind, in order to prove and retain its significant role in society.

## References

Abt, Jefrey. The Origins of the Public Museum. // A Companion to Museum Studies / Macdonald, S. (ed). Blackwell Publishing Ltd, 2006, 115-134

Baker, T.; et al. Collaborative History - Creating (and Fostering) a Wiki Community. // *Museums and the Web 2009* / Trant J.; Bearman D. (eds). Toronto: Archives & Museum Informatics, *2009*, http://www.archimuse.com/mw2009/papers/baker/baker/.html (August 13, 2009)

Benjamin, Walter. The Work of Art in the Age of Mechanical Reproduction. // Illuminations – Essays and Reflections / Arendt, H. (ed). New York: Schocken Books, 2007, 217 - 251

Bennett, Tony. The Birth of the Museum. London: Routledge, 1997

Fyfe, Gordon. Sociology and the Social Aspects of Museums. // A Companion to Museum Studies / Macdonald, S. (ed). Blackwell Publishing Ltd, 2006, 33 – 49

Gates, J., Case Study: New World Blogging within a Traditional Museum Setting. // *Museums and the Web 2007* / Trant J.; Bearman D. (eds.). Toronto: Archives & Museum Informatics, 2007. http://www.archimuse.com/mw2007/papers/gates/gates.html (August 13, 2009)

Graham, Janna; Yasin Shadya. Reframing Participation in the Museum: A Syncopated Discussion. // Museums after Modernism: strategies of Engagement / Pollock, G.; Zeman J. (eds). Blackwell Publishing Ltd, 2007, 157-172

Hooper-Greenhill, Eilean. Museums and the Shaping of Knowledge. London: Routledge, 1995

Hooper-Greenhill, Eilean. Museums and the Interpretation of Visual Culture. London: Routledge, 2004

Methven, D.; Hart, T. Organisational Change for the On-line World – Steering the Good Ship Museum Victoria. // *Museums and the Web 2009*/ Trant, J.; Bearman D. (eds). Toronto: Archives & Museum Informatics, 2009. http://www.archimuse.com/mw2009/papers/ methven/ methven.html (June 14, 2009)

Müller, Klaus. Museums and Virtuality. // Curator. 1(2002), 45; 22-33

O’Reilly, Tim. “What is Web 2.0.” O’Reilly Media. September 30, 2005. http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (June 20, 2009)

Parry, Ross. Recoding the Museum – Digital Heritage and the Technologies of Change. London: Routledge, 2007

Šola, Tomislav. Marketing u muzejima. Zagreb : Hrvatsko muzejsko društvo, 2001

Trant, Jennifer. Social Classification and Folksonomy in Art Museums: early data from the steve.museum tagger prototype. http://dlist.sir.arizona.edu/1728/01/trant-asist-CR-steve-0611.pdf (July 27, 2009)

Witcomb, Adrea. Re-Imagining the Museum – Beyond the Mausoleum. London: Routledge, 2003

# Digital Heritage Archiving in India: A Case Study of Panjab University Library, Chandigarh

Anita Chhatwal
Librarian, Sri Guru Gobind Singh College,
Sector-26, Chandigarh-160019, India , Phone: 91-172- 2792754, 91- 9855101044
anita_chhatwal@yahoo.co.in

Preet Kanwal
Sr. Lecturer, Department of Computer Science, Sri Guru Gobind Singh College,
Sector-26, Chandigarh-160019, India , Phone: 91-172- 2664628, 91-9814611365
preetkc@indiatimes.com, preetkc@hotmail.com

Payare Lal
University Assistant Librarian, Panjab University,
Sector-14, Chandigarh-160014, India , Phone: 91-172- 2534556, 91-9417870507
payarelal_1@hotmail.com

**Summary**

*The Indian libraries are undergoing dramatic transformation by getting converted into digital libraries thereby fulfilling the objective of preserving national heritage and culture and making it globally accessible.*
*Purpose: The aim of the paper is to call for digitization and preservation of manuscripts in library area.*
*Methodology: This paper highlights the procedure of digitization undertaken by Panjab University Library to preserve the manuscripts and provides a case study.*
*Findings: This paper explores that the only way out for preservation and access to manuscripts is digitization*
*Implications: Suggestions for best possible preservation access and strategy is provided.*
*Originality: Section 1: Discusses the methods of preserving the manuscripts*
*Section 2: Explores the concept and need of digitization*
*Section 3: Highlights the digitization policy and criteria for selecting the documents for digitization.*
*Section 4: Provides valuable insight into the infrastructure required for digitization*
*Section 5: The digital library initiatives taken in India are explored.*
*Section 6: A Case Study of the initiative of Panjab University Library in digitizing the manuscripts with an objective to preserve, conserve and upgrade the manuscripts through digitization and digital preservation.*
*Section 7: Conclusion/Suggestions*

**Key words:** Manuscripts, Archives, Methods of Preservation, Digitization, Panjab University Library, Digitization Policy, Digitization Initiatives

## Introduction

The preservation of the documentary heritage has to be seen in the broader context of managing what we have inherited from the past in a way which will allow us to hand it over to the future. Documentary preservation makes sense only when we take accountability of the preservation of the objects, built-in environment and create landscapes for the same. But we cannot preserve everything, and indeed should not seek to do so. As a society needs and desires change, the political and social expression of its understanding of, and need for, its inheritance also changes. In responding to change, each generation makes its own contribution to the heritage of the future. How we decide what must be preserved, and how we preserve it in a meaningful way, is the question which lies at the heart of preservation management.

## Methods of Preserving the Manuscripts

Physical preservation of manuscripts is an intricate procedure. Indian paper manuscripts may last four hundred years while palm leaf manuscripts may, under the best of environment, last seven hundred years. The various methods for preserving the manuscripts in India are:

- Photography, especially microfilming, and photocopying (xeroxing) but they are recommended if the documents are to be preserved for only a few decades. Manuscripts could also get lost during microfilming. Scanners are comparatively time-consuming, thus can damage the deteriorating manuscripts.
- One of the best techniques for digitization could be first to microfilm the manuscript, then medical or high-scan it through high-definition film scanner.
- Digital still cameras are another method but it is costly. A few pages could be copied and then these images had to be downloaded into a computer before other pages could be copied.
- DV format camcorder was introduced in June 1988 to digitize the manuscripts directly.
- In 1999 a simple and easy to learn and use digital still cameras was launched which met the requirements for in-house digital copying.

## Concept of Digitization

The process of converting information or input data in any physical form such as a print, images, photographs, video tapes, etc into machine readable digital form of computer processing is known as digitization. Digitization is the dominant means for comprehensive dissemination of information.

Taking into account the variety of alternatives for preservation of manuscripts alongwith the expansion and widespread applications of ICTs and networks, it has been universally acknowledged that digitization of library materials is cost effective and thriving improvement in storage, preservation, search, retrieval, dissemination and ensure efficient usage of information in the age of information technology. According to Yerkley (1996) digital libraries are electronic libraries in which a large number of geographically distributed users can access the contents of large and diverse repositories of electronic objects.

## Need of Digitization

Information explosion has lead to constraints in finances and manpower in the libraries all over the globe and it is unfeasible for any library to acquire whatever has been published worldwide, thus giving way to the introduction of a new concept of automation, e-resource sharing and networking which has improved and accelerate the working of the library. Resultantly the reliability, productivity augmented, consequently saving the time of users and staff. It is then, the notion of digitization evolved in the field of libraries in three stages:

Traditional Libraries - Automated Libraries
Automated Libraries - E-Libraries
E-Libraries - Digital Libraries / Digitization
Digital Libraries - Virtual Libraries

In India, large numbers of university and college libraries have introduced diigitization in their institutions. The need for initiating digitization arises:

- To take advantage of the ICT facilities for e-sharing of resources worldwide
- To access resources from remote areas
- To access the information and digital resources 24X7 anywhere, anytime instantly.
- Multiple number of users can access the information simultaneously
- Simple and easy search and retrieval techniques
- Cost effective
- Obstacle of time, space has been reduced
- Rapid and flawless access to geographically distributed
- User friendly display of information
- Any number of copies can be generated with the help of digitization

## Digitization Policy and Criteria for Selecting the Documents for Digitization

Vogt O'Connor (2000) recognized three segments in the selection of documents namely, nomination, evaluation, and prioritization. Legal aspects and stakeholder concerns are talked about and also made accessible the checklist for the

appraisal of resources, including factors such as contributor limitations, condition of materials, and the legitimacy of the items.
Similarly, De Stefano (2000) accord prime importance to copyright issue. Although sometimes, it is not possible for libraries to obtain copyright permission, resultantly the project gets crumbed.
Smith (2001) recommends that the purpose of digitizing the documents should be lucid as to whether it is for conservation, or some other purpose.
The Digitization Policy for India should be extensively within the Information Policy only. While framing the digitization policy, one has to take into consideration certain factors keeping in view the dissemination of information and services:

- Education and enduring learning for general public
- To enhance the information access for participation in the socio-economic field.
- Vocational training and employability.
- Cultural heritage preservation by acquiring knowledge about the conventional set up
- Historical substantiation and history.

The following is the guideline for selecting the documents and framing the policy for digitization:

- Selection Principle – Those contents which are important in terms of intellectual implication, distinctiveness, relevance and as per demands of the users. As most of the content in India is available in varied formats and media, therefore, it is also significant to consider the same.
- While undergoing the process of digitizing the documents for long-term preservation, certain set standards and guidelines have to be followed for the same
- Quality perspectives for digitization, access and preservation.
- As India is having a diverse culture, numerous numbers of languages and scripts, there is a need to generate suitable metadata as per specified standards for the access of such diverse type of documents in order to gratify the information needs of the users.
- OCR facility as per specified standards for Indian languages may be developed
- Keeping in view the information explosion and large number of information resources, the issues like Intellectual Property Rights (IPR), piracy problems, copyright issues and other legal aspects have to be taken into consideration while formulating digitization policy.
- After the process of digitization, the procedure for preserving the original documents have to be specified.

- Expenditure associated with digitization process, recurring, non-recurring, costs associated with obligatory infrastructure expansion for the development of digital libraries
- Trained and qualified human resources for the implementation of digitization and preservation process in the library.
- National Repository of Indian digital material may be formed.

## Infrastructure Required for the Digitization of Manuscripts

For the digitization of Manuscripts, the following infrastructure is required:

### Hardware

- Computers with Pentium IV, Dual 2 Core, PCI Bus for information flow, Ethernet for transfer of data, RAM to load, reload or create digital image of different size and colors.
- Storage Devices like Hard Disk Drive, Removable Hard Disk Drive for backup of digital objects and storage, Optical Drive, Digital Audio Tape for archiving and retrieving the data.
- Monitors for sharpness and lucidity of colors are vital to create professional looking digital images
- Digitization Devices like Scanners used to digitize photographs, artwork and slides. Digital Cameras are also required to capture the images for downloading them to the computers
- Output Devices such as Printers, Modem, CD Writer etc

### Software

The software with the following facility needs to be installed for the purpose of digitization:

- That which could edit images
- That which has a page layout programmes for amalgamating text and graphics
- That which has a file transfer efficacy to share files between computers
- That which has a file translation programmes to translate files from graphics to text and from text to graphics
- That which has a facility of file compression.

## Digital Library Initiatives Taken in India

Digitization in Indian Libraries is still in infancy stage but gaining prominence in the field of information processing, digitizing, preserving, disseminating and accessing. In this context, it can be said that the application of digital technologies to preserve the cultural heritage in Indian libraries is entirely new concept as it is a intricate process of experimentation with achievements and disappointments. At institutional, organizational and national level, a number of

digital library initiatives, some booming and some making momentous growth have been taken in India as detailed in Table 1.

Table 1: Digitization Initiative in India

| Digital Library Initiative | Initiated By | Funded By | Website |
|---|---|---|---|
| Digital Library of India (DLI) | IISc (Indian Institute of Science) | Ministry of Communication and Information Technology | http://www.dli.ernet.in |
| Nalanda Digital Library | National Institute of Technology (NIT) Calicut | All India Council of Technical Education (AICTE) | http://www.nalanda.nitc.ac.in |
| Archives of Indian Labour: Integrated Labour History Research Programme | V.V.Giri National Labour Institute and Association of Indian Labour Historians | --- | http://www.indialabourarchives.org |
| Indian Institute of Science | NCSI | --- | http://vidya-mapak.ncsi.iisc.ernet.in/cgi-bin/library |
| Kalasampada: Digital Library-Resource for Indian Cultural Heritage (DLRICH) | Indira Gandhi National Centre for Arts (IGNCA) | Ministry of Communication and Information Technology (MCIT) | http://www.ignca.gov.in/dlrich/ |
| Mobile e-Library | C-DAC Noida | Ministry of Communication and Information Technology (MCIT) | http://mobilelibrary.cdacnoida.in |
| Traditional Knowledge Digital Library (TKDL) | National Institute of Science Communication and Information Resources (NISCAIR) | Department of Indian Systems of Medicine and Homoeopathy (ISM&H) | http://www.tkdl.res.in |
| National Science Digital Library (NSDL) | National Institute of Science Communication and Information Resources (NISCAIR) | --- | http://www.niscair.res.in |
| Down the Memory Lane | Central Secretariat Library | Ministry of Culture | http://csl.nic.in |
| Digitization of Manuscripts | National Mission for Manuscripts | Ministry of Culture | http://namami.nic.in |

## Initiative of Panjab University Library, Chandigarh

### Historical Background of Panjab University, Chandigarh

Panjab University was established in the year 1882 in Lahore and after the partition of India and Pakistan, the library was shifted to Shimla. In the year 1955-

56, it was moved to its present campus in Chandigarh. The library building was formally inaugurated in the year 1963. Since that time the library has progressed in all ways and shifted from a manual system to fully automated one. The library introduced computers for the first time in mid 1990 and in the year 1996, the scenario was changed with the introduction of integrated system, connected to the campus network and subsequently, possessed numerous facilities like telefax, e-mail, internet, Online Public Access Catalogue, multimedia, CD-ROM databases, e-books and e-journals etc. The library has a rich collection of more than 7 lakh volumes and 600 Periodicals. The digital library alongwith the facility of e-resources both online as well as offline has been created.

**Manuscript Collection**

Collection of Manuscripts in Panjab University Library is rare and important. In order to preserve its heritage, the Panjab University Library commenced digitization of its rare collection in the year 2004 and took the decision to open an archive for the upkeep of its numerous collections. Because of the cultural and historical importance and its implication, the magnitude of such holdings can be well ascertained. There are total of 1493 manuscripts available in various languages like Hindi, Urdu, Persian, Punjabi, Sanskrit and Sharda Script wrapping extensive range of subjects for instance, Persian, Court Etiquette, Poetry, Writings of the Sikh Gurus and other translations of eminent personalities as listed below. Government Reports and other general archival trends are also component of the holdings.

List of Manuscripts Available in Panjab University Library

- Mutiny Records.
- Writings on different tribes of North-Eastern States & Andaman Nicobar by different English authors.
- Reports on the resultes of scientific voyages (H.M.S. Challenger).
- Educational Records published by the Govt. of India since colonial days.
- Reports and Surveys of the flora and fauna of the British India.
- Atlases. (both historical and Geograhical)
- Imperial Gazetteer of India.
- Natural history of plants.
- Books on Art, Architecture and Painting.
- Wrtings of the Viceroys and Governor-Generals.
- Biographies.
- English Factory Records.
- Religious literature pertaining to temple, Gurudwara and Mosque.
- English literature on Shakespeare and George Bernard Shaw.
- Books on Sanskrit and Hindi literature, history.
- Books written by medieval writers and travellers.

- Laboratory results from different laboratories of India on scientific subjects and many more...

**Panjab University Library Initiative**

The Panjab University Library started digitization of its collection in the year 2003 as per the guidelines provided by National Manuscripts Mission; (NMM) established by Department of Culture, Government of India with an objective to preserve, conserve and upgrade the manuscripts through digitization and digitize preservation. In order to harness the knowledge embedded in the Manuscripts and to preserve the cultural heritage of our national the digitization process was commenced. Both national and international users make use of the manuscripts for the research purpose. This use increased manifold with the onset of digitization. Although the digitized collection of manuscripts is presently not available on Panjab University website, still efforts are being made to put them on web using D-space software.

## Conclusion

The accountability of making the digital technology successful rests upon librarians, policy makers, educationists, technical personnel, and institutions as well. Individual organization cannot make an adequate amount of effort and accomplish the desired results. That's why organizations / institutions have to work together in synchronization so as to prepare appropriate guidelines for constructive and sustainable digitization programmes.

Digitization is a new conception that is gaining eminence in India and lot of literature on this theme and other issues are mushrooming and it is to be seen that how such imperative issues are being tackled by the library professionals. In order to survive in this world of competition, it is obligatory to recognize and welcome such advances with an unbolt mentality. The digitization process is undertaken by a good number of the Indian libraries these days for preservation, conservation and 24X7 accessibility.

Digitization is the existing area of investigation in this day and age, as it offers high-impact research opportunities for researchers in library and information science field and many librarians and library and information science departments are focusing on it.

In view of the fact that Indian information professionals have currently understood that information is supreme, the Government of India is taking necessary steps for the development of telecommunications and other ICT facilities to make IT based Information access veracity, thus there can be noteworthy enhancement in the excellence of dissemination of information.

The Government of India is initiating efforts to preserve its cultural heritage by formulating policies and strategies at the National Level. The liability lies with the National Informatic Centre, National Library, National Archives and many other individual libraries and information centres across the Nation and over-

seas. Since the manuscripts are scattered in different libraries, museums and archives all over the country, hence, it is the accountability of each of the separate institutions to preserve their cultural heritage, that is manuscripts, with the contemporary digital technology and that technology is called digitization.

## References

Arora, Jagdish (2004), Network-enabled digitized collection at the Central Library, IIT Delhi. *International Information and Library Review,* 36, 1-11.

Arms, William Y. (2000), Editorial: Digital libraries for distance education, *D-Lib Magazine*, 6(10). (http://www.dlib.org/dlib/october00/10editorial.html; retrieved on December 20, 2008)

Bhargava, V. and Vergiya, A. (2000), Preserving information content of old documents some issues, *Annals of Library Science and documentation,* 47,1, 1-4.

Dugdale,D. and Dugdale,C. (2000), Growing an electric library: Resources, utility, marketing and policies, *Journal of Documentation*, 56,6, 644-659.

Gaur, R C. (2003), Rethinking the Indian digital divide: The present state of digitization in Indian management libraries. *International Information & Library Review,* 35, 189-203.

Gertz, J. (2000), Selection for preservation in the digital age: an overview, *Library Resources & Technical Services,* 44, 2, 97-104.

Gupta, C.B. and Haider, S.H. (1995), Conservation practices in Ancient India, *Conservation of Cultural Property in India*, 28, 36-43.

India, Department of Culture (2002), National Mission for Manuscript, Project Document, 36.

Vyas, S.D. and Singh, D.K., Digital libraries: problems, issues and challenges, *In* Proceedings of the National Conference on Information Management in e-libraries, 26-27 February 2002 edited by S. Parthan and V.K.J. Jeevan, New Delhi, Allied Publishers

# USING OPEN-SOURCE SOLUTIONS IN CULTURAL HERITAGE

# Open Source in Art: Originality, Art Process and Digital Preservation

Boris Čučković, student
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
bcuckovi@ffzg.hr, boris.cuckovic@gmail.com

Hrvoje Stančić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
hrvoje.stancic@zg.t-com.hr

**Summary**

*Art formed in the digital age has not yet been sufficiently scientifically studied. The problems of representing artworks made with the help of digital technologies are considerably (inter)connected with the need for a detailed elaboration of stylistic, technical, typological and cultural phenomena associated with such forms of contemporary visual expression. This paper will discuss one uncharted segment of this area which was conceived by integrating open source principles of development and distribution of software into the creation of artworks.*
*The emphasis is set on innovations and alterations which open source introduces in the field of digital art, especially in the categories of author and the original. Through the selected examples the authors examine the possibilities of observing structure and creation of artwork that open source enables. While considering the technical innovations, we will also discuss the continuing and evolving tendencies inherent to art, such as transformation from artwork into art process. The authors offer recommendations on the means of storing, saving and communicating these specific art forms. In their research the authors apply an interdisciplinary approach which includes methodologies of art history and visual communications as well as information sciences.*

**Key words**: art, open source, digital art, author, original, art history, visual communications, information sciences, preservation, museology, heritage

> *"Computers are bringing about a situation that's like the invention of harmony. Subroutines are like chords. No one would think of keeping a chord to himself. You'd give it to anyone who wanted it. You'd welcome alterations of it. Subroutines are altered by a single punch. We're getting music made by man himself, not just one man."* – John Cage, 1969.

## Introduction

When the prehistoric man first started laying pigments of color on the walls of his caves or on the surfaces of rocks in the open, originating the adventure called art that follows mankind still today, he chose freely the material of his new activity out of the nature which surrounded him. Everything that was available to our ancestors was theirs to use. Art, like human beings, traveled a long way from then. One of the latest and most promising fields of contemporary art is certainly the digital art. If we were to apply the circumstances that determine the digital environment on the creative individual from the prehistoric beginning of our story, amongst other problems, we would find him puzzled with the inscription "Trial" over his painting and a required serial key or credit card forms to view the real picture. Not every tool digital artists can find is immediately and completely available to them. And that is nothing unusual. The history of art is in a considerable way determined by the commercial availability of specific materials, for example, the availability of a paint color for a painter or a marble type for a sculptor. The same can be said for computer programs used to produce digital artwork.

Implementing open source principles of development and distribution of software into digital art completely changes the stated characteristics of this medium, enabling free usage of tools necessary for this type of artistic production. Open source as a principle dictates complete access to software source code, resulting with inability to charge it. Several different programmers can cooperate and work on one code and, with the help of Internet, products can be made through the public collaboration where no one charges for his or her contribution to the software development. Anyone can use the resulting software for free. If open source implementation is considered according to the division of digital art into 'art which uses digital technology as a tool' and 'art which uses digital technology as a medium'[1], we can conclude that it does not only modify artwork production but also radically changes some basic categories of artwork, such as authorship and originality. Therefore, the influence of open source principles can be determined on two levels, both illustrated by the example which was also the impulse for the research for this paper – the first animated *open movie* project named "Elephants Dream"[2]. The first level is utilization of open source programs in the creation of every single element used in an artwork, in this case every object used in the digitally animated movie. The second is that the movie itself should be open source, meaning that its every element is publicly available and that everyone who wants to and knows how to can work on it with the open source software. This level will interest us

[1] Categorisation from: Paul, Christiane. Digital Art, London: Thames & Hudson, 2008.

[2] The project was first named "The Orange Project", and then renamed to "Elephants Dream" according to way Dutch children stories suddenly end, http://orange.blender.org/ (30th July 2009).

in this paper, which does not aim to make an overview of such artworks but to study the alterations which open source makes in comprehension of artwork itself, the new options it enables, as well as to discuss those tendencies which are inherent to art and which potentate art for a new development in the open source environment.

Picture 1: A scene from *Elephants Dream* animated movie from 2006. It showed that Blender and other open source tools can match visual quality with the commercial solutions in the field of 3D animation.

Though mainly researching unexplored (and not yet emphasized) field of digital art, the paper also tries to open a new possibility of observation of digital art and its classification. Focusing on a concept which is a product of a digital environment and information age, in this case the open source, sets a grounding for creation of a future classification that comes from the nature and the specifics of the digital medium instead of putting digital artwork into drawers made by some older branches of history of art. In that manner, for example, Bruce Wands sorts digital art into: digital imaging, digital sculpture, digital installation, performance, music and sound art etc[3]. This and similar systematizations have been extremely useful in the difficult task of exploring this relatively new and certainly dynamic area. The attempt to open a new possibility by creating a different focus is not confronting to the current views. Instead it strives to upgrade and relate to them, creating a wider and more applicable framework for elaborating digital art. The concept of open source chosen in this paper can associate to certain examples from very different categories where it equally stands as a characteristic element of distinction, such

[3] Wands, Bruce. Art of the Digital Age. London: Thames & Hudson, 2006.

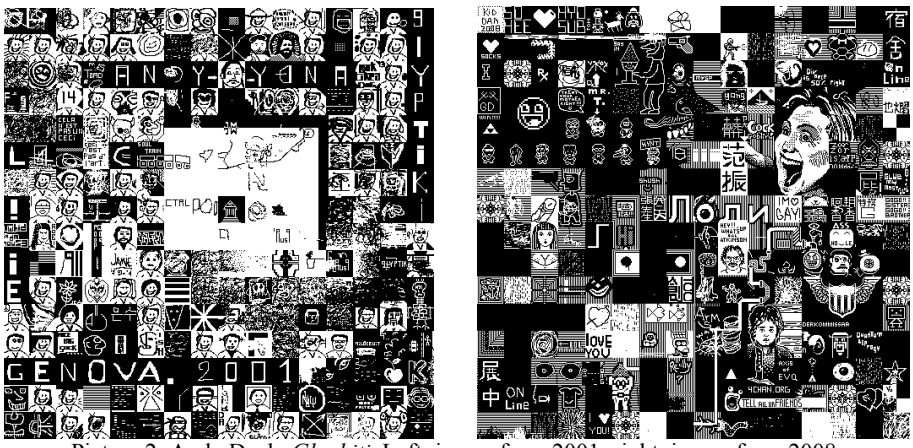
as digital animation, software art, net.art, digital printing or flash art. Such selection of artwork will uncover their related attributes much more authentic to the context of reasoning in which they have been conceived then is apparent while they are split up into several different areas.
The interdisciplinary approach applied in this paper will also show a necessity of cooperation between the art history and the information science, working on a subject that is certainly a field of contemporary art, and works of art that are unthinkable without a concept addressed by the information science. Presentation of these works also require fresh museological solutions that are grounded in conclusions of a scientific analysis of contemporary art practice, which, on the other hand, must not exclude the software specifics that produce the very essence of the open source phenomenon.

## Open Source Artwork

We have established what implementation of the open source software means for a digital artist, primarily in terms of availability of different programs used without financial or temporal limitations. But what does the open source approach to artwork mean for the observer (visitor, viewer)? In fact, it should be noticed that, in a certain degree, it wipes the boundary between the observer and the author by providing the observer with the same authority over the artwork the author had while making it. If we consider the mentioned example of the "Elephants Dream" animated movie once again, it would mean that, considering the complete accessibility of all 3D objects, textures and sounds used in the movie on the internet, the viewer can rearrange or upgrade a movie in any way he or she wants to, or compose a completely new creation out of the same elements. Still, in this case the original artwork is signed, which implies that every new viewer-author creates his or hers own, new original. They are connected by a common starting point and the open source concept without which the artwork would not have been created. There are also examples in which the observers are intervening in the same original made by the initiating author, as can be seen from the case of "Glyphiti"[4] project by Andy Deck from 2001. and 2006. The author created an image available on the internet, split into smaller units (called "glyphs") which the visitor could select and then make desired alterations to them. The complete initial image is constantly changing by the actions of the visitors, so that each new visitor is not *looking* at the image, one and unique, in a traditional sense, but *watching a live image*, with all its phases and shifts happening in real time.
The matter of discussion is, therefore, a unique original in constant transformation. Interventions in works of art have, of course, been seen before the appearance of digital art as the conceptual art is abundant with such options. Still,

[4] Deck, Andy. Glyphiti. Versions from years 2001 and 2006, http://artcontext.net/glyphiti/index.php (last access: 28 July 2009).

there is another level of open source implementation present in this artwork – one very different from the situations in which the terms *author* and *original* have been found throughout contemporary art. The author, Andy Deck, has determined the characteristics of this digital work of art and of course its very concept through the process of programming. He declared the size and colors the visitors can use – black and white. However, he allowed visitors the possibility to change the source code of this digital artwork. By literally declaring: "…if you don't like the options given to you, please revise the source code. Copy it. Steal it. Share it. Print it. Pretend it's yours. I don't care", Deck is provoking the possibility of changing the prescribed interventions into the artwork, its very concept. Unlike (merely) encouraging the activity of the observer of a given artwork, he is also promoting him into a coauthor. Such coauthor can then modify the rules of intervening into an artwork which govern the actions of the visitors remaining in the role of active observers. These two categories exist side by side and it is clear that we do not only have an upgrade of observer activities but also a change of role. Programming of digital artwork's source code is of structural importance for the finite original. For instance, an analogy would be clay modeling for a bronze sculpture model – undoubtedly key process for the final form of an artwork. The founder's work is correspondent to the web browser reading[5] the code and displaying the image on the visitors screen. The interventions to the Glyphiti image that the users make would be *en pair* with the, if the sculptor allows it, coloring of the cast sculpture or adding certain elements to it, a hat for instance.




Picture 2: Andy Deck. *Glyphiti*. Left: image from 2001; right: image from 2008.

[5] In programming the term "interpreting" would be more appropriate, but the term "reading" is selected instead to avoid conflict with the art history terminology.

It is emphasized that the color is determined in the source code[6]. The change of color, for example, black into red, is intruding the concept of the work, like swapping the cold palette with a warm colored one before the painter begins his work. If we take a step back we might also claim that this possibility is also a concept. However, if we look wider we will find that this concept did not come directly from the author of an artwork, and neither did its name. In the digital world that concept has a familiar term – the open source[7]. It is a concept of an entire community which has public creation as a principle. In the case of the Glyphiti project it is perfectly clear who is the author, just as is the possibility of co-authorship for anyone who wants to participate and has a minimum knowledge in informatics to do so. If we return from the conceptual level back to the very image, we can establish that its space is also intended for collaboration and group work. In so doing, the co-authors working on an image are not necessarily aware of each other, and are not obliged to know each other, which is nothing unusual for a digital environment. Therefore, a completely open source digital work of art is immersed into public which is forming it on all stated levels.

## Art Tendencies in Open Source Environment

Area opened by the open source concept has provided grounding for developing some already existent art tendencies, as well as the creation of certain new characteristics that would not have been formed without it inside the frame of digital art. During the twentieth century art*work* has in many different ways transformed into an art *process*, whether it is in forms of artistic expression with immanent temporal dimension, such as performance or happening, or in innovations of "timeless" artistic fields, such as those made by Jean Tinguely in the sculpture with his works of limited time duration. One characteristic of these aspirations is the determined time span of a process. In Tinguely's works, "Homage to New York" (1960.) and "Study for an End of the World No. 2" (1962.), the sculpture is existent until its own mechanism destroys it, and in every performance or happening it is possible afterwards, and sometimes even in advance, to determine its duration. The "Glyphiti" project is already detected as a process running in real time, and it can be added that this aspiration has lost its need for temporal determination when found in an open source environment. Because of the open source availability the image is subject to constant modifications. Theoretically, the open source art process available on the internet does not have to end. "Art is never finished, only abandoned" said Da Vinci. The continuation of such process is not dependent upon the natural limitations of the author, performer or observer, weather conditions or day and night cycle be-

[6] The source code is written in JavaScript.

[7] The term has been widely promoted after a summit in April 1998 organized by Tim Riley under the name of "Freeware Summit", later referred to as the "Open Source Summit".

cause it is always available in different parts of the world through the Internet. If the dynamic of morphing is great enough, the observer is greeted by a new form upon every new visit, thus raising interest of the public. The possibility of achieving such open source art process that has a continuation of morphing (instead of temporal determination) is real, which is proven by the fact that open source exists, functions and integrates creative people even on a much more commercially demanding area of software development. Furthermore, these solutions can be dominant in their, often highly sophisticated, areas like, for instance, the Apache web server[8] in its domain. The technological and social possibilities are there. The challenge lies in creating a process of adequate quality and involvement.

Such creations could also intensify an interesting possibility of form that changes context independent of the actions of the original author. The elements of a digital artwork that are originally used in creation of one artwork, or even a whole artwork itself, can be found in another artwork of a completely different character. In a certain sense, this is an extension of a postmodern tendency to quote, and its subsequent recontextualization. If an art process of open code available on the Internet outlives a certain period of some visual style domination it will continue to change according to style applied by the future visitors and co-authors. Theoretically, this process does not need to have an end, and it realizes an artwork adopting to change of context, social conditions, and spirit of an age. From the historic perspective, it might also be concluded that a unique artwork is spanning across several different periods. Notable is an analogy with a drama screenplay divergently adapted during the course of history.

If the open source license does not prevent it[9], the public work can also be used in a work of commercial or private function. The characters of "Elephants Dream" could be found in some advertising campaign, with the only condition of providing attribution to the original project. This possibility works two ways – a commercially successful digital artist can also contribute to the open source community. Joshua Davis has worked for brands such as Nokia, Nike and Diesel, and he was also amongst the first to offer open source flash files over his webpage Praystation.com. He is significant for observation of the concept of originality in the digital art because he tries to restore its uniqueness. It has been lost by vast possibilities of multiplication in digital art, both on the level of the "final original" and on the level leading to its concretization (like the digital model of a sculpture that can cease to be unique by a simple copy-paste

[8] Lerner, Josh; Tirole, Jean. Some Simple Economics of Open Source // The Journal of Industrial Economics. Vol. 50, No. 2 (Jun 2002), pp. 197.

[9] Often, the open source licenses do not approve commercial usage, like the Open Art License version 1.0, subsection 2: *The reuse is not for profit*. http://www.three.org/openart/license/index.html (last access: 3 August 2009).

method[10]). Davis creates series of several thousand commercial posters that are changed before printing by an algorithm. His model, digital template for print, has in its code the instruction of uniqueness.
The strength of the open source idea has also affected the content of artworks. Narrative specifics and open ending of "Elephants Dream" cannot be left unnoticed. In it, the older character, Proog, is trying to explain to the younger one, Emo, the abstract Machine they are found in. We can comprehend it as a metaphor of any idea or a concept. He is doing so by *forcing* his view of the idea, instead of *sharing* it, which results in a physical conflict. The open ending is a call to the public to join in and make new versions of the movie. Open source concept by itself is also politically and economically provocative and could thus be dearly used in the art world. Perhaps the best example of this orientation of open source art is the project CarnivorePE by Alex Galloway and the RSG (Radical Software Group). It is inspired by the DCS1000 software used by FBI for surveillance of e-mail and communications, previously known as Carnivore. The RSG's Personal Edition open source version, instead of collecting information about the suspects, is transforming electronic information into vibrant images and sound, generating art instead of incriminating evidence[11]. Digital artists are using client-server principle, creating clients which produce the desired effects from information given by CarnivorePE server. It is left to the artist to interpret the information through their clients. The Guernica[12] client is turning the web into a dystopic world of oil pumps and rockets, while the Amalgamatmosphere[13] client is creating a live and vivid vision of network activities.

## New Challenges of Communicating and Preserving Digital Art

Institutional preservation and communication of digital art is a difficult task because the museum institutions have grown on a *white cube* model of functioning, while the new challenges call for a *wired cube* approach. It is important to have in mind all the analyzed specifics of open source artwork while considering these problems. Orientation to the Internet has made art easily accessible outside institutional channels, communicating directly to the audience without the need of taking a conventional journey through museums or galleries to their visitors. On the first glance it might appear that just as open source blurs the

[10] Čučković, Boris. Razmatranje skulpture ostvarene digitalnom tehnologijom // Symposium "Original u skulpturi", Galerija Antuna Augustinčića. Klanjec. 4-6 June 2008 (in print).

[11] Mirapul, Matthew. Cybersnooping For Sounds & Images, Not Suspects. // New York Times, 1 October 2001, online edition: http://www.nytimes.com/2001/10/01/arts/design/01ARTS.html (last access: 3 August 2009).

[12] Creative duo Entropy8Zuper!; http://entropy8zuper.org/guernica (last access: 3 August 2009).

[13] Davis, Joshua. Hall, Brandon. Shapeshifter; http://ps3.praystation.com/pound/assets/2001/11-20-2001/index.html (last access: 3 August 2009).

boundary between the author and the observer, it does so between the author and the curator. Authors usually create their own exhibition space on the Internet in the form of a web page or a web site. Still, that boundary is not completely lost and there is a great need for an institutional framework solving the problems of preserving these artworks. One of the important efforts in this direction has been made by the Solomon R. Guggenheim Museum in New York by creating an archive of new media art – Variable Media Network[14].

Considering the close bonds of digital art with rapidly advancing technological achievements, making obsolescence of current solutions very probable, institutional aid in preserving certain characteristics of open source art is necessary. Perhaps the best example of this need is preserving the possibility of running an indefinitely long art process that is endangered by outdating of software and hardware environment it is made on. This concept is surpassing the limited duration of medium, especially in the circumstances of constant web browsers', and their plug-in, development. Same files which are readable today can became unreadable tomorrow, and protocols making them accessible can be replaced by new ones[15]. The solution to this problem Mark Tribe sees in applying four methods[16]: *documentation* (screen captures, artist diagrams, installation instructions and statements), *migration* (updating work to accommodate newer technology and file formats), *emulation* (running projects through additional software that allows them to work on newer hardware), and *recreation* (remaking the artwork for a new technical environment). In the Guggenheim Museum the authors are entitled to choose the modes of migration or recreation of their code in the future.

But for open source art, the communication between authors collaborating on the development of an idea is just as important as the communication between the author, his work and the observer. The Open Art Network[17] has, therefore, developed principles enhancing this communication, and also working preven-

[14] Variable Media Network. http://www.variablemedia.net/ (last access 5 August 2009).

[15] For further discussion on digital preservation issues see:

Thibodeau, Kenneth. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years // The State of Digital Preservation: An International Perspective. Washington, D.C.: Council on Library and Information Resources (CLIR). July 2002, pp. 4-31, http://www.clir.org/pubs/reports/pub107/pub107.pdf (last access: 25 October 2004).

Stančić, Hrvoje. Arhivsko gradivo u elektroničkom obliku: mogućnosti zaštite i očuvanja na dulji vremenski rok // Arhivski vjesnik. 2006, No. 49, pp. 107-121.

Stančić, Hrvoje, Očuvanje elektroničkih informacijskih objekata: arhivi, knjižnice, muzeji – zajednička koncepcija // Katić, Tinka (ed.). 7. seminar Arhivi, knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture. Zagreb: Hrvatsko knjižničarsko društvo. 2004, pp. 26-35.

[16] Wands, Bruce. Art of the Digital Age. London: Thames & Hudson, 2006, pp. 206.

[17] Open Art Network. http://www.three.org/openart/ (last access: 5 August 2009).

tively on certain problems of saving and (potential) future usage of the code. The basic principle is that program should have a modular structure, so it could be easily changed and its components used in other projects, and that the code should not be procedural, but object-oriented. Unintelligible code is not contributing to the open source community, so detailed comments following the code are welcomed. Furthermore, for the open source movement a search engine of such meta-programming, helping artist beginners in programming, would be a great progress.

The main principle of preserving the original open source art process should contain the requirement of preserving the availability of its code to the public so it can continue to live on that level. This alone will not be satisfactory enough without emphasizing the importance of documenting the process as well. In determined time intervals the phases of such artwork should be recorded, so that the answer of the public to the concept is also preserved. A good example would be the documentation of the Glyphiti image in the form of a .gif stream.

The influence of the open source on the very institution of the museum is also interesting. OSMOSA[18], an open source museum of open source art, exists in the virtual world of Second Life. Anyone can add, modify or remove objects from OSMOSA. Likewise, anyone can do the same with the very elements of the museum building. It certainly represents a challenging environment but it completely follows the terms of open source principles.

## Conclusion

Implementation of open source principles of software development has radically changed the important categories of digital artwork. Enabling a never before seen level of co-authorship, it has included the public into the act of creation much more than it was the case with active observers of artwork invited to do interventions in it. A completely available code allows structural and conceptual modifications of digital artwork, and the redefined term of *the original* not only does not stream towards uniqueness (as is the case with the rest of digital art), but also invokes and provokes multiplications and its own usage in realization of other visions and ideas. That is certainly a stimulation for specific tendencies of contemporary art that have found new possibilities inside the open source framework, such as achieving artwork as a process, or creating an artwork that is adopting to the change of context by the public itself, without the acts of its originator. Preservation and communication of these complex concepts will be a demanding mission of conservation and museology of the twenty first century. The main task of the scientific elaboration of the digital art lies in examining and discussing the phenomenon in accordance to the way a specific concept was

[18] Open Source Museum of Open Source Art. http://osmosa.blogspot.com/ (last access: 6 August 2009).

formed, never neglecting the categorization according to characteristic branches of contemporary art, but also, if there is a need, not running from the creation of a new classification based upon creative principles underlying a certain group of digital works of art.

Evading the dominant corporative principle of developing the digital possibilities, the open source approach is relying on public creation and idea sharing to accomplish the designated goals, benefiting the initiator of the project as well as the whole community. This seemingly utopian idea is functioning, a fact backed up not only by software development success, but also by digital artwork made through open source collaboration presented in this paper. The prerequisites have been set for creation of a public work of art that is formed by public itself, or in terms with the John Cage quote from the beginning of this paper – an art made by man himself, not just one man. All these possibilities are a great challenge, and also a glimpse of an interesting and creative future of artistic creation.

## References

Čučković, Boris. Razmatranje skulpture ostvarene digitalnom tehnologijom // Symposium "Original u skulpturi", Galerija Antuna Augustinčića. Klanjec. 4-6 June 2008 (in print)

Davis, Joshua. Hall, Brandon. Shapeshifter; http://ps3.praystation.com/pound/assets/2001/11-20-2001/index.html (last access: 3 August 2009)

Deck, Andy. Glyphiti. http://artcontext.net/glyphiti/index.php (last access: 28 July 2009)

Entropy8Zuper!. http://entropy8zuper.org/guernica (last access: 3 August 2009)

Lerner, Josh; Tirole, Jean. Some Simple Economics of Open Source // *The Journal of Industrial Economics*. Vol. 50, No. 2 (Jun 2002), pp. 197-234

Mirapul, Matthew. Cybersnooping For Sounds & Images, Not Suspects. // *New York Times*, 1 October 2001. http://www.nytimes.com/2001/10/01/arts/design/01ARTS.html (last access: 3 August 2009)

Open Art License version 1.0. http://www.three.org/openart/license/index.html (last access: 3 August 2009)

Open Source Museum of Open Source Art. http://osmosa.blogspot.com/ (last access: 6 August 2009)

Paul, Christiane. Digital Art. London: Thames & Hudson, 2008

Stančić, Hrvoje. Arhivsko gradivo u elektroničkom obliku: mogućnosti zaštite i očuvanja na dulji vremenski rok // Arhivski vjesnik. 2006, No. 49, pp. 107-121

Stančić, Hrvoje, Očuvanje elektroničkih informacijskih objekata: arhivi, knjižnice, muzeji – zajednička koncepcija // Katić, Tinka (ed.). 7. seminar Arhivi, knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture. Zagreb: Hrvatsko knjižničarsko društvo. 2004, pp. 26-35

The Orange Project. http://orange.blender.org/ (last access: 30 July 2009)

Thibodeau, Kenneth. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years // The State of Digital Preservation: An International Perspective. Washington, D.C.: Council on Library and Information Resources (CLIR). July 2002, pp. 4-31, http://www.clir.org/pubs/reports/pub107/pub107.pdf (last access: 25 October 2004)

Wands, Bruce. Art of the Digital Age. London: Thames & Hudson, 2006

# Open Access in Croatia: a Study of Authors' Perceptions

Ivana Hebrang Grgić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
ivana.grgic@ffzg.hr

**Summary**

*The most important participants of scientific communication are scientists. They are both producers and users of scientific information. Publishers, libraries, scientific associations and institutions have to ensure transmission of information but scientists are in the centre of scientific communication.*
*Open Access (OA) to peer-reviewed scientific information is relatively new form of scientific communication and is accepted by a number of worlds' scientists, universities, libraries and publishers. Numerous surveys of scientists' involvement in the open access movement have been published throughout the world, but Croatian scientists' perceptions have never been examined. In this paper, a study of Croatian scientists' perceptions of Open Access movement will be presented. The sample will consist of over 300 Croatian scientists who had published their papers in the last issues of Croatian scientific journals published on the Hrčak portal (portal of Croatian scientific journals) in 2008. Anonymous web questionnaire will be send to the scientists. They will be asked about their publishing experiences – how many scientific papers they publish yearly, are the journals they publish in OA journals, what are reasons for publishing (or not publishing) in OA journals, have they ever self-archived their scientific paper, and if yes - why and where. They will also be asked how they deal with copyright issues while self-archiving.*
*Results will be analyzed altogether as well as according to scientific field.*
*Results will show whether Croatian scientists are familiar with the Open Access movement, whether they practice submitting articles to OA journals and whether they self-archive their papers. Some proposals about future development of Open Access in Croatia will be presented.*

**Key words:** OA journals, OA repositories, scientific communication

## Introduction

Scientists are the reason why scientific communication exists – using scientific information they produce new information. Publishers and librarians also play very important role, but they would not have it without scientists. As Open Access (OA), during last decade, evolved into a global movement, researches of its influence on scientific communication have been made. We can talk about three kinds of researches. The aim of the earliest researches was to investigate effect of open access on citation impact. The second phase was investigating the reasons for higher citation impact of open access articles. The third group of researches was about participants of scientific communication, their role in open access movement, their habits and concerns about the OA. Those researches are mainly conducted by publishers, journal editors, organizations involved in promotion of OA or scientists who are interested in OA movement, its development and its future. We are here interested in the last kind of researches, specially the researches of authors of scientific papers published in scientific journals.

In 2004, Cozzarelli and co-authors surveyed more than 200 corresponding authors of accepted papers in journal Proceedings of the National Academy of Science (PNAS).[1] About half of the respondents were in favour of the open access option, willing to pay a surcharge to make their article freely available online at the time of publication.

In 2006, Swan and Brown published the results of their international cross-disciplinary study on OA.[2] The study had more than 1200 respondents and was focused on self-archiving. Almost half of the respondents have self-archived at least one article during the last three years (2003-2005). Self-archiving activity was greatest amongst those who published the largest number of papers, and only 20% of authors found some difficulties while depositing an article in a repository for the first time. It is also interesting that only 10% of authors knew of the SHERPA/RoMEO list of publisher permissions policies with respect to self-archiving. Swan and Brown found out that there were still a substantial proportion of authors unaware of the possibility of providing open access to their work by self-archiving.

Schroter and Tite made an electronic survey of almost 500 authors of research papers submitted in 2004 to three medical journals.[3] Less than 50% of the respondents reported that they were familiar with the term Open Access. Au-

[1] Cozzarelli, N. R.; Fulton, K. R., Sullenberger, D. M. Results of a PNAS author survey on an open access option for publication. 2004. http://www.pnas.org/content/101/5/1111.full (1-2-2009)

[2] Swan, A.; Brown, S. Open access self-archiving : an author study. Truro : Key Perspectives Limited, 2005. http://cogprints.org/4385 (1-2-2009)

[3] Schroter, S.; Tite, L. Open access publishing and author-pays business models : a study of authors' knowledge and perceptions. // Journal of the Royal Society of Medicine. 99 (2006), 141-148

thors concluded that, at that time, Open Access policies had had little impact on authors´ decision of where to submit papers.
Studies similar to those described above have never been conducted in Croatia, among Croatian authors of scientific papers.

## Aim, methodology and hypothesis of the survey

In March 2009 an anonymous online questionnaire was sent to 345 Croatian scientists who had published scientific papers in the last issues of Croatian scientific journals that were available on the Hrčak portal (central portal of Croatian scientific journals) at that time.[4] Response rate was 170 (49.3%).
The purpose of the survey was to find out about Croatian scientists' perceptions of Open Access movement, concerning OA to their own scientific papers. Authors were asked about their publishing experiences – how many scientific papers they publish yearly, do they publish in OA journals, what are the reasons for publishing (or not publishing) in OA journals, have they ever self-archived their scientific paper, and if yes - why and where. They were also asked how they dealt with copyright issues while self-archiving.
Results will be analyzed altogether as well as according to scientific fields.
Formdesk website forms were used for creating an online, ten questions questionnaire. Formdesk offers various useful features, such as auto responds by e-mail, statistics, simple and advanced filters, results download, secure data transfer and many others. The questionnaire was available during two weeks period in March and April 2009 at: http://fd8.formdesk.com/grgic/oa.[5]
Before starting this cross-disciplinary survey of Croatian scientists' perspectives, following hypothesis were set up:

- there is a large acceptance of OA journals among the Croatian scientists from the fields of natural sciences and biomedicine and health (those fields have the longest tradition of OA acceptance in the global scientific community);
- Croatian scientists support free access to scientific information;
- Croatian scientists rarely self-archive their papers;
- the majority of Croatian scientists are aware of the importance and the possibilities of Open Access;
- Croatian scientists do not know enough about ownership of copyright of their published papers.

## Results and discussion

The analysis will show the answers to all the 10 questions. In some questions, correlations will be shown (e. g. scientific field – publishing in OA journals).

[4] Hrcak : portal znanstvenih casopisa Republike Hrvatske. www.hrcak.hr (3-6-2009)

[5] Otvoreni pristup. 25. 3. 2009. http://fd8.formdesk.com/grgic/oa (5-4-2009)

**Question 1 - Scientific field**

The majority of respondents are from humanities (46, or 27%) and social sciences (41, or 24,1%). Here we have to notice that the sample was chosen from the journals on the Hrčak portal and there are more journals from humanities and social sciences than from other fields. Table 1 shows answers to the question.

Table 1: Scientific field

| Scientific field | No of respondents | Percentage |
|---|---|---|
| Humanities | 46 | 27.0 |
| Social sciences | 41 | 24.1 |
| Natural sciences | 28 | 16.5 |
| Biomedicine and health | 19 | 11.2 |
| Technical sciences | 18 | 10.6 |
| Biotechnical sciences | 18 | 10.6 |
| **Total** | **170** | **100.0** |

**Questions 2 and 3 - How many scientific papers per year do you publish in Croatian/foreign scientific journals?**

Types of the questions were multiply choice, select one. The answers to the questions are analysed according to the scientific field and the results are shown in Table 2.

Authors from humanities and social sciences publish more articles in Croatian journals. The most productive authors in foreign journals are those from biomedicine and health. Six out of 19 authors from that field publish three of more articles per year in foreign journals.

Table 2: Number of scientific papers published in Croatian and foreign journals according to scientific field (Cro=Croatian; for=foreign)

| Scientific field | 0 | | 1-2 | | 3 or more | | No answer | |
|---|---|---|---|---|---|---|---|---|
| | Cro | for | Cro | for | Cro | for | Cro | for |
| Natural sciences | 1 | 5 | 25 | 19 | 2 | 3 | 0 | 1 |
| Technical sciences | 0 | 7 | 14 | 11 | 3 | 0 | 1 | 0 |
| Biomedicine & health | 1 | 3 | 13 | 10 | 4 | 6 | 1 | 0 |
| Biotechnical sciences | 1 | 2 | 11 | 11 | 6 | 5 | 0 | 0 |
| Social sciences | 0 | 19 | 29 | 19 | 12 | 3 | 0 | 0 |
| Humanities | 2 | 26 | 30 | 16 | 13 | 1 | 1 | 3 |
| **Total** | **5** | **62** | **122** | **86** | **40** | **18** | **3** | **4** |

**Question 4 - How many of the journals you publish in are OA journals?**

The answers to that question should show if the authors are aware of free accessibility of the journals they publish in. We assumed that the majority of authors would know the answer to the question. 20 authors (11.8%) think that none of the journals they publish in is OA journal. 93 authors (54.7%) answered that

one or two journals are OA journals, 37 authors (21.7%) answered that more than two journals are OA journals, 19 authors (11.2%) did not know the answer and one author (0.66%) did not answer the question. Results are shown in Table 3.

Table 3: Publishing in OA journals – frequency

| Scientific field | 0 | 1-2 | 3 or more | Don't know | No answer | Total |
|---|---|---|---|---|---|---|
| Natural sciences | 3 | 16 | 5 | 3 | 1 | **28** |
| Technical sciences | 3 | 10 | 4 | 1 | 0 | **18** |
| Biomedicine & health | 1 | 11 | 6 | 1 | 0 | **19** |
| Biotechnical sciences | 0 | 15 | 3 | 0 | 0 | **18** |
| Social sciences | 7 | 17 | 9 | 8 | 0 | **41** |
| Humanities | 6 | 24 | 10 | 6 | 0 | **46** |
| **Total** | **20** | **93** | **37** | **19** | **1** | **170** |

**Question 5 - What are your reasons for publishing in OA journals?**

The purpose of the question was finding out the most common reasons for publishing in OA journals. Type of the question was multiply choice, select many. 160 respondents (94,1%) answered the question. The most common reason for publishing in OA journals for Croatian authors is support of free access to scientific information and the second important reason is higher impact of articles published in OA journals. All the other reasons are shown in the Table 4.

Table 4: Reasons for publishing in OA journals

| Reasons for publishing in OA journal | No. | Percentage |
|---|---|---|
| Support of free access to scientific information | 88 | 55.0 |
| Higher impact of the article | 59 | 36.9 |
| No reason | 55 | 34.3 |
| Short submission/acceptance process | 35 | 21.9 |
| Reputation of the journal | 28 | 17.5 |
| Higher impact factor of the journal | 28 | 17.5 |

**Question 6 - If you do not submit papers to OA journals, what are your reasons?**

120 respondents (70.6%) answered the question. The majority of respondents do not have reasons for not publishing in OA journals. We can conclude that they either do publish in OA journals or do not know enough about OA journals.

All the other reasons for not submitting papers to OA journals are shown in the Table 5.

Table 5: Reasons for not publishing in OA journals

| Reasons for not publishing in OA journals | No | Percentage |
|---|---|---|
| No reason | 98 | 81.7 |
| Lower citation impact | 14 | 11.7 |
| Can't find appropriate OA journal | 9 | 7.5 |
| Lower quality of OA articles | 5 | 4.2 |
| Bad reputation of OA journals | 4 | 3.3 |
| Do not support free access to scientific information | 3 | 2.5 |
| Problems with permanent access to OA journals | 1 | 0.8 |

**Question 7 - Have you ever self-archived your scientific paper?**
This question opens the issue of OA repositories. Our presumption was that Croatian scientists do not self-archive their papers. 93 (54.7%) answers to the question were negative and 77 (45.3%) affirmative. It is not clear from those answers whether respondents understand the concept of self-archiving. That problem will be emphasized in the next two questions.

**Question 8 - Where have you self-archived your scientific paper(s)?**
Although 77 respondents answered affirmatively to the previous question, there were 83 answers to this question. 57 respondents answered that they self-archived their papers in Croatian scientific bibliography (CROSBI). Here we have to point that CROSBI has the possibility of archiving full-texts, but it primarily stores metadata about Croatian scientists' papers published from 1997 to the present. Today there are more then 190,000 records in the bibliography, 2,000 with full-text available. When answering the eighth question, there is a possibility that some of 57 respondents do not distinguish between archiving metadata and full-text self-archiving. Other respondents have self-archived their papers on their own web sites, on institutional web sites or in a repository (institutional or other), as shown in the Table 6.

Table 6: Location of self-archived articles

| Location of self-archiving | No. | Percentage |
|---|---|---|
| CROSBI | 57 | 68.7 |
| Institutional web site | 21 | 25.3 |
| Repository other than institutional | 13 | 15.7 |
| Their own web sites | 10 | 12.1 |
| Institutional repository | 7 | 8.4 |

**Question 9 - If you have self-archived your paper, which version have you archived?**
The purpose of this question was to find out if Croatian scientists self-archive preprints, postprints or both. We presumed that the authors mostly self-archive

postprints (peer-reviewed versions accepted for publication), especially when archiving papers published in Croatian journals (because they are non-profit and they do not have financial reasons for not allowing self-archiving). And indeed, 76 out of 82 respondents who answered to the question (92.7%) had self-archived postprints. However, are Croatian scientists aware that publishers are the owners of copyright? We will try to find it out in the next question.

**Question 10 - Have you asked publishers' permission to self-archive?**
Answers to the question are supposed to show Croatian scientists' awareness of copyright issues. According to the answers to the previous question, we could conclude that they do not think that they have to ask publishers' permission to self-archive. The presumption is correct again – 61 out of 83 respondents (73.5%) have never asked publishers' permission.

## Conclusion

Some of Croatian scientists do not know enough about the OA movement and are not aware of all the potential benefits of OA. They are afraid of lower citation impact and lower OA journals quality. While analyzing the answers, there is always a question – do respondents know the meaning of the term Open Access? 24% of them do not know whether the journals they publish in are OA journals, some are not sure what self-archiving is, and if they are, they do not ask publishers' permission for self-archiving postprints.
Regarding the longest tradition of OA acceptance by the world's scientists in the fields of natural sciences and biomedicine, we presumed that Croatian scientist from those fields would also accept the movement more then their colleagues from other fields. That was not proved by the survey. The only field where all the scientists publish at least one article in an OA journal per year is the field of biotechnical sciences, and they are all aware of it.
Less than 50% of respondents have self-archived their papers and, as was mentioned earlier, there is a question: do they know the meaning of the term "self-archiving"? Some of them answer that they self-archive in CROSBI, and CROSBI is a bibliography, not a repository (although it offers the possibility of attaching full-texts of articles).
Copyright issues are not important to our respondents. We can find similarity to Swan and Brown's 2006 survey where majority of respondents did not know about publisher permissions policies and where there were a substantial proportion of authors unaware of the possibility of self-archiving.[6]
Croatian scientists know more about OA journals than about OA repositories. The main reason is the Hrčak portal that popularizes OA journals in Croatia. Hrčak is a national portal, supported by the Ministry of science (one condition

[6] Swan, A.; Brown, S. Open access self-archiving : an author study. Truro : Key Perspectives Limited, 2005. http://cogprints.org/4385 (1-2-2009)

for financing scientific journal is its presence at the portal). There is no similar project that would popularize and encourage OA repositories in Croatia.

Open access movement is not totally unknown among Croatian scientists, they are maybe not always sure about the exact definition of OA, but they do support free access to scientific information. They want their work to be visible and they surely have nothing against increasing citation impact of the journals they publish in. As we mentioned earlier, the point of this survey was not finding out about Croatian scientists’ perceptions of OA to non-Croatian scientific papers. This could be a matter of another survey.

At the end, we can pose the question about the necessity of open access in a peripheral scientific community such as Croatian. Is the country too small? Are Croatian scientists’ papers accessible regardless OA because Croatian journals are non-commercial? That could also be the subject to some future surveys and discussions.

## References

Cozzarelli, N. R.; Fulton, K. R.; Sullenberger, D. M. Results of a PNAS author survey on an open access option for publication. 2004. http://www.pnas.org/content/101/5/1111.full (1-2-2009)

Hrčak : portal znanstvenih casopisa Republike Hrvatske. http://www.hrcak.hr (3-6-2009)

Otvoreni pristup. 25. 3. 2009. http//fd8.formdesk.com/grgic/oa (5-4-2009)

Schroter, S.; Tite, L. Open access publishing and author-pays business models : a study of authors' knowledge and perceptions. // Journal of the Royal Society of Medicine. 99 (2006), 141-148

Swan, A.; Brown, S. Open access self-archiving : an author study. Truro : Key Perspectives Limited, 2005. http://cogprints.org/4385 (1-2-2009)

# Open vs. Proprietary Source Software in Croatia

Nikolaj Lazić
Department of Phonetics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
nlazic@ffzg.hr

Mihaela Banek Zorica
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mbanek@ffzg.hr

Jasmin Klindžić
Department of informatics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
jklind@ffzg.hr

**Summary**

*This paper shows differences between open source software and proprietary source software from the Croatian user perspective. These confronting approaches have their own positive and negative aspects which are viewed through present 2009 financial crisis and the need for lifelong learning projects promoted by the Croatian Ministry of Science, Education and Sports. User survey was conducted at the Faculty of Humanities and Social Sciences, University of Zagreb.*

**Key words:** open source, education

## Introduction

There are three different properties of the software that we can distinguish. Those could be understood as a kind of metadata for a program as they say nothing about the purpose of the program itself but only in which cases it can be used. Or one could think of these as legal data for usage stating rights of the author and user of the program.

First distinction we have to make is open source and proprietary source (sometimes also called closed source) software. We mark something as proprietary when the source code is kept secret and never presented to the public. In oppo-

site we have open source software which is released to the public thus enabling end user of the software to see how that particular program works.
The second distinction we make will be commercial and non-commercial software. Commercial software aims to make money from the use of the software either before the user uses it of after a trial period. Non-commercial software does not require payment for usage of the software at any time.
Third distinction that can be made only for open source software and that is free and non-free software by the definition of Free Software Foundation[1]. Free software is, in essence, defined as one that is open source and cannot be included in proprietary software.
One can try to "categorise" existing software according to these properties: Microsoft Office is a proprietary source, non-free, commercial software; OpenOffice.org is an open source, free, non-commercial software; IBM Lotus Symphony is proprietary, non-free, non-commercial software.
Having in mind 2009 financial crisis and constant shortage of funds in complete world economy we wanted to test the usage of software according to open versus proprietary source software (bearing in mind that this would usually also mean commercial versus non-commercial software) among students at the Faculty of Humanities and Social Sciences.
Hypothesis was that users would choose open source or non-commercial software because this would help local economy. If one does not have to pay for software then this money could be spent for personal education or for tutorials in local institutions. This approach would keep the funds inside the country and push local economy.

## Methodology of the survey

Survey was done using online survey software[2] and presented to students of Information sciences and students of Phonetics. 88 students have participated in the survey and anonymously submitted the questionnaire.
Each student was presented with four groups of questions according to the purpose of the software: office packages, photo editing, vector diagrams and sound editing. For each program they were presented with four possible answers: I've never heard of the program; I've heard of it but I have never used it; I've used it but it did not meet my demands; I'm still using it.
The participants were asked not to browse on the web for the names of programs. They were also asked to try to finish the survey as fast as they can (around 2 minutes).

[1] Categories of Free and Non-Free Software - GNU Project - Free Software Foundation (FSF), 2009

[2] LimeSurvey.org, 2009

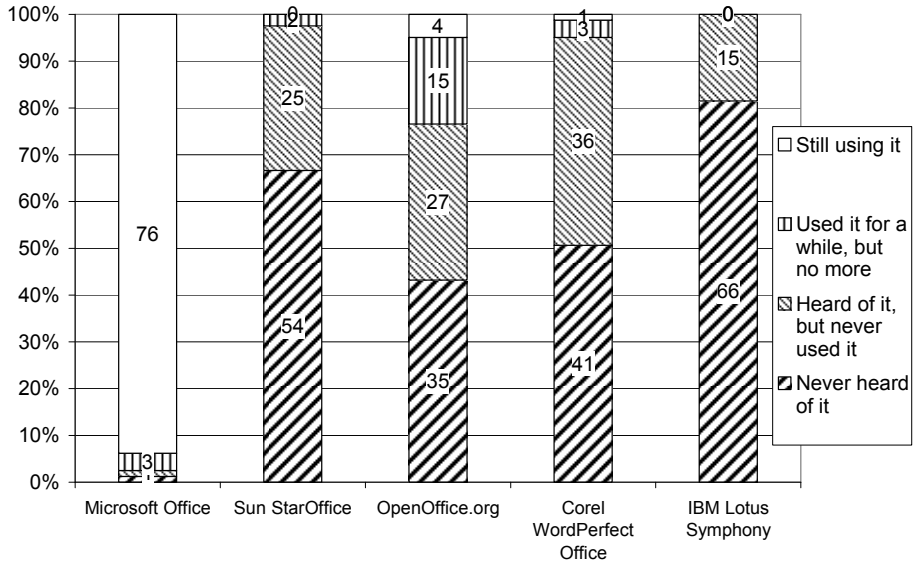
The results were then collected and each answer counted. The results were again grouped according to the purpose of the software in question and displayed as a chart.

## Results and discussion

81 of the students completed the survey and only those results are presented in following charts.

Chart 1 shows that more than 80% of participants have never heard of IBM Lotus Symphony package and more than 90% use Microsoft Office. Other results show that more than 40% of participants have never heard of other office packages then Microsoft Office.

Chart 1. Usage of different office packages (n=81)




These results were also surprising for photo editing software (see Chart 2). More than 50% of participants use Photoshop and Microsoft Paint and more than 70% have never heard of GIMP.

Looking at Chart 3 we can see that more than 75% of participants have never used any vector graphic program. Cairo library was put in the questionnaire just to test if anybody has ever heard of this 2D drawing library used in many open source projects.

Another thing tested was sound editing. More than 50% of participants use Praat (Praat, 1998) for sound editing. This was not a surprise because students of Phonetics use it on regular basis and students of Information sciences are familiarised with the program through an elective course.

Chart 2. Usage of different photo editing software (n=81)

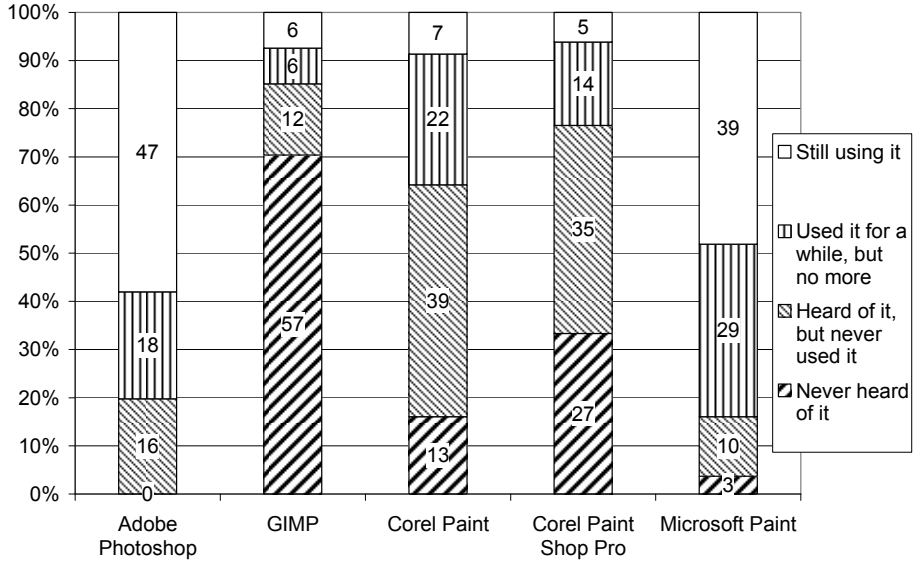



Chart 3. Usage of different programs for making vector diagrams and images (n=81)

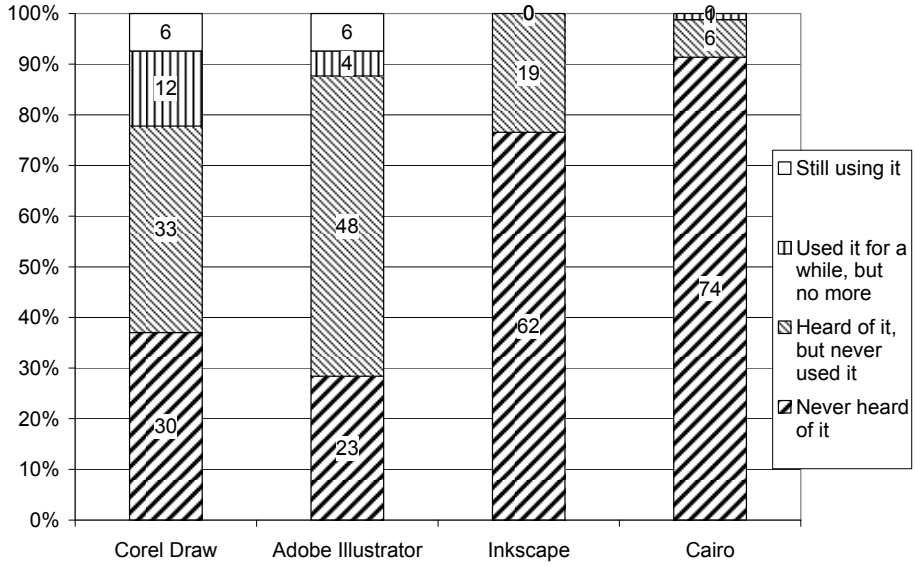
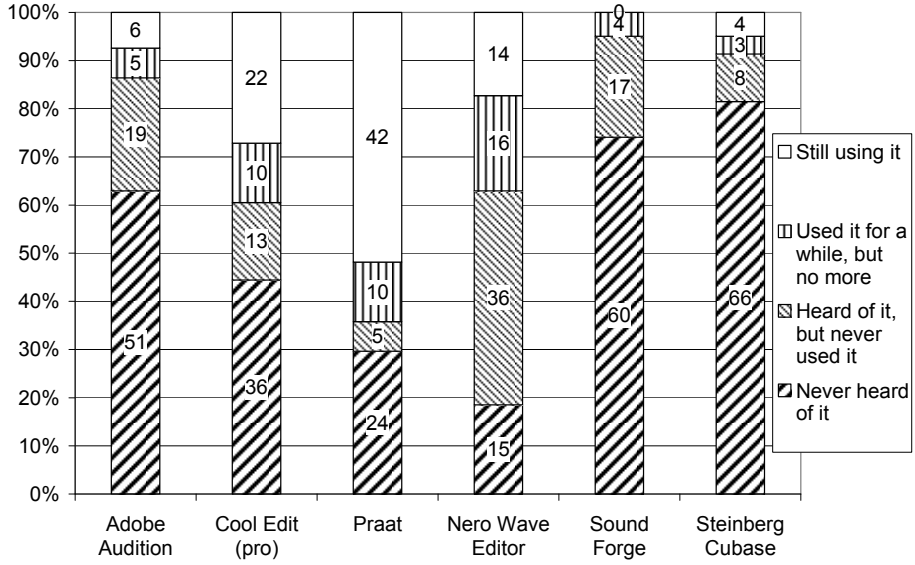


Chart 4. Usage of different sound editing programs (n=81).




Examining the whole situation we can see that participants tend to use commercial software in favour of non-commercial regardless of financial situation. Participants started using open source and free program for sound editing during a course held at our institution, and they kept using it afterwards.

The problem that arose was: why were participants only using commercial, proprietary software?

We had to backtrack through their education process.

For that data we examined the Croatian National Educational Standard (CNES) and National Curriculum Framework for Preschool Education and Elementary and High School Education[3] and Curricula for Elementary School[4] (both documents are only in Croatian language). These documents show that computer sciences courses in elementary school should "provide introduction to information and communication technology". There is no mention of what type of software they should use (in regard to its license) or any specific software. There is even a list of activities: editing photos on computer, editing text, programming languages (LOGO or some procedural language). As there is no mention of proprietary, commercial software there is no reason why pupils should not use non-commercial and even open source software for every task in their computer science courses.

[3] Nacionalni okvirni kurikulum za predškolski odgoj i opće obvezno obrazovanje u osnovnoj i srednjoj školi.

[4] Nastavni plan i program za osnovnu školu.

The answer was somewhere else. Each school (elementary, secondary) has a freedom to make their curricula the way they feel fit. We examined two curricula of secondary schools: school in Pazin[5] and school in Grubišno Polje[6] (both documents are only in Croatian language). Each school presents its curricula and declare that pupils will learn to use Microsoft Windows and Microsoft Office. Sentence from National Curricula stating that pupils will learn to edit digital photographs is translated to sentence: pupils will learn to use Photoshop.
This analysis shows that there is possibility for introduction of open source or non-commercial software to schools (according to CNES) but teachers are not willing to do that even though this would lead to higher level of computer literacy due to free availability to open source software.
Another example is the European CDL that lists its curricula as: text processing, spreadsheets, presentations etc… which is, at the end, also translated into: Microsoft Word, Excel and PowerPoint. Croatian ECDL[7] organisation does the same. There is no official mention of commercial software but it is there as default.
Looking at Chart 3 and data for Praat leads to a feasible solution: university teachers should present different choices to students and provide open source or non-commercial alternatives to commercial ones. Those students will eventually become teachers and teach in primary or high schools.
From economic point of view we should invest in our people and in our knowledge (as usually stated: "knowledge society" is something we should aim at). Using open source software can provide jobs at the same way as commercial one: there is always someone in need of knowledge and support in using software, either open source or proprietary one.
If we look at the present situation we can see that schools and universities "produce" people that can only use proprietary and commercial software. For that reason they have to buy the same software for home and their employers have to provide the same software for them at work place. At each instance we have to buy something that our country does not produce.

## Conclusion

At present time students at the Faculty of Humanities and Social Sciences are unfamiliar with open source software. The reason is exposure to only proprietary and commercial software throughout their education.
Students presented with open source programs continue to use it.
Open source and non-commercial software should be promoted in educational institutions for economical if not for any other reason.

[5] Gimnazija i Strukovna škola Jurja Dobrile Pazin: Školski kurikulum.

[6] Kurikulum Srednje škole Bartola Kašića Grubišno Polje za školsku godinu 2008/09.

[7] ECDL Hrvatska.

## References

Categories of Free and Non-Free Software - GNU Project - Free Software Foundation (FSF). 22.06.2009. http://www.gnu.org/philosophy/categories.html (01.07.2009.)

ECDL Hrvatska. http://www.ecdl.hr/ (01.07.2009.)

Gimnazija i Strukovna škola Jurja Dobrile Pazin: Školski kurikulum. 12.08.2008. http://www.gssjd.hr/wp-content/uploads/2009/02/skolski_kurikulum_radni_dokument_oblikovano_kraj.pdf (25.03.2009.)

Kurikulum Srednje škole Bartola Kašića Grubišno Polje za školsku godinu 2008./09. http://www.ss-bkasica-grubisnopolje.skole.hr/upload/ss-bkasica-grubisnopolje/newsattach/216/Kurikulum%20za%202008.-09..doc (23.03.2009.)

LimeSurvey.org - THE Survey software - free and open source. http://www.limesurvey.org/ (01.07.2009.)

Nacionalni okvirni kurikulum za predškolski odgoj i opće obvezno obrazovanje u osnovnoj i srednjoj školi. Zagreb : Ministarstvo znanosti, obrazovanja i športa, 2008. http://public.mzos.hr/lgs.axd?t=16&id=14170 (01.07.2009.)

Nastavni plan i program za osnovnu školu. Zagreb : Ministarstvo znanosti, obrazovanja i športa, 2006. http://public.mzos.hr/lgs.axd?t=16&id=14181 (22.03.2009.)

Boersma, Paul; Weenink, David. Praat: doing phonetics by computer. http://www.praat.org/ (23.03.2009.)

# Pedro Meyer's Retrospective Exhibition, Heresies – Bringing to Life an Innovative Model of Museum Presentation of Photography

Nataša Ivančević
Museum of Modern and Contemporary Art
Dolac 1/II, Rijeka, Croatia
natasa.ivancevic@mmsu.hr

**Summary**

*The Museum of Modern and Contemporary Art from Rijeka, among 100 museums in the world was invited to participate in the international project Heresies, a retrospective exhibition of four decades' work of one of the world's most innovative photographers, Pedro Meyer. The aim of this project was to construct an immense retrospective involving the publication of his images via specially created web pages. In October 2008, it was simultaneously opened in 65 museums around the world as an attempt to bring a new model of museum presentation of photography into life. Pedro Meyer's personal innovations in the field of digital photography include the creation of the first CD ROM that combined sound and visuals, first digital prints ever made, and the creation of the famous on-line photographic forum http://www.zonezero.com. With the Heresies exhibition, Meyer has directed his visionary view on the concept of the museum exhibition of photography, asking the question what do they look like today and how can they be remodeled for the future. Pedro Meyer posted 300 selected images in a private section of the web site pedromeyer.com. Participating curators have chosen between 10 and 90 of these images. Any print selected for inclusion in the Heresies exhibition became the property of the museum's permanent collection, as a gift from Pedro Meyer. Some of the printable images on the web site have been linked to audio files containing the artist's comments, and could have been downloaded by the museum for use in audio guides or visitor provided iPods. Pedromeyer.com is a website that serves as a living collection of his works, and on-line portfolio. Furthermore it was an interactive site where participants and visitors could fully take part. Meyer's new and heretic paradigm of a photographic exhibition includes: creative collaboration of curators and artists, global networking of 65 museums participating in the Heresies program, enhancement of investigative capacity and museum holdings, and educational programs for the iPod generation.*

**Key words:** retrospective exhibition, international project, digital photography, website, database, on line, iPod.

## Introduction

The transition from mechanical (analogue) into digital era, the rapid development of digital technologies and the transfer of information have reflected on changes made in the production of art at the end of the 20th century. Analogue photography, although still present, has been continuously suppressed by digital photography. It has become a photographic technique for the masses because of its availability, ease of use and distribution, lower cost, and it gives the user freedom to modify reality using computer programs (Adobe Photoshop and others). The Mexican photographer Pedro Meyer is known for his captivating and provocative photographs, but also for his pioneering work in digital technology. Meyer's personal innovations in the field of digital media includes the creation of the first CD ROM that combined sound and visuals (in 1991), the first digital prints ever made, and the creation of the famous on-line photographic forum http://www.zonezero.com[1] in 1993. In the 60's he worked with documentary photography, and since then many photographic cycles have evolved. When in 2002 Alejandro Castellanos, director of the Mexico City – based *Centro de la Imagen* (The Image Center) invited him to present a summary of the last five decades of his photographic work, he realized that the public would never have the opportunity to see his complete photographic opus, not even the greater part of his 80,000 archived photographs. He then started to think about the most appropriate way to reach a wider audience. He had to think of a new way of exhibiting, and the knowledge he gained by creating and maintaining the ZoneZero website helped him in this. Thus he created the idea of an international project named "Heresis" which after many years of preparation was held in October 2008 in 65 museums around the world.

### The role of Pedro Meyer

Pedro Meyer is one of the pioneers and most recognized representatives of contemporary photography, and one of the most widely recognized Mexican photographers. Pedro Meyer is the author of numerous exhibitions, lecturer in different academic institutions of the world.[2] He is the founder and president of the Mexican Council of Photography and the organizer of the first three Colloquiums, Biennials of Latin American Photography (1977-1983). His photo-

[1] ZoneZero won an Internet award from Encyclopædia Britannica.

[2] Mexico, USA, Germany, Argentina, Spain, Ecuador, and Sweden

graphs are part of museum holdings of many renowned museums of the world.[3] He has received numerous international prizes and awards.[4]

Throughout his life, Meyer has documented significant social events, such as the Student Movement of 1968, the Avándaro Rock Festival, the guerrilla in Nicaragua, and the tragic 1985 Earthquake in Mexico City. He also produced a major photo essay on Pemex, covered the presidential campaign of Miguel de la Madrid, and has done photographic portraits of celebrities of the cultural world, politicians, and of ordinary people during his many trips.[5] Christian Caujolle points out, in the introduction text of the book *Heresies* : "It is surprising or at least disconcerting: one of the most brilliant representatives of documentary photography in Latin America is also the person who has overtaken others and appropriated the historic mutations of images through the digital system to create disturbing images, rehabilitating photo-montage, using technology to continue expressing himself which constitutes the key to a curious, transformative work."[6] Meyer's photography consistently tests the borders of truth, fiction, and reality. With the development of digital photography in the early nineties, Meyer evolved from a documentary photographer of the so-called "real photography" into a digital documentarist who often combines photographic elements from different periods and spaces in order to achieve a different or a *higher* truth. Meyer's often expressed opinion is that all photographs – digitally manipulated or not – are equally *true* and *untrue*. This has been called heretic among orthodox documentary photographers, hence the title *Heresies*. In the 1990s he is the one who promoted the transition from analogue to digital photography, with project like *Truths and Fictions, I Photograph to Remember,* and *ZoneZero*. Despite resistance by "classical" photographers, he managed to gain recognition for digital photography to become acknowledged as a technique for art photography. To globally popularize and make photography accessible to a wide audience he created the on-line photographic gallery ZoneZero. It has been one of the most frequented web pages on the Internet. In its on-line galleries, ZoneZero today hosts more than a thousand renowned international photogra-

[3] New York Museum of Modern Art, The Victoria and Albert Museum London, The Musée National D'art Moderne Centre Georges Pompidou Paris, The International Center of Photography New York, George Eastman House, The California Museum of Photography, Tucson's Center for Creative, Havana's Casa de Las Américas and Centro Studie e Archivo della Comunicazione dell' Universitá Parma, Italy

[4] Guggenheim, 1987, Cultura Cita di Anghiari, 1985, National Endowment for the Arts, Mexican Photography Biennales

[5] http://www.pedromeyer.com/biography/biography.html

[6] Caujolle, Christian. Presentation // Heresies, Pedro Meyer. Madrid : The Pedro Meyer Foundation Lunwerg Editores, 2008. page 3

phers, and the pages are visited by more than 500,000 visitors a month and more than 5.5 million a year.[7]

***Heresies* project**

The fact that a large number of his photographs will never be seen by public motivated Meyer to think of a new way of museum presentation of photographs. He explains the creation of the project in his book Heresies: "I realized that I need to use the experience I had accumulated on the Internet in the years we had spent producing the ZoneZero website… We concluded that the only feasible solution would be to scan all the images and arrange them in a data base so that the person in charge of curating a particular topic could carry out the selection in the comfort of his own working space… Linking the on-line exhibition to the physical display would open up a host of possibilities. Thus, while a museum can display large-scale printed works, the Internet is capable of showing a far greater number of images and reaching a much wider audience."[8] In this way the public could search for photographs which are not exhibited in an on-line database. The problem of making available a large photographic archive was finally solved. A special program was developed for grouping, tagging and choosing photographs. Each digital file is tagged and has a caption which contains the name of the photograph, the dimensions of the negative, the technique (b/w, colour), the year and place where it was taken, the year of digitalization, and information about the digital modification. When that hard work was finished for the exhibition in The Image Center, it was decided that this worked out model should be offered to other museums around the world with defined conditions. During 2006 he invited one hundred museums from all over the world to participate in his project *Heresies* – retrospective exhibition of four decades' work. Museum of Modern and Contemporary Art from Rijeka was the only invited museum from the south-eastern European region, and the only one which had the opportunity to present this outstanding exhibition in the region. This fact gave encouragement to the museum staff because it acknowledged the international recognition of this institution. In October 2008, it was simultaneously opened in 65 museums around the world as an attempt to bring an innovative model of museum presentation of photography into life[9].

[7] During 2007 ZoneZero wins the following awards: "CNET Best of the Web" NET Magazine – awards ZoneZero, as one of the top all time 100 sites and one of the top 5 art web sites. Luckman Interactive gives ZoneZero a "Luckman Five-Star winner." NET Guide- gives 5 stars to ZoneZero and put it in its Platinum category. Izvor:
http://www.pedromeyer.com/biography/websites_en.html

[8] Meyer, Pedro. Behind the scenes and a list of acknowledgements // Heresies, Pedro Meyer. Madrid : The Pedro Meyer Foundation Lunwerg Editores, 2008. page 22

[9] List of participating museums available on:
http://www.pedromeyer.com/museums_list/museums.php?idiom=EN

**Conditions**

The invited institution received a written memo in which Meyer explained the project and defined the conditions which every participant had to meet:

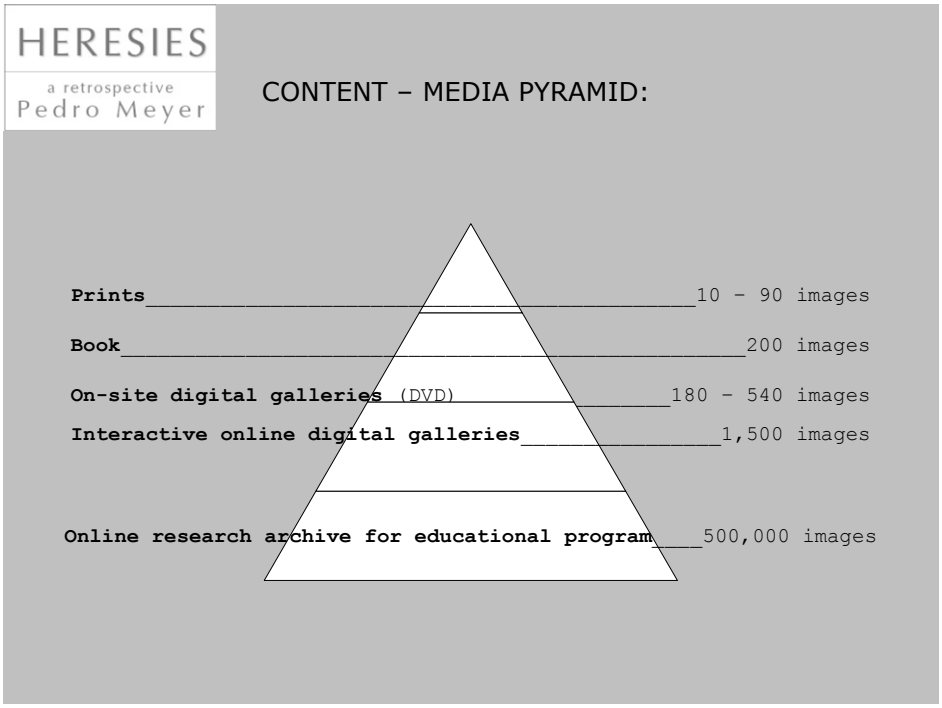
1. The opening of the *Heresies* exhibit must take place between October 6 and October 12, 2008.
2. *Heresies* must remain on exhibit in the museum for no less than 3 weeks.
3. All large-format archival prints delivered to the museum and hung in the *Heresies* exhibit will become the property of the museum's permanent collection at the conclusion of the exhibit. These prints may not be sold, given away or otherwise disposed of by the museum until October 12, 2058. Any print delivered but not hung must either be a. returned to Zone Zero at the museum's cost prior to the opening of the exhibit, or b. purchased by the museum for $6,000 USA. No reproduction rights to any photograph delivered to the museum in any form will be transferred to the museum. These rights must be negotiated separately. No more than 3,000 prints are available for the entire *Heresies* exhibit program. These will be allocated on a first-come, first-serve basis and solely at the discretion of the exhibit organizers.
4. Museum also agree to display at least three "digital galleries" on 1-3 DVD players attached to 1-3 large-screen television displays or video projectors. Museums agree to provide a DVD player / screen for each DVD provided. DVDs will be tailored to each museum's configuration.
5. Museums agree to post suitable directions, provided by Western Arts Management crediting exhibit sponsors and directing visitors to pedromeyer.com and to zonezero.com to view on-line interactive galleries.

**Innovations**

What is the innovative essence in the new model of the museum presentation of photography? The characteristics of *Heresies* set it apart from other photo projects that came before it. The aim of this project was to construct an immense retrospective involving the publication of his images, as well as more than 60 museums around the world that have joined in this global retrospective.

On the media pyramid we can see all elements the Heresies project included: Pedro Meyer posted 300 selected images in a private section of the web site pedromeyer.com. Participating curators have chosen between 10 and 90 of these images to hang in print form for the duration of the Heresies exhibition.

Picture 1. Media pyramid




The large-format archival-quality prints were delivered to the museums no less than five weeks prior to the opening. Each photo is labeled by a certificate which guarantees that the print of the current work of art has been supervised by the artist himself and carried out by the master printer Carlos Escobar in HP Designjet Z series printers with HP Vivera inks that last 200 years, in 310 grs. acid-free cotton Hahnemühle paper. The printer produced approximately 2,500 images requested by the museums, all of which were printed at the photographer's studio in Coyoacán in Mexico City and subsequently delivered to 65 venues all over the world.

Participating museums have been charged a shipping and handling fee of $995 USA. Any print selected for inclusion in the Heresies exhibition became the property of the museum's permanent collection at the conclusion of the exhibit at no additional cost. The curator of the Rijeka exhibition Nataša Ivančević, who leads the photography collection in cooperation with the consultant for photography Helena Srakočić have chosen 90 black and white and colour images from different Meyer's cycles, which are now part of the museum collection of photography. Some of the printable image on web site has been linked to a one- to two-minute audio file containing the artist's comments about that par-

ticular image. Audio files could have been downloaded by the museum for use in audio guides or visitor provided iPods.

Picture 2. Heresies newsletter number 6 with a certificate which labels each photograph

HERESIES NEWSLETTER #6

5/28/08

PEDRO MEYER - HERESIES
Authenticity Certificate

Master Pedro Meyer certifies that this photographic image is a work of himself, to whom all author's copyrights belong and which is part of the retrospective called Heresies. Pedro Meyer also certifies the veracity of the following card and technical specifications:

Title:
Date:
Location:
Print date:
Printed at Pedro Meyer's studio, Coyoacan, Mexico.

The print of the current work of art has been supervised by the artist himself and carried out by the master printer Carlos Escobar in HP Designjet Z series printers, with HP Vivera inks that last 200 years in 310 grs. free acid cotton Hahnemühle paper.

Pedro Meyer

Every photo you receive is certificated.

We are glad ti inform how your selected Heresies pictures will be delivered to you.

1. Every photograph will be packed in a thin plastic bag to protect the work against possible abrasion or humidity during delivery.

2. Inside the bag, in addition to the photo, there will be a certificate as the one shown above, containing all the information regarding the image.

3. To avoid any kind of problems with materials that could be either acidic or acid-free, we recommend that you don't glue or attach the certificate directly to the picture.

The Museum Coordination Team.

Nadia Baram Johan Trujillo Valeria Vega

1

In addition, as part of the project, Editorial House Lunwerg, has published the book Heresies separately, with over three hundred photographs selected by photographer Francisco Mata, as well as texts and reflections on the work of Pedro Meyer. The book is more than a catalogue of an exhibit, because no exhibit, being all of them different, concurs fully with its content. Each museum took its own path when selecting the works it would exhibit. The printed book features an audio visitor's guide that can be listened to on-line, so that the reader has the opportunity of hearing a personal interpretation of the content of his work by the author himself. Also, the electronic version of the printed book in PDF format will be downloadable without cost[10]. This will dissolve the habitual barrier that renders books unaffordable for many individuals.[11] Heresies

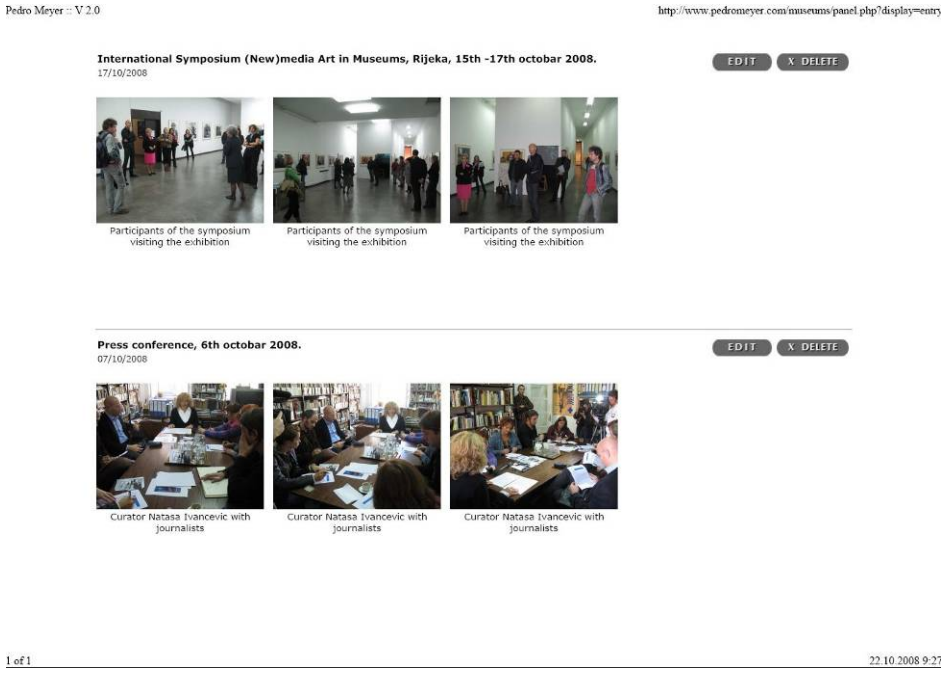
[10] http://www.pedromeyer.com/book/

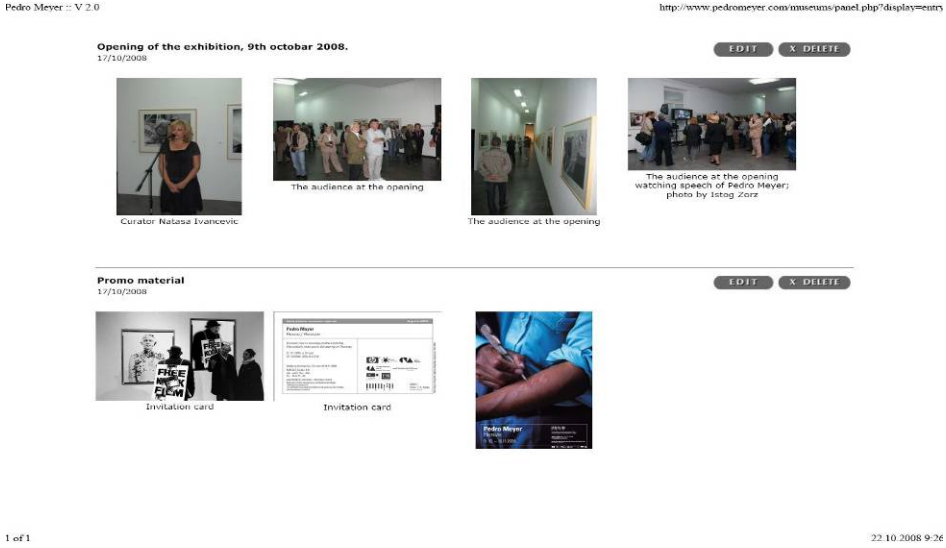
[11] http://www.pedromeyer.com/museums/press.php?idiom=EN

book was published one month prior to the opening. Museums purchased copies of the book from the publisher for sale at the museum.
Pedromeyer.com is a website that serves as a living collection of his works, and on-line portfolio. However, the database is the website's most important feature for it contains the entire body of Pedro Meyer's work. The objective of this database is to generate an on-line collection of images serving people in the field of research, communication, and knowledge. Furthermore it was an interactive site where participants and visitors could fully take part. In addition to the printable images, Meyer posted 23 "digital galleries" on website corresponding to the 23 subjects into which his work was divided for this project. Each digital gallery contained between 12 and 80 of Meyer's photographs selected by a leading editor or curator. In the Rijeka exhibition it has been presented as five digital galleries in a continuous loop. All digital galleries posted on web site were available for viewing by visitors, which means that they could investigate through 1.500 images.
Pedro Meyer and his staff have developed different educational programs which were available on web site. During the preparation period his main assistant and life partner Nadia Baram was sending newsletters with instructions and information necessary for the successful preparation of the exhibition, to everybody included in the project. Museums staff were invited to communicate between each other, and to post all relevant data on the specially designed link on web site. Each museum was invited to download photos related to the preparation and opening of the exhibition, and to all activities during the preparation and duration of the exhibition. Participating countries were: Brazil, Chile, Columbia, Ecuador, Mexico, Uruguay, Cuba, USA, Bangladesh, China, India, Pakistan, Singapore, Croatia, Italy, Slovakia, Spain and Australia.[12]
Since the opening was all over the world during the beginning of October, Meyer could not attend all the openings. Instead of being present physically, he screened himself and sent via internet, a short movie with a message of greeting in English[13] and welcomed his visitors virtually all over the world.

[12] The list of museums which participated in the project is available in the web site: http://www.pedromeyer.com/museums_list/museums.php?idiom=EN
[13] Type: QuickTime Movie, 18,7 MB, available on: http://www.pedromeyer.com/museums/press.php?idiom=EN

3. The documentation about the *Heresies* exhibition in Rijeka on its specially designed link on the web site



4. Photograph from the opening of the *Heresies* exhibition in Rijeka’s Museum of Modern and Contemporary Art

## Conclusion

With *Heresies*, Meyer has directed his visionary view at the concept of the museum exhibition of photography, asking the question, what they look like today and how they can be remodeled for the future. In the era of financial restrictions and the redefinition of the museum’s basic role, Meyer’s new and heretic paradigms of a photographic exhibition included creative collaboration of curators and artists, a worldwide networking of 65 museums participating in the Heresies program, enhanced research and collection-building capacity for museums and educational programs for the iPod generation.

## References

Heresies, Pedro Meyer. Madrid : The Pedro Meyer Foundation Lunwerg Editores
http://www.zonezero.com
http://www.pedromeyer.com

# LANGUAGE TECHNOLOGIES

# First Steps Toward Developing a System for Terminology Extraction

Petra Bago
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
pbago@ffzg.hr

Damir Boras
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
E-mail: dboras@ffzg.hr

Nikola Ljubešić
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
E-mail: nljubesi@ffzg.hr

**Summary**

*The aim of this paper is to describe first steps in developing a system for terminology extraction. First a data sample is built from synopses of doctoral theses at the Faculty of Humanities and Social Sciences, University of Zagreb, accepted in the period from 2004 to 2009 written mostly in Croatian language. Data sample consists of 420 documents and 338,706 tokens. A small sample was manually tagged for terminology to be used in an initial experiment. The approach for terminology extraction is knowledge-driven and consists of differential analysis of reference and domain-specific corpora. Specific method used is log-likelihood ratio test. Experiment deals with different reference corpora and linguistic pre-processing. First results are promising. Further research guidelines are discussed.*

**Key words:** terminology extraction, data sample, log-likelihood ratio test

## Introduction

Text mining deals with information detection in natural language texts. One area of text mining is called terminology extraction which applies to (semi)automatic extraction of technical terms of a specific domain. A list of

technical terms is a requirement for e.g. specialized dictionaries. This list can help to understand an area of expertise. There are two methods to approach this problem: statistics and linguistics. Statistical method is concerned with the idea of differential analysis, which is to find a correlation between specialized lexicon and general lexicon. Linguistic methods process the text mostly on morphological and syntactic level finding proper term candidates. Hybrid methods combine these two approaches (Witschel, 2004).

## Building a data sample

Documents collected for Synopsis corpus were downloaded from official web pages of the Faculty of Humanities and Social Sciences (Online arhiva dokumenata, 2009). Documents contained 420 synopses of doctoral theses accepted in the period from 2004 to 2009. They were exclusively digital texts in .doc format with a mostly uniform structure, which made it easier to import it into a database. Importing was done manually into a database management system Access 2003.

Table 1 shows the elements of the synopses with the number of synopses not containing the specific element.

Table 1: Elements of synopses

| Name of element | Number of synopses with empty part (%) |
|---|---|
| title | 0 (0.00%) |
| introduction | 325 (77.38%) |
| theoretical background | 1 (0.02%) |
| narrower field of work | 1 (0.02%) |
| aims and problems of research | 1 (0.02%) |
| methodology | 6 (1.43%) |
| expected scientific and/or practical contribution | 17 (4.05%) |
| structure of thesis | 326 (77.62%) |

### Processing

After importing data into the database, Synopsis corpus was verticalized i.e. tokenized. The token rule states that a token is a constant array of letter characters, wherewith the digits and punctuation are eliminated.

Synopsis corpus was semi-automatically lemmatized. Using several specialized databases helped detect a number of tokens and matching lemmas with its word category, while the rest was lemmatized by hand. Databases used for lemmatization are following: lexical database of the Croatian literary language (Kržak, 1985), Croatian Frequency Dictionary (Moguš, 1999), a database of surnames (Boras, 2003) and a database of settlements.

Finally, by tokenizing and lemmatizing, two new columns were added to the Synopsis corpus: lemma of a particular word and its word category.

## Corpus analysis

Synopsis corpus comprises 420 synopses of doctoral theses at the Faculty of Humanities and Social Sciences, University of Zagreb, accepted in the period from 2004 to 2009 written mostly in Croatian language. 305 synopses fall under the field of humanities (72.62%), while the rest of 115 fall under the field of social sciences.

Corpus has 338,706 tokens, of which 98.84% (334,799) are written in Croatian, while the rest of 1.16% (3,907) is written in other languages[1].The average size is 806.44 of tokens per document.

Corpus has 45,788 types, of which 95.08% are Croatian, while the rest of 4.92% (2,254) are in other languages. The average number of types per document is 51.32.

In Synopsis corpus one can find 338,706 tokens and 45,788 types, which makes a type-token ratio of 0.135. Researching on a corpus consisting of documents from the field of finances, (Tadić, 2003) detected that the type-token ratio for that corpus is 0.05. Comparing it to Croatian Frequency Dictionary (Moguš, 1999) where it is 0.119, they gave a possible explanation of why it is unusually high: "… the vocabulary in the field of finances shows less variation in inflection as well as limited number of different lexical entries than the general vocabulary" (Tadić, 2003). If we consider that argument to be true, the opposite statement would be an explanation of why type-token ratio for Synopsis corpus is lower than the one of Croatian Frequency Dictionary. This should not be a surprise if we keep in mind various subfields of humanities and social sciences (Table 2).

Table 2: Subfields of humanities and social sciences

| Field of humanities | Field of social sciences |
|---|---|
| Philosophy | Political science |
| Philology | Information sciences |
| History | Sociology |
| Art history | Psychology |
| Science of art | Science of education |
| Archaeology | |
| Ethnology and anthropology | |

## Initial experiment

The idea behind the initial experiment is to get a feel for the data and the terminology extraction problem in general.

[1] Languages other than Croatian that can be found in Synopsis corpus: English, Latin, German, Italian, French, Portuguese, Hungarian, Slovenian, Czech, Polish, Serbian, Romanian, Slovak, Greek, Old English, Dutch, Ikavian Croatian, Spanish, Istro-Romanian, Middle High German, Bosnian, Kajkavian Croatian, Turkish and Swedish.

The sample which was manually tagged and used as a gold standard is rather small. It consists of only one article which has 671 tokens. The sample is tagged by only one person so no interannotator agreement can be computed. There is also just this small tagged sample meaning that there is no possibility of having a development and an additional testing corpus which would make the methodology more accurate.
The sample was tagged in a straightforward fashion - the sample is verticalised and the rows containing a terminus or part of a multiword terminus are given an additional column with the value 1. Other tokens are given the value 0. Since in the corpus preprocessing lemmatization and part-of-speech tagging are performed, this information is also provided in the sample in the form of two additional columns.
The frequency of specific syntactic patterns is shown in Table 3. The data shows that most frequent patterns are the simple ones. Nevertheless, in such a small sample highly complex patterns also occur. One example showing very clearly the syntactic complexity of the text is the following: “... postmodernom ili postindustrijskom, a kod nas i postsocijalističkom društvu.” This phrase contains actually three terms: “postmoderno društvo”, “postindustrijsko društvo” and “postsocijalističko društvo”. It is very common in the whole sample that more terms share a common head in the noun phrase. Because of this syntactic complexity, in this experiment we will try to locate only tokens that are terms or just part of terms, and not their whole phrases. One of the obvious reasons for this is the lack of syntactic language tools for Croatian language.

Table 3: Frequency of specific syntactic patterns in tagged sample (N – noun, A – adjective, C – conjunction, x – not part of the term, N(g) – noun in genitive form, A(g) – adjective in genitive form)

| **syntactic pattern** | **frequency** |
|---|---|
| N | 11 |
| AN | 8 |
| A | 4 |
| NA(g)N(g) | 3 |
| ANCN | 3 |
| ACAN | 2 |
| AxAxxxxAN | 1 |
| NN(g)CN(g) | 1 |
| NN(g) | 1 |

The method used to identify tokens that are possible termini or parts of multiword termini is the log-likelihood ratio test introduced by Dunning (Dunning, 1993). This method is chosen as the first to be experimented on because of its popularity in the differential analysis community (Kiss, 2002; Witschel, 2005; Kuhn, 2009). The log-likelihood ratio compares two statistical hypotheses - the zero hypothesis that the token distribution in the corpus of interest and a well

balanced reference corpus is the same, and the alternative hypothesis - that they are not. In the Dunning log-likelihood ratio test the binomial distribution is used. The binomial likelihood of a token is computed as

$$L(p,k,n) = p^k (1-p)^{n-k}$$

with

$$p = \frac{k}{n}$$

where k is the token frequency and n the size of the corpus. The logarithm of the likelihood ratio is computed as

$$-2\log\lambda = 2[\log L(p_1,k_1,n_1) + \log L(p_2,k_2,n_2) - \log L(p,k_1,n_1) - \log L(p,k_2,n_2)]$$

with

$$p = \frac{k_1 + k_2}{n_1 + n_2}$$

The higher the -2logλ the more significant is the difference between the term frequencies. If $p_1$ is greater than $p_2$ for a specific token, than the -2logλ value shows how more important the token is for the first corpus and vice versa.
There are three different reference corpora used in the research - the corpus described in this paper and a small and a large newspaper corpus.
The large newspaper corpus is built from the on-line version of the daily newspaper Vjesnik. The initial size of the corpus is 746,683 tokens. Numerals and interpunctions are not included in the corpus. The corpus is also verticalized and additional information like the lemma and part-of-speech are added as separate columns. This reference corpus was morphosintactically tagged in a different manner than the corpus described in this paper. A trigram statistical tagger (Agić, 2006) was used and no additional human intervention was undertaken. This reference corpus is called “Vjesnik1”.
The small newspaper corpus is just a subset of the large newspaper corpus. It consists of 70,000 tokens. In this research it is called “Vjesnik2”.
The corpus described in this paper used as a reference corpus consists of 338,035 tokens. It includes all documents but the one used as the gold standard. This corpus is also verticalized and lemma and part-of-speech information is also present. This reference corpus is called “Synopsis”.

The possible advantage of this reference corpus to the Vjesnik reference corpora might be that the lemmatization and POS tagging method was identical as in the gold standard. The disadvantage could be the non-representativeness of this corpus. The newspaper corpus is also not really a well-balanced reference corpus, but is probably nearer to that idea than this one.
The free parameter that has to be optimized when using the log-likelihood ratio test is the result of the test. The -2log$\lambda$ value will be optimized concerning evaluation measures computed by comparing the gold standard and the result the method produces. Normally, an additional free parameter would be the minimum frequency of a token, but in this experiment this parameter will be fixed to 1. One of the arguments for doing so is the small size of the sample the experiments are performed on.
Three evaluation measures are computed on the classical measures of precision and recall - $F_{0.5}$, $F_1$ and $F_2$. The parameter optimization is performed concerning the $F_2$ measure. The reason for that is the most frequent usage of terminology extraction methods. Mostly the output is given to human specialists and therefore recall is more important than precision.
In this experiment baselines are considered random results. This means that when identifying terminology in the source without any POS-filtering, the probability of finding a terminus randomly is 70 divided by 671, i.e. 10.43%. This also means that on average every tenth token is a terminus.
The first experiment uses all three reference corpora. As features it uses plain lowercase tokens. In all cases the -2log$\lambda$ is optimized concerning the $F_2$ measure. In all experiments the -2log$\lambda$ measure takes values in range from 1 to 15 with step 1. The baseline is 0.104. The results are shown in Table 4.

Table 4: Evaluation measures regarding the reference corpus (RC) when using tokens as features

| RC | precision | recall | $F_{0.5}$ | $F_1$ | $F_2$ | -2log$\lambda$ |
|---|---|---|---|---|---|---|
| Vjesnik1 | 0.183 | 0.757 | 0.215 | 0.294 | 0.465 | 7 |
| Vjesnik2 | **0.194** | **0.757** | 0.228 | 0.309 | **0.479** | 7 |
| Synopsis | 0.180 | 0.743 | 0.212 | 0.290 | 0.457 | 4 |

The different experiment layouts show pretty similar results. The only significant difference is the -2log$\lambda$ optimal measure. When using any version of the Vjesnik reference corpus, it is 7 and, when using the Synopsis corpus, it is 4. The reason for that is probably the greater similarity between the gold standard and the Synopsis reference corpus. Interesting is also that the smaller newspaper reference corpus did not lower the result; on the contrary, it improved it, but not significantly. The reason for that can, of course, be also pure coincidence, i.e. the content of the smaller corpus.
In general all reference corpora show a significant improvement in comparison to the random baseline.

The distribution of part of speech in the gold standard and the optimal result in the previous experiment (Vjesnik2 as reference corpus and -2log$\lambda$=7) is shown in Table 5. As expected, the gold standard consists only of nouns and adjectives with exception of the conjunction "i" ("and"), since this conjunction is used where terms share the same head and human annotator considered it part of the multi-term noun phrase. The fact that the result consists also of other parts of speech (especially verbs) indicates the potential usefulness of a POS filter that will be introduced later in the experiment.

Table 5: Distribution of part of speech in the gold standard and the optimal result in the first experiment

| **part of speech** | **gold standard** | | **result** | |
|---|---|---|---|---|
| | type | token | type | token |
| noun | 35 | 39 | 97 | 144 |
| adjective | 22 | 25 | 71 | 84 |
| verb | 0 | 0 | 17 | 27 |
| conjunction | 1 | 6 | 0 | 0 |
| pronoun | 0 | 0 | 2 | 6 |
| number | 0 | 0 | 3 | 3 |
| abbreviation | 0 | 0 | 3 | 3 |

The second experiment has a similar layout to the first experiment, it just uses lemmata as features and not tokens. The baseline of this experiment is the same as in the previous case 0.104. The results are shown in Table 6.

Table 6: Evaluation measures regarding the reference corpus (RC) when using lemmata as features

| **RC** | **precision** | **recall** | **$F_{0.5}$** | **$F_1$** | **$F_2$** | **-2log$\lambda$** |
|---|---|---|---|---|---|---|
| Vjesnik1 | 0.118 | 0.514 | 0.139 | 0.191 | 0.307 | 1 |
| Vjesnik2 | **0.125** | **0.600** | 0.148 | 0.206 | **0.340** | 1 |
| Synopsis | 0.152 | 0.486 | 0.176 | 0.231 | 0.337 | 2 |

The data show a rather surprising result – a decline in all three reference corpora when using lemmata as features in comparison to using tokens. A possible explanation is that morphological normalization added less information than it was written in specific word forms. Interestingly, the smaller newspaper reference corpus secures a tight win in this experiment again. Second best is the synopsis corpus. The reason for that is probably the fact that lemmatization in the Synopsis reference corpus and the sample was realized with the same method while the Vjesnik corpus was lemmatized by a different method using different language resources. Optimal -2log$\lambda$ is in all cases very low. The reason for that is the unification done by lemmatization, namely the number of different values in the sample is now much lower.

The third experiment introduces a POS filter. Namely, only nouns and adjectives are allowed as results. This method should improve the results since al-

most all termini are or consist of only adjectives and nouns. The random baseline for this experiment is higher since now candidate termini are only nouns and adjectives That means that only 253 nouns and 134 adjectives, ie. 387 tokens are termini candidates. The probability of picking a terminus on random is 70/387, ie. 18.1%. The results are shown in Table 7.

Table 7: Evaluation measures regarding the reference corpus (RC) when using a POS filter and tokens as features

| RC | precision | recall | $F_{0.5}$ | $F_1$ | $F_2$ | -2log$\lambda$ |
|---|---|---|---|---|---|---|
| Vjesnik1 | 0.220 | 0.813 | 0.258 | 0.347 | 0.528 | 7 |
| Vjesnik2 | **0.205** | **0.891** | 0.242 | 0.333 | **0.534** | 5 |
| Synopsis | 0.211 | 0.859 | 0.248 | 0.338 | 0.532 | 3 |

These results show, as presumed, a significant improvement in comparison to the previous methods. Again, the winner is the Vjesnik2 reference corpus. In this method, the -2log$\lambda$ is slightly lower than when not applying a POS filter. Interestingly, the improvement of the POS filter is not too big. The reason is that the log-likelihood ratio test does a pretty good job in identifying primarily nouns and adjectives. The presumption is that the distribution of other part-of-speech entities is rather constant. In Table 8 the distribution of part of speech of the optimal results of the first experiment in comparison to the distribution on the whole reference corpus.

Table 8: Comparison of POS distributions in the result of the first experiment and the Vjesnik1 reference corpus

| part of speech | reference corpus | result | difference |
|---|---|---|---|
| noun | 0.384 | 0.56 | +45.8% |
| adjective | 0.274 | 0.32 | +16.8% |
| verb | 0.101 | 0.10 | -1.0% |
| other | 0.242 | 0.02 | -91.7% |

The results show that amost 90% of the tokens in the result are nouns and adjectives. In the newspaper reference corpus they make some 55% of all the tokens. Verbs are rather constant. Nouns and adjectives gain in the probability mass from other parts of speech. The conclusion is that other parts of speech are equally distributed over different samples. Nouns are mostly differently distributed. Adjectives take the second place. Verbs do not show any difference in the probability mass.

## Further research

Further research will include a bigger tagged sample. This sample, namely, contains only 671 tokens.

Different document sizes will be included in the research. For differential analysis the length of the domain-specific corpus is of great importance.

Experimenting with different text complexity will also be of interest. The doctoral synopses texts are very complex which was shown by the high type-token ratio. This sample is especially syntactically complex. That fact would make the process of finding syntactic cues for termini identification very hard. Samples will also be annotated by more annotators. That will provide us with the measure of interannotator agreement.
In further research the methodology of using distinct development and testing samples will be followed.
Further experiments will be conducted concerning the size and content of reference corpora.
The minimum frequency criterion for document features will also be included.
More methods of differential analysis will also be experimented with.

## Conclusion

This paper describes the process of building a data sample for terminology extraction and an initial research on the data.
The data sample consists of 420 documents and 338,706 tokens. The type-token ratio is high which indicates complex vocabulary. The sample is syntactically particularly complex.
At this point just a small portion of the sample is tagged. This part of the sample is used as a gold standard for the initial research.
An interesting result of the research is that a smaller newspaper reference corpus yields better results than the two other corpora. Additional research is necessary to inspect the reasons for such results.
When using lemmata as document features, results were consistently worse. We assume that more information was lost by not including tokens than information was gained by including lemmata. A combination of both features could further improve results.
The POS filter improves the results significantly by choosing only nouns and adjectives as candidate termini. When not using the POS filter, nouns and adjectives are chosen more often than by chance. This leads to the conclusion that they differ between corpora more than verbs and, especially, other parts of speech. Nouns differ more than adjectives.
In general, the investigated methods achieve significantly better results than the random baseline.
Further research will include a bigger and more versatile gold standard, different reference corpora, more annotators and a more complex methodology.

## References

Agić, Željko; Tadić, Marko. Evaluating Morphosyntactic Tagging of Croatian Texts. // *Proceedings of the 5th International Conference on Language Resources and Evaluation* / Genova: ELRA, 2006.

Boras, Damir; Mikelić, Nives; Lauc, Davor. Leksička flektivna baza podataka hrvatskih imena i prezimena. // *Modeli znanja i obrada prirodnoga jezika.* / Zagreb: Zavod za informacijske znanosti, 2003, 219-237

Dunning, Ted. Accurate Methods for the Statistics of Surprise and Coincidence. // *Computational Linguistics.* 10 (1993), 1; 61-74

Kiss, Tibor; Strunk, Jan. Scaled log likelihood ratios for the detection of abbreviations in text corpora. // *Proceedings of the 19th International Conference on Computational Linguistics* / ACL, 2002, 1-5

Kržak, Miroslav; Boras, Damir. Rječnička baza hrvatskog književnog jezika = Lexical Data Base of the Croatian Literary Language. // *Informatologia Yugoslavica*. 17 (1985), 3 4; 223-242.

Kuhm, Adrian. Automatic labeling of software components and their evolution using log-likelihood ratio of word frequencies in source code. // *Mining Software Repositories, 2009. MSR '09. 6th IEEE International Working Conference on* /2009, 175-178

Moguš, Milan; Bratanić, Maja; Tadić, Marko: Hrvatski čestotni rječnik. Zagreb : Školska knjiga, Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu, 1999.

Online arhiva dokumenata. 17.07.2009. http://www.ffzg.hr/dokument/index.php?cid=1801 (20.07.2009.)

Tadić, Marko; Šojat, Krešimir. Finding Multiword Term Candidates in Croatian. // *Proceedings of Information Extraction for Slavic Languages 2003 Workshop* / Borovets : BAS, 2003, 102-107.

Witschel, Hans Friedrich. Terminology Extraction and Automatic Indexing -- Comparison and Qualitative Evaluation of Methods. // *Proceedings of Terminology and Knowledge Engineering (TKE)* / 2005

Witschel, Hans Friedrich. Text, Wörter, Morpheme – Möglichkeiten einer automatischen Terminologie-Extraction. (Diploma thesis) Leipzig, 2004.

Original scientific paper

# Automatic Keyphrase Extraction from Croatian Newspaper Articles

Renee Ahel
Faculty of Electrical Engineering and Computing,
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
renee.ahel@gmail.com

Bojana Dalbelo Bašić
Faculty of Electrical Engineering and Computing,
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
bojana.dalbelo@fer.hr

Jan Šnajder
Faculty of Electrical Engineering and Computing,
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
jan.snajder@fer.hr

**Summary**

*Keyphrases provide a way to summarize documents and enable cross-category retrieval. The paper describes a robust system for automatic keyphrase extraction from newspaper articles in Croatian language. Keyphrase candidates are generated based on linguistic and statistical features, and naïve Bayes classifier is used to select the best keyphrases among the candidates. A prediction model is built using training documents with human-assigned keyphrases. System performance is measured on a corpus of newspaper articles, by comparing the automatically extracted keyphrases with those assigned by professional indexers. In absence of comparable results, we consider our results to be of modest performance.*

**Key words:** keyphrase extraction, naïve Bayes classifier, Croatian language

## Introduction

In recent decade, the number of digital documents is growing exponentially, caused by ever-growing use of computers in every aspect of human endeavor. Along with that growth, the need for efficient search, indexing, and categorization over those documents becomes ever more important. A practical way of

summarizing document contents is to assign keyphrases. Keyphrases describe the content of the document, thus enabling the user to decide whether a particular document is relevant for his or her information need, without reading the entire document. Unfortunately, most authors assign keyphrases to their documents only when they are compelled to do so. Manual assignment of keyphrases is a tedious task, especially considering the ever-growing amount of documents. Thus, there is a great need for means of automatic keyphrase assignment.

Keyphrase assignment methods can be divided into two categories: *keyphrase assignment* and *keyphrase extraction*. Both methods revolve around the same fundamental problem of selecting the best keyphrases, and in both methods this problem is tackled as a machine learning problem – a chosen algorithm learns the "good keyphrase" concept using documents with manually pre-assigned keyphrases. Automatic keyphrase assignment is the closest to human keyphrase assignment in that keyphrases are chosen from a predefined thesaurus (a limited, possibly hierarchically organized set of possible keyphrases). Keyphrase extraction generates keyphrases from document text, rather than from a thesaurus. Since authors may assign keyphrases that do not exist in document text, the performance of extraction methods tends to be lower than that of assignment methods. Turney (2000) argues that 65%–90% of author assigned keyphrases appear verbatim in document text, implying a rather satisfactory upper bound on recall at 90%.

To the best of our knowledge, no previous keyphrase extraction or assignment approaches have been developed for Croatian language. This paper deals with keyphrase extraction method only, and applies it to the Croatian newspaper articles. For a more efficient keyphrase extraction, problems relating to inflectional morphological complexity of the Croatian language will also be addressed.

The rest of the paper is structured as follows. In Section 2 an overview of related work for keyword extraction for English language is given. Section 3 describes the four phases of our keyphrase extraction algorithm, while Section 4 shows the results of experimental evaluation. Section 5 concludes the paper and outlines future work.

## Related work

The majority of work on keyword extraction deals with the English language. Turney (2000) used Quinlan's C4.5 decision tree algorithm combined with bagging technique, as well as his own algorithm called GenEx. GenEx algorithm consists of a parameterized keyphrase extraction algorithm paired with Genitor genetic algorithm for parameter optimization. He applies GenEx on several corpora, ranging from e-mails and scientific papers to user and institutional web pages. Several variations of the C4.5 algorithm are applied by varying the number of trees and the proportion of sampled positives. Algorithm C4.5 with 50 bagged trees and 1% positive candidate sampling was shown to perform best,

while GenEx algorithm has been shown to consistently outperform C4.5 on all tested corpora.

Witten et al. (1999) used a naïve Bayes classifier on a subset of corpora that was used by Turney as well as their own CSTR (Computer Science Technical Reports) corpus. Their so-called KEA algorithm outperforms Turney's on some corpora.

KEA++ is an improved version of the KEA algorithm devised by Medelyan and Witten (2006). KEA++ performs thesaurus-based term expansion using the Agrovoc thesaurus. On documents from agricultural domain, KEA++ significantly outperformed KEA.

Hulth (2003) used a rule induction system combined with bagging technique and POS tags, n-grams, and NP-chunks based methods on a corpus of scientific papers from the Inspec database. Results indicated that POS tags method outperforms other methods, while additional combination of all three methods helps reduce false positives.

Ercan and Cicekli (2007) used the C4.5 algorithm combined with bagging technique. They calculated candidate features using lexical chains and WordNet ontology. Their results indicated that the use of lexical chain-based features improves keyword extraction.

Work reported in this paper is inspired by KEA (Witten et al. 1999), mainly because this algorithm has shown good performance using a rather simple set of features. We improve on KEA's candidate generation and feature selection mechanisms, and additionally use POS tag filtering similar to that of Hulth (2003). Turney's approach is less adequate in our case because the GenEx algorithm was tailored for the English language, while C4.5 was in fact shown to perform worse than KEA. Other abovementioned approaches are yet unfeasible due to the lack of appropriate linguistic resources for Croatian language.

## Keyphrase extraction

In this section, we describe the four phases of our keyphrase extraction algorithm: preprocessing, candidate generation, feature calculation, and learning and classification. The algorithm operates on documents that consist of a title, text body, pre-assigned keyphrases, and pre-assigned set of categories. Additional information source that we use is a predefined category taxonomy.

### Pre-processing

Each individual document is pre-processed as follows. Document text is first divided into sections determined by the so-called *phrase boundaries* (punctuation marks, brackets, and non-printable characters such as new line, new paragraph, etc.). Phrase boundaries limit parts of document text within which keyphrase candidates are generated. Afterwards, text is tokenized based on regular expressions; tokens not consisting of at least one letter or digit are discarded and all tokens are case-folded. Tokens are lemmatized using the automatically ac-

quired inflectional lexicon (Šnajder et al. 2008). Because in our case lemmatization does not disambiguate homographs, each token may have one or more lemmas assigned. Along with each lemma, a POS tag is assigned using the same lemmatizer. If the lemmatizer does not recognize a token, the original value of the token is assigned as the lemma, and a special POS tag "F" (lemmatization failed) is assigned. The set of possible POS tags is: N (noun), A (adjective), V (verb), X (stopword – pronoun, conjunction, preposition, interjection, number, particle), F (lemmatization failed). Stopwords are detected by comparing against a fixed list of stopwords for Croatian language. The same processing that is applied to document text body is performed on the pre-assigned keyphrases, as well as the title of the document. Additionally, for each token from the document text we check if it appears in (1) a name of a category from category taxonomy or (2) the document title.

**Candidate generation**

Next step is the keyphrase candidate generation, for which we use the previously determined phrase boundaries. Candidate is generated as a sequence of one, two, or three consecutive tokens within neighboring phrase boundaries. Since every token has one or more lemmas assigned, for each candidate all possible lemma combinations are generated, along with their POS tag combinations (POS patterns).

After all lemmatized forms for a candidate are generated, a POS pattern filter is applied. Only those candidates having at least one lemmatized form whose POS pattern passes the POS filter are retained; for these candidates, only lemma forms passing the POS filter are stored. POS patterns used in the POS filter are taken from (Petrović et al. 2009), where they were used to detect collocations; preliminary experiments have shown that these patterns will also be suitable for keyphrase extraction. POS filter patterns are given in Table 1.

Table 1: POS patterns used in POS filter

| |
|---|
| N, F |
| AN, NN, NF, FN |
| NXN, NAN, AAN, ANN, NNN, ANF, NXF, AFN, FNN, FXN, NFN, NNF |

For each candidate that passed the POS filter, some further processing is performed. It is determined if the candidate appears in the category taxonomy; a candidate is considered to appear in the category taxonomy if at least one of its tokens appears in the category taxonomy. Similarly, a candidate is considered to appear in document title if at least one of its tokens appears in the title. Position of candidate appearance within document text is determined as the position of the first candidate token in the document.

Since candidates are generated within phrase boundaries, it is possible that on document level multiple instances of the same candidate are generated. Candidate instances are resolved using lemmatized forms of the candidates, otherwise

different inflectional forms of the same candidate would not count as distinct candidates.
In our case, since every candidate can have more than one lemmatized form, two candidates are considered to be instances of the same candidate if they share at least one common lemmatized form. All subsequent instances are replaced by the candidate instance that appears first in a document.
After candidate instances have been resolved, a match with pre-assigned keyphrases is performed. It is again based on comparison of corresponding lemmatized forms, using the abovementioned matching principle.
Note that only one lemmatized form match is sufficient to determine a match. Also note that pre-assigned keyphrases do not have to pass the POS filter, whereas the candidates do. Thus, some pre-assigned candidates will never match to a potential candidate if that candidate did not pass the POS filter. Nevertheless, insisting that candidates must pass the POS filter is a deliberate trade-off in which a small number of false negatives is traded for a more significant reduction in the number of false positives.

**Feature calculation**
After keyphrase candidates have been generated, each candidate is described using the following four features:

- relative first appearance in document text ($First_R$),
- *TFxIDF* value,
- a value that indicates whether the candidate appears in category taxonomy (*IsInCategory*),
- a value that indicates whether a candidate appears in document title (*IsInTitle*).

Relative first appearance in document text is determined using the following formula:

$$first_R(C) = \frac{first(C,D)}{size(D)},$$

where *first(P,D)* represents the position of first appearance of candidate *C* in text of the document *D*, and *size(D)* represents the total length of document *D*, measured in tokens.
*TFxIDF* is calculated as:

$$TFxIDF = \frac{freq(C,D)}{size(D)} \times \log_2 \frac{N}{df(C)},$$

where *freq(C,D)* represents the frequency of the candidate *C* in document *D*, value *df(C)* represents the number of documents in the entire corpus where candidate *C* appears, and *N* is the total number of documents.
The motivation behind the *IsInCategory* and *IsInTitle* features is that candiates are likely to be more descriptive – and thus better keyphrases – if they appear in category taxonomy or document titles, respectively.

Relative first appearance and *TFxIDF* are quantitative variables, which can pose a problem for the naïve Bayes classifier. Moreover, Witten et al. (1999) argue that discretizing quantitative variables improves prediction performance. We employ two means of discretization: percentile discretization (feature values are ordered ascending and divided into 100 bins containing equal numbers of candidates) and entropy-based discretization using the Minimum Description Length principle proposed by Fayyad and Irani (1993).

**Learning and ranking**

Application of the naïve Bayes classifier to the problem of keyphrase extraction using described features amounts to calculation of two posterior probabilities for each candidate *C*:

$$p(key \mid First_R(C), TFxIDF(C), IsInCategory(C), IsInTitle(C)) = p(key)p(First_R(C) \mid key)p(TFxIDF(C) \mid key)p(IsInCategory(C) \mid key) p(IsInTitle(C) \mid key) \quad (1)$$

$$p(\neg key \mid First_R(C), TFxIDF(C), IsInCategory(C), IsInTitle(C)) = p(\neg key)p(First_R(C) \mid \neg key)p(TFxIDF(C) \mid \neg key)p(IsInCategory(C) \mid \neg key) p(IsInTitle(C) \mid \neg key) \quad (2)$$

i.e., the probability that, given its feature values, candidate *C* is a keyphrase or is not a keyphrase, respectively. Value *p(key)* is the *a priori* probability that any given candidate is a keyphrase. Value $p(First_R(C)|key)$ is the conditional probability that a specific feature value $First_R(C)$ will appear, given that candidate *C* is a keyphrase; and similarly for other features. Formulae (1) and (2) derive from the Bayes theorem under the independence events assumption, which – despite being often theoretically unjustified – has been found to work well in practice.

The learning process consists of estimating, for each feature value, the probabilities necessary to calculate posterior probabilities given by (1) and (2). To this end, we use the so-called m-probability estimate (Mitchell 1997).

The ranking process for candidate *C* consists of calculating the posterior probabilities according to (1) and (2). Calculated probabilities are then normalized by their sum to unit scale. Candidates from the same document are ordered descending by the normalized value of (1) and descending by *TFxIDF*. Similar candidates are then removed from ordered list of candidates; two candidates are considered similar if they overlap in text, and in this case the lower-ranked of the two candidates is removed from the list. Resulting ranked list is the final list

of extracted keyphrases. From this list we take the *N* top-ranked candidates as the document keyphrases.

## Corpora

Two corpora of Croatian newspaper articles have been made available by the Croatian News Agency (HINA): *October 4457* and *January 4532*. Documents in both corpora have pre-assigned keyphrases; from these phrases, we filtered out those that do not appear in text. Basic statistics for the two corpora is summarized in Table 2.

Table 2: Basic statistics for corpora *October 4457* and *January 4532*

| | **October 4457** | **January 4532** |
|---|---|---|
| Number of documents | 3905 | 4521 |
| Average document length | 525 | 335 |
| Average number of candidates per document | 235 | 153 |
| Average number of keyphrases per document | 2 | 2 |

### Deficiencies of the available corpora

Human indexers have assigned keyphrases to documents regardless of whether the phrases appear in document text. In our case – because we are addressing the task of keyword extraction rather than keyword assignment – we have filtered out the keyphrases that do not appear in document text. This reduced the total number of keyphrases by 57%. Discarding keyphrases that do not appear in text inevitably deteriorates the quality of the training corpora. In some cases, more descriptive keyphrases will be removed, while the less descriptive keyphrases remain, merely because they appear in document text. Consequently, some meaningful automatically extracted keyphrases will not be matched against less descriptive pre-assigned keyphrases.

Another issue worth mentioning is that, in order to objectively judge system's performance, an evaluation of inter-annotator agreement should be carried out. We leave this important issue for future work.

Inconsistency is yet another curious characteristic of the described corpora, reflected by the fact that 63% of all keyphrases have been assigned to a single document. Not only does this deteriorate the quality of the training corpora, but it also raises doubts about the usefulness of the pre-assigned keyphrases as a means for cross-category search.

### Experimental training set

From the described corpora we have compiled an experimental dataset for training and evaluation. The set consists of 200 newspaper articles with pre-assigned keyphrases that appear in document text. On average, there are 6.5 such keyphrases per document and 370 keyphrase candidates per document. Because current version of our system generates candidates only up to length three, pre-assigned keyphrases longer than that were discarded from the training set; this amounts to 1.7% of the pre-assigned keyphrases.

## Results and discussion

In the three experiments that follow, performance is measured in terms of the $F_1$ measure with 10-fold cross validation on the training set. Because cross validation is used, results are expressed as a mean value of all ten iterations.

First experiment compares the performance of two basic configurations, each using a different discretization method. Results for some selected numbers of extracted keyphrases are given in Table 3. Performance rises steeply until seven extracted keyphrases, and more slowly afterwards. Best performance is achieved for 10 keyphrases with MDL discretization and for 12 keyphrases with percentile discretization. Results show that MDL discretization consistently outperforms percentile discretization.

Table 3: Performance of percentile discretization compared to MDL discretization

| | Extracted keyphrases | Precision (%) | Recall (%) | $F_1$ (%) |
|---|---|---|---|---|
| **MDL** | 1 | **22.0** | 3.4 | 5.9 |
| | 7 | 13.4 | 14.5 | 13.9 |
| | 10 | 12.5 | 19.3 | **15.1** |
| | 15 | 10.4 | **24.1** | 14.5 |
| **Percentile** | 1 | **18.5** | 2.9 | 5.0 |
| | 7 | 12.1 | 13.1 | 12.6 |
| | 12 | 10.5 | 19.5 | **13.7** |
| | 15 | 9.4 | **21.8** | 13.1 |

The second experiment examines the effect of additional POS filtering of the candidates. Prior to this, we have carried out an analysis of POS patterns of all candidates from corpora *October 4457* and *January 4532* (cf. Section 4). Results of this analysis, given in Table 4, reveal that by filtering out the candidates that do not match the POS patterns N, AN, NN, NXN we can discard 30% negative candidates, while discarding only 7.5% positive candidates.

Table 4: POS patterns of keyphrase/not keyphrase candidates

| POS pattern | Keyphrase (%) | Not a keyphrase (%) | Total (%) |
|---|---|---|---|
| N | 49.77 | 38.75 | 38.87 |
| AN | 28.17 | 13.26 | 13.42 |
| NN | 11.58 | 10.41 | 10.42 |
| F | 1.91 | 9.72 | 9.64 |
| NXN | 2.98 | 7.57 | 7.52 |
| NAN | 1.56 | 3.22 | 3.2 |
| Other | 4.02 | 17.06 | 16.92 |

Performance of two basic configurations combined with the described POS filtering is shown in Table 5. Results reveal that additional POS filtering consistently improves performance. E.g., for 10 extracted keyphrases with MDL discretization the improvement is 2.1%, while for 12 extracted keyphrases with percentile discretization the improvement is 1.7%.

Table 5: Performance with additional POS filtering of the candidates

| | Extracted keyphrases | Precision (%) | Recall (%) | $F_1$ (%) | % change $F_1$ |
|---|---|---|---|---|---|
| **MDL + POS filtering** | 1 | **23.5** | 4.0 | 6.9 | 14 |
| | 7 | 14.9 | 17.9 | 16.3 | 14 |
| | 10 | 13.7 | **23.4** | **17.2** | 12 |
| **Percentile + POS filtering** | 1 | **18.0** | 3.1 | 5.3 | 6 |
| | 7 | 13.2 | 15.9 | 14.4 | 13 |
| | 12 | 11.4 | **23.5** | **15.4** | 11 |

The best overall result from the above experiments is obtained for 10 keyphrases using MDL discretization with POS filtering. Examples of keyphrases extracted from two documents using that configuration are given in Table 6 (true positives are underlined).

Third experiment is an ablation study: we measure the influence of each feature on performance by holding out one feature and doing keyphrase extraction using the remaining features. Experiment is carried out using the previously established best configuration (MDL + POS, 10 keyphrases). Results are given in Table 7.

Except for *IsInTitle* feature, which seems not to contribute to the performance, all other features seem to positively affect the performance. The improvement is, however, statistically significant at the 0.05 level on for the *TFxIDF* feature.

Table 6: Extracted keyphrases compared to pre-assigned keyphrases

| Document title | Pre-assigned keyphrases | Extracted keyphrases |
|---|---|---|
| *Lijevo-desna nerazumi-jevanja* | *ljevica*<br>*politički život*<br>*vrijednosti*<br>*socijalna država*<br>*socijaldemokrati*<br>*liberali*<br>*socijalna politika*<br>*crkveni socijalni nauk*<br>*ekonomska politika*<br>*suverenitet*<br>*zakonitost* | *poimanju*<br>*suvereniteta*<br>*stranke*<br>*politika*<br>*različitim poimanjem*<br>*predizbornu kampanju*<br>*zakonitosti*<br>*liberale*<br>*socijaldemokrate*<br>*konzervativce* |
| *EU lansirala petogodišnji plan sigurnosti* | *akcijski plan*<br>*sigurnost*<br>*imigracijska politika*<br>*pravosuđe*<br>*granična kontrola*<br>*međunarodna suradnja* | *petogodišnji plan*<br>*sigurnosti*<br>*akcijski plan*<br>*imigracijske politike*<br>*pravosuđa i sigurnosti*<br>*europsku sigurnost*<br>*suradnje između zemalja*<br>*terorizma*<br>*sigurnost granica*<br>*slobode i sigurnosti* |

Table 7: Influence of each feature on performance of the best configuration (MDL + POS, 10 keyphrases)

| | Precision (%) | Recall (%) | $F_1$ (%) | % change $F_1$ |
|---|---|---|---|---|
| **IsInCategory** | 13.4 | 22.9 | 16.9 | -2 |
| **IsInTitle** | **13.7** | **23.5** | **17.3** | 1 |
| **First$_R$** | 12.5 | 21.4 | 15.7 | -9 |
| **TFxIDF** | 10.5 | 18.0 | 13.2 | -23 |

## Conclusion

Keyphrase extraction is a practical way to summarize document contents. Doing it manually on a large number of documents is a tedious task. In this paper, an algorithm for keyphrase extraction in Croatian language has been described and evaluated. In absence of comparable results, we consider our results to be of modest performance. This is mainly due to performance measure being based on pre-assigned keyphrases – though these were assigned by professional indexers, we suspect that the inter-annotator agreement might have been low. Furthermore, we have determined that documents from training corpora had inconsistently assigned keyphrases, which certainly negatively affects the performance. Nevertheless, results suggest that despite these deficiencies, it is possible to improve the extraction performance.

Future work can be focused on determining a more suitable performance measure, improving the quality of training corpora, as well as applying methods for dealing with the class disproportion problem.

## Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia and under the Grant 036-1300646-1986. The authors are grateful to the Croatian News Agency (HINA) for making available the newspaper corpus.

## References

Ercan, Gonenc; Cicekli, Ilyas. Using Lexical Chains for Keyword Extraction. // *Information Processing and Management.* 43 (2007), 6; 1705–1714

Fayyad, Usama M.; Irani, Keki B. Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. // *Proceedings of the 11th International Joint Conference on Artificial Intellingence (IJCAI'93)* / Bajcsy, R. (ed.). Chambery, France : Morgan Kaufmann, 1993, 1022–1027

Frank, Eibe; Paynter, Gordon W.; Witten, Ian H.; Gutwin, Carl. Domain-specific Keyphrase Extraction. // *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99)* / Dean, T. (ed.). Stockholm, Sweden : Morgan Kaufmann, 1999, 668–673

Hulth, Annette. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. // *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)* / Sapporo, Japan : ACL Publications, 2003, 216–223

Hulth, Annette. Reducing False Positives by Expert Combination in Automatic Keyword Indexing. // *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'03)* / Borovets, Bulgaria : 2003, 216 – 223

Medelyan, Olena; Witten, Ian H. Thesaurus Based Automatic Keyphrase Indexing. // *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* / Chapel Hill, NC, USA : ACM Press, 2006, 296–297

Medelyan, Olena; Witten, Ian H. Thesaurus-based Index Term Extraction for Agricultural Documents. 09.05.2005. http://www.cs.waikato.ac.nz/~ihw/papers/05-OM-IHW-Agri-Docs.pdf (05.07.2009.)

Mitchell, Tom M. Machine Learning. New York : McGraw – Hill, 1997

Petrović, Saša; Šnajder, Jan; Dalbelo Bašić, Bojana. Extending Lexical Association Measures for Collocation Extraction. // *Computer Speech and Language.* (2009); doi:10.1016/j.csl.2009.06.001

Šnajder, Jan; Dalbelo Bašić, Bojana; Tadić, Marko. Automatic Acquisition of Inflectional Lexica for Morphological Normalisation. // *Information Processing and Management*. 44 (2008), 5; 1720–1731

Turney, Peter D. Learning to Extract Keyphrases from Text. // *Information Retrieval*. 4 (2000), 2; 303–336

Witten, Ian H.; Paynter, Gordon W.; Frank, Eibe; Gutwin, Carl. Kea: Practical Automatic Keyword Extraction. // *Proceedings of Digital Libraries 99 (DL '99)* / Rowe, N.; Fox, E. A. (ed.). Berkeley, CA, USA : ACM Press, 1999, 254–255

Original scientific paper

# Comparative Analysis of Automatic Term and Collocation Extraction

Sanja Seljan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
sanja.seljan@ffzg.hr

Bojana Dalbelo Bašić, Jan Šnajder, Davor Delač
Faculty of Electrical Engineering and Computing,
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
bojana.dalbelo@fer.hr, jan.snajder@fer.hr, davor.delac@fer.hr

Matija Šamec-Gjurin, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
msamecgj@ffzg.hr

Dina Crnec
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
dcrnec@ffzg.hr

**Summary**
*Monolingual and multilingual terminology and collocation bases, covering a specific domain, used independently or integrated with other resources, have become a valuable electronic resource. Building of such resources could be assisted by automatic term extraction tools, combining statistical and linguistic approaches.*
*In this paper, the research on term extraction from monolingual corpus is presented. The corpus consists of publicly accessible English legislative documents. In the paper, results of two hybrid approaches are compared: extraction using the TermeX tool and an automatic statistical extraction procedure followed by linguistic filtering through the open source linguistic engineering tool. The results have been elaborated through statistical measures of precision, recall, and F-measure.*

**Key words**: automatic extraction, term and collocation base, English language, evaluation metrics

## Introduction

Term and collocation resources have become useful tool in business, education, and research. Building of such resources, both monolingual and multilingual, can be greatly facilitated by using existing tools for automatic term extraction. Although such tools – especially those combining statistical and linguistic approaches – certainly do require human intervention, they nonetheless can substantially speed up the process. At present times, when Croatia is approaching the EU and undergoing a period of intensive international written communication, use of electronic resources (monolingual and multilingual dictionaries, terminology and collocation bases) could be of a considerable help in the translation work. As the most frequently translated language pair is English-Croatian and vice versa, this paper presents the pilot project of the monolingual term extraction from the English legislative documents, in future to be followed by the Croatian language counter pair. Such resources covering a specific domain may be used in machine translation and computer-assisted translation, information retrieval, building of multilingual bases, glossaries, thesauri, document indexing, and in creation of semantic networks.

According to the Sager's list of requirements (Love, 2000), the *term* should relate directly to the concept and express it clearly. It should be lexically systematic, not overlap in meaning with other terms, and independent from the context. It should not convey unnecessary information and it should not be pleonastic. The terms should conform to the general rules of word-formation, be capable of providing derivatives, and preserve the original transcription.

On the other hand, according to (Manning and Schütze, 2002) *collocations* contain two or more consecutive words expressing a conventional way of saying things and therefore appear more frequently near each other (e.g., *in general, King of England, freeze up*). Collocations are characterized by non-compositionality (meaning of the collocation can not be predicted from the meaning of the parts), non-substitutability (components can not be substituted), and non-modifiability (not modified through additional lexical material of grammatical transformations). Collocations are considered to be a subset of *multi-word expressions* that constitute arbitrary conventional associations of words within a particular syntactic configuration (Wehrli et. al, 2009). In *multi-word units* the component words include meaningful units (e.g., *Knight of the Round Table*). Although multi-word units are composed of two or more orthographic words (linked by dash, conjunction, or blank), they are treated as a single grammatical unit. Multi-word (MW) units can include foreign expressions (e.g., *ad hoc*), prepositions (e.g. *freeze up, depend on*), adverbs (e.g., *of course*), idiomatic noun constructions (e.g., *know how, per cent*), expressions (e.g., *well being*), as on BNC (British National Corpus) web-page.

The term extraction process, by which a list of term candidates is generated, generally includes two phases (Harris et al., 2003; Thurmair, 2003):

(i) *term extraction* (*term acquisition*), which amounts to identification of term candidates in a corpus, and
(ii) *term recognition*, which refers to verification with a pre-defined list created by an expert in order to identify the (un)known terms.

In this paper, the research on term and collocation extraction from monolingual corpus is presented. The corpus consists of English legislative documents publicly accessible at the EUR-Lex web page[1] providing direct free access to European Union law and the appropriate Croatian translation available at TAIEX–CCVista[2] – Technical Assistance and Information Exchange, containing translations of legal acts of the EU. In the paper the results of two approaches to term extraction are compared: (i) extraction using the TermeX tool, developed at Faculty of Electrical Engineering and Computing, Knowledge Technologies Laboratory and (ii) the automatic statistical extraction procedure followed by linguistic filtering through the open source linguistic engineering tool. The results have been elaborated through statistical measures of precision, recall, and the F-measure.

## Related work

There are several collocation extraction tools available today. One of the first collocation extraction tools is the Xtract (Smadja, 1991; Smadja, 1993). Xtract tries to detect collocations based on association measures (AMs), predictive relations, and phrasal templates. Collocate (Barlow, 2004) is a commercial tool that offers collocation extraction based on PMI and Log Likelihood association measures. A span of up to twelve words (12-gram) can be extracted using PMI, whereas Log Likelihood can be used only to extract 2-grams. Collocate does not use morphological normalization such as lemmatization, but is capable of processing previously POS-tagged corpora.

Another tool, presented in (Seretan, Nerima and Wehrli, 2004), is an advanced collocation extractor designed for computer aided translation. It differs from the aforementioned tools in that it focuses on syntactic analysis combined with AMs. The TermeX tool used in this work differs from the above-mentioned tools in that it provides a much wider range of AMs to choose from: as much as fourteen different AMs for extraction of 2-, 3-, 4-grams are provided. To improve extraction performance, TermeX uses morphological normalization, POS filtering, and filtering by frequencies.

Much of the work on the usability of extraction tools, hybrid approaches, and their integration into machine translation and information retrieval systems has been discussed by Thurmair (2003) Building of bilingual lexicons by extracting bilingual entries from aligned bilingual text using bidirectional transfer has been

[1] http://eur-lex.europa.eu/en/index.htm

[2] http://ccvista.taiex.be

discussed by Turcato(1998). According to Wehrli et al. (2009), "collocations could present a particular problem for machine translation, because of their frequency, their different morpho-syntactic properties, and long-distance dependencies." The ITS-2 system presented by Wehrli et al. (2009) is a large-scale translation system relying on a detailed linguistic analysis provided by the parser, which exploits monolingual lexicons. In their research, the transfer system is used to produce information-rich phrase-structure representation related to the predicate-argument structure, identifying multi-word expressions such as idioms and collocations. Extraction of collocations is made by a hybrid method, combining syntactic information from the parser with statistical methods for detection of typical constructions in the corpus. Another two hybrid models (Daille 1996; Izuha 2001) for term extraction from parallel bilingual text used linguistic various statistical scores for ranking. Extraction of multi-word expressions for the Croatian language have been presented in (Bekavac and Tadić, 2008).

According to the previous research, best results are obtained using hybrid approaches. In this research two types of hybrid approaches will be presented and compared. The first model uses statistical lexical association measures (AMs) combined with POS filtering and morphological normalization.

The second approach extracts the terms in two steps. It first uses statistical extraction regardless of the length of n-grams, filtered by a predefined frequency threshold and a stop-words list. In the second step, the list of potential candidates is fed through language dependant local grammars in the NooJ engineering tool, combined with its high-priority dictionary for disambiguation.

## Research

### Resources

This research includes a selection of ten different types of EU legislation documents related to the EU activities: three Council Decisions, one Commission Decision, one Decision of the European Central Bank, three Council Regulations, and two Commission Regulations – in total amounting to about 20,000 words. The documents have been translated from the original Croatian legislation. The texts have been revised and used for creation of a term and collocation base. Extraction process was made by two independent groups of researchers using:

- TermeX tool (Delač, 2009) developed at the Faculty of Electrical Engineering and Computing in Zagreb, Knowledge Technologies Laboratory;
- a statistically-based term extraction tool SDL Multi Term Extract and a linguistically-based environment NooJ (Silberztein, 2004) developed at the University Franche-Comté Paris, France.

For the purpose of evaluation, a reference list has been created by the human experts, representing the gold standard of terms typical for EU legislation vocabulary.

**Tools**

*Approach A*

The first approach uses the TermeX tool (Delač, 2009), a tool for construction of terminology lexica with possible applications in NLP. Collocation extraction in TermeX is based on association measures (AMs), statistical measures that provide information on how likely it is for an n-gram (the sequence of *n* words) to be a collocation. Extraction is done by creating ranked lists of n-grams based on their AM value. This way terms that are most likely to be a collocation become top ranked. TermeX implements fourteen AMs, based on Pointwise Mutual Information (PMI), Dice, and Chi-square. Implemented measures are extensions of the corresponding bigram measures for n-grams spanning up to four words as described in (Petrovic, 2009)0. In order to improve collocation extraction, TermeX implements POS filtering and morphological normalization to better cope with morphological complexity of natural languages.

In TermeX, a terminology lexicon is created by selecting collocations from lists of automatically extracted collocation candidates. Building of a single lexicon is referred to as a *project;* multiple corpora can be processed simultaneously as parts of a single project. For the purpose of this experiment, TermeX was first run on the *Acquis communautaire* corpus to gather the complete statistical data, after which the terms from the ten selected documents were extracted. The AMs used in this experiment were PMI for 2-grams and heuristic measures described in (Petrovic, 2009)0 for 3-grams and 4-grams. It should be noted that the AMs and POS filters used by TermeX are optimized for the extraction of noun phrases rather than verb phrases.

*Approach B*

The second type of research started from the language independent statistically-based approach with a predefined frequency threshold using SDL Multi Term Extract. It offered a number of term candidates and probable translations, both presented in a term candidate list on a user-friendly graphic interface. After validating terms and their translations, it was possible to export them to MultiTerm XML or a tab-delimited format. This list was then filtered from stop-words.

In the next step, the specialized language dependent tool NooJ was used. NooJ is a linguistic engineering platform providing tools for the formalization of language phenomena at different levels: orthography, morphology, lexicon, syntax, and semantics. It therefore includes large-coverage dictionaries and grammars, and parses corpora in real time. Its linguistic engine is multilingual and there are a dozen of modules for different languages available, as well as a dozen more being prepared. NooJ processes texts and corpora in numerous file formats (varying from HTML, PDF, and MS Office to XML documents). NooJ issues sophisticated queries in order to produce various results (i.e., concordances, statistical analysis, or information extracts). In this research, statistically ob-

tained lists were filtered by 36 types of regular expressions (local grammars) in order to identify word combinations that match certain POS patterns. For the purpose of disambiguation, an additional pre-compiled filter dictionary in NooJ was set up at high priority level, after which the linguistic analysis was performed.

**Lists**

*Reference list*

The reference list contains 470 terms and collocations, excluding unigrams. The terms in the list vary from bodies' titles, functions' titles, documentation and common phrases, introductory and operative clauses, etc. Creation of a reference list is a rather difficult task, aiming to cover a specific domain, but balancing between lexical coverage, adequacy for the domain, and inclusion of typical expressions.

The reference list in this case study contains

- terms as semantic units in canonical forms (*acquiring company, annual account, applicant country*),
- collocations chosen because of their frequency at the pragmatic level as “preferred ways of expressing things”, according to Thurmair (2003) (*adopt a resolution, decided as follows, entry into force, for the purpose of, having regard to*), names and abbreviations (*Economic and Monetary Union EMU, European Union EU, European Central Bank ECB*), and embedded terms relevant for the domain (*crime prevention, crime prevention bodies, national crime prevention measures*).

While terms are mainly noun phrases (346 out of 470), collocations also contain many verbal phrases. Distribution of n-grams in the reference list is presented in Table 1.

Table 1: Number of n-grams in the reference list

| N-grams | 2-grams | 3-grams | 4-grams | 5-grams and more |
|---|---|---|---|---|
| Total: 430 | 119 | 138 | 98 | 75 |

*List A*

The list extracted with TermeX consists of 1816 terms. TermeX uses POS filters tuned to extract noun phrases consisting of two, three, and four words. Of the 1816 extracted terms, 758 consist of two words, 679 terms consist of three words, and 379 consist of four words. Most of the extracted terms are indeed semantically full noun phrases, mostly named entities and compound nouns.

*List B*

Using a language-independent statistically-based SDL Multi Term Extract tool, a list of terms has been obtained. This list was filtered by the list of stop-words, eliminating words such as determiners, pronouns, prepositions, conjunctions,

etc. that appear at the beginning or at the end positions of candidates. The number of extracted term candidates, with frequency threshold set to 4, was 369. This list included not only semantically full terms, but also meaningless sequences of words or unfinished terms, requiring for, e.g., a noun, past participle, or a prepositional phrase, but extracted because of their frequency. These lists also contain terms that embed a noun and a number (e.g., Directive 68/151/EEC), which should not be included in the term base. Therefore, these lists contain considerable number of meaningless candidates, which would not pass the linguistic test. In the next step, 36 local grammars have been applied on the statistical list, containing mostly <A><N> and <N><N> candidates, followed by <N><PREP><N>, <A><V>+<G><N> (G for gerundive, i.e., verb in gerundive form), <N><CONJ><N> and <N><A>, as presented in Table 2. Percentage of local grammars is presented as one example per match. Because many of the lexical items were polysemous meanings, a new dictionary in NooJ was compiled and set up at high priority level. After linguistic filtering, a list of 512 term candidates has been created. The reason for bigger number (512 after linguistic filtering comparing to 369 candidates after statistical analysis) lies in the extraction of embedded terms (e.g., after pure statistical approach the term *applicable to public limited-liability companies in the Member* was extracted while after linguistic approach the following candidates were identified: *public limited, limited-liability, liability companies*; the term *Counterfeit Analysis Centre* extracted in statistical analysis was identified after linguistic analysis as *Counterfeit Analysis, Counterfeit Analysis Centre, Analysis Centre, Analysis Centres*).

Table 2. Local grammars

| **Regular expressions (local grammars)** | |
|---|---|
| | 1 example per match |
| **Most common** | <A><N> **31%** |
| | <N><N> **30%** |
| | <N><PREP><N> 17% |
| | <A><V>+<G><N> 12% |
| | <N><CONJ><N> 6% |
| | <N><A> 6% |
| **Least common** | <N><CONJ><A><N><N> |
| | <V>(<DET>)<N><N> |
| | <V><A><N> |
| | <A><N><A><N> |
| | <A><N><P><N> |

## Results

### Statistical Analysis

Statistical analysis is performed via measures of precision, recall and the F-measure, by comparing the lists of terms extracted by the two tools against the terms form the reference list.

*Recall* is defined as the proportion between valid computer extracted terms and expert extracted terms (the reference list), although it is hard to define the relevant set in the reference list regarding the quality and the quantity. The perfect recall score of 100% indicates that all valid terms were extracted, but does not say anything about the fact how many irrelevant terms were also extracted.

*Precision* is defined as the proportion between valid computer extracted terms and all computer extracted terms. As precision reflects the noise, it is also possible to have certain amount of false positives, i.e., terms that are extracted by the tool, but not included in the reference list. The perfect precision score of 100% indicates that every extracted term was relevant, but does not at all indicate whether all relevant items were extracted.

*F-measure* (or F-score) allows adjusting the relationship between recall and precision. The F-measure is the weighted harmonic mean between precision and recall.

Table 3. Results of extraction evaluation

| | **List A** | **List B** |
|---|---|---|
| No. of terms | 1816 | 508 |
| Valid terms | 202 | 234 |
| Precision (%) | 11.56 | 47.37 |
| Recall (%) | 42.98 | 49.79 |
| F1 (%) | 18.22 | 48.55 |

True positives were calculated by taking into account the inflectional variants of terms: a simple suffix stripping procedure was applied to conflate the inflectional variants to a single canonical form. Note that a more sophisticated morphological normalisation procedure (such as lemmatisation) was not required in this case: suffix stripping did not introduce any ambiguity and linguistic validity of norms was not required. Moreover, when comparing two terms, the determiners were ignored, so that, for instance, *adopt a decision* and *adopt decision* would be considered as match.

The results are shown in Table 3. The results for list A are rather unsatisfactory, while for list B they are modest. The number of terms common to both lists is 355. The low recall for list A can be traced down to the fact that TermeX tool does not extract verb phrases nor does it extract terms consisting of more than four words. If such terms are removed from the reference list, recall reaches up to 77.47%.

Results of the list B could be improved by lemmatization in order to have expressions in canonical forms, definition of upper/lower cases, precision of determiner in collocations, and by more detailed local grammars.
In the lists, there are also a number of false positives, i.e., terms and collocations not found in the reference list. We plan to address this issue as part of future work.

## Conclusion

In this paper the results of two hybrid approaches to automatic term extraction were evaluated and compared. Human-created term and collocation lists differ from automatically created lists, mostly because of human knowledge, experience, and intuition that is involved in deciding whether a certain candidate can or can not be a term or a collocation.
Results show that extracted terms cover the specific domain in question and may serve to complement the dictionary, but there is certainly space for improvement.

Automatic extraction combined with human intervention may give usable results. We believe that the direction that should be taken is the fine-tuning of human criteria (when compiling the reference list) and application of hybrid models for automatic extraction.

## Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grants No. 130-1300646-0909 and 036-1300646-1986.

## References

Aoughlis, Farida. A Computer Science Electronic Dictionary for NooJ. Lecture Notes in Computer Science, 2007. pp. 341-351

Barlow, M.: Collocate 1.0: Locating collocations and terminology. TX:Athelstan, 2004.

Bekavac, B.; Tadić, M. (2008) A Generic Method for Multi Word Extraction from Wikipedia. Proceeings of the ITI 2008 30th Int. Conf. on Information Technology Interfaces, 2008.

Daille, Béatrice. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pages 49–66. MIT Press, Cambridge, Massachusetts, 1996.

Delač, Davor; Krleža, Zoran; Dalbelo Bašić, Bojana; Šnajder, Jan; Šarić, Frane : TermeX: A Tool for Collocation Extraction. Lecture Notes in Computer Science (Computational Linguistics and Intelligent Text Processing, 2009; 149-157.

Dillinger, M. Dictionary Development Workflow for MT: Design and Management. MT Summit VIII, 2001.

Drouin, Patrick. *Acquisition automatique de termes : l'utilisation des pivots lexicaux spécialisés*. Thèse de doctorat présenté à l'Université de Montréal, Montréal, 2002. http://www.olst.umontreal.ca/pdf/DrouinPhD2002.pdf

Drouin, P. Term extraction using non-technical corpora as a point of leverage. In *Terminology*, 2003, vol. 9, no 1, p. 99-117. http://www.olst.umontreal.ca/pdf/ Terminology_2003.pdf, 9.6. 2008.)

Harris, M.R.; Savova, G. K.; Johnson, T.M.; Chute, C.G. (2003) A term extraction tool for expanding content in the domain of functioning, disability, and health: proof of concept. J Biomed Inform.; 36(4-5) 250-9, 2003.

Izuha, T. Machine Translation Using Bilingual Term Entries Extracted from Parallel Texts (R). MT Summit VIII, 2001.

Keller, Frank. Evaluation: Connectionist and Statistical Language Processing http://homepages. inf.ed.ac.uk/keller/teaching/internet/lecture_evaluation.pdf, 9.6.2008.)

Love, Stacy. *Benchmarking the performance of Two Automated Term-extraction systems: LOGOS and ATAO*. Mémoire de maîtrise, Université de Montréal, 2000. http://www.olst. umontreal.ca/pdf/memoirelove.pdf, 6.6.2008.)

L’Homme, Marie-Claude and Hee Sook Bae. A Methodology for Developing Multilingual Resources for Terminology. In *Proceeding of LREC 2006. Language Resources and Evaluation, 2006*. http://www.olst.umontreal.ca/pdf/LREC-2006-Lhomme-bae.pdf, 9.6. 2008.)

Manning, Chiristopher. D.; Schütze, H. Foundations of Statistical Natural Language Processing. MIT, 2002.

NooJ http://www.nooj4nlp.net

Petrović, S., ·Šnajder, J., Dalbelo Bašić, B.: Extending lexical association measures for collocation extraction. Computer, Speech and Language, 2009 (doi:10.1016/j.csl.2009.06.001.).

SDL MultiTerm Extract http://www.sdl.com/en/products/products-index/multiterm.asp

Smadja, F.: Retrieving collocations from text: Xtract. In: Proceedings of 31th Annual Meeting of the Association for Computational Linguistics. Vol. 19, 1993, 143-177

Smadja, F.: From n-grams to collocations: An evaluation of Xtract. In: Proceedings of 29th Annual Meeting of the Association for Computational Linguistics, 1991. 279-284

Seretan, V., Nerima, L., Wehrli, E.: A tool for multi-word collocation extraction and visualization in multilingual corpora. Proceedings of EURALEX Congress, 2004.

Silberztein, M. NooJ: A Cooperative, Object-Oriented Architecture for NLP. In : INTEX pour la Linguistique et le traitement automatique des langues. Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté, 2004.

Tadić, M., Šojat, K.: Finding multiword term candidates in Croatian. Proceedings of Information Extraction for Slavic Languages 2003 Workshop IESL, 2003, 102-107

Thurmair, Gregor. Making Term Extraction Tools Usable. Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, EAMT-CLAW 2003.

Turcato, Davide. Automatically Creating bilingual Lexicons for Machine Translation from Bilingual Text. Proceedings of the 17th international conference on Computational linguistics, vol. 2, 1998, 1299 - 1306

Vintar, Špela. Extracting terminological collocations from parallel corpus. 5th EAMT Workshop, 2000.

Wehrli, E.; Seretan, V.; Nerima, L.; Russo, L. Collocations in a Rule-Based MT System: A Case Study Evaluation of Their Translation Adequacy. Proceedings of the 13th Annual Conference of the EAMT, 2009, 128–135

Zienlinski, D.; Safar, Y.R. (2005) Research meets practice: t-survey 2005: An online survey on terminology extraction and terminology management.

Original scientific paper

# Biological and Cognitive Plausibility in Connectionist Networks for Language Modelling

Maja Anđel
Department for German Studies
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
mandel@ffzg.hr

**Summary**

*If we want to explain cognitive processes with means of connectionist networks, these networks have to correspond with cognitive systems and their underlying biological mechanisms in different respects.*
*The question of biological and cognitive plausibility of connectionist models arises from two different aspects – first, from the aspect of biology – on one hand, one has to have a fair understanding of biological mechanisms and cognitive mechanisms in order to represent them in a model, and on the other hand there is the aspect of modeling – one has to know how to construct a model to represent precisely what we are aiming at. Computer power and modeling techniques have improved dramatically in recent 20 years, so the plausibility problem is being addressed in more adequate ways as well. Connectionist models are often used for representing different aspects of natural language. Their biological plausibility had sometimes been questioned in the past. Today, the field of computational neuroscience offers several acceptable possibilities of modeling higher cognitive functions, and language is among them.*
*This paper brings a presentation of some existing connectionist networks modeling natural language. The question of their explanatory power and plausibility in terms of biological and cognitive systems they are representing is discussed.*

**Key words:** connectionism, computational neuroscience, language

## Introduction

The development of computers in the second half of the twentieth century and new insights in functioning of the human brain, which developed simultaneously, brought many researchers to start thinking of brain as of an extremely sophisticated information processing device. This idea (often labelled as *computer-metaphor*) led further to the emergence of the so-called subsymbolic models of cognitive processes. The first wave of enthusiasm subsided after recognizing limits of perceptrons (Minsky, Papert 1969), but after a while, network

architectures were improved (Rumelhart, McClelland, PDP research group 1986) and computers became much faster and more efficient, resulting in an increased interest in artificial neural networks.

## From Early Connectionism to Computational Cognitive Neuroscience

*Connectionism* (term first used in Feldman, Ballard 1982) is

> "[...] an approach to artificial intelligence (AI) that developed out of attempts to understand how the human brain works at the neural level and, in particular, how people learn and remember. (For that reason, this approach is sometimes referred to as neuronlike computing.)"[1]

The term itself refers to the fact that small computing units (neuronlike units) are interconnected through a range of connections whose weights have to be adjusted during the learning process in order to satisfy all the constraints imposed on the network by the learning task.

After 20 years, models as metaphors are closer to their original inspiration, brain, and the term *computational neuroscience*, although not new (cf. Sejnowski, Koch, Churchland 1988), is sometimes preferred, in order to emphasize the increasing similarity:

> "[...] computational neuroscience makes systematic use of mathematical analysis and function of living brains, building on earlier work in both neural modeling and biological control theory." (Arbib, 2002:11)

From this definition one can see that there are two possible directions for a researcher to choose from: on one side, putting more weight to mathematical analysis, striving for better computational efficiency – which is useful in robotics, or, on the other side, giving more attention to the imitation of biological structures, which leads to building models of cognitive processes. However, recent development has shown that adhering to biologically plausible models does not necessarily mean neglecting computational efficiency, as will be shown in further text.

## Biology as inspiration

It is often pointed out that connectionist networks are inspired by biological networks of neurons, and therefore they are also known as artificial neural networks. In the beginnings of connectionism, neurons and populations of neurons were represented as simple computational units, and connections between them as relatively simple mathematical functions. Due to this reductionism, connectionist models have often been criticized in different respects (Fodor, Pylyshin 1988, Pinker, Prince 1988).

[1] Citation from Britannica online, accessed on August 13, 2009.

However, as the time goes on, the mathematical functions and algorithms have become more sophisticated and more adjusted to known biological processes and functions, so the focus of criticism has shifted towards other domains. Recently, their capability to represent cognitive processes without using symbol manipulation has been questioned (cf. Marcus 2003). This debate (symbolic vs. subsymbolic processing) still goes on without an answer (cf. Christiansen, Chater 1999), so there is no intention to give the answer in this paper either, but rather to show what connectionist models are capable of doing today in explaining psycholinguistic processes.

## Connectionism and language (history)

Language was one of the cognitive domains covered in the most important volumes dealing with parallel processing in the eighties – the Parallel Distributed Processing by Rumelhart, McClelland and the PDP group (1986). The position of linguistic phenomena in these volumes indicates the importance of language for modelers of cognitive processes. Today, however, the focus has somewhat shifted more towards visual processing and memory. Models of linguistic processing are still present, but their proportion seems to have decreased.

The models presented here are by far not all that deserve to be presented for their importance in development of connectionist networks. They were selected due to the fact that the themes they cover – language morphology, represented by models of English past tense acquisition, and language syntax, represented by models of thematic role assignment – are the themes that appear repeatedly in connectionist models. Nevertheless, all those models bring something new in terms of architecture or algorithms, and therefore they are different in respect to their biological plausibility. Because of their importance for the architecture of linguistic models, descriptions of Elmans models (Elman 1990) are also added.

## Models

### First models

One of the most debated models in the history of connectionism is the model of English past tense acquisition by Rumelhart and McClelland (1986). With the perceptron learning algorithm (Rosenblatt 1962) they trained the network to connect 430 root verbs with their past tense forms. For the process of learning itself, it was important to imitate the U-shaped curve observed by researchers of child language acquisition by that time (e.g. Brown 1973). With no means to represent the temporal sequence of phonemes, they used a system of so called wickelfeatures, where three subsequent phonemes were encoded as one input block, followed by another block of three phonemes where the phonemes are moved by one and so on. They started with training the network on 10 verbs, and after that they proceeded with all remaining 420 verbs. The procedure yielded the desired learning outcome, with the network being able to connect most of the verbs with their correct past tense form and exhibiting an U-shaped

form of the learning curve (showing the process of overregularization), but the procedure itself (training the network first on a very small number of verbs, than suddenly increasing the training set) as well as the system of wickelfeatures were often pointed out as being problematic (Pinker, Prince 1988).
McClelland and Kawamoto (1986) model of assigning thematic roles to words within sentences, i.e. understanding their meaning along with their syntactic structure shows that networks are capable of learning concepts greater than words - sentences. Sentences are presented to the network as strings of words, which are in turn represented as subsets of microfeatures. The architecture is similar to those from Rumelhart and McClelland (1986) – simple two-layered, feed-forward perceptron.
The first models showed that there is a good reason to believe that connectionist models are indeed capable of approximating at least some aspects of human cognitive processes.
From the point of view of biological plausibility, few questions were posed at the time. The first goal – cognitively plausible models – seemed possible to achieve, and the second – biological plausibility – was yet to come.

**Some important models of natural language in the past**
Elman (1990) proposes a method for encoding temporal sequences, called simple recurrent network (SRN). Instead of encoding sequences in a spatial manner (as seen e.g. in wickelfeatures introduced by Rumelhart and McClelland (1986)), he introduces an additional layer to a feedforward network. The new layer "memorizes" the current step of the system and feeds its contents back into the hidden layer along with the contents of the next step. In this way, the network takes into account what it had learnt in the past - a temporal sequence of elements. Elman validates the method by testing it on four tasks, three of them of linguistic nature - learning a letter sequence, learning to recognize word boundaries, learning to categorize words depending on syntactic features of simple sentences. The network performs successfully on all three tasks. In the first one, it has to learn three short pseudowords - *ba*, *dii* and *guu*. It successfully learns to predict vowels, because they always appear after same consonants. It also learns the length of sequences, since they are fixed for every word. In the second task, the network learns 14 pseudowords of different length. The more elements the network obtains for recognizing the word, the better its prediction for the next element (letter). When the word ends, the networks prediction for the next element is again inaccurate, because it is never sure what the next (randomly picked) word will be. One can say that the network has successfully learned to recognize word boundaries. In the third task, short sentences (two or three words, represented as sparse localist 31-bit vectors) were presented to the network. This time, the network learns to predict possible word(s) to follow and the likelihood of their occurrence. In addition, by analyzing the internal representations for each word in the hidden layer, Elman shows that the

network has correctly organized the words into semantic categories, based only on statistical data - their co-occurrence, sequential order and context. However, Elman finds that the categories are not always distinct and clear - some category boundaries seem to be "soft" and implicit.

Trying to improve results obtained earlier by Rumelhart and McClelland (1986) in modeling English past tense, Plunkett and Marchman (1993; 1991) adjust the training procedure in order to make it more similar to the actual input received by children acquiring language. They also start with a smaller set of verbs (initially 20, eventually 500), but the increase in number of verbs is gradual. New verbs are added only after all previous have been acquired. They also try to find out the critical extent of irregularities in the training set that might have an impact on learning and explore the possible differences in sizes of the hidden layer. Their network is a feed-forward perceptron with a hidden layer, using the backpropagation algorithm. Using the new training regime, that was more plausible from the cognitive point of view (more realistic in terms of language acquisition) they obtained better results than Rumelhart and McClelland initially – the U-shaped learning occurred without manipulations of the training set. Many more models on same topic with similar outcomes were made at that time or somewhat later (Daugherty, Seidenberg 1992; Hare, Elman 1992; Hoeffner 1997).

In a recurrent network for sentence processing and decoding, St. John and McClelland (1991) model a system that makes internal representations of entire sentences with their syntactic and semantic properties, similar to cognitive sentence frames made by speakers of natural languages. The network (called the Sentence Gestalt network) could recognize thematic roles for words within sentences, even when they were ambiguous – the network could distinguish thematic roles for words, depending on their variable semantic role within sentences (for possible critique on Sentence Gestalt cf. Plaut, Kello 1999). With four hidden and context layers, the architecture of their model is somewhat more complex than usual for simple recurrent networks. The difference in word representation/encoding is also important, because due to the criticism of Kawamoto and McClelland (1986) model, they avoid to encode words as subsets of microfeatures and use localist representations instead (Waskan 2001).

As much as all the described models represent further improvements for models' similarity with cognitive processes they aim to describe, they all basically rely on the same principle of backpropagation, which was heavily debated for its incompatibility with real biological mechanisms. Furthermore, the question of representations (localist or distributed) was raised – distributed representations, such as microfeatures used by McClelland and Kawamoto (1986), were said to disclose to the network too much information that it should extract (learn) from the data on its own. On the other hand side, the localist representations do not correspond with the biological reality of data processing.

As one can see, in the nineties the models were gradually improving when it comes to their performance in representing cognitive processes. However, their closeness to their biological ideal remained under question mark, with the widely used backpropagation algorithm and localist representations (and some other features).

**Today**

In the last decade, there were several proposals to improve the problem of biological plausibility. One of them was postulated by O'Reilly (1998). He described six principles that should be followed in order to achieve greater biological plausibility: The first principle, the biological realism, should be the central theme of the cognitive modeling in general. Thereafter, all models should be "constrained and informed by the biological properties" of the brain (O'Reilly 1998:456). It is not enough that models imitate cognitive processes, but they should do so by respecting the biological properties of the brain.

The first principle is followed by three more that describe the ideal network architecture: distributed representations, inhibitory competition and bidirectional activation propagation. It is believed that the cortex uses distributed (more neurons are activated at the same time in order to represent a concept), rather than localist representations (one neuron-one concept), so the networks should do the same. Inhibitory competition between neurons assures that in the process of learning only the most strongly excited activations remain active (and the less excited are inhibited), and therefore enables the network to make fine differentiations between concepts. In other words, it enables the network to successfully distinguish between (even similar) concepts. The bidirectional activation propagation makes it possible to the network to function both bottom-up and top-down, as it is case in human cognitive processes. In the example of reading, we recognize words by recognizing letters (bottom-up), but we can also read a word even if we fail to positively identify one of the letters – if we manage to recognize the entire word, it will help us fill in the gap (top-down). The two remaining principles refer to the learning process – the error-driven task learning and the Hebbian model learning. In the past, the error-driven task learning (supervised learning) was most often implemented as the error backpropagation (Rumelhart, Hinton, Williams 1986). This procedure was criticized, because similar process does not exist in neurobiology. An alternative to it has been proposed as early as in 1987 (GeneRec by Hinton, McClelland 1987), but it was not widely used until 1996 (Leabra, as an improvement of GeneRec, by O'Reilly 1996). It is about settling the network weights in two phases – in the first phase, the network's guess is propagated up to the output layer, and in the second phase the expected outcome (teacher signal) is propagated. Without any backpropagation the difference between the two signals is computed, which represents the network's error, and the weights are adjusted according to this error. The more the network is informed about its errors, the better its guesses be-

come – it learns. Further constraining an error-driven network by Hebbian learning can facilitate the learning and improve the network's results.

In his paper, O'Reilly (1999) discusses all of these principles and the fact, that they are not often combined in models, which makes models less biologically plausible; he acknowledges the fact that some of the principles seem to be in conflict, but also offers ways to overcome the difficulties – not by simplifying, but rather by combining all (or most of) the principles described above.

In addition, he proposes a new algorithm, called Leabra, in which all of his principles were implemented, the biological realism above all:

> "[…] the algorithms they [O'Reilly, Munakata 2000] introduce are constrained by biologically realistic principles, and the resulting models thus incorporate detailed assumptions about such things as membrane potentials, leak currents, or spiking rates." (Cleeremans, Destrebecqz 2003)

O'Reilly and Munakata (2000) describe their models of English past tense and Sentence gestalt, using the Leabra algorithm (as replications of Hoeffner 1997 and St. John and McClelland 1990, respectively). For the English past tense model, they criticize older models for the fact, that it in all of them there is no clear distinction between two levels of analysis that both influence the U-shaped learning – the mechanistic level (mechanic properties of the model itself) and the environmental level (structure of input data). Relying only on the Leabra algorithm, no context layers in the network, their model is mapping from semantics of the words to their phonological shape. Their results are even more consistent with data for human speakers, compared to usual backpropagation models.

The success of replicating the Sentence Gestalt model was comparable to the original, managing to satisfy multiple constraints set by syntax and semantic simultaneously. Some advantages, however, were observed: learning was much faster; Hebbian learning and inhibitory competition, added by the Leabra algorithm made the model more real neuron-like.

Rosa (2004) compared the performance of two connectionist networks trying to solve the same task on the same set of sentences. The task consisted in assigning the thematic roles to words within sentences, fed to the network one at a time. The words were encoded as subsets of distributed microfeatures. The only difference between the networks were their learning algorithms - the first one used the backpropagation algorithm on a simple recurrent network of the Elman type (recurrent connections to the context layer from the hidden layer), whereas the second used the Leabra algorithm, with no need for additional context layers. The training corpus consisted of 364 different sentences as combinations of 30 nouns and 13 verbs, with some verbs allowing for more than one semantic interpretation:

*The man hit something.*
*The stone hit something.*

Thematic roles to formal subjects of these two sentences are different – *man* is the agent of the action, and *stone* is its cause. The ability of the network to distinguish between the two shows the network's deeper understanding of syntactic and semantic relations within the sentence.
Comparing the overall performance of the two networks, Rosa concludes that the network using the Leabra algorithm is not only more acceptable from the biological point of view, but also more computationally efficient.

## Conclusion

Although scientists are very much intrigued by psycholinguistic and neurolinguistic phenomena, connectionist models, or computational models of cognitive (linguistic) processing are not very numerous. There are many more models trying to explain our processing of visual stimuli or memory, but the language models are not very common, despite their position in the early modeling literature. Most of the existing models deal with phonetic and/or phonological phenomena, widely used for robotic purposes (voice recognition, simple text/commands recognition). Only few tackle the language in its complexity (such as in Rohde 2002)
Today, the field of the computational neuroscience offers acceptable ways of modeling and exploring higher-level cognitive processes, and can therefore give us valuable insights in the core of many (psycho)linguistic processes.
The art of connectionist networks, their architecture and algorithms have evolved since 1986, so that many of their shortcomings pointed out by watchful observers are no longer valid (such as in Fodor, Pylyshin 1988; Pinker, Prince 1988). All models described in the previous chapter make use of a new algorithm that is more consistent with biological properties of neurons and populations of neurons. Given the fact that Leabra is not the only algorithm of this kind (see O'Reilly 1998), but was only chosen to demonstrate the facts, one can say that connectionist models today are indeed much closer to the biological systems that they were inspired by, than they were only two decades ago.

## References

Arbib, Michael A.. The handbook of brain theory and neural networks. Cambridge, MA: MIT Press, 2002

Brown, Roger. A first language. Cambridge, MA: Harvard University Press, 1973

Christiansen, Morten; Chater, Nick. Connectionist natural language processing. The state of the art. *// Cognitive Science.* 23 (1999), 4; 417-437

Cleeremans, Axel; Destrebecqz, Arnaud. Harder, Better, Stronger, Faster: A review of Computational Explorations in Cognitive Neuroscience. // *European Journal of Cognitive Psychology*. 15 (2003), 3; 474-477

Daugherty, Kim; Seidenberg, Mark S.. Rules or connections? The past tense revisited. // *The Proceedings of the 14th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992, 259-264

Elman, Jeffrey L. Finding structure in time. // *Cognitive Science.* 14 (1990); 179-211

Feldman, Jerome A.; Ballard, Dana H.. Connectionist models and their properties. // *Cognitive Science.* 6 (1982), 3; 205-254

Fodor, Jerry A.; Pylyshin, Zenon W.. Connectionism and Cognitive Architecture: A Critical Analysis. // *Cognition.* 28 (1988); 3–71

Hare, Mary; Elman, Jeffrey L.. A connectionist account of English inflectional morphology: Evidence from language change. // *The Proceedings of the 14th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992, 265-270

Hinton, Geoffrey E.; McClelland, James L.. Learning Representations by Recirculation. // *Neural Information Processing Systems /* Anderson, D. Z. (ed.). New York: American Institute of Physics, 1987, 358-366

Hoeffner, James H.. Are rules a thing of the past? A single mechanism account of English past tense acquisition and processing. Unpublished doctoral dissertation, 1997

Marcus, Gary F.. The Algebraic Mind: Integrating Connectionism and Cognitive Science. Cambridge, MA: MIT Press, 2003

McClelland, James. L.; Kawamoto, Alan H.. Mechanisms of sentence processing: Assigning roles to constituents of sentences. Chapter 19. // *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* / McClelland, J. L.; Rumelhart, D. E.; the PDP research group (ed.). Cambridge, MA: MIT Press, 1986, 272-325

Minski, Marvin L.; Papert, Seymour A.. Perceptrons: An Introduction to Computational Geometry. Cambridge, MA.: MIT Press, 1969

O'Reilly, Randall C.. Biologically Plausible Error-driven Learning using Local Activation Differences: The Generalized Recirculation Algorithm. // *Neural Computation.* 8 (1996), 5; 895-938

O'Reilly, Randall C.. Six principles for biologically based computational models of cortical cognition. // *Trends in Cognitive Sciences*. 2 (1998), 11; 455-462

O'Reilly, Randall C.; Munakata, Yuko. Computational explorations in cognitive science. Cambridge, MA: MIT Press, 2000

Pinker, Steven; Prince, Alan. On Language and Connectionism: Analysis of a parallel distributed processing model of language acquisition. // *Cognition.* 28 (1988); 73-193

Plaut, David C.; Kello, Christopher T.. The Emergence of Phonology From the Interplay of Speech Comprehension and Production: A Distributed Connectionist Approach. // *Emergence of Language* / MacWhinney, B. (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, 1999, 381-415

Plunkett, Kim; Marchman, Virginia. U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. // *Cognition.* 38 (1991); 43–102

Plunkett, Kim; Marchman, Virginia. From rote learning to system building. // *Cognition.*, 48 (1993); 21–69

Rohde, Douglas L. T.. A Connectionist Model of Sentence Comprehension and Production. PhD thesis, School of Computer Science. Pittsburgh, PA: Carnegie Mellon University, 2002

Rosa, Jose L. G.. A Biologically Motivated and Computationally Efficient Natural Language Processor. // *Lecture Notes in Computer Science 2972: Advances in Artificial Intelligence: Proceedings of the Third Mexican International Conference on Artificial Intelligence.* / Monroy, R. et al. (eds.). Heidelberg: Springer Verlag, 2004, 390-399

Rosenblatt, Frank. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. New York: Spartan, 1962

Rumelhart, David E.; Hinton, Geoffrey L.; Williams, R.. Learning internal representations by error propagation. // *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I.* / McClelland, J. L.; Rumelhart, D. E.; the PDP research group (ed.). Cambridge, MA: MIT Press, 1986, 318-368

Rumelhart, David E.; McClelland, James L.; PDP research group. Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press, 1986

Rumelhart, David E.; McClelland, James L.. On learning the past tenses of English verbs. Chapter 18. // *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* / McClelland, J. L.; Rumelhart, D. E.; the PDP research group (ed.). Cambridge, MA: MIT Press, 1986, 216-271

Sejnowski, Terrence J.; Koch, Christoph; Churchland, Patricia S.. Computational Neuroscience. // *Science, New Series*. 241 (1988). 4871; 1299-1306

St. John, Mark F.; McClelland, James L.. Learning and applying contextual constraints in sentence comprehension. // *Artificial intelligence*. 46 (1990); 217-257

Waskan, Jonathan A.. A critique of connectionist semantics. // *Connection Science*. 13 (2001). 3; 277-292

# Vocabulary Entry of Neologism
## A Lexicographical Project aided with NLP Application

Catherine Dahlberg
School of Foreign Studies
Nanjing University, Nanjing, China
22 Hankou Road, Nanjing, China 210093
nanwai@hotmail.com

Tracy Qian Liu
School of Foreign Studies
Nanjing University, Nanjing, China
22 Hankou Road, Nanjing, China 210093

Carolyn Fangya Chen
School of Foreign Studies
Nanjing University, Nanjing, China
22 Hankou Road, Nanjing, China 210093

**Summary**

*This paper describes a lexicographical project utilizing natural language processing tools, where the entries of a glossary are identified by Wordsmith and HC2009YLCL v. 3.0 from a corpus segmented by ICTCLAS 3.0.*
*The researchers of this project identified two important issues: 1) the entry test as canonized by Chinese lexicographers and 2) the position of neologism in language development. The researchers made a detailed analysis of the nature of neologism and offered a new taxonomy. In addition, a brief evaluation of technology application is included.*

**Key words:** creative language use, neologism, NLP processing, novelty detection, lexicography

## Introduction

Internet has greatly promoted language change. A neologism used on the Internet is spread almost instantly to readers who are miles away from the physical location of the creation of this word. In today's wired world, neologism appearing on the web also enters verbal communications in people's real life, a phenomenon itself suggesting the driving urge of man to copy creative expres-

sions. In China1, Internet, since its introduction in 1995, has provided rapid sharing of information to 40 million registered users. It has greatly amplified the influence of neologism, by giving web browsers a chance to quote a neologist unlimited times to an infinite number of audience who might adapt intuitively this very expression at various degrees to suit certain contexts. In this process, a neologism gradually transforms into a real-life linguistic being.

## Literature review

Lexicographers have come to understand the difference between evidential and general dictionary. But in China, where the concept of evidential dictionaries is under appreciated, linguists perceive neologism as an uncomfortable variant threatening the stability and comfort of *their* language system. Yang blames the calque expressions in Chinese as having caused confusion in use and deformation of Chinese language (2007: 121–122). He criticizes phrases that frequent TV and cover stories for their violating the structural order of Chinese and challenging established norms (ibid.).

To minimize the impact of variants to their linguistic comfort, scholars established canons in lexicography, which will be consistently referred to in this paper as the entry test. The entry test examines three things: 1) the frequency of usage, 2) the scope of users, and 3) the resistance to change of a phrase. Canonism, *prima facie*, offers a nice operational guidance on lexicography … until one asks specific questions. For example, how do people know if a word, say W, is frequently used? Su and Huang explain that the words are frequent because a number of dictionaries have listed them (2006). The researchers feel sorry for such "frequency" view and wish to provide some background information about Chinese lexicography before moving on to topics on corpus linguistics. The pre-corpora lexicographers had no access to statistical evidence of a particular word except by relying on their own assumption. Now statistical evidence is available, but lexicographers still cling to the old practice. The dictionary data from Su and Huang offer no creditability at all, for they could have been either a perception or coincidence. Worse is the case if some lexicographers simply trusted other scholars' judgement by copying them.

Corpus linguistics offers a better understanding of usage frequency. Meyer explains that computational linguists base their analysis of language on "real data", that is, on evidence provided by corpus and retrieved records of language usage (2002: xiii). The question for Chinese linguists is what kind of corpus to use? Meyer explains why balanced large corpora are the solution (2002: 15). His idea is followed by the Centre for Chinese Linguistics of Peking University

[1] People's Daily (story 2008/11/3) claims 1990 as the year of China's initial Internet connection, downloaded 2009/7/2 from http://english.people.com.cn/90002/95607/6526583.html. Note that the 1990-er Internet was in fact privileged technology designated for academic and research purposes. Access for general China was made available in 1995 by China Telecom.

in its 477-million-character CCL corpus. Unfortunately, pertinent to this project, sorting neologism records from among millions of CCL data is a problem. Other linguists have built smaller corpora for specialized purposes. Su and Huang mention annual dictionaries made available by synchronized corpora (2007). Meyer cites linguists' idea of a monitor corpus (2002: 15).

The researchers in this project also found literature discussions on the legitimacy of language change. Aitchison, who examines language from the socio-historical perspective, emphasizes the importance of careful study of language development, particularly the descriptive approach that examines the "frayed edges" of language and gives enough credit to expressions which conflict with grammar (1997: 104). She concludes that new expressions are essentially a reflection of the changing face of the world (ibid.).

This digital glossary aims to detect neologism actively used in journalism, and the first question is how does it get into media? Yuan summarizes channels of neologism in Chinese as 1) loan words, 2) borrowing from dialects, 3) new meanings injected into an existing word, and 4) coinage (2005). Dai asserts that the adoption of neologism in media is motivated by the need to communicate with the users of neologism (2004: 69). The researchers are not quite content with their explanation. It is necessary to further develop the ontology of neologism, which is a question that may be properly answered by discussions on: 1) the taxonomy of neologism and 2) how people respond to neologism.

## Project methodology

Raw data of entertainment news dated between 2005 and 2009 are collected and this corpus is then segmented by ICTCLAS 3.0, so that a Wordsmith key word list is created, lemmas of which then serve as key words for HC2009YLCL v. 3.0 concordancing to identify neologism entries for the glossary.

There are three major steps in this project: data collection, entry list building, and glossary compiling.

Step 1: Data collection

1.1. Internet portals

The researchers identified a number of portals which pool various sources of news service as the data source. They are http://www.yahoo.com.cn, http://www.longhoo.net, http://www.sina.com, http://www.sohu.com, and http://www.tom.com.

1.2. Data extraction

Each researcher is assigned with the collection of 200 sentences, and a total of 573 sentences from 60 articles are actually extracted. A sentence is defined as a body of text within two consecutive periods. The concern about copy right issues is addressed by ratio control and random sequence. Here is the operational routine established in this project: first, the researcher selects a news story

which contains at least 30 sentences; second, by clicking the “page down” button, the researcher positions the cursor to the last period in the article; third, going backward, the researcher extracts a running sentence into the corpus; fourth, the researcher places the cursor at where she had stopped in step three and randomly selects a period that is above it. By repeating step three, the researcher collects 10 sentences from an article without actually copying the story, thus gaining some legitimacy of corpus for research purposes.

Using a period as the sentence boundary is thought to simplify the process. However, the 32,000 characters in the raw data still came as a surprise, which prompted the researchers to examine the style of entertainment news. In the section titled “Findings”, the researchers will report related details.

data treatment

For easy analysis, the researchers regard each character as a lemma in this project. Theoretically, a neologism is a (multi-syllable) word, but the researchers encountered the problem of physical spacing of words in Chinese text, whose iconic layout lacks a visible boundary. To make things worse, Chinese is full of parsing ambiguity. Yu and Zhu provided many examples of ambiguous sentences which are pre-NLP formatted (2006).

Previous research by Chinese computer scientists focused on word segmentation. To illustrate the features of Chinese text, a pair of raw and treated sentences is concordanced in Wordsmith for a lemma. Table 1 is the result.

Table 1. Concord results of “chao”

| sentence | treat | Concord |
|---|---|---|
| 超女张靓颖已在歌坛摸爬滚打一年有余，但这次她与周笔畅、刘亦菲、弦子、薛之谦入围的 却是最佳新人奖。 | no | 0 |
| 超/v 女/nz 张靓颖/nr 已/d 在/p 歌坛/n 摸爬滚打/vl 一/m | seg | 超/v 女/nz 张靓颖/nr |

ICTCLAS is selected as the segmenter in this project. Before turning to it, the researchers had tried out a number of tools including the MS Word and HC2009YLCL, but dismissed their application for various reasons.

ICTCLAS provides a combined segmenting and tagging treatment, where the bonus POS tagging does not interfere with the project. Proofreading for any segmenting errors is supplemented, and in the case of ambiguity, manual reseg was done according to linguistic sense. In this way, the researchers collected a processed corpus of journalism.

Step 2: Entry list building

2.1. Wordsmith Wordlist

Wordsmith is used to generate a list of words from the segmented corpus. In pilot studies, the researchers identified that the ideal parameters for neologism are length between 2 and 4, and frequency under 3. A two-syllable list was

made by following the low frequency parameter (between 2 and 3), and a few other lists were made at varying parameter values.

2.2. Wordsmith Key Word

The two-syllable list was used as reference and 0.5% setting was used to make a Key Word list which contains a modest amount of neologism.

2.3. Supplement list

Software novelty detection in the previous two steps did not reach the full potential of the corpus. Fortunately the research team consists of experts in language, who possess excellent memory of and impressive vocabulary in neologism. So human knowledge is tapped into to make a quality entry list.

Earlier in step 1.3, ICTCLAS made a noticeable amount of mistakes. The researchers collected information on a discrepancy between the claimed 98.45% accuracy rate and the actual 95% accuracy in lemma processing. This discrepancy suggests an incompatibility between ICTCLAS and neologism. The researchers initially tried to use NLP errors as a sign of neologism. But further research suggested that this thought was over simplistic. Details on this part will be provided in the section titled "NLP evaluation".

HC2009YLCL concordancing replaced corpus proofreading. The short list made available in step 2.2 provided an initial amount of seed lemmas. The flexibility of Chinese is taken advantage of, so that a seed is concordanced to identify additional neologism. Before going too far, the researchers wish to offer an example to explain the flexibility of the collocation power in Chinese. In a two-syllable word, *daxue*, *-xue-* is a free lemma that can be combined with another lemma to form a new word *xuexiao*. The researchers thought that a list of seeds which possess good neologism productivity can help identify enough glossary entries if queried in HC2009YLCL.

HC2009YLCL proves an efficient method of novelty detection as it returns more and better results than Wordsmith. It concordances sentences[2], and a maximum of eleven concords can be recalled in a search.

Step 3: Glossary compiling

The digital glossary is saved as an Excel file. There are a few sections in this glossary, and the major sections are the index page and the entry section.

3.1. Index page

The index page is designed to provide glossary users with access to quick lookup. Lemmas, as listed on the Key Word list and the supplement list, are called the seeds on the index page, since they link a group of new words. A few

[2] Two periods are set as the boundary of a sentence.

miscellaneous records of neologism where no lemma can be found as their common index is assigned to a number.

3.2. Entry section

There are over 220 entries. For easier use, the entry section is broken down into three alphabetically listed entry sheets. Next to each entry is a horizontal array of its definition, grammar explanation and example phrases. The researchers gave each entry a definition based on their linguistic knowledge and the understanding gained through corpus lookup.

## Findings

### Low frequency of neologism

Wordsmith Key Word indicates a tendency of low frequency of neologism distribution in the corpus. The researchers followed a frequency curve from 7 to 3 before capturing enough records from the corpus.

Interestingly, HC2009YLCL also shows a tendency of low frequency. For example, among a total of 43 concordances of men, only 11 are neologism. In most cases, there was only a single record of a neologism.

Relating findings in NLP application, the researchers conclude that low frequency words should be used in neologism search in a corpus.

### Concentration in seeds

Despite the above finding of low frequency, a small group of lemmas show a tendency of high productivity of neologism, as is noticed by the researchers. Table 2 lists a few productive phrases in the corpus.

Table 2. Lemmas with high concentration of neologism

| | *fensi* | *pinpai* | *qijian* | *quan* | *shanzhai* | *shijian* | *yiren* | *zu* |
|---|---|---|---|---|---|---|---|---|
| $C_{all}$ | 12 | 9 | 9 | 17 | 23 | 15 | 19 | 13 |
| $C_n$ | 12 | 9 | 7 | 17 | 17 | 15 | 19 | 11 |

Note: $C_{all}$=all results, $C_n$=result of neologism concordancing

### NLP evaluation

Despite its consistency, ICTCLAS is not suitable for neologism detection. The researchers suspect that the conflict between neologism and grammar (rules on which the ICTCLAS is built) is probably caused by dated linguistic information on the developers' part. In this project, POS and segmenting of inicinym[3] is a big challenge to technology. Below are a few examples of mistakes.

[3] This is a term given to a specific type of neologism. Discussion on inicinym is provided in the section titled "On neologism".

| | | | |
|---|---|---|---|
| NLP tagging error: | yule/v | quan/n | 1 (12) |
| | yule/n | quan/v | 2 (12) |
| NLP segmenting error: | fei/b | lin/ng | 2 (2) |

As mentioned in the methodology, the researchers had hoped that all NLP mistakes can be contributed to neologism. But the truth is, NLP developed with a dated grammar system simply fails to recognize unfamiliar data, whether they are neologism or not. Among the NLP mistakes, a majority is not caused by neologism. In general, ICTCLAS had a difficult time processing person names (PN), film titles, and some proper names. The following example shows a mistake in segmenting a PN and a POS error of a conventional word.

Fengxiao/nr gang/d niandu/n da/d zuo/v 《/wkz ye/tg yan/vg 》/wky
Tr. Feng Xiaogang's masterpiece of the year "*Yeyan*"

In this phrase, Fengxiaogang (a PN) is segmented into two parts, where Fengxiao is considered as a PN, and gang is tagged as an adverb. Two words after this, dazuo (segmented into two parts) is tagged as adverb+verb, but the correct tagging is adj.+noun. To give some credit to technology, the confusion is caused by the ambiguity of lemmas of gang, da, and zuo. But mistakes like this show that NLP technology still needs improvement.

The film title in this example (yeyan) shows another issue of ambiguity. The machine's tagging as [temporal]+verb is correct, but this word has a dual tagging as [temporal]+noun, in which case human intelligence has a better application than NLP technology to offer a sensible solution.

NLP processing is one thing, and the novelty detection needed in this project is another. Inicinyms obviously challenged ICTCLAS. For example, xiuchang is correctly segmented, but NLP is not able to distinguish a neologism of an isonym from a conventional use. Another type of neologism, the structural frame, can be a challenge, too. In the section titled "On neologism", the researchers will explain why.

When comparing ICTCLAS with HC2009YLCL, the researchers are impressed by the latter for its smart processing of PN and neologism. A possible explanation for that is the correlation between the make and year of a technology and the language development in vocabulary and grammar.

## Related discussions

### On entry test

The entry test is meant to reject neologism, but it no longer fits lexicography. Here is an example to illustrate why the frequency canon fails to address the reality. *Xiu* has over 46,000 records in the CCL corpus, but only 61 of them are in the neologism sense. If the canon of frequency is followed, obviously *xiu* as a neologism does not deserve academic attention. Or does it?

Generally speaking, there is a correlation between time and frequency. Old words have high frequency as time rewards them with an accumulated occurrence. Neologism, on the contrary, naturally lacks enough time to get established. As in the case of xiu, the time factor guarantees a much higher frequency of the conventional isonym than neologism. The researchers object using frequency as an entry test. Instead, they propose giving neologism a separate status and making entry test to focus on the features of neologism.

As for the canons of general audience and stability, the researchers think they are impractical. A Neologism, by definition, is a new word. Internet may have facilitated the encountering of neologism, but the scope of its usage is probably still small. Again, neologism should be studies as a separate subject from conventional words. The scope of a conventional word belongs to the past, but that of a neologism to the future. It will be too early to assess the scope of usage in the case of a neologism if only the current (or to be exact, the initial) popularity is examined. Besides, a careful lexicographer should conduct a full range check of the scope of all the words for his or her dictionary, instead of applying the entry test only to neologism as currently practised.

Then what is the perspective that may suit the reality of a neologism? Suppose A is a new word and B old. Stability of B can be easily measured by looking into its past. But no one can predict the future of A, so there is no way to compare the stability of A with that of B. There is, instead, a possibility to monitor the user profile of A. If A initially appeared as a blog expression and soon gained momentum through repeated online quoting, and finally if it successfully incarnates in some kind of traditional fields, for example, in prints, on the radio, and in speeches, it is safe to assume that the scope of usage of this word has undertaken an expansion. The researchers of this project are interested in the process in which a neologism duplicates from its initial virtual identity into actual usages in business, in news reports, and in daily expressions.

Unfortunately, some people are simply determined to neglect neologism, especially in the case of an established one. Xiu, if looked from the temporal perspective, should no longer be considered as new. Many years ago, Lou described expressions of xiu which had been actively used in Taiwan (1993). In the next decade, established knowledge of xiu never led to dictionary entry. In this glossary, the researchers have identified a number of xiu-group entries, such as gerenxiu, xiuziji, and xiuchang. Since 1993, this word has expanded its scope in meaning. The researchers can identify at least two types of neologism: a) a transliteration of “show” (as explained by Lou (1993)), and b) a fixed expression meaning new performers in sports or entertainment industry.

Is it possible to acknowledge the historical position of neologism? First the researchers would like to establish this argument: newspaper is a public institution and it does not favour the usage of neologism. The media presence of neologism simply reflects the updated face of language at the time of being. If news reporters choose to use a neologism, it means there is a perception that this

word is understood by a large audience. Thus being used in journalism reflects an established status of a neologism in the society. The researchers argue that access to journalism should be considered as some kind of entry test, since media is not the source, but the channel of the circulation of a neologism. The researchers further argue that stability should be replaced by usability. Many words in a dictionary are simply "dead". Why not clean up some dead entries to make room for neologism?

**On neologism**

There are two types of neologism, one is a coinage, the other is a new use of an existing phrase when the speaker has created a new meaning for it. The researchers would like to describe the latter as a meaning injection, since it seems that the A-sense meaning is injected from outside. For easy reference, the researchers name a neologism created this way an inicinym.

What exactly is a new word? There is a paradox of neologism. One has to possess a repertoire of B to ground the comparison where A can be recognized as new. Now this is exactly how the paradox exists: a new word is no longer new once it is known. In the case of neologism, while bringing the perception of the creativity in a word, the encounter of A takes away its newness.

The complexity of this paradox can be further examined from various angles. The researchers will take you to a speculative discussion on the route of the circulation of an inicinym First, the researchers put the factor of individuality aside and only examine the time factor. Suppose a homogeneous group p encounters A. At a given time t, its subgroup, p1 completes the encountering ahead of the rest, p2. As a result, p1 recognizes A, but p2 does not.

Second, the time factor is replaced by the growth factor. A reaches p which is a mixture of grown-ups g and children c. The repertoire of B is available to g but not to c. As a result, despite simultaneous encounter, g sees A as a neologism, but c acquires this word as a conventional expression.

And finally, personal preference is introduced. Still in a t-instant model among every member of a group e. Subgroup e1 welcomes A, subgroup e2 lacks sensitivity to vocabulary, and subgroup e3 rejects A. When the newspaper starts to be filled with A phrases, e2 may be gradually assimilated into e1, but e3 will not. Then how does A propagate among its users? This is a complicated issue so difficult yet so interesting that we have to research into.

Here are additional thoughts on neologism. Besides asserting that low frequency is the nature of new expressions, the researchers argue that neologism is designed to enter the body of language. Neologism suits the need of people in a way that no other words can. The researchers think that the use of neologism is caused by the urge for creativity which is a unique function of intelligence. A neologist enjoys making new words, and such a manipulation of language is an integral aspect of intelligence. The researchers think the fact that a neologism

gains popularity suggests that people in general appreciate creativity and new perspectives. It is part of human nature to adopt a new word once it is created. After all, language is man's tool to index and explore the world. A neologism means a new outlook of the world.

Neologism calls for better solutions in novelty detection. Ambiguity in the case of isonyms makes it very hard for machine to recognize neologism. Shanzhai is either 1) a conventional word (whereas it is a compound) or 2) an inicinym (a single word). NLP correctly segments it in a running text, but it makes no attempt to POS tag its structure accordingly. Inicinyms in this project were hand picked to make up for the missing NLP available for neologism detection.

Another challenge to NLP is pattern extraction. The researchers call neologism identified in this way structural frame. Though not mentioned in the literature, structural frame exists as a third form of neologism, which is observed in the media. What is it? Simply put, it is a pattern that has been extracted by human intelligence to be used for word formation. For example, hen+[adj.]+hen+[adj.2]. A hen+[adj.] is a conventional phrase, and no grammar ever prohibits putting two of them together, but doubling a hen- structure is a neologism. In this word, the second adjective is always in a two-syllable format. Users of neologism believe that it was created by an actress whose name is Zhong Xintong (Ajiao)4. The latest use of it is the commercial of Xtep "The athlete is henkuhenqiangda, his rival is henruohenkelian".

## Conclusion

Lexicographers should give credit to web language and new words used in the media. The entry test contradicts with the nature of neologism. A more reasonable approach to a specialized dictionary should be a corpus-based methodology that is developed with evidential references. NLP application can be utilized to improve the efficiency in a project.

Neologism is a complicated phenomenon of intelligent creativity including coinage, inicinym, and structural frame. The flexibility in neologism is a challenge to NLP technology. Current programs developed for Chinese NLP are adequate in conventional tasks, but they lack competence in novelty detection. Human detection of neologism relies on encyclopaedic knowledge of lexemes, sensitivity to creative collocation, abstraction of "rules" of expression construal, and calibration of semantic deviant. NLP application should take human-specific factors into consideration to enhance the utility of neologism detection.

Academic attention should be paid to the spontaneous linguistic changes caused by free online publishing and easy content sharing. Though it conflicts with language norms, neologism deserves detailed study by linguists.

[4] Her speech of "*henshahentianzhen*" (很傻很天真 tr.: I made a fool of myself. I was too naïve.) soon gained popularity in 2008.

This glossary is compiled with such philosophy in mind, that novelty in language usage claims its position in the vast spectrum of human history. In this sense, this glossary is designed not to serve the general public but to describe current language development. This project hopes to capture the nature of neologism. The researchers call for serious academic considerations of neologism and the changing face of language.

**Additional information**

1. ICTCLAS is developed by the Institute of Computing Technology of the Chinese Academy of Sciences. The researchers used this technology to segment raw data in a corpus. ICTCLAS is based on HHMM (Hierarchical Hidden Markov Model). It is able to identify the boundary of a chunkable item in a near-human-perception fashion. It can process 1 million characters at the same time.
2. HC2009YLCL is developed by Nan Yanfei (Cheng Nanchang), of the College of Literature of Guangxi University for Nationalities. The researchers used this technology to concordance a corpus. Frequency count is a useful utility of this program, where the boundary of a phrase is automatically identified. It can process 60,000 characters at the same time.

## References

Aitchison, Jean. Language Change. Tr. by Xu Jiazhen. Beijing: Yuwen Press. 1997.

Catoni, Bruno. Lexical Resources for Automatic Translation of Constructed Neologisms. //LREC, Marrakech, Morocco. 2008. Retrieved June 12, 2009 from http://www.lrec-conf.org/proceedings/lrec2008/pdf/247_paper.pdf

Cui, Shiqi; Liu, Qun; Meng, Yao; Yu, Hao; Fumihito, Nishino. New Word Detection based on Large Scale Corpus. // Journal of Computer Research and Development. 43 (5), (2006). p. 927–32.

Dai, Qingxia (Ed.). Introduction to Sociolinguistics. Beijing: Commerce Press. 2004.

Janssen, Maarten. Lexical vs. Dictionary Database. // COMPLEX, Budapest, Hungary. 2005. Retrived June 10, 2009 from http://maarten.janssenweb.net/Papers/COMPLEX2005-mjanssen.pdf

Janssen, Maarten. NeoTrack: semiautomatic neologism detection. // APL, Lisboa, Portugal. 2005. Retrieved June 14, 2009 from http://maarten.janssenweb.net/Papers.

Kilgarriff, Adam. What computers can and cannot do for lexicography. Retrieved June 12, 2009 from http://www.kilgarriff.co.uk/Publications/2003-K-AsialexKeynote.doc

Lee, Min-Jeh; Huang, Chien-Kang; Chien, Lee-Feng. Automatic Construction of a Bilingual Dictionary for Spoken Language Processing Applications. // Oriental COCOSDA99, Taipei:Academia Sinica. p. 37–40.

Lou, Chengzhao. On *Xiu* and more. // Chinese Translators Journal. 5(3), 1993. p. 45–6.

Meyer, Charles. English Corpus Linguistics. Port Chester, NY, USA: Cambridge University Press. 2002.

Su, Xinchun; Huang, Qiqing. Development of Neologism and the Canons of Prescriptive Dictionary Compiling. // Lexicographical Studies. (3), 2003. p. 106–13, 15.

Sun, Maosong; Huang, Changning; Fang, Jie. Quantified Research in the Collocation of Chinese. //Zhongguo Yuwen [Chinese]. 256 (1),1997, p. 29–38

Sun, Honglin. Features of Collocation in Discourse. //1998 Zhongwen Xinxichuli Guoji Huiyi Lunwenji [Proceedings of the International Symposium on the Digitalized Processing of Chinese 1998]. Huang, Changning (Ed.). Beijing: Tsinghua University Press. p. 230–6.

Wen, Duanzheng. On Dialect and Popular Sayings. Originally published in Linguistic Research, 30 (1), (1989). Selected Papers on Linguistics by Wen Duanzheng. Shanghai: Shanghai Lexicographical Publishing House. 2003. p. 339–50.

Yang, Xipeng. Research on Foreign-Origin Vocabularies in Chinese. Shanghai: People’s Press of Shanghai. 2007.

Yu, Shiwen; Zhu, Xuefeng; Li, Feng. Design of a Lexicon Database of Contemporary Chinese and its Application. // Chinese Teaching in the World. (2), 1999. p. 39–46.

Yu, Shiwen; Zhu, Xuefeng. Yuwen Xiandaihua yu Hanyu Xinxichulijishu [Modern Technology applied to Chinese and Chinese Digitalization]. // Yuwen Xiandaihua Luncong [Papers on the Modernization of Chinese], vol. 6. Su Peicheng (ed.). Beijing: Yuwen Press. 2006. p. 176–89.

Yuan, Jiheng. Mechanism of Chinese New Words. // Journal of Kaifeng Institute of Education. 25 (1), 2005, p. 32–3.

Professional paper

# Slovenian Biographical Lexicon – From a Digital Edition to an On-Line Application

Jan Jona Javoršek
Department of Experimental Particle Physics
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
jan.javorsek@ijs.si

Petra Vide Ogrin
Slovenian Academy of Sciences and Arts, Library
Novi trg 3, Ljubljana, Slovenia
petra.vide@zrc-sazu.si

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
tomaz.erjavec@ijs.si

**Summary**

*We presents the digitization of Slovenian Biographical Lexicon (SBL), an extensive publication that used bio-bibliographical methods to provide synthetic assessments of work and significance of historical figures on the basis of primary sources. SBL has been out of print for a long time, but the publication has been seen as an important resource for encyclopaedic and reference editions and research in the Slovenian humanities, social sciences and history of the natural sciences. Therefore, Slovenian Academy of Sciences and Arts (SASA) and the Scientific Research Centre of the SASA decided to produce a freely available on-line digital re-edition of SBL. In the process of digitalization, manually corrected OCR has been semi-automatically converted to XML-based Text Encoding Initiative format (TEI P5). Its extensive annotation vocabulary, notably from the biographical and prosopographical modules, has been used to mark-up as much data as possible. The resulting XML document has become the data resource of an online digital repository based on Fedora Commons platform, where we implemented an infrastructure of XML processing methods on top of native relationships and a Lucene/SOLR based search engine to produce a full-fledged web application and search engine with browser, metadata and web application interfaces.*

**Key words:** digital library, encyclopaedia, TEI-XML markup, document repository, Fedora Commons, XSLT workflows, search, Lucene, SOLR

## Introduction

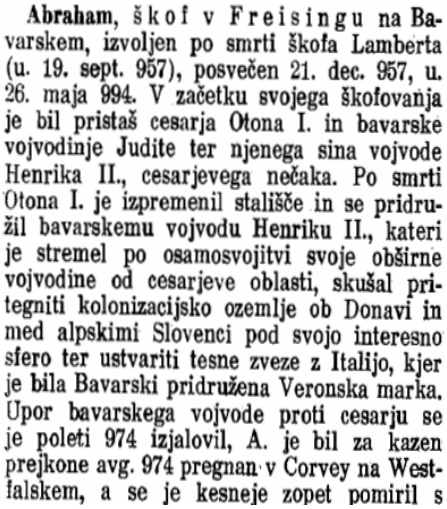
The Slovenian Bibliographical Lexicon (SBL, 1925-1991) has been conceived as a publication that was to give an accurate picture of Slovenia’s cultural life, from its beginnings up to the contemporary time by including everybody who participated in the cultural development, either of Slovenian origin, born on Slovenian soil or influencing Slovenian cultural life.

This broad aim resulted in the proposed list of 2,335 names, mostly from the fields of humanities and social sciences, proposed by the original editorial board. In its long history, this list has been changed and expanded: at first due to the need to include other disciplines and areas, such as the natural sciences, later due to changes in perspective, especially the WW2, when participants in recent had to be included and the focus has been shifted to include “increasing development of natural sciences, modern technologies and their applications, as part of the spiritual superstructure” (SBL, vol. 15, 1991). In spite of several eliminations from the original list, the final publication in 1991 comprised as many as 5,031 biographical entries, with more than 5,100 persons covered. Since the publication was published sequentially, it is important to consider that the criteria for different published volumes have therefore varied significantly.

The second aim of the publication was to be both informative and exhaustive, so much substantial information had to be included in rather short articles (with several longer exceptions) and it has been decided that the data in the articles had to be been checked against relevant historical materials and pre-existing publications., e.g. biographical and other dates are always compared to dates in registers and other primary documents, literary citations are compared with originals, sources are cited at the end of the articles and the publication includes an index of all person names that appear in the articles and a list of abbreviations.

As a result, SBL contains a surprising amount of high-quality information and references and remains to this day a precious resource for encyclopaedic and reference editions and research in the Slovenian humanities, social sciences and history of the natural sciences. It had, however, two severe drawbacks: the original edition has quickly become difficult to find and the information, once published, has never been updated, so much care has to be taken to consider the time of publication of each article. The present project of digitalization of SBL has been started by the Slovenian Academy of Sciences and Arts (SASA) and the Scientific Research Centre of the SASA to make this precious resource available again, this time in the form of a freely accessible on-line edition, and has been based on previous similar projects (cf. Erjavec, Ogrin, 2005). We aim to describe here the steps taken on the path from the original publication towards a fully interactive, searchable and cross-indexed on-line edition. The first

steps of the process, from digitalization using OCR and manual revision to semi-automatic encoding and mark-up in the form of Text Encoding Initiative XML document (TEI P5), will be summarized (see Ogrin, Erjavec, 2007 for a more detailed treatment of this topic) before we consider the methodology and implementation of and on-line digital repository for the digital edition that can function as a flexible web application. We will present several possibilities offered by our implementation, some of which remain for further experiments. In closing, we wish to present some possible solutions to the hardest problems encountered in this work, from normalization of data and abbreviation expansion to the question of treatment of the original in view of the necessities of every day users who wish to have access to updated information. It is worth noting at this point that the digital edition in current form is available for testing and the reader can follow the presentation at the following URL: http://nl.ijs.si/fedora/sbl/.



**Abraham**, škof v Freisingu na Bavarskem, izvoljen po smrti škofa Lamberta (u. 19. sept. 957), posvečen 21. dec. 957, u. 26. maja 994. V začetku svojega škofovanja je bil pristaš cesarja Otona I. in bavarske vojvodinje Judite ter njenega sina vojvode Henrika II., cesarjevega nečaka. Po smrti Otona I. je izpremenil stališče in se pridružil bavarskemu vojvodu Henriku II., kateri je stremel po osamosvojitvi svoje obširne vojvodine od cesarjeve oblasti, skušal pritegniti kolonizacijsko ozemlje ob Donavi in med alpskimi Slovenci pod svojo interesno sfero ter ustvariti tesne zveze z Italijo, kjer je bila Bavarski pridružena Veronska marka. Upor bavarskega vojvode proti cesarju se je poleti 974 izjalovil, A. je bil za kazen prejkone avg. 974 pregnan v Corvey na Westfalskem, a se je kesneje zopet pomiril s

```
<div>
  <listPerson>
    <person n="main">
      <sex value="1"/>
      <persName>
        <name>Abraham</name>
        <roleName type="eccl">škof</roleName>
      </persName>
      <occupation>duhovnik</occupation>
      <death>
        <date when="0994-05-26">26. maja 994</date>
      </death>
    </person>
    <person n="author">
      <sex value="1"/>
      <persName key="M. Kos.">
        <surname>Kos</surname>
        <forename>Milko</forename>
      </persName>
    </person>
  </listPerson>
  <p>Abraham, škof v Freisingu na Bavarskem, izvoljen
po smrti škofa Lamberta (u. 19. sept. 957), posvečen 21.
```

Figure 1: An SBL article excerpt with its TEI XML encoding

## Text Encoding Initiative: the Data Source

Encoding of SBL is based on open standards and software, in particular the Text Encoding Initiative P5 Guidelines, since TEI has been used as the encoding standard for the digitalization of the Slovenian Bibliographic Lexicon.

TEI produced recommendations or guidelines for the creation and processing of electronic texts for better interchange and integration of scholarly textual data in all languages and from all periods (Burnard, 1988). We used the latest edition of TEI Guidelines, TEI P5, finalized in 2007 (Burnard, Bauman, 2007), since it provides important new encoding features, including new support for manuscript descriptions, multimedia and graphics, stand-off annotation, representa-

tion of data pertaining to people and places and improved specifications for encoding textual alternatives. Additionally, TEI P5 takes advantage of the power of XML schema languages, so that other XML tag-sets, such as MathML or SVG, can now be referenced from within a TEI document and a TEI document can be embedded within other types of XML documents, such as METS and MODS records (Burnard, Bauman, 2007), which turned out to be crucial for the implementation of our on-line repository, since this makes TEI a well-behaved XML citizen, able to take part in any, however complex, XML processing chains and composite documents.

## XML Data Source Structure

The vast majority of SBL articles present information on the life and actions of a single person, while some describe well known families, detailing life and work of several members of the family. An article usually starts with the name of the person or the family, its variants, mostly those used towards the end of their life or the most generally used, followed by a chronological summary of the person’s life and activity, including birth, death, locations, occupations, activities etc. An article may consist of one (usually) or many paragraphs, depending on the exhaustiveness of the article, and is written in dense language, using abbreviations wherever possible, ending with a brief bibliography and other materials relevant to the person, such as portraits or photographs.

The text of the articles has been digitized and manually revised to fix OCR errors before it was automatically converted into the basic TEI-XML format. In the next stage, segments of text that needed to be marked up but could not be identified automatically had to be tagged manually, in particular with details such as different variants of names (linguistic and orographic variations, married names, ecclesiastic names and titles, pseudonyms, complex name parts in the case of foreign names and names with denotation of nobility etc.), making the process slow and error-prone. Since the original data was not normalized, considerable effort had to be spent to achieve high quality TEI XML mark up, and some work with data normalization is still ongoing. (The major aspects of this conversion process have been reported in more detail in Vide Ogrin, Erjavec 2007). In this manner, essential information about the subject of the article and its bibliographical section have been encoded with special purpose elements from TEI P5 biographical and prosopographical modules, representing each article as a  element starting with a <listPerson> element, which contains the detailed information on the subject of the article (names, sex, birth/death date and location, locations of activities, occupations or activities etc.), but also meta data, (volume and year of the first publication, author of the article, revision status etc.). This element is followed with one or more <p> elements with the article text and a <listBibl> element with the extracted bibliographical data.

Obviously, the actual structure to a certain extent depends on the information of that particular article, and so the type and number and elements varies considerably (ie. marriage, ordination, exile, further education, number of occupations, residence, active period etc.). This makes any mapping into a more formal structure, i.e. a relational database, at least awkward.
There is a number of further details that could be extracted from the text but meticulous manual intervention would be required to achieve suitable accuracy. The most important of these are activities undergone by the person, encodable in the <occupation> tag or tags, and locations and times of these activities, encodable by the <floruit> tag.

## Anatomy of an XML-based Document Repository

We have evaluated a number of possible platforms to serve as the base of an on-line web edition of SBL with integrated query and search tools. One possibility considered was PhiloLogic[1], a full-text search, analysis, and retrieval tool developed by the ARTFL Project[2] and the Digital Library Development Center at the University of Chicago. PhiloLogic uses an abstract representation of document structure, projecting the XML data into sets of related database tables, so that the application can search document structure and refine word searching by using the XML structure (Cooney et al., 2007). However, in the end we have choosen the Fedora Commons platform (see The Fedora Project[3]), an extensible framework for storage, management and dissemination of complex objects and object relationships implemented as a portable Java web application (Lagoze et al, 2006). Fedora Commons became the repository on top of which we created a digital library of bibliographical articles with browsing, searching and querying interfaces: a digital library that is presented as an on-line web service and application.
Fedora Commons represents its digital objects as a collection of data streams, where each document is specified as an XML document (Fedora Commons has a native format, called FOXML, but also supports METS). Data streams can be of different types: formally, they can be created as embedded XML documents, as managed independent files in the repository (used for binary files, i.e. images, PDF documents and similar) or as external URI-specified documents. In addition, each object has a number of infrastructure-supporting data streams, all in the form of embedded XML documents. Among them, there is a Dublin Core data stream to contain object meta-data, a RDF data stream to declare inter-object relationships, and internal FOXML revision specifications to allow tracking of object history.

[1] http://philologic.uchicago.edu/

[2] http://humanities.uchicago.edu/orgs/ARTFL/

[3] http://fedora-commons.org/

Furthermore, Fedora Commons objects have dissemination methods (analogues to object or class methods in object-oriented systems), implemented as web application interfaces to objects and their contents (both REST and SOAP interfaces are supported). Since version 3 of the platform, a new Content Model Architecture has been introduced under which dissemination methods are specified with three special objects types: Content Model objects specify available methods and necessary data streams for the dissemination methods they declare, Service Definition objects define a web API for dissemination methods and Service Deployment objects use WSDL (web service definition language) to specify the actual web application API calls necessary to execute a dissemination method request (cf. Fedora Commons Content Model Architecture documentation for version 3[4]).

All the required information, such as the necessary data for each dissemination method call, supported data-types and the manner of invocation to produce the result, are specified in the form of embedded XML documents in the three types of objects, which are otherwise structured as any other object in the repository. To add dissemination methods from one or several different Content Models, it is therefore sufficient to add a special relationship to an object, refering to the Content Model in question. Its dissemination methods will become available under the access URI of the object, contained in a path element of of the Content Model’s name. Furthermore, even the core Fedora Commons features, such as object introspection and direct data stream access, are implemented in this way using the default Content Model.

This extensible and standards-based Content Model Architecture in combination with a number of web services, namely the SAXON XSLT processor, an image manipulation library, a RDF query interface to the object relationship RDF store, a simple search interface to Dublin Core meta-data and object properties and the Fedora Generic Search interface to a number of optional search engines, provide an infrastructure for development of rich application interfaces and complex multi-layered digital repositories using standard technologies and XML workflows.

## From XML Datasources and Workflows to an Online Application

In the Fedora Commons framework, each dissemination method is realized as a web application call (using REST of SOAP methods) with a number of arguments, usually one or several of the object’s data streams. But a data stream can have the form of an URI-specified data stream, referin to another dissemination method and thus resulting in a chain of processing calls. While this approach can be used with binary data, i.e. to apply a number of transformations to an image, it is usually used to create an XML workflow. Such XML workflows

[4] http://fedora-commons.org/confluence/display/FCR30/Content+Model+Architecture

have become the backbone of our application since they allow us to poll together XML data from several sources, such as object data streams and object relationship query results, to form the final XML response, usually in the form of an XHTML rendering in the users browser.
Essentially, there are two kind of data objects in our application: collections and articles. Collections use inter-object relationships to represent different views of the data to the user: these are the top-level object, containing all the other views, letter-objects and volume-objects that allow browsing alphabetically through lists of articles or browsing through the articles in the units of their publications, and search-result objects that take a search query and represent its results as a collection of objects.
Articles are much simpler – in essence they transform the TEI data to an XHTML-encoded web page. However, the resulting page contains a number of context-dependent links, including facilities such as a search interface, browsing links (previous, next), links to instant search queries that provide article lists representing i.e. members of the same occupation class, members of the same generation or all the contemporaries of the subject of the article, but there are also other links, such as a link to all the articles by the same author, accessed via the author name, etc.

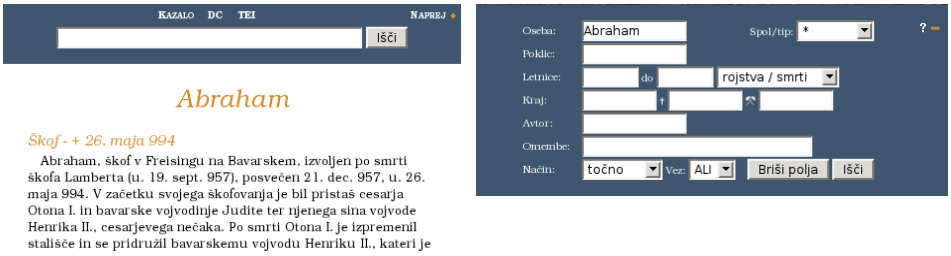



Figure 3: A SBL Article rendering next to the advanced SBL search form

At the same time, all the objects provide direct links to their TEI XML source and their meta data in Dublin Core encoding, transformed from the TEI header-like structures in <listPerson> element of each article (cf. Miller, Brickley 2001 and Liddy et al., 2002). This seemingly trivial feature is actually essential: it means that it is possible, through a public API, to access all of the original TEI document data and all of the meta data, structured in a standard-compliant way. With the combination with platform-integrated support for the Open Archives Initiative Metadata Harvesting Protocol (OAI-MHP), this makes our digital repository very easy to integrate with other systems (cf. Benjamin and Siberski, 2002, Ward 2004).
There is a number of simpler objects, mostly renderings of parts of TEI header information, that simply convey meta data about the whole collection in an accessible form – but they all obey the same logic and use the same infrastructure.

This architecture, in spite of its relative simplicity, has allowed us to construct a flexible and efficient user interface. In general, all the context information has become clickable or otherwise accessible through simple links, making the browsing interface extremely powerful. But the true power of the implementation comes from its search interface.

## Search and Query Interface

Fedora Commons provides a simple integrated search system, capable of simple searches on Dublin Core metada ta and object properties, but a much more powerful system, Fedora Generic Search, is available. The power of this system derives from the fact that is simply provides native Fedora Commons interfaces between an external search system and Fedora Commons API.

In our application, we have chosen to implement Fedora Generic Search on top of SOLR, a search system based on Apache Lucene search and indexing library. In this set-up, Lucene library can use Fedora Commons API to index the document contents, using specially crafted rules (in the form of XSLT stylesheets) to break up the documents in a number of searchable text fields, and the repository gains a search interface with full power of Lucene query language (cf. Hatcher, Gospodnetic 2004), while SOLR takes over the interface and formatting of query results in an easy-to parse XML list.

The power of the interface has been most useful while crafting special queries, such as the context links for different SBL article features, but due to the complexity involved in the use of the flexible Lucene query language, this is hardly the optimal solution for the average user of the system.

To solve this problem, we have implemented two use search interfaces: the simple search is targeted by default at the most often requested fields, namely the subjects’ names, places of their birth and death, and their occupation descriptions. This interface in fact accepts the full Lucene syntax, so it can be used both for simple searches by general users and for complex queries by advanced users.

The secondary interface can be accessed by a click on an expansion button. It presents a form with several fields that easily enable an average user to compose even fairly complex queries, selecting different indexed fields and even using advanced features such as full-text searches, proximity searches, number or date ranges, fuzzy searches etc. This interface is implemented by a secondary script that parses the form and converts its data into a single Lucene query string. Our initial testing with users has been reasonably successful: it makes it trivial to find, for example, female writers, who lived and worked in the period between 1830-1860 in Ljubljana, or priests, who were also philosophers and born in Maribor etc.

The same system can be used for experimental research on the data, especially if one wants to analyse the particularities of the original SBL publication. It is now easy to compare, for example, the average length of articles with the num-

ber of articles contributed by an author (showing the difference between regular contributors and specialists for narrow fields), to plot the average length of life by year of birth or by occupation (and find extreme cases, such as ‘ exceptionally short life spans for heroes of WW2 and revolution). In fact, a number of such queries with graphic interfaces, together with usual on-line features, for example a listing births and deaths on the current date, are being included in the current version of the application. Obviously, this opens up possibilities for further research that falls outside the scope of the present paper.

## Conclusions

The project of the digital edition of SBL has reached production stage, where a complete digital edition is hosted in an on-line digital library and an on-line user interface is available. The main goal of the project has been been achieved: the valuable reference information captured in the lexicon is now again available to the research community and general public, this time enhanced with cross-linking, context information and an advanced query and retrieval system that facilitates it use.

But there are several objectives that are still being worked on. Primarily, we would like to extract and encode further extents of information, since we now have a functional framework thank enables us to use the information, once marked up, to provide further features. The foremost points are the two crucial pieces of information: the subject’s name and occupation, representing the two identificators used to mos often find subjects of interest. The work of manual annoation of name parts with further corrections in spelling and markup of names is being finished at the time of this writing. At the same time, we are working on normalization of occupation specifications (there are more than 1500 different strings for occupations or activities in the SBL articles) and introducing a simple taxonomy to enable meaningful grouping and searching.

Furthermore, we have plans to normalize names of places and annotate them with geographic identifiers. Since we also want to mark-up any names mentioned in the article texts, we are planning to develop a Named Entity Recognition (NER) tool, (Jackson, Moulinier, 2002; Bekavac, 2002) for Slovenian, and we have gathered substantial databases of names and places for this purpose. This task is further complicated by the fact that also Slovenian inflected forms need to be normalised, and the many abbreviated entities in the texts have to be disambiguated, expanded and properly inflected.

In closing, we are happy to report that our system will be used by ongoing and new project in the field of bibliographic publications in Slovenia and will likely become the platform for the digital publication of Slovenian Biographical Lexicon 2 and for the Slovenian Bibliographical Hub which is to integrate most available resources in this domain.

## References

Bekavac, Božo. Strojno obilježavanje hrvatskih tekstova – stanje i perspektive // *Suvremena lingvistika.* 53-54 (2002), 173-182. http://nl.ijs.si/et/tmp/matija/BB_disertacija_verT9.pdf (2007-09-14)

Benjamin, Wolfgang Nejdl, Siberski, Wolf , 2002: OAI-P2P: A Peer-to-Peer Network for Open Archives, v Workshop on Distributed Computing Architectures for Digital Libraries - ICPP2002.

Burnard, Lou. Encoding standards for the electronic edition // Znanstvene izdaje in elektronski medij: razprave / Ogrin, M. (ed.). Ljubljana: Založba ZRC, ZRC SAZU, 2005, 25-42

Burnard, Lou. Report of Workshop on Text Encoding Guidelines // *Literary & Linguistic Computing*. 3 (1988), 131-133

Burnard, Lou; Bauman, Syd. TEI P5:Guidelines for Electronic Text Encoding and Interchange (TEI P5). Text Encoding Initiative Consortium. HTML Version. Oxford, 2007. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ (2009-08-19)

Cankar, Izidor et al. (eds). Slovenski biografski leksikon. Ljubljana: Slovenska akademija znanosti in umetnosti, 1925-1991

Cooney, Charles M. et al.. Extending PhiloLogic. April 2007.

http://www.digitalhumanities.org/dh2007/abstracts/xhtml.xq?id=175 (2007-09-09)

Erjavec, Tomaž; Ogrin, Matija. Digital Critical Editions of Slovenian Literature: an Application of Collaborative Work Using Open Standards. // *From Author to Reader: Challenges for the Digital Content Chain: proceedings of the 9th ICCC International Conference on Electronic* Publishing, Arenberg Castle / Dobreva, M.; Engelen, J. (eds.). Leuven: Peeters, 2005, 151-156

Hatcher, Erik, Gospodnetic, Otis 2004: Lucene in Action. New York: Manning Publications.

Jackson, Peter; Moulinier, Isabelle. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. Amsterdam: John Benjamins, 2002, 180-185

Kahn, Robert; Wilensky, Robert. A framework for distributed digital object services. Working Paper, 1995. cnri.dlib/tn95-01, http:// www.cnri.reston.va.us/k-w.html

Lagoze, Carl; Payette, Sandy; Shin, Edwin; Wilper, Chris. “Fedora: an architecture for complex objects and their relationships”, *International Journal on Digital Libraries*, vol. V6, no. 2, 2006, 124-138.

Liddy, Elizabeth D., Allen, Eileen, Harwell, Sarah, Corieri, Susan, Yilmazel, Ozgur, Ozgencil, Ercan N., Diekema, Anne, Mccracken, Nancy, Silverstein, Joanne in Sutton, Joanne, 2002: Automatic metadata generation & evaluation. V SIGIR ‘02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM Press, 401-402.

Miller, Eric, Brickley, Dan, 2001: Expressing Simple Dublin Core in RDF/XML, Dublin Core Metadata Initiative Proposed Recommendation.

The Fedora Project: An Open-Source Digital Repository Management System, http://fedora-commons.org/

Vide Ogrin, Petra; Erjavec, Tomaž: Towards a Digital Edition of the Slovenian Biographical Lexicon. In: *The Future of Information Sciences: Digital Information and Heritage : Proceedings of the 1st International Conference The Future of Information Sciences - INFuture 2007* / Seljan, Sanja ; Stančić, Hrvoje (ur.). Zagreb : Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu, 2007. 115-124.

Ward, Jewel, 2004: Unqualified Dublin Core usage in OAI-PMH data providers. OCLC Systems & Services, 20(1), 40-47.

Professional paper

# Multilingual Multimedia Thesaurus for Conservation and Restoration – Collaborative Networked Model of Construction

Lucijana Leoni
University of Dubrovnik
Ćira Carića 4, Dubrovnik, Croatia
lleoni@unidu.hr

**Summary**

*Standardization and systematization of terminology within the area of cultural heritage preservation in Croatia will be done by a construction of a model of multilingual multimedia thesaurus through collaboration of experts for conservation and restoration implying the participation of different competences in a community of practice in a network based environment by use of information technologies.*
*A model of controlled dynamic vocabulary with hierarchically connected concepts will be constructed. The source language for thesaurus construction will be Italian, based on three basic relationships terms: equivalence, hierarchical and associative. The number of terms will be limited to cca. 500 preferred terms subdivided into hierarchies. The hypertextual and hypermedial thesaurus would be placed on a web for online browsing and upgrading. The multilingual lexical organization followed by multimedia presentations could be a benefit for the development of conservation and restoration profession in the field of cultural heritage preservation and for joining future European cultural projects. It could also play a significant role in the mobility initiatives aimed at study, conservation, restoration and the appreciation of historic and artistic heritage.*

**Key words:** terminology, cultural heritage preservation, multilingual thesaurus, multimedia, collaborative system, community of practice, open network

## Introduction

Guidelines for professional development in the area of cultural heritage preservation in Croatia underline the need for standardization and systematization of terminology within this interdisciplinary field. This could be done by a construction of a model of multilingual multimedia thesaurus using a collaborative system of a community of practice in a networked open environment. A model of controlled dynamic vocabulary with hierarchically connected concepts will be constructed through collaboration of experts by use of information technolo-

gies. The hypertextual and hypermedial thesaurus will be placed on a web for online browsing and upgrading.

## Theoretical background

Theoretical background consists of the works of key authors in the field of thesaurus construction, international guidelines from the area of documentation of cultural heritage, the relevant ISO standards and existing thesauri of cultural heritage as well as previous experience and achievements in establishing a unified nomenclature and classification in Croatia.

The current historical moment is strongly influenced by the possibilities offered by new technologies, especially networks for research, dissemination and data archiving. Today it is possible to manage extensive data banks and thanks to multimedia environment it is possible to create a basis for context reflection. With numerous initiatives the European Commission encourages the linking of certain targets (usually to overcome language barriers within the EU) with the optimal use of information technologies. Among the initiatives are projects such as: Argos (Art and Restoration Glossary Operating System) financed under INFO2000, then PlasterArchitecture, (http://www.palazzospinelli.org/plaster/) and LMCR (Lessico multilingue tecnico-scientifico di conservazione e restauro) which are co-financed within the framework of EU Culture 2000.

Here are also examples of Paul Getty Foundation AAT - Art & Architecture Thesaurus, Altarpieces - illustrated basic terminology, and Conservation and Art Materials Dictionary Museum of Fine Arts, Boston (http://www.mfa.org/conservation) and others, and finally The multilingual thesaurus attached to the HEREIN project (http://thesaurus.european-heritage.net/sdx/herein/thesaurus/introduction.xsp).

## Area of research

Research and analysis in the field of language of cultural heritage preservation refer to several different disciplines, each of which has an important role in the approach to the project of restoration and conservation of cultural heritage artefacts: the history of art, knowledge of techniques and materials for conservation and restoration of various artefacts, natural sciences such as chemistry and biology, etc. In case when language refers to different art techniques, their specific instruments and materials, research is linked to a variety of traditions, whose use of lexical meaning can take different values with respect to the period, the context and the geographical area of use. This complexity too is part of the specific cultural heritage to be preserved and transmitted.

In such a complex area demanding a variety of competences a starting point for building a Multilingual Thesaurus will be the analysis and comparison of existing relevant specialized texts from Italian and Croatian-speaking areas. Because professional terminology is characterized by higher levels of accuracy, its analysis and systematization seek conceptual professional knowledge. Thor-

ough inquiry into the relevant literature and existing thesauri, in co-operation between experts of this interdisciplinary field the structure of a database for compiling the Thesaurus will be determined.

## Aim of research

The basic objective of this work is the creation of a model of multilingual multimedia Thesaurus in the field of art conservation and restoration with the use of a collaborative web-based environment that will provide an interdisciplinary approach to research and analysis of specialized texts where cooperation between cultural heritage professionals, IT and language experts will be possible. Thesaurus should support terminological uniformity and the correct interpretation of professional texts and manuals.

IT tools would enable the creation of an online Thesaurus that, although incomplete, would have the quality of practical applicability. The web-based network would be used as a laboratory for future building and updating a more exhaustive thesaurus which will start up continuous interaction between students, professionals and researchers in the field of art conservation and restoration from the Italian and Croatian contexts. Such a thesaurus would also help to overcome communicational and intercultural obstacles within this specific context.

In the introductory part a brief historical overview to the present day of the Italian and Croatian literature and lexicography in the field of art conservation and restoration will be shown. On the basis of relevant professional texts intercultural differences in art conservation-restoration profession between Croatian and Italian-speaking areas will be identified, all with the goal of standardization of terminology for the Croatian language in the field of art conservation and restoration.

The construction of a Multilingual multimedia Thesaurus will be started up through the collaborative system by creating a permanent network between institutions in the field of art conservation and restoration with the aim of developing, upgrading and homogeneous use of terminology.

Thesaurus would eventually contribute to the promotion of exchange of traditions, experiences, techniques, tools and different methods of art conservation and restoration workshops and schools, between Italy and Croatia.

Fundamental research review starts with the following hypothesis:

- The terminology of art conservation and restoration for Croatian language has not been systematically studied nor standardized
- Collaborative network tools provide a quality interdisciplinary approach and create a more precise contextual Multilingual Thesaurus construction.
- A construction of a Multilingual multimedia Thesaurus allows more accurate contextual understanding and lexical acquisition.
- Multilingual multimedia database contributes to consistency of translations of specialized texts.

## Methodological procedures

Matter to be used in this research consists of relevant textual and, when possible, multimedia records from the theory of art conservation and restoration, both in Italian and Croatian, as of studies and manuals on the techniques and materials used in the preservation and restoration of tangible artifacts of cultural heritage. Methodology: the source language for thesaurus construction will be Italian, based on three basic relationships terms: equivalence, hierarchical and associative,. The number of terms will be limited to 500 preferred terms subdivided into hierarchies. Comparative analysis of existing sources, data collection, translation, deductive methods of collecting terms, all in collaboration with a group of experts, inductive method of inclusion of new names, photographic collecting, descriptive methods for certain objects or procedures, collection of samples, statistical methods of processing the results, prescriptive methods through the proposal after analyzing the status, etc.

Research will be conducted in collaboration with students of undergraduate and graduate studies of the Department of Arts and restoration at the University of Dubrovnik, who study art conservation and restoration both in Croatian and Italian language, as well as in cooperation with their teachers and mentors in the profession, from the Croatian and Italian-speaking areas. Cooperation will be started up in an open network environment. Specialized articles and texts from the selected context will be analyzed with the aim of defining the relevant terminology.

Lexical equivalents between these two languages will be defined through comparaison and translation. Equivalents, descriptors and definitions in croatian language will be imported in the thesaurus, and where it will be possible, with the accompanying media and iconographic material that contribute to explanations of each term.

In order to demonstrate the usefulness of this model of Thesaurus construction research will be carried out on possibilities of thesaurus construction with a control group of students of Italian language in Croatia who have no knowledge of art conservation and restoration. In order to examine the consistency of professional texts translation with students of art conservation and restoration and a control group of students a model of multilingual multimedia thesaurus will be used when translating technical texts.

## Expected scientific and practical contribution

Considering that in Croatia the terminology in the field of art conservation and restoration has not been systematically investigated a model of Multilingual multimedia Thesaurus of art conservation and restoration (with Croatian - Italian equivalents) based on models of some of the existing multilingual Thesaurus would be set up. Such a project would be in accordance with the guidelines of modern language technology projects in which the emphasis is on the development of the multimedia interface as a means for communication between hu-

mans and computers, as on the development of semantic-based system that would allow access to stored data.
We expect this study to provide new insights into the Croatian terminology in the field of art conservation and restoration, which would improve communication between experts and researchers in this area and enable participation in future European projects of cultural heritage preservation. We also hope to get insight into the degree of connection of collaborative tools and possibilities of creating Multilingual multimedia Thesaurus. These results could also contribute to illuminate the role of the use of e-technologies in the process of professional lexical acquisition.

## Draft structure of a thesis

The work will consist of two parts: the first part will bring the actuality of a subject choice and selection of topics and will elaborate the organization of knowledge in the field of conservation and restoration of tangible cultural heritage, in the second part will appear a model Thesaurus with about 500 lemma in the form of alphabetical printing and hierarchical display.
Art conservation and restoration profession and legal background to these activities in Croatia and Italy will be presented. Review of relevant Croatian-Italian experiences in this area, as well as in the area of the establishment of computer databases and information systems of cultural heritage will be exposed.
A thesaurus as a controlled terminological dictionary will be analyzed followed by representation of Thesauri in Italian language in the field of art conservation and restoration and the world's leading multilingual Thesauri for cultural heritage.
The need for construction of a thesaurus of art conservation and restoration for Croatian language will be exposed and the collaborative approach within the community practice for its modeling will be described.
The results of research conducted in the area of lexical acquisition in developing and applying a thesaurus in a network environment and application of Thesaurus in translating technical texts in the field of art conservation and restoration.

## Conclusion

The network of relevant institutions and experts for conservation and restoration implying the participation of different competences will be organized with the scope to build a thesaurus for conservation and restoration with scientifically detailed definitions subscribed by two languages, Italian and Croatian, to decrease the terminological confusion around the subject. The hypertextual and hypermedia thesaurus would be placed on a web for online browsing and upgrading. The results of the model of Thesaurus construction could be continued to achieve the systematic accomplishment of a complete multilingual Thesaurus of conservation and restoration. The multilingual lexical organisation followed

by multimedia presentations could be a benefit for the development of conservation and restoration profession, for the cultural heritage preservation and for joining future European cultural projects. It could also play a significant role in the mobility initiatives aimed at study, conservation, restoration and the appreciation of historic and artistic heritage.

## References


Amato, Giuseppe. Debole, Franca, Peters, Carol. and Savino, Pasquale. *The MultiMatch Prototype: Multilingual/Multimedia Search for Cultural Heritage Objects.* 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14-19, 2008. Proceedings

Aitchison, Jean; Gilchrist, Alan; Bawden, David. *Thesaurus construction and use: a practical manual*. 4th ed. London : Aslib, 2000

Baldinucci, Filippo. *Vocabolario toscano dell'arte e del disegno*, Firenze, 1691.

Bašić, D., Dovedan, Z.; Raffaelli, I., Seljan, S., Tadić, M. Computational Linguistic Models and Language Technologies for Croatian. Information Technology Interfaces, ITI. Cavtat, 2007. IEEE Str. 521-528

Boras, Damir; Tadić, Marko. *Dva značajna projekta izgradnje računalnih resursa za hrvatski jezik* // Thesaurus Archigymnasii, Zbornik radova u prigodi 400. godišnjice Klasične gimnazije u Zagrebu / Koprek, Ivan (ur.). Zagreb : Klasična gimnazija u Zagrebu, 2007.

Brandi, Cesare. *Teoria del restauro*, Einaudi, 2000.

Bratanić, M. (1991). *Rječnik i kultura*. Zagreb: Filozofski fakultet, Odsjek za opću lingvistiku i orijentalne Studije

Broughton, Vanda. *Essential thesaurus construction*, Facet Publishing, London, 2006.

Burnett, John. *An introduction to terminologies for decorative art, social history, and the history of science. // Terminology for museums /* ed. Roberts, D. Andrew. Cambridge : The Museum Documentation Association, 1990.

Cabré, M. T. *Terminology, Methods and Applications*. A'dam, Philadelphia, Benjamins Publ, 1999.

Cennino Cennini, *Il libro dell'arte, Knjiga o umjetnosti*, Institut za povijest umjetnosti, Zagreb, 2007.

Chapman, R. L. (ur.) (1992). *Roget.s International Thesaurus*. Fifth Edition. New York: HarperCollins Publishers.

Cvrčanin, Milica, *Tezaurus za likovne umetnosti : metodologija izrade*, Narodna biblioteka Srbije, 1983

D’Agostino Laura, Mercalli, Marica, (a cura di) A scuola di restauro - Le migliori tesi degli allievi dell’Istituto Centrale per il Restauro e dell’Opificio delle Pietre Dure negli anni 2003-2005, Roma: Gangemi Editore, 2007.

Faldi, Manfredi; Paolini, Claudio. Tecniche fotografiche per la documentazione delle opere d'arte, Firenze: Edizioni Palazzo Spinelli, 1987.

Križaj, Lana. *Tezaurus spomeničkih vrsta: podatkovni standardi u inventarima graditeljske baštine* / pred bolonjski magistarski rad. Zagreb : Filozofski fakultet, 2006 2006, 204 str. Voditelj: Maroević, Ivo.

Lasić-Lazić, J., *Znanje o znanju*, Zagreb : Zavod za informacijske studije Odsjeka za informacijske znanosti, 1996.

Marlow J., Clough, P.D. and Dance K. *Multilingual Needs Of Cultural Heritage Web Site Visitors: A Case Study Of Tate Online*. International Conference on Image and Analysis and Processing, ICIAP 2007, Modena, Italia, September 2007

Maroević, Ivo. *Uvod u muzeologiju*. Zagreb: Zavod za informacijeske studije, 1993.

Maroević, Ivo. *Muzeologija i znanost u virtualnom okruženju.* // 2. i 3. seminar Arhivi,

knjižnice, muzeji : mogućnosti suradnje u okruženju globalne informacijske infrastrukture / uredile Mirna Willer i Tinka Katić. Zagreb : Hrvatsko muzejsko društvo, 2000.

Nikolić-Hoyt, Anja. *Potential Applications of Conceptually Arranged Thesauruses for Translation Purposes* // Woerterbuch und Uebersetzung / Vida Jesenšek, Alja Lipavic Oštir (ur.). Hildesheim Zurich New York : Georg Olms Verlag, 2008. Str. 308-322.

Nikolić-Hoyt, Anja. *Izrada tezaurusa hrvatskoga jezika*. // Suvremena lingvistika 53-54. 1-2. (2002(2004)) , 53-54; 73-84

Nikolić-Hoyt, Anja. *Iz povijesti višejezičnih tezaurusa*. // Filologija. 38-39 (2002.(2003.)) ; 101-114

Nikolić-Hoyt, Anja. *Odgovarajući obrazac kategorizacije izvanjezičnoga univerzuma kao polazišna faza u izradi tezaurusa*. // Filologija. 30-31 (1998) , 1-2; 91-104

Paolini, Claudio. *L'arte del legno in Italia. Storia tecniche e restauro: un repertorio bibliografico*, Firenze: Edizioni Palazzo Spinelli, 1993.

Paolini, Claudio. *Repertorio bibliografico dell'oreficeria italiana: Storia tecniche e restauro,* Firenze: Edizioni Palazzo Spinelli, 1994.

Paolini, Claudio; Faldi, Manfredi. *Glossario delle tecniche pittoriche e del restauro*, Firenze: Edizioni Palazzo Spinelli, 1999.

Paolini, Claudio; Faldi, Manfredi. *Glossario delle tecniche artistiche e del restauro*, Firenze: Edizioni Palazzo Spinelli, 2005.

Paolini, Claudio. *Glossario delle malte e degli intonaci da rivestimento, decorazione plastica e supporto pittorico*, Firenze: Edizioni Palazzo Spinelli, 2001.

Riccio, Angela.(a cura di) *Chimica del restauro,* Venezia: Marsilio ed., 1984.

Roberts, Helene E. *Naming, defining, ordering : an evolving and never-ending process. //*
*Terminology for museums /* ed. by D. Andrew Roberts. Cambridge : The Museum Documentation Association, 1990.

Roger T. Bell, *Translation and translating*, Theory and practice, Longman, 1991.

Sager, J.C. *A Practical Course in Terminology Processing*. Amsterdam: J. Benjamins, 1990.

Salmon, Jilly. *E-tivities*, London and New York: RoutledgeFalmer, 2006.

Scarpa, F., *La traduzione specializzata,*. Milano: Hoepli, 2001.

Seljan, Sanja; Gaspar, Angelina. *Primjena prevoditeljskih alata u EU i potreba za hrvatskim tehnologijama.* Jezicna politika i jezicna stvarnost / Language Policy and Language Reality. Zagreb: HDPL, 2009. 617-625.

Seljan, Sanja; Agić, Željko; Tadić, Marko. Evaluating Sentence Alignment on Croatian-English Parallel Corpora // Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages. Zagreb: Croatian Language Technologies Society, 2009. 101-108

Seljan, Sanja; Gašpar, Angelina; Pavuna, Damir. Sentence Alignment as the Basis For Translation Memory Database. // INFuture2007-The Future of Information Sciences: Digital Information and Heritage. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, 2007. Str. 299-311

Seljan, Sanja. Information Technology in Machine Translation and in e-Language Learning of Croatian. Current Research in Information Science and Technologies: Multidisciplinary Approaches to Global Information Systems. 1st International Conference on Multidisciplinary Information Sciences & Technologies. 1st International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006. Open Institute of Knowledge, Badajoz, Spain, 2006, vol. II. Str. 359-363

Seljan, Sanja; Pavuna, Damir. Why Machine-Assisted Translation (MAT) Tools for Croatian? // Proceeding of 28th International Information Technology Interfaces Conference – ITI, 2006. Str. 469-475

Stančić, Hrvoje; Seljan, Sanja; Lasić-Lazić, Jadranka. *Digitisation and Language Technologies in the Learning Process of Information Sciences – Approaching the EU Standards* // Proceed-

ings of the 4th International Conference on Quality Management in the Systems of Education and Training : CIMQUSEF, 2007.

Štorga, Mario. *Model rječnika za računalnu razmjenu informacija u distribuiranom razvoju proizvoda /* doktorska disertacija. Zagreb : Fakultet strojarstva i brodogradnje, 23.11. 2005.

Šulek, Bogoslav. *Hrvatsko-njemačko-talijanski rječnik znanstvenog nazivlja*. Zagreb: Globus, 1990.

Tuđman, M. *Struktura kulturne informacije.* Zagreb: Zavod za kulturu Hrvatske, 1983.

Tuđman, M. *Prikazalište znanja*. Zagreb : Hrvatska sveučilišna naklada, 2004.

Tomi Kauppinen, Kimmo Puputti, Panu Paakkarinen, Heini Kuittinen, Jari Väätäinen and Eero Hyvönen: Learning and Visualizing Cultural Heritage Connections between Places on the Semantic Web. *Proceedings of the Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLeS2009), The 6th Annual European Semantic Web Conference (ESWC2009)*, May 31 - June 4, 2009. bib pdf

Urbanija, Jože, *Metodologija izrade tezaurusa,* Zagreb: Naklada Nediljko Dominović, 2005.

Vokić, Denis. *Lakiranje umjetničkih slika,* Zagreb: Kontura, 1996.

Vokić, Denis. *Preventivno konzerviranje slika, polikromiranog drva i mješovitih zbirki,* Zagreb: K-R Centar i Hrvatsko restauratorsko društvo, 2007.

Vokić, Denis (Priredio) *Smjernice konzervatorsko-restauratorskog rada,* Zagreb: K-R Centar i Hrvatsko restauratorsko društvo, 2007.

Wenger, Etienne; Richard McDermott, William Snyder*, Cultivating communities of practice A guide to managing knowledge,* Harvard Business School Press, 2002

Zlodi, Goran. *Muzejska vizualna dokumentacija u digitalnom obliku* . Zagreb : Muzejski dokumentacijski centar, 2004 (u obliku knjige publicirani magistarski rad). Zlodi, Goran. CIDOC-ove Međunarodne smjernice za podatke o muzejskom predmetu i Dublin Core : problemi i perspektive. // 4. seminar Arhivi, knji.nice, muzeji: mogućnosti suradnje u okruženju globalne informacijske infrastrukture. / uredile Mirna Willer i Tinka Katić. Zagreb : Hrvatsko knjižničarsko društvo, 2001.

# Improvements of Dictionaries – Suggestions by Evroterm

Miran Željko
Secretariat General of the Republic of Slovenia, Translation Division,
Gregorčičeva 27A, SI-1000 Ljubljana, Slovenia
miran.zeljko@gmail.com

**Summary**

*The paper presents some possibilities for improving electronic dictionaries from a translator's point of view. Dictionaries, glossaries, terminology databases and corpora are a translator's basic tools. The existing Slovenian electronic dictionaries are based on book dictionaries – data from the books were transformed into computer software using the path of least resistance. However, electronic dictionaries can provide more functions than books, e.g.: full-text search, fuzzy search, terminological analysis, corpus as a source of collocations, dynamically linked dictionary and corpus, and continuous improvement of a dictionary instead of new dictionary projects every few decades. The Evroterm terminology database is presented as a practical example of the proposed improvements.*

**Key words:** terminology database, on-line dictionary, corpus, continuous improvement, full-text search, terminology analyser, Evroterm

## 1. Introduction

Dictionaries in Slovenia are generally first published as books and subsequently transformed into electronic form. The newest and the most extensive English-Slovene dictionary (Krek, 2005-2006) has not yet reached this stage at the time of writing this paper (http://slovarji.dzs.si/dokumenti/Dokument.asp?id=2).

In the future, it will be necessary to change this procedure completely: a book and software are two completely different products and should be designed and made separately. Dictionaries in books are useful for bibliophiles, while translators need electronic dictionaries and during the development of such dictionaries it is necessary to make full use of IT capabilities.

During the development of the Evroterm terminology database (term base) we added some features to it that make it much more useful than ordinary term bases. I believe that at least some of these features (if not all) could also be applied to dictionaries.

## 2. Improving dictionaries[1]

### 2.1. Full-text search

The first electronic dictionaries were books transformed into electronic form, e.g., the contents of *Veliki angleško-slovenski slovar* (Grad 1997) on CD is the same as the book under that title. The main advantage (together with some minor additions) is a faster search because the user does not need to turn pages. What is missing is the most useful improvement: instead of the search being limited to English headwords, a search of Slovene translations of words and collocations could be added.

### 2.2. Fuzzy search

We occasionally make errors when writing – sometimes because of mistyping and sometimes because we have the wrong spelling of a particular word in our mind. In a word processor, a spell checker gives a warning when it encounters an error. When an electronic dictionary does not find a word typed in, it would be user-friendly to show hits similar to the search word.

### 2.3. Corpora

A dictionary and a corpus seem two completely different products: words in a dictionary are sorted alphabetically, while a corpus is a disordered collection of data and the user gets some kind of a sorted output only when the search results are listed. However, a dictionary and a corpus are much more similar than they seem at the first glance. Suppose we have a glossary as the simplest form of dictionary and, in this glossary, one word in the source language (SL) corresponds to one word in the target language (TL) – then this is the simplest form of a corpus. On the other hand, in a large enough bilingual corpus we could find all the words from the glossary, the only obstacle being that we would have to find the mapping between the words (there is more on this topic in Vintar, 2003); a corpus can therefore be regarded as a sort of glossary with a large amount of noise.

Dictionaries and corpora are usually treated as two separate entities. On the web, e.g., there are *Slovar slovenskega knjižnega jezika* (SSKJ – Dictionary of Slovene Literary Language) and the *Nova beseda* (New Word) corpus, which incorporates Slovene literature. It seems natural that the software would list links to examples from Slovene literature when presenting a list of hits from SSKJ; the user would thus see how a particular word is used in the literary language. However, on-line and CD versions of SSKJ are the same as the book form.

The number of examples in a book is limited due to the nature of the medium: depending on paper size and thickness, it is possible to use a book if it contains

[1] In this paper I use the term "dictionary" for dictionaries and similar tools (e.g. terminology databases).

up to about 2000 pages. If a dictionary is too voluminous, it becomes too heavy. This problem can be overcome by publishing a dictionary in several volumes but in practice, we quickly hit a limit. In computer media, this limitation is higher by several orders of magnitude: e.g. (Grad 1997) contains almost 1400 pages with 5000 to 6000 characters on each page. By multiplying these numbers we get between 7 and 8.4 million of characters – which is just 1% of a CD capacity! And a CD is an old-fashioned medium by today's standards.

One of the basic arts of making a dictionary is therefore the selection of suitable examples of use (more on this in Drstvenšek 2003). Several problems can be encountered during this process:

- each author has limited knowledge, so some examples are missing in the dictionary (mostly newer collocations);
- the author may have wanted to prove some hypothesis and thus only selected examples that support his ideas;
- authors of dictionaries are usually people with several decades of experience – so dictionaries may contain words and collocations that are rarely used in modern texts.

These problems can be overcome if a corpus is compiled and a dictionary is designed so that the software searches the corpus directly. If the corpus data are not deliberately biased, the user should obtain the actual data on word use. There are errors in corpora due to the large volume of data but it is usually possible to find a rule from a larger set of hits.

### 2.4. Corpus linked with a dictionary

In book dictionaries, examples of use are static (there is no other possibility). In an electronic dictionary, it makes sense to create a dynamic link between a headword and examples of use; the link is established through search.

From a translator's point of view, a simplified expression of this is that a bilingual dictionary entry consists of three parts (Figure 1):

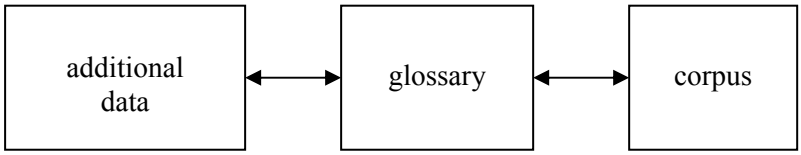
- glossary (headword in SL and TL);
- additional information about headword (depending on the part of speech, language, volume of a dictionary and target users); data that should be in a term base are listed in the ISO 12616 standard;
- examples of use – the simplest way is to use a corpus for this purpose; additional data may also be listed here, e.g., reliability of translation, subject, source, full text containing this word/term, etc.

Two search modes exist in such a system:

1. glossary → additional data → corpus: the user usually does not need all the data; so it makes sense for information to be displayed to him in layers: after making a search he first gets a list of headwords from the glossary. By clicking a particular item he obtains additional data, and with the next click he obtains examples of use from the corpus. The output de-

pends on the dictionary design (more complex operations can be performed in a dictionary that is installed on the user's PC than in an on-line dictionary), purpose of the dictionary, volume of data, etc.

Figure 1: dictionary elements and their relations




2. glossary → corpus → additional data: any dictionary contains a limited volume of words: a general dictionary contains too few technical terms and general terms are missing in a technical dictionary, so it always happens that some terms cannot be found in a dictionary. Some terms probably exist in a corpus of adequate size, so it makes sense to search in another direction. Again, the data are presented to the user in layers: the inputted word is first searched in the glossary and subsequently in the corpus. The user will find the meaning of the unknown word from its vicinity. If the search word exists in the glossary, the user should be able to obtain additional data from it.

From a translator's perspective, a dictionary linked with corpus has the following advantages:

- more search options;
- higher probability that the meaning of a word will be found;
- more data on the search word;
- faster and simpler search.

The disadvantage of this approach is the use of a corpus. A good corpus contains several tens of millions of words. It is not possible to check all of the corpus data because this would be too expensive, so corpora as a rule contain more errors than dictionaries. Users must be aware that if a particular hit deviates from the rest, it is possible that it is not an exception but an error and, in such a case, the data should be checked in another source.

An additional benefit can be obtained by using a multilingual corpus – if the meaning is not clear from a translation from one language into another, then an additional language may help in clarifying the term.

### 2.5. Terminological analysis of a text

A search by entering one word/term into the entry field originates from a search in a book. It is useful if the user wants to find meanings of just a few terms. The real life of a technical translator is entirely different: a translator often gets the

following instructions: "When translating the text, use the terminology from our glossary". If the glossary contains several hundred (or even thousand) terms that are continuously updated and supplied by various translators, it is impossible to know which terms are in the glossary. In such cases, computer-assisted translation tools offer basic help, (e.g., "Translate terms" in SDL Trados TWB). However, the translator needs more: the software should analyse the original text, mark the terms that exist in the term base and, by clicking them, the user should obtain corresponding terminology and corpus data.

**2.6. Presentation on a screen**

A dictionary or corpus search usually produces a large volume of output. It is necessary to put these data in order and to present them so that the user quickly finds what he is interested in. In addition to a sensible layout, colours are very helpful: various data should be in various colours, less important data should be plain black. The aid of colours when using a dictionary can be seen from Amebis dictionaries.

Corpus output is more user friendly if individual units are clearly separated from each other and if the search word is coloured (if it is just bold (as in the SVEZ-IJS corpus) it is more difficult to find it on a screen). In a parallel bilingual corpus, it is easier to find the word in SL and TL if the two segments are parallel to each other. In a sequential output, visual aligning takes longer because the eye has to cover larger distances (this can be seen if the SVEZ-IJS corpus is compared with Evrokorpus). A coloured translation of the search word (if found in the glossary) is an additional help to the user.

**2.7. Continuous improvement**

In the past, a group of people made a dictionary. It was re-printed several times and it was on sale without changes for several years or even decades.

A different approach can be used in electronic dictionaries: the basic version of a dictionary is made as before. Every dictionary has errors and deficiencies, no matter how much effort has been put into its production. Large expenses are associated with changes of a book dictionary. In electronic dictionaries, the problems are solved by the very nature of the media: the cost of a CD is much lower than the book-printing cost; nevertheless, a problem remains because there are several versions of the dictionary on the market. Everything can be simplified by using the Internet: if a dictionary is on the web, the data only have to be updated at one location and each user has access to the newest version. People do not carry Internet-connected computers with them all the time, so it is good to make the dictionary accessible to mobile-phone users, too.

Manufacturing companies have been using the Deming principle of continuous improvements for several decades and this idea can also be applied to the development of dictionaries.

*2.7.1. Improvement of contents*

Updating consists of several tasks:

- correction of errors;
- adding new headwords;
- adding new meanings to existing headwords;
- marking or removing obsolete terms.

Users of dictionary will point out the most obvious errors. The question remains, which headwords to add to an existing dictionary.

One possibility is to use as large a corpus as possible, calculate word frequency and add the most frequent words. Lönneker 2004 suggests that a corpus of literary works should be used as a source of new terms with this approach. What about technical terms?

Another possibility is even simpler: those words are added that users did not find in the existing dictionary. Jakopin 2004 suggests that a web server’s log file could be analysed for this purpose. This may be difficult to do on servers with heavy traffic, because log files grow very quickly. A better solution is for the search program to write unknown words into a special file. The most frequent words from this list are the first candidates for addition to the dictionary.

It rarely occurs to us that dictionary data should be reviewed in order to find and remove obsolete words; this subject is covered, e.g., in Brookes 2004.

*2.7.2. Technical improvements*

Most of the routine update procedures (conversion of data, transfer of data between servers, statistical processing) can be automated. Update frequency depends on the volume of new or changed data within a time unit; an update can be performed monthly, weekly or even daily. In addition to changes to the dictionary contents, software features can also be changed. New functions are available to all users from the moment they are implemented.

It is true that this imposes additional costs. However, the value of the dictionary is much higher because it always contains up-to-date information. High starting costs arise only with the initial preparation of the dictionary.

**2.8. Copyright**

Much more work is required to compile a new dictionary than to write a novel, so the price of the first is much higher. Because of this, the problem of unauthorised copying arises more often. Dictionaries on CDs are protected in such a way that they can only be installed on one disk, but this protection can often be overcome. And if we stick to the publisher’s rules, we may encounter other types of problems:

- suppose I have a desktop PC and a notebook, but I use only one at a time: with this type of protection I need two licenses;

- suppose my hard disk (with the dictionary installed) breaks down and the data cannot be restored;
- or even worse: suppose my PC gets stolen.

In the latter two cases, it is probably possible to obtain another CD with proof of purchase, but some time is lost for this operation and the user will be without a dictionary for some days. All these problems occur because the license is attributed to a PC instead to a person.

If the publisher does not publish a dictionary on CD and stores everything on a web server instead, there appears to be even less protection (protection by username/password is not serious protection because people share passwords).

However, there is professional protection available; banks use it for on-line access by their customers, and government uses it for communication with citizens when transferring sensitive data (e.g., tax data): a digital certificate.

A simplified description of digital certificate protection: a company that wants to limit access to its dictionary must obtain a digital certificate for its server, while a user of the dictionary (client) must obtain a digital certificate for his browser. The client pays a yearly subscription, which is much cheaper than the cost of a dictionary, and he then has on-line access to the dictionary for a specified period.

It is possible that users would share digital certificates but this possibility is rather theoretical, because a user of a borrowed bank certificate would have access to all on-line bank services, and the user of a borrowed government certificate would have access to all personal government-related data. I believe that the volume of false-identity-dictionary-access frauds would be much lower than the volume of unauthorised copied CDs.

There are several advantages when transferring a dictionary to the web and using digital certificates for access protection:

- the publisher of the dictionary maintains data and software in one location;
- all users have access to the most recent dictionary version;
- there is no more production and distribution of CDs;
- dictionary-use license is limited to a person, not to a PC. If the user has several devices and uses only one at a time (e.g., at home, in job, notebook, mobile phone), he can legally access the dictionary from any of them. If he has a copy of the digital certificate, there is no problem if he has to change the PC;
- the user has to pay a much lower initial charge than when buying a CD, so there are more potential customers.

The disadvantage of this approach is that the dictionary is accessible only on-line; but more and more people have full-time Internet access today, which makes this solution ever more applicable.

## 3. A practical example: Evroterm

A term base that uses the improvements mentioned in section 2 (with the exception of limited access) is Evroterm combined with Evrokorpus and Terminator (terminology analyser).

The term base contains terms in 15 languages (the emphasis is on English and Slovene terms – there are more than 100,000 terms in these languages).

The corpus side of the database consists of several corpora: there are five bilingual corpora (English, French, German, Italian and Spanish paired with Slovene as the second language) and one 22-lingual corpus.

The Eur-Lex database is used for access to full-text data.

Modern software has many functions (and many of them are never used), so some functions are hidden deeply in the menu system. Google made an important improvement in this field: with the exception of some lines above and below the entry field the screen is practically empty. On the other hand, terminology experts need additional functions in order to limit the volume of output. It is therefore possible to use either simple or advanced search in Evroterm and Evrokorpus.

In simple search, the user enters the search term (a word, part of a word or several words that can be combined with wildcards) into the entry field and clicks the search button. As a result, he obtains a list of hits in all languages. If there are no hits, the program writes a warning and switches to fuzzy search. If there are no hits even in this case, the program searches Evrokorpus directly. If there are no hits even in the corpus, the program makes a search in IATE (EU term base).

If the user does get a list of hits then he gets details about the first term on the list. Details about other terms on the list are available just by clicking them. If terms on this detailed output exist in the corpus, they are clickable and lead to corpus output. If the corpus segment has been published in the Eur-Lex database or in a database of international treaties that Slovenia has concluded with other countries, a link is provided to the full text of this document. By clicking a Celex number, the user gets a monolingual output, and if he clicks another language at the top of the page, he gets a bilingual output.

If he uses the mobile-phone version, he gets only the basic data.

In advanced search, the user can define:

- SL and one or more TLs,
- one or more fields,
- word-match pattern,
- the type of output.

Corpus search is similar: the program first checks whether the search term exists in the glossary. If it does, its translation is listed with a link to additional data. Afterwards, hits from the corpus are shown. The search terms that were found in the corpus are coloured blue on the output screen. If the program finds

a translation of the search term in the corpus output, the translation is also coloured blue. Every corpus unit also has revision stage data appended to it. The output is sorted so that the user first gets the hits of the highest quality. As before, the user can define specific search criteria in the advanced search.
In Termacor (multilingual terminology combined with multilingual corpus) there is just one user interface for both terminology and corpus search and the user selects the type of output (terminology, corpus or both). The user selects one SL and any number of TLs. If there are up to five TLs, the output is parallel, otherwise the output is sequential.
If the user wants to use the terminology analyser (Terminator), he just copies the text into the text box (word(s), sentence(s) or complete text), selects SL and starts processing. On the output page all terms that exist in the term base are converted to Evroterm links. If input text is bilingual ("Trados segmented"), each sentence will be presented in a separate cell table; the idea is to simplify the search of new terms. The terminology analyser has several functions:

- translators use it to check and use existing terminology;
- terminologists use it to check the glossaries supplied by translators and to find new terms in existing translations.

More than 50,000 searches are performed every workday in the term base.
Wolfgang Teubert finished his paper (Teubert 1999) with the idea that the user of a dictionary should be able to check the corpus data himself – instead of obtaining filtered data by lexicographers. A dictionary is much more complex system than a term base – but it is necessary to start somewhere and it can be said that Evroterm in combination with Evrokorpus is a step in this direction.

## Conclusion

On the basis of the development of the Evroterm term base and Evrokorpus bilingual corpus, the paper presents some possibilities of how to design electronic dictionaries to overcome book limitations:

- searching in both languages in a bi-lingual dictionary;
- full-text search;
- fuzzy search;
- division of dictionary into three parts: glossary, additional data and examples of use;
- independent development of these three parts;
- use of a corpus for retrieving examples of use;
- a corpus as a supplement to dictionary data and a glossary as a supplement to corpus data;
- terminological analysis of a text to be translated;
- continuous improvements of software and data;
- dictionary copyright protection with digital certificates.

## References

Amebis dictionaries: http://www.amebis.si (access date: 10 August 2009)

Brookes, Ian. Painting the Fort Bridge: Coping with Obsolescence in a Monolingual English Dictionary. // *Proceedings of the Eleventh Euralex International Congress*. Lorient: France: Euralex 2004, pp. 221-231.

Deming cycle of continuous improvements: http://www.hci.com.au/hcisite3/toolkit/pdcacycl.htm (access date: 10 August 2009)

Drstvenšek, Nina. Vloga besedilnega korpusa pri postavitvi geselskega članka v enojezičnem slovarju. // *Jezik in slovstvo*, 48/5, 2003, pp. 65-81.

ELAN, SVEZ-IJS and TRANS corpora: http://nl2.ijs.si/index-bi.html (access date: 10 August 2009)

Eur-Lex: http://eur-lex.europa.eu/ (access date: 10 August 2009)

Evrokorpus: http://evrokorpus.gov.si/index.php?jezik=angl (access date: 10 August 2009)

Evroterm: http://evroterm.gov.si/index.php?jezik=angl (access date: 10 August 2009)

Grad, Anton, Škerlj, Ružena, Vitorovič, Nada. Veliki angleško-slovenski slovar. Ljubljana: DZS. 1997

IATE: http://iate.europa.eu/ (access date: 10 August 2009)

ISO 12616. Translation-oriented terminography. ISO, Geneva. 2002

Jakopin, Primož, Lönneker, Birte. Query-driven Dictionary Enhancement. // *Proceedings of the Eleventh Euralex International Congress*. Lorient: France: Euralex 2004, pp. 273-284.

Krek, Simon (ed.). Veliki angleško-slovenski slovar Oxford. Ljubljana: DZS. 2005-2006

Lönneker, Birte, Rozman, Katarina. Online SLO-DE-SLO: spletni slovensko-nemški in nemško-slovenski slovar. // *Zbornik 7. mednarodne multikonference Informacijska družba IS 2004*, book B: Language technologies. T. Erjavec, J. Gros (ed.). 2004, pp. 56-63.

Nova beseda corpus: http://bos.zrc-sazu.si/s_beseda.html (access date: 10 August 2009)

Slovar slovenskega knjižnega jezika: http://bos.zrc-sazu.si/sskj.html (access date: 10 August 2009)

Termacor: http://evrokorpus.gov.si/k2/index.php?jezik=angl (access date: 10 August 2009)

Terminator: http://evroterm.gov.si/x/indexe.html (access date: 10 August 2009)

Teubert, Wolfgang, 1999: Korpuslinguistik und Lexikographie. Deutsche Sprache 4/99. 1999.

Vintar, Špela. Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov: doktorska disertacija. Ljubljana. [COBISS.SI-ID 21981538]. 2003

# Thesauri Usage in Information Retrieval Systems: Example of LISTA and ERIC Database Thesaurus

Kristina Feldvari
Departmant of Information Sciences, Faculty of Philosphy in Osijek
Lorenza Jägera 9, Osijek, Croatia
kfeldvari@ffos.hr

**Summary**

*This paper offers some thoughts on the usage of thesaurus in information retrieval with special reference to information retrieval systems like databases. A thesaurus is an example of controlled vocabulary and an important aid in subject analysis. Controlled vocabularies are used to describe the subject within knowledge organization systems, where the sole purpose of a vocabulary control is to achieve a consistency of subject description and facilitation of information retrieval. We survey some approaches to this question in literature and give two examples of usage of thesaurus in the following databases: the Thesaurus of ERIC Descriptors and LISTA thesaurus. These thesauri are described along with their functions and display in database.*

**Key words:** thesaurus, information retrieval systems, Thesaurus of ERIC Descriptors, LISTA thesaurus

## Introduction

Main research purpose is presentation of thesauri, their function and role in facilitation of information retrieval as well as their function and importance in information retrieval systems. A thesaurus is an example of controlled vocabulary and an important aid in subject analysis. It controls the vocabulary and is formed in a way that facilitates seeking and marking within a specific subject area. It actually has a place at both ends of the information access process, at both storage and retrieval.
Main research questions are: 1) How do IR systems (data bases) use thesauri?
2) Which functions do thesauri support? 3) How are thesauri displayed in data bases? We used the following methods: literature overview, analysis and the comparison of the data.

## Purposes of controlled vocabularies

A keyword search for information on a particular subject performed on the World Wide Web may retrieve thousands of irrelevant documents[1]. According

[1] Svenonius, Elaine. Intelektualne osnove organizacije informacija. Lokve: Benja, 2005. Page 125

to Lancaster the major defect of the Internet as an information source, apart from its size, is the fact that it lacks any form of quality control.[2] We can try to solve this problem by using a subject language that incorporates measures designed to improve retrieval of the desired information. Usage of subject languages to retrieve information provides a value-added quality, which, in the case of highly refined languages, can transform information into knowledge.[3]
A subject language is used to describe what the document is about. The main purpose it serves are primary those of the collocation of documents that have the same information content and the navigation of the users. To achieve the collocation objective, the language must be designed so as to facilitate the retrieval of all and only relevant documents[4]. This is estimated by the twin measures of precision[5] and recall[6].
We can name five main purposes that controlled vocabularies serve:
1) Translation: provide a means for converting the natural language of authors, indexers, and users into a vocabulary that can be used for indexing and retrieval. 2) Consistency: promote uniformity in term format and in the assignment of terms. 3) Indication of relationships: Indicate semantic relationships among terms. 4) Label and browse: provide consistent and clear hierarchies in a navigation system to help users locate desired content objects. 5) Retrieval: serve as a searching aid in locating content objects.[7]
The last purpose is the our research question and will be elaborated in the rest of this paper.

## What is thesaurus?

### Definition

Librarian Lexicon defines thesaurus as a vocabulary of key words, i.e., a standardized set of terms and phrases authorized for use in an indexing system to describe a subject area or information domain.[8] If we take a look at definitions of some authors, e.g. D. Bawden we can notice that he offered definition of the-

[2] Lancaster, F. W. Do indexing and abstracting have a future? // Anales de documentation. 2003, 6; page 137

[3] Svenonius, Elaine. Ibid. Page 125

[4] Ibid, page 126

[5] Recall (R) is the proportion of relevant material retrieved and precision (P) is the proportion of retrieved material that is relevant.

[6] Lancaster, F.W. Indexing and abstracting in theory and practice. Compaing Illions: University of Illinois, 1998. Page 3-4

[7] ANSI/NISO Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. 2005. http://eric.de.gov/ERICWebPortal/resources/html/help/Z39-19-2005.pdf (2009-07-22)

[8] Bibliotekarski leksikon. Beograd: Nolit, 1984. Page 186

saurus purpose. He accents that in information retrieval thesaurus limits and controls the diversity of natural languages by offering an expression that should be use for each concept[9] while M.L. Nielsen says that thesaurus is a well-known and well-established tool in information retrieval that is used to guide indexing and retrieval based on controlled as well as natural language indexing.[10]

**Thesaurus deficiencies- user comprehension and usage**

J.Greenberg in her article stresses out three questions when she talks about user thesaurus comprehension: thesaurus interface design, processing options and end-user warrant. Current thesaurus-supported systems often fail to adequately highlight the thesaurus search option. Information systems may include the word ¨thesaurus¨ on a navigation bar or as a hypertext button, but the explanation of how this feature can assist with the selection of search terms may be hidden. Additionally, systems that include the thesaurus often provide confusing interfaces. They use thesaural identifiers like ¨BT¨ and ¨NT¨ which may not be clear to a user. In this study it is also concluded that if we give a basic thesaurus introduction, users will indicate a desire to employ these tools and also that users favor either interactive or a combination of automatic and interactive thesaurus processing compared to completely automatic processing.[11]

Second question ¨end-user warrant¨ is explained by M. Bates. She proposes that matching and lead-in terminology should be made available for information searchers to help them in their search process. Such an end-user thesaurus would recognize the many variants, informal terms and other terms that users actually input when searching. The thesaurus would be designed to link directly with whatever database the searcher wanted to use, so that the searcher could be led to the ¨legitimate¨ indexing terms.[12]

## Thesaurus and IR systems

Information retrieval is defined as the process of searching a collection of documents, using the term document in its widest sense, in order to identify those documents which deal with a specific subject. The success is determined

[9] Bawden, David. Tezaurusi: nova postignuća. // Vjesnik bibliotekara Hrvatske. 44 (2001), 1-4; page 183

[10] Nielsen, Lykke M. Thesaurus construction: key issue and selected reading. // The thesaurus: review, renaissance, and revision / Roe ,Sandra K. ; and Thomas ,Alan R., (ed.). Binghamton: The Haworth Information Press, 2004., page 58

[11] Ibis, page 15-16.

[12] Bates, M.J. Task force recommendation 2.3 research and design review: improving user access to library catalog and portal information: final report (version 3). 2003. http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf (2009-09-18)

by the accuracy of data retrieved. It is the recall and the precision which attempt to measure the effectiveness.[13]

The information retrieval has changed dramatically in recent years, with the immense increase in availability of searchable full text and the increasing availability of powerful engines for searching the text. Today it is beginning to seem as if all information is available in full text.[14] However, there are many problems (ambiguity, synonyms, etc.) that indicate otherwise. Full text searching will always be valuable for browsing in any size of file but in large files, controlled language access searching will always support efficient retrieval.[15] Therefore, we can conclude that thesauri and indexing are required in facilitating information retrieval. Electronic thesaurus versions have strengthened its role as a search aid. Many operational systems accessible via the internet have incorporated thesauri in their interface as a part of their browsing and searching facilities.[16] Any of these types of system could produce better results by taking advantage of the presence of thesaurus. Most information professionals also point to the value-add of thesauri to justify the cost of traditional databases.[17]

Currently, most large-scale IR systems in general use consist of an indexed document database and a static thesaurus of terms and simple relationships. There are many such thesauri already in existence, designed in the first instance as printed documents to be consulted by human searchers, and there is an international standard setting out detailed rules for their compilation.[18]

An information retrieval thesaurus should, ideally, serve many purposes in information origination, storage and retrieval. Some of the more important applications of the thesaurus in such an environment are listed by Eugene Wall: 1) To serve as a term authority for indexers, so that only "acceptable" terms are employed by indexers. 2) To enable indexers quickly to find the "right" term to signify a concept in mind-"right" in the sense that the term must not only con-

[13] Muddamalle, Manikya Rao. Natural Language versus Controlled Vocabulary in Information Retrieval: A Case Study in Soil Mechanics. 1998. http://nlp.korea.ac.kr/new/seminar/2001spring/research/%5BMuddamalle98%5DNaturalLanguageVSControlledVocInIR.pdf (2009-07-22)

[14] Milstead, Jessica L. Use of Thesauri in the Full-Text Environment. 1998. http://www.bayside-indexing.com/Milstead/useof.htm (2009-07-22)

[15] Batty, David. WWW - Wealth, Weariness or Waste: Controlled Vocabulary and Thesauri in Support of Online Information Access. // D-Lib Magazine. 4 (1998), 10; page 2 http://www.dlib.org/dlib/november98/11batty.html (2009-07-22)

[16] Sihvonen, Anne; Vakkari, Pertti. Subject knowledge improves interactive query expansion assisted by a thesaurus. // Journal of Documentation. 60 (2004), 6; page 674

[17] Ojala, M. Finding and using the magic words: Keywords, thesauri and free text search. // Online. 31 (2007), 4; page 42

[18] Jones, Susan...[et al.]. Interactive thesaurus navigation: Intelligence rules OK? // Journal of the American Society for Information Science. 46 (1995), 1; page 53

note the proper concept but also must be appropriately specific (or general) with respect to the information being indexed. 3) To serve as a means of validating the results of the indexing effort, from the viewpoint of correctness of spelling, to insure that non-preferred synonyms are not employed by indexers, and to "flag" any terms newly required (in the indexer's judgment) by the system. 4) To enable the addition of cross-references between terms in any publication issue-periodic or cumulative-and to validate such cross-references to guarantee against circularity and ¨blindness¨ 5) To enable appropriate formulation of queries put to either printed or computerized indexes. 6) To provide a starting point for other systems which require a vocabulary significantly similar to the one encompassed by the thesaurus at hand. 7) To encourage consistent use of terminology by authors, abstractors, and other originators of information.[19]

## ERIC database thesaurus

The Education Resources Information Center database is sponsored by the U.S. Department of Education to provide extensive access to educational-related literature. ERIC provides coverage of journal articles, conferences, meetings, government documents, theses, dissertations, reports, audiovisual media, bibliographies, directories, books and monographs. We can search ERIC using keywords or using descriptors from Thesaurus. Searching by keywords requires matching the exact words found in a record, while searching by descriptors allows location of the records indexed by subject, regardless of the terminology the author may have used.[20] The Thesaurus of ERIC Descriptors contains an alphabetical listing of terms used for indexing and searching in the ERIC database. This word-by-word alphabetical display provides a variety of information for each descriptor.[21] Except alphabetic search we can also use browsing the Thesaurus by 41 categories. Of course, there is a possibility of combining the selected descriptors with Boolean operators (basic and advanced search) to refine retrieval.[22]

### Cross-references and relations between descriptors

Seven types of cross- references are used: Scope Note (SN), Use For (UF) and Use (USE) references, Narrower Terms (NT), Broader Terms (BT), Related Terms (RT) and Parenthetical Qualifiers.[23]

[19] Wall, Eugene. Symbiotic development of thesauri and information systems: A case history // Journal of the American Society for Information Science. 26 (1975), 2; pages 71-72

[20] ProQuest: ERIC. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[21] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[22] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[23] Ibid.

*Scope Note (SN)*= brief statement of the intended usage of a descriptor. It may be used to clarify an ambiguous term or to restrict the usage of a term.[24]

*Example*:

INFORMATION RETRIEVAL

SN Techniques used to recover specific information from large quantities of stored data.[25]

*Use For (UF)* and *USE (USE)*= terms we consider to be equivalent (equal or almost equal by the meaning) we can combine to the category of equivalence so that equivalent expressions match only one term. Equivalence relations direct synonyms and pseudosynonyms of specific term to appropriate descriptor. For these relations we use UF and USE references.[26]
The UF reference is employed generally to solve problems of synonymy occurring in natural language. Terms following the UF notation are not used in indexing. They most often represent either (1) synonymous or variant forms of the main term, or (2) specific terms that, for purposes of storage and retrieval, are indexed under a more general term. Years listed in parentheses indicate the time period during which the term was used in indexing. It provides useful information for searching older printed indexes, or computer files that have not been updated.[27]

*Example:*

BIBLIOGRAPHIC DATABASES

UF Bibliographic Records (2004); Bibliographic Utilities (2004)[28]

The USE reference, the mandatory reciprocal of the UF, refers an indexer or searcher from a no usable (non indexable) term to the preferred indexable term or terms.[29]

[24] Craven, Tim. Thesaurus construction. 2008. http://publish.uwo.ca/~craven/677/thesaur/main00.htm. (2009-07-22)

[25] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[26] Urbanija, Jože. Ibid. Page 27

[27] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[28] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[29] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

*Example:*

KINESCOPES
USE Films

*Narrower Terms (NT)* and *Broader Terms (BT)*= These indicate the existence of a hierarchical relationship between a class and its subclasses. In a hierarchical relation, one term is viewed as being "above" another term because it is broader in scope. Narrower terms are included in the broader class represented by the main entry. The Broader Term (BT) is the mandatory reciprocal of the NT. Broader Terms include as a subclass the concept represented by the main (narrower) term.[30]

*Example:*

SCHOOL CULTURE
BT Culture; Organizational Culture

*Example:*

RECREATIONAL ACTIVITIES
NT Playground Activities; Recreational Reading[31]

*Related Terms (RT)*= Associative relations express the analogy (not equivalence) between concepts. These kinds of relations are used for not hierarchical semantic relations in the thesaurus.[32]

*Example:*

ALCOHOLISM
RT Addictive Behaviour; Alcohol Education; Antisocial Behaviour; Behaviour Disorders; Drug Addiction; Fetal Alcohol Syndrome; Physical Health; Special Health Problems[33]

*Parenthetical Qualifiers*= A Parenthetical Qualifier is used to identify a particular indexable meaning of a homograph. In other words, it discriminates between terms (either Descriptors or USE references) that might otherwise be confused with each other. Examples include LETTERS (ALPHABET) and LETTERS (CORRESPONDENCE). The Qualifier is considered an integral part of

[30] Urbanija, Jože. Ibid. Page 28

[31] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

[32] Urbanija, Jože. Ibid. Page 31

[33] ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

the Descriptor and must be used with the Descriptor in indexing and searching.[34]

## LISTA database thesaurus

LISTA (Library, Information Science & Technology Abstracts) is a bibliographic database made available by EBSCO. The database offers searchable cited references, alerts functionality, author profiles and online tutorials. It provides coverage on subjects such as librarianship, classification, cataloguing, online information retrieval and information management. The thesauri in both the LISTA and LISTA with Full Text databases include 6,800 terms, 2,700 of which are preferred terms.[35] We can browse LISTA thesaurus by choosing tree type of displays: *Term begins with* displays a browsable alphabetical list, *Term contains* displays all the subject descriptors that contain requested term, whether it's the first word or not, and other terms to which requested term is related and *Relevancy ranked* displays the exact match to requested term first, if one exists, followed by subject terms "in order of relevance." As well as it is in the Thesaurus of ERIC Descriptors, here we can also combine two or more descriptors using Boolean operators to refine retrieval. Unlike the the Thesaurus of ERIC Descriptors in LISTA thesaurus there is no browsing by category list, just alphabetic list.. Another obvious difference from the Thesaurus of ERIC Descriptors is that cross-reference USE (USE) appears in all displays and there is possibility of ¨exploding¨ the term.[36]

### Cross-references and relations between descriptors

Since we have explained types of relations between descriptors on the example of the Thesaurus of ERIC Descriptors, in the rest of the paper will be given only examples of that relations in LISTA thesaurus.

Six types of cross- references are used: Scope Note (SN), Use For (UF) and Use (USE) references, Narrower Terms (NT), Broader Terms (BT), and Related Terms (RT).

*Scope Note (SN) example:*

ACADEMIC librarians

SN Here are entered works on librarians who manage and maintain college and university libraries.

[34] ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

[35] EBSCO publishing: Customer Success Center: LISTA. 2009. http://www.ebscohost.com/customerSuccess/default.php?id=7 (2009-03-08)

[36] Library, Information Science & Technology Thesaurus. 2009. http://web.ebscohost.com/ehost/thesaurus?vid=2&hid=8&sid=bb81003c-cc48-4dfd-8a97-593c9d9ec7a8%40sessionmgr10 (2009-03-08)

*Use (USE) example:*
INFORMATION centres
USE INFORMATION services
*Use For (UF) example:*
PUBLIC domain (Copyright law)
UF COPYRIGHT-- Public domain
*Narrower Term (NT) example:*
COMPUTER FILES
NT COMPUTER programs; DATABASES; IMAGE files; TEXT files
*Broader Terms (BT) example:*
TELEGRAPH
BT TELECOMMUNICATION
*Related Terms (RT) example:*
SCHOLARLY publishing
RT ACADEMIC writing; CONFERENCE proceedings; MONOGRAPHIC series; SCHOLARLY periodicals; UNIVERSITY presses[37]

## Conclusion

Today, when information sources are growing enormously, there is a need for more effective information retrieval. Although in each database we have possibility to use Boolean operators to refine our retrieval it seems that is not enough. This is because of linguistic problems that can occur. Thesaurus copes with these problems very well so we can conclude that this tool is vital retrieval tool in databases. The main problem that remains is limited users' thesauri comprehension. If we could correct and overhaul these problems thesauri would probably be more useful for users during IR processes.

## References

ANSI/NISO Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. 2005. http://eric.de.gov/ERICWebPortal/resources/html/help/Z39-19-2005.pdf (2009-07-22)

Bates, M.J. Task force recommendation 2.3 research and design rewiew: improving user acess to library catalog and portal information: final report (version 3). 2003. http://www.loc.gov/catdir/bibcontrol/2.3BatesReport6-03.doc.pdf (2009-09-18)

Batty, David. WWW - Wealth, Weariness or Waste: Controlled Vocabulary and Thesauri in Support of Online Information Access. // *D-Lib Magazine*. 4 (1998), 10; pages 1-6 www.dlib.org/dlib/november98/11batty.html (2009-07-22)

Bawden, David. Tezaurusi: nova postignuća. // *Vjesnik bibliotekara Hrvatske*. 44 (2001), 1-4; pages 182-187

Bibliotekarski leksikon. Beograd: Nolit, 1984.

Craven, Tim. Thesaurus construction. 2008. http://publish.uwo.ca/~craven/677/thesaur/main00.htm. (2009-07-22)

[37] Ibid.

EBSCO publishing: Customer Success Center: LISTA. 2009. http://www.ebscohost.com/customerSuccess/default.php?id=7 (2009-03-08)

ERIC: Education Resources Information Center: Search the Thesaurus. http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Thesaurus&_nfls=false (2009-08-01)

Greenberg, J. User comprehension and searching with information retrieval thesauri // The thesaurus: review, renaissance, and revision / Roe ,Sandra K. ; and Thomas ,Alan R., (ed.). Binghamton: The Haworth Information Press, 2004., page 103-120.

Jones, Susan.[et al.]. Interactive thesaurus navigation: Intelligence rules OK? // Journal of the American Society for Information Science. 46 (1995), 1; pages 52-59

Lancaster, F. W. Do indexing and abstracting have a future? // Anales de documentation. 2003, 6; pages 137-144

Lancaster, F.W. Indexing and abstracting in theory and practice. Compaing Illions: University of Illinois, 1998.

Leščić, Jelica. O tezaurusu načela, izradba, struktura: pregled. // Vjesnik bibliotekara Hrvatske 44 (2001), 1-4; pages 172-181

Library, Information Science & Technology Thesaurus. 2009. http://web.ebscohost.com/ehost/thesaurus?vid=2&hid=8&sid=bb81003c-cc48-4dfd-8a97-593c9d9ec7a8%40sessionmgr10 (2009-03-08)

Milstead, Jessica L. Use of Thesauri in the Full-Text Environment. 1998. http://www.bayside-indexing.com/Milstead/useof.htm (2009-07-22)

Muddamalle, Manikya Rao. Natural Language versus Controlled Vocabulary in Information Retrieval: A Case Study in Soil Mechanics. 1998. http://nlp.korea.ac.kr/new/seminar/2001spring/research/%5BMuddamalle98%5DNaturalLanguageVSControlledVocInIR.pdf (2009-07-22)

Nielsen, Lykke M. Thesaurus construction: key issue and selected reading. // The thesaurus: review, renaissance, and revision / Roe ,Sandra K. ; and Thomas ,Alan R., (ed.). Binghamton: The Haworth Information Press, 2004., pages 57-74

Ojala, M. Finding and using the magic words: Keywords, thesauri and free text search. // *Online*. 31 (2007), 4; pages 40-42

ProQuest: ERIC. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

ProQuest: ERIC Thesaurus. 2009. http://www.csa.com/factsheets/eric-set-c.php (2009-08-01)

Sihvonen, Anne; Vakkari, Pertti. Subject knowledge improves interactive query expansion assisted by a thesaurus. // *Journal of Documentation*. 60 (2004), 6; pages 673-690

Svenonius, Elaine. Intelektualne osnove organizacije informacija. Lokve: Benja, 2005.

Urbanija, Jože. Metodologija izrade tezaurusa. Zagreb: Naklada Nediljko Dominović, 2005.

Wall, Eugene. Symbiotic development of thesauri and information systems: A case history // *Journal of the American Society for Information Science*. 26 (1975), 2; pages 71-79

Original scientific paper

# Tagset Reductions in Morphosyntactic Tagging of Croatian Texts

Željko Agić [*], Marko Tadić [**], Zdravko Dovedan [*]
[*] Department of Information Sciences, [**] Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{zeljko.agic, marko.tadic, zdravko.dovedan}@ffzg.hr

**Summary**

*Morphosyntactic tagging of Croatian texts is performed with stochastic taggers by using a language model built on a manually annotated corpus implementing the Multext East version 3 specifications for Croatian. Tagging accuracy in this framework is basically predefined, i.e. proportionally dependent of two things: the size of the training corpus and the number of different morphosyntactic tags encompassed by that corpus. Being that the 100 kw Croatia Weekly newspaper corpus by definition makes a rather small language model in terms of stochastic tagging of free domain texts, the paper presents an approach dealing with tagset reductions. Several meaningful subsets of the Croatian Multext-East version 3 morphosyntactic tagset specifications are created and applied on Croatian texts with the CroTag stochastic tagger, measuring overall tagging accuracy and F1-measures. Obtained results are discussed in terms of applying different reductions in different natural language processing systems and specific tasks defined by specific user requirements.*

**Keywords:** morphosyntactic tagging, part-of-speech tagging, stochastic tagger, Multext East tagset, tagset reductions, Croatian language

## Introduction

A typical usage cycle for a majority of stochastic morphosyntactic taggers found today consists of sequentially applying two procedures: the training procedure and tagging procedure. The training procedure takes a previously annotated training corpus of a certain language as input, which it derives into an output language model readable by the tagging procedure. The tagging procedure is fed afterwards with unseen sentences of that language and it uses the language model in order to assign the most probable tags to word forms in the input sentences. Types of these language models and assignment algorithms vary in state-of-the art solutions: from hidden Markov models (Brants 2000; Halácsy et al. 2007) and support vector machines (Giménez and Márquez 2004) to cyclic

dependency networks (Toutanova et al. 2003) and bidirectional perceptron learning (Shen et al. 2007). The tagging accuracy of these methods peeks between 96 and 98 percent on the task of tagging English. Due to such high scores on English, the morphosyntactic tagging task is often considered as a closed or resolved issue in the computational linguistics and natural language processing communities. However, when using these procedures in tagging languages other than English, namely highly inflectional languages such as Czech, Croatian, Slovene and other Slavic languages, the tagging accuracy decreases (cf. Agić et al. 2008a and 2008b) to a point from which the given task does not seem as resolved as it did from the viewpoint of English language.

There are basically two issues that emerge when focusing on Slavic languages rather than English: the size of available corpora and the size of the tagset. On one side, rich morphology demands a more complex tagset in order to describe all the morphosyntactic phenomena. For example, the Penn Treebank is tagged using only 36 morphosyntactic tags (or part-of-speech tags, as it is perhaps better suited in this case), while the experiment with tagging Croatian texts using the TnT tagger (Agić and Tadić, 2006) utilized around 900 different morphosyntactic tags out of the overall 1475 tags that occur in the Croatian Morphological Lexicon (Tadić and Fulgosi 2003, Tadić 2005). And on the other side, lesser spread languages such as Croatian usually do not have at their disposal the person-months required to develop large manually annotated corpora such as, e.g., the Prague Dependency Treebank (Böhmová et al. 2003) for Czech. Even though the 100 Mw Croatian National Corpus does exist (Tadić 2002; Tadić 2006), only its minor part, the Croatia Weekly 100 kw subcorpus was manually annotated with morphosyntactic tags in order to train and experiment with stochastic taggers.

There are basically two separate approaches to improving morphosyntactic tagging accuracy that can be found in the field today:

1. Combining various taggers with each other or with other available language resources and language processing tools. For example, (Rupnik et al. 2008) combines a hidden Markov model tagger with a support vector machine tagger in the task of tagging Slovene, while (Sjöbergh 2003) utilized seven different taggers that implemented six different stochastic tagging paradigms in order to raise overall tagging accuracy for Swedish. For Croatian, an approach with combining the existing hidden Markov model tagger CroTag and the Croatian Morphological Lexicon was undertaken (Agić et al. 2008b), based on the experience of the HunPos tagger of Hungarian texts (Halácsy et al. 2006 and 2007). These approaches are said to either create hybrid taggers – such is the case with CroTag and HunPos when coupled with inflectional lexica – or voting taggers, using additional stochastic for deciding on the best of outputs provided by different taggers, hoping for a divergence of those towards the actual solution. Voting taggers are considered to have an advantage over hybrid taggers when adaptability to various languages is re-

quired, while hybrid taggers are usually more finely tuned for tagging a single specific language.

2. Manipulating the language model. These approaches mainly focus on reducing the tagset to a size desirably comparable to that of the e.g. Penn Treebank in order to downgrade the tagging problem for a given rich morphology language to that of tagging English. Reducing the tagset targets the language model directly, as stochastic taggers are based on counting occurrences of tags in the training corpus: the lower the overall tag count, the finer grained their distributions in the resulting language model. Notable approaches include the so-called tiered tagging approach (Tufiş 1999, Tufiş and Dragomirescu 2004), which compresses or maps the actual tagset into a hidden layer of tags with which the tagging is performed. The real tags are afterwards restored from the hidden layer using a lexicon and a set of handwritten rules. The approach has been shown to work well with different tagging paradigms (cf. Ceauşu 2006). The idea of tiered tagging can be traced back to (Brants 1995), a similar approach that did not yield significant improvements over the baseline tagging accuracy, unlike the tiered tagging approach.

In hindsight, all of these approaches are strictly scientific and task-oriented, as they aspire towards the ideal solution of approaching 100% tagging accuracy for a given language (or any language) while using the full morphosyntactic tagset for that language. However, keeping in mind that morphosyntactic taggers are generally not utilized as standalone applications, but rather as one of many modules in assembling larger natural language processing systems such as named entity recognizers or document classifiers, it should be considered – and this is of special importance for processing languages with sparse language resources and tools, such as Croatian – when and how to reduce the complexity of the tagging task in terms of user- or system-specific requirements. This paper investigates the specific user-oriented approach in which the full morphosyntactic tagset used for tagging Croatian corpora is mapped or split into several meaningful subsets from which the prospective user can choose a language model that is best suited for a specific natural language processing task.

Further sections of the paper describe this generally set research plan in more detail, including short descriptions of the corpus and tagger used in the experiment, along with the setup of the experiment itself. Results are afterwards discussed along with future work plans in the ending section.

## Experiment

In the task of reducing a full morphosyntactic tagset into subsets for tagging Croatian texts, three modules must be observed in more detail: the tagger, the corpus from which the language model of the tagger is constructed and finally the tagset itself. The first two modules – the tagger and the corpus – are thoroughly described in previous publications (Agić and Tadić 2006) and (Agić et

al. 2008b) and therefore we present them here in a short overview, focusing afterwards on the morphosyntactic tagset.

The Croatia Weekly 100 Kw manually tagged newspaper corpus (the CW100 corpus further in the text) consists of articles extracted from seven issues of the Croatia Weekly newspaper, which has been published from 1998 to 2000 by the Croatian Institute for Information and Culture (HIKZ). This 100 Kw corpus is a part of Croatian side of the Croatian-English Parallel Corpus (CW corpus) described in detail in (Tadić 2000). The CW100 corpus was pre-tagged using the Multext-East version 3 morphosyntactic specifications (Erjavec 2004) on top of the XCES corpus encoding standard. The whole CW corpus was in fact built in two separate processing stages, as described in (Tadić 2000): firstly, the raw text data was automatically converted into XML format and afterwards tokenized in order to be semi-automatically tagged using full Multext-East version 3 tagset by matching the CW100 corpus and the Croatian Morphological Lexicon (Tadić and Fulgosi 2003, Tadić 2005) at unigram level via the Croatian Lemmatization Server (http://hml.ffzg.hr). The corpus consists of exactly 118529 word forms in 4626 different sentences, tagged by 896 different morphosyntactic tags. Nouns make for a majority of corpus word forms (approximately 30%), followed by verbs (~15%) and adjectives (~12%) which is in fact a predictable distribution for a newspaper corpus.

CroTag is a hybrid tagger consisting of two modules: the second order hidden Markov model training and tagging module (often called the trigram tagger, even though hidden Markov model tagging and trigram tagging are not necessarily the same procedures) and the inflectional lexicon module for boosting the tagger accuracy on unknown word forms. Its description is given in (Agić et al. 2008b) and error analysis provided in (Agić et al. 2009). The tagger uses the second order Viterbi algorithm with beam search to do the actual tagging, while language model sparseness is handled by linear interpolation smoothing at model building time and suffix tries with successive abstraction at runtime, i.e. upon encountering unknown and unhandled word forms. Its accuracy is obviously input dependent as it is a stochastic tagger: it yields an overall accuracy score of approximately 85 percent on a test corpus containing approximately 15 percent unknown word forms. Accuracy rises when decreasing the number of unknown word forms to ~95% correctly assigned tags with ~5% unknown word forms. With such figures, CroTag can be considered a state-of-the-art morphosyntactic tagger.

As mentioned before, Croatian texts are tagged using morphosyntactic tags from the Multext-East version 3 tagset specification for Croatian. As described in detail in (Agić et al. 2009), the tagset is positional, with each of the positions inside tags representing a single morphosyntactic category using different alphabetical characters for denoting different category values. For example a tag Ncmsn would denote a {Noun, common, masculine, singular, nominative} token. Position zero always represents part of speech information (PoS), while

other tag positions represent morphosyntactic categories and their values belonging to this part of speech (MSD). Querying the database backend of the Croatian Lemmatization Server (Tadić 2005) revealed a total of 1475 different Multext-East v3 morphosyntactic tags that are currently instantiated from this tagset in the Croatian Morphological Lexicon, i.e. on approximately 110.000 different lemmas and more than 4 million corresponding word forms.

Table 1. Properties of reduced tagsets on the CW100 corpus

| Reduction | Type | Number of tags |
|---|---|---|
| $subset_0$ | Full Multext-East v3 tagset | 896 |
| $subset_1$ | Removes all MSD information for all non-inflective parts of speech and numerals | 800 |
| $subset_2$ | Removes all MSD information for all non-inflective parts ($subset_1$) of speech, numerals and verbs | 739 |
| $subset_3$ | Uses $subset_2$ and removes all other MSD information except gender, number and case on nouns, pronouns and adjectives and type on nouns | 243 |
| $subset_4$ | Uses $subset_3$ and removes information on case from all remaining MSD information | 48 |
| $subset_5$ | Uses $subset_4$ and removes information on gender and number from remaining MSD information | 15 |
| $subset_6$ | Part of speech information only | 13 |

Now that the modules are presented, tagset reductions must be introduced. Each of the reductions made for this experiment introduces another tagset, i.e. a specific subset of the full Multext-East v3 for Croatian. Obvious enough, the subsets will always impose fewer tags on the corpus than the original tagset. They will be named as subseti, the subscript i indicating depth of the reduction: the higher the index, the stricter the reduction and fewer the number of tags in the subset. Overview of the reductions is given in table 1 and a more elaborate description follows the table.

The first reduction in the table is not a reduction at all: subset0 represents the full tagset and is provided as a reference point or baseline figure. Similar to that, subset6 is a trivial reduction in which all information except the one about the part of speech is discarded. The reductions that can be found in between these upper and lower bounds are designed considering two viewpoints: the error analysis for CroTag in (Agić et al. 2009) and some basic intuition on system- and user-requirements. Namely, the above-mentioned experiment found that approximately 85 percent of all tagging errors occur on nouns, adjectives, pronouns and verbs and that approximately 50 percent of these are, in fact, incorrect assignments of case values. Therefore, the subsets are constructed by first dropping all the information on morphosyntactic categories of non-inflective parts of speech and verbs, eliminating the noise and focusing the analysis on the most difficult categories of the most difficultly tagged inflective parts of speech:

adjectives, nouns and pronouns. In addition, type and degree are stripped from adjectives and type and person from pronouns. Furthermore, case is stripped from these three parts of speech in subset4 and gender and number in subset5, leaving only morphosyntactic category of type for nouns (reminder: a noun can be common or proper and type denotes this). A common guideline for these reductions, besides the error analysis, was – as mentioned before – intuition on user and system requirements. This basically means that amount of information carried by a morphosyntactic category was considered from an average user and system viewpoint. From this perspective, it could be argued that, for example, information on noun type (common or proper) encodes more information – and in addition, information that is more valuable to the natural language processing system or its user – than information on noun case (nominative, genitive, etc.). As an illustration of this argument, consider a named entity detection and classification (NERC) system such as (Bekavac and Tadić 2007). In order to implement a normalization feature that would normalize various types of named entities occurring in the text to their normal (singular, nominative) form, one would require a morphosyntactic tagger able to correctly discriminate between common and proper nouns and male and female gender than e.g. between cases of adjectives and pronouns. Otherwise, the user might end up with a system that would normalize the entity Ive Sanadera as Iva Sanader (female) rather than the obvious choice Ivo Sanader (male) for example. Avoiding or encountering such an error in this framework depends exclusively on morphosyntactic tagging module and hence the intuition that led to these specific tagset reductions.

The data in table 1 is self-explanatory. However, it is rather interesting to note that maintaining gender, number and case for adjectives, nouns and pronouns and type for nouns and removing all other information from the tags induces a serious drop in the number of tags from subset2 to subset3. Removing case information expectedly reflects in overall tag numbers roughly as division of subset3 cardinality by seven as there are seven distinct cases in the Croatian language. The gaps in tag-space between subset2 and subset3 and also subset3 and subset4 should by all means be noted as they indicate there are many other options than only these presented in this paper. All of them should be considered for detailed sub-tagset design on basis of specific user or system requirements.

The experiment setup was also taken from the (Agić et al. 2009) experiment with CroTag error analysis. More specifically, the CW100 corpus is split into ten different parts, equal in number of sentences contained. Nine parts are used for creating the language model for the tagger and the tenth is always used for validating that model. The training sets had ca 106.676 tokens on average (average 23.426 types), while the testing sets had average 11.852 tokens (average 4.638 types). All counts and results are tenfold cross-validated. This procedure is repeated for each of the reduced tagsets subseti. Overall tagging accuracy is provided for the subsets and separate F1-measures are given on adjectives,

nouns and pronouns, i.e. the most difficult parts of speech for tagging Croatian texts. The following section provides experiment results and discussion.

## Discussions of results

The results of the experiment are presented in condensed form by tables 2 and 3. Table 2 provides information on overall tagging accuracy achieved by the CroTag tagger on all the tagset reductions. For each of these subsets, the tagger was first trained on 90 percent of the CW100 corpus – the full tagset of the corpus reduced beforehand, corresponding to the subset in question – and then tested on the remaining 10 percent. The procedure was repeated ten times for each of the subsets, i.e. it was tenfold cross-validated. In the table, overall accuracy is given as a function of the number of different morphosyntactic tags found in each subset (see table 1). The tagging accuracy itself is presented by stating the average accuracies for each of the reduced tagsets, followed by their 95 percent confidence intervals. The table is accompanied by a simple histogram in figure 1 in order to indicate the functional dependency between the number of tags and overall tagging accuracy.

Table 2. Overall tagging accuracy with reduced tagsets

| Reduction | Number of tags | Accuracy |
|---|---|---|
| $subset_0$ | 896 | 84.80±1.62 |
| $subset_1$ | 800 | 85.35±1.86 |
| $subset_2$ | 739 | 85.77±1.76 |
| $subset_3$ | 243 | 86.18±1.94 |
| $subset_4$ | 48 | 90.35±1.69 |
| $subset_5$ | 15 | 96.02±1.00 |
| $subset_6$ | 13 | 96.23±0.97 |

Both table and figure indicate an expected behaviour of the stochastic tagger: accuracy steadily rises with the decrease of the tagset size. More precisely, this dependency is expected due to the sparseness issue in the contextual probability matrices of second order hidden Markov model taggers (cf. Agić et al. 2008a). However, with respect to goals of this experiment, it should be noted that the decrease in tagset size gained when moving from subset2 to subset3 – amounting to a difference of 496 morphosyntactic tags – is shown here to provide only a slight gain of 0.41 percent in tagging accuracy while dropping 195 tags when moving from subset3 to subset4 caused the tagger to be a substantial 4.17 percent more accurate. Moreover, moving from subset4 to subset5, thereby dropping 33 tags also resulted in a substantial accuracy increase of 5.67 percent (i.e. accurately tagging 1 or 2 more word forms in a sentence with 25 word forms!), indicating that the stochastic tagger gains more accuracy when decreasing in the region of smaller tagsets. Therefore, tagset design should be approached with

caution between these margins when keeping in mind overall goals of specific natural language processing system design.

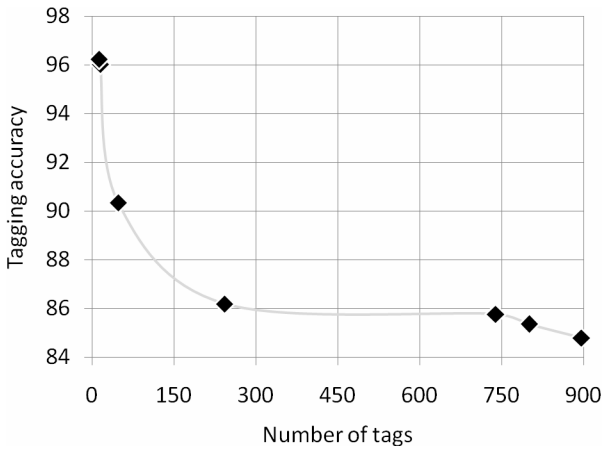



Figure 1. Tagging accuracy as a function of tagset size

Table 3 provides F1-measures on the most difficultly tagged parts of speech in Croatian: adjectives, nouns and pronouns. Recall and precision are left out of the table for conciseness and also because they were so narrowly tied with each other, thus rendering them uninteresting.

Table 3. F1-measures on adjectives, nouns and pronouns

| | $subset_0$ | $subset_1$ | $subset_2$ | $subset_3$ | $subset_4$ | $subset_5$ | $subset_6$ |
|---|---|---|---|---|---|---|---|
| **Adj** | 0.64±0.04 | 0.63±0.04 | 0.63±0.04 | 0.65±0.05 | 0.74±0.05 | 0.92±0.02 | 0.91±0.03 |
| **Noun** | 0.79±0.03 | 0.78±0.03 | 0.78±0.04 | 0.78±0.04 | 0.86±0.03 | 0.95±0.01 | 0.97±0.01 |
| **Pro** | 0.76±0.03 | 0.75±0.04 | 0.75±0.05 | 0.76±0.05 | 0.87±0.04 | 0.99±0.01 | 0.99±0.01 |

As in previous experiments with tagging Croatian texts, adjectives are shown to be the most difficult of Croatian parts of speech, followed by pronouns and nouns. As with the previous table, notable accuracy increases can be seen between subset3 and subset4 and also subset4 and subset5 on all three parts of speech. Consulting the descriptions of reductions in table 1, it is clear that the first increase occurs when these parts of speech are stripped of the category of case, shown in (Agić et al. 2009) to be the most difficultly tagged category in Croatian. The other increase occurs when subset5 virtually becomes a part-of-speech-only tagset, removing information on gender and number and keeping only the type of nouns.

## Conclusions and future work

Using the CroTag stochastic morphosyntactic tagger and the Croatia Weekly 100 kw manually tagged corpus of Croatian, this experiment has shown how tagset design or, more specifically, tagset reductions influence the accuracy of morphosyntactic tagging of Croatian texts. Its results may be used in other, more elaborate sub-tagset designs based on the Multext-East version 3 tagset specifications, with respect to overall goals of the resulting system and the requirements of the end user.

## Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776 and 130-1300646-0645.

## References

Agić, Željko; Tadić, Marko. Evaluating Morphosyntactic Tagging of Croatian Texts // *Proceedings of the 5th International Conference on Language Resources and Evaluation* / ELRA, Genoa-Paris, 2006.

Agić, Željko; Tadić, Marko; Dovedan, Zdravko. Investigating Language Independence in HMM PoS/MSD-Tagging // *Proceedings of the 30th International Conference on Information Technology Interfaces* / Lužar-Stiffler, Vesna; Hljuz Dobrić, Vesna; Bekić, Zoran (ed.). Zagreb, SRCE University Computer Centre, University of Zagreb, 2008. pp 657-662.

Agić, Željko; Tadić, Marko; Dovedan, Zdravko. Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. // *Informatica*. 32 (2008), 4; pp. 445-451.

Agić, Željko; Tadić, Marko; Dovedan, Zdravko. Error Analysis in Croatian Morphosyntactic Tagging // *Proceedings of the 31st International Conference on Information Technology Interfaces* / Lužar-Stiffler, Vesna; Jarec, Iva; Bekić, Zoran (ed.). Zagreb, SRCE University Computer Centre, University of Zagreb, 2009. pp. 521-526.

Bekavac, Božo; Tadić, Marko. Implementation of Croatian NERC system // *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007, Special Theme: Information Extraction and Enabling Technologies* / Piskorski, Jakub; Tanev, Hristo; Pouliquen, Bruno; Steinberger, Ralf (ed.). Prague, ACL, 2007. pp. 11-18.

Böhmová, A.; Hajič, J.; Hajičova, E.; Hladká, B. The Prague Dependency Treebank: A Three-Level Annotation Scenario. // *Treebanks: Building and Using Parsed Corpora* / A. Abeillé (ed.), Kluwer, 2003, pp. 103-127.

Brants, Thorsten. Tagset Reduction Without Information Loss. // *Proceedings of ACL-95 student session* / Association for Computational Linguistics, 1995. pp. 287-289.

Brants, Thorsten. TnT - a statistical part-of-speech tagger // *Proceedings of ANLP 2000*.

Ceaușu, Alexandru. Maximum Entropy Tiered Tagging // *Proceedings of the 11th ESSLLI Student Session* / Janneke Huitink and Sophia Katrenko (ed.), June 20, 2006, Malaga, Spain, pp. 173-179.

Erjavec, Tomaž. Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora // *Proceedings of the Fourth International Conference on Language Resources and Evaluation* / ELRA, Lisbon-Paris 2004, pp. 1535-1538.

Giménez, J.; Márquez, L. SVMTool: A general POS tagger generator based on Support Vector Machines // *Proceedings of the 4th International Conference on Language Resources and Evaluation* / Lisbon, Portugal, 2004.

Halácsy, P., Kornai, A., Oravecz, C., Trón, V., Varga, D. Using a morphological analyzer in high precision POS tagging of Hungarian. // *Proceedings of 5th Conference on Language Resources and Evaluation* / ELRA, 2006, pp. 2245-2248.

Halácsy, P., Kornai, A., Oravecz, C. HunPos - an open source trigram tagger // *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* / Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 209-212.

Rupnik, Jan; Grčar, Miha; Erjavec, Tomaž. Improving morphosyntactic tagging of Slovene by tagger combination. // *Proceedings of the Slovenian KDD conference – SiKDD 2008.* / Ljubljana, Slovenia, 2008.

Shen, L.; Satta, G.; Joshi, A. Guided learning for bidirectional sequence classification. // *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* / Prague, Czech Republic, 2007. pp. 760-767.

Sjöbergh, J. Combining POS-taggers for improved accuracy on Swedish text // *NoDaLiDa 2003, 14th Nordic Conference on Computational Linguistics*. / Reykjavik, 2003.

Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech // *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*. / Prague, Czech Republic. Association for Computational Linguistics, 2007.

Toutanova, K.; Klein, D.; Manning, C.D.; Yoram Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network // *Proceedings of HLT-NAACL 2003* / pp. 252-259.

Tadić, Marko. Building the Croatian-English Parallel Corpus // *Proceedings of the Second International Conference on Language Resources and Evaluation* / ELRA, Paris-Athens 2000, pp. 523-530.

Tadić, Marko. Building the Croatian National Corpus. // *Proceedings of LREC 2002* / ELRA, Pariz-Las Palmas 2002, Vol. II, str. 441-446.

Tadić, Marko; Fulgosi, Sanja. Building the Croatian Morphological Lexicon // *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages* / Budapest, ACL, 2003. pp. 41-46.

Tadić, Marko. The Croatian Lemmatization Server // *Southern Journal of Linguistics*. 29 (2005), 1/2; pp. 206-217.

Tadić, Marko. Developing the Croatian National Corpus and Beyond // *Contributions to the Science of Text and Language. Word Length Studies and Related Issues* / Grzybek Peter (ur.), Kluwer, Dordrecht 2006, str. 295-300.

Tufiş, Dan. Tiered Tagging and Combined Classifiers // *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692* / F. Jelinek, E. Nöth (eds.), Springer, 1999, pp. 28-33.

Tufiş, Dan, Dragomirescu, Liviu. Tiered Tagging Revisited. // *Proceedings of the 4th LREC Conference* / Lisbon, Portugal, 2004, pp. 39-42.

Original scientific paper

# An Optimization of Command History Search

Vladimir Mateljan
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
vladimir.mateljan@gmail.com

Krunoslav Peter
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
kruno_peter@yahoo.com

Vedran Juričić
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
vedran.juricic@gmail.com

**Summary**

*A command that a user issues at the command prompt is a string encoded with certain character encoding. This character string can be stored into an appropriate data structure in order to be documented or reused. Additional functionality of the command line to store entered commands, along with ability to list, edit and re-execute previously entered commands, is called command history. The paper suggests an optimization of command history search based on proposed grammar of the command language.*

**Key words:** command, command line, formal language, string, command history, grammar, algorithm, search, optimization, grep, awk

## Introduction

This paper focuses on commands as words of a formal language and the search for those commands in command history. A grammar of the command language is proposed; this grammar defines the command as a word of a formal language that consists of singleton words named "single-words," separated by spaces. A general search algorithm reads commands from command history and matches commands that contain search string. Four examples of command history implementations, in various operating systems, follow the description of the command language. There are also two examples of command history search by using software tools. The problem is how to find only commands that contain single-words equal to search string. An optimization of command history search is based on abstraction "a command as a sequence of single-words, separated by

spaces." Modified general search algorithm reads command from command history, parses the command string to divide it up into single-words and then matches any single-word to find one that is equal to search string.

## The command line

A *command line*, as a way of interacting with software system, is provided by operating systems shells, programming languages that have interactive mode, query processors of the database management systems, or other interactive software systems. The command *prompt* is a visual component of the command line[1]; it tells the user that he can enter a command. A blinking *cursor* following the command prompt indicates the place where next entered *character* will appear (Example 1).

The command line has the *syntax* and *semantics*. In short, the syntax defines rules for writing commands; the semantics defines what commands do. A *command interpreter* parses a command issued by the user and coordinates what happens between the user and the software system (Peek et al, 1998, 5).

Example 1: The command line with the prompt ('**#**'), the cursor ('**_**'), and the command *ls*[2] ('`ls -al`')

```
# ls -al_
```

An additional functionality of command line called the *command history* is to keep in appropriate data structure a history of executed commands along with ability to list, edit and re-execute previously entered commands.

## The command as a word of a formal language

The command is the *word* of a formal language; it is a *string* of characters (Klint, 1985, 5) over an *alphabet* that a user types in the command line. Let $A$ be a finite set called the alphabet (Abiteboul, 1995, 13); the command $c$ over an alphabet $A$ is a finite sequence $a_1...a_n$, where $a_i \in A$, $1 \leq i \leq n$, $n \geq 0$. Its length is $n$. Empty string is denoted by $\varepsilon$ and its length is 0 (Dovedan, 2003, 15). The alphabet is implemented as the *character encoding* in software system and the string is a finite sequence of characters.

[1] The command line can also be implemented as an input field in the graphic user interface.

[2] The *ls* command is well known command of the operating system Unix that lists the content of a directory; option '*–a'* is used to list the content for a directory and '*–l'* to create a vertical list of a directory's content (Toporek, 2003, 87).

When the user enters a command, he performs the *concatenation* of characters (or strings). "If $x$ and $y$ are strings, then the concatenation of $x$ and $y$, written $xy$, is the string formed by appending $y$ to $x$ (Aho et al, 1986, 92). The concatenation and substring selection are the most primitive operations on strings (Klint, 1985, 5).

If $x$, $y$, and $z$ are words over alphabet $A$, a concatenation of this words is a word $xyz$. Strings $x$, $y$, $z$, $xy$, and $yz$ are *substrings* of string $xyz$. Concept of the substring is important for the focus of this paper, because the command history can be searched for commands that contain a certain substring.

The set of all words over alphabet $A$ is denoted by $A^*$ (Abiteboul, 1995, 13). The *formal language* $L$ is a subset of $A^*$ ($L \subseteq A^*$). A *command language* is also a formal language over an alphabet.

The *grammar* is a formal system for generating words of a formal language (Dovedan, 2003, 19). A *context-free grammar* $G_c$ that defines the command language $L(G_c)$, is a 4-tuple $G_c(N, A, S, P)$, where:

- $N$ is a finite set of *non-terminal* symbols *{Z, W}*;
- $A$ is a finite set of *terminal symbols* (characters), disjoint from $N$, an alphabet that contains lowercase letters *{a, b, c, d,...}*, numbers *{0, 1, 2, 3,...}*, space *{' '}*, and other symbols *{'.', '|', '>', ...}*;
- $S$ is a *start symbol*, a distinguished symbol from $N$;
- $P$ is a finite set of *productions* of the form $\alpha \rightarrow \beta$, called Backus-Naur form (BNF) (Dovedan, 2003, 35), where is $\alpha \in N$ and $\beta \in (N \cup T)^*$:

$S \rightarrow \varepsilon$
$Z \rightarrow$ ' ' *(space)*
$W \rightarrow$ *{a | b | c | ... | 0 | 1 | 2 | ... | .| > | $ | * | ...}*
$S \rightarrow \varepsilon \mid W \mid W$ *[{{Z}W}]*

Non-terminal symbol $Z$ represents white space (blank). $W$ stands for *single-word* – a sequence of characters without any spaces. *{Z}W* is the concatenation of one or more spaces and a single-word.

Presented grammar $G_c$ of the command language $L(G_c)$ defines the language consisting of all words in $A^*$ that can be derived from the start symbol $S$ by repeated applications of the productions; a word of $L(G_c)$ can be:

- an empty string ($\varepsilon$),
- a single-word ($W$),
- or a string of single-words separated by one or more spaces ($W$ *[{{Z}W}]*).

Words of the command language $L(G_c)$ correspond to the commands of the operating systems *Unix*, *DOS*[3], *Microsoft Windows XP* and *Vista* (Example 2). Operating system's commands can be simple, single-word entries or more complex. The general syntax for commands is (Mateljan et al, 2007, 448):

*command [{option}] [{filename}]*

So, by using proposed grammar $G_c$, commands can be generated as words of a formal language where single-word is:

- a command (without any arguments, options, piping, and redirection),
- an optional argument or of the command (e. g. a filename (without any spaces), a name of the variable, or a name of the constant),
- an option of the command (e. g. '`/p`' or '`-l`'),
- an operator (e. g. redirection operator '>' or adding operator '+').

Example 2: Commands *dir* and *more*[4] of the operating systems DOS and Microsoft Windows XP/Vista

```
dir
dir *.txt
dir | more
dir *.txt > dir.txt
```

Words of $L(G_c)$ also conform to the commands of the programming language *Tcl*[5] (Example 3). Tcl has interactive mode that gives a user an ability to learn individual Tcl commands. A *Tcl script* consists of one or more commands. Each command consists of one or more single-words, where the first single-word is the name of the command and additional words are arguments to that command. Single-words are separated by spaces or tabs (Ousterhout, 1994, 29). So, the general syntax for Tcl commands is:

*command [{argument}]*

[3] *DOS* (short for Disk Operating System) is an operating system copyrighted by Microsoft in 1979 and initially written by Tim Paterson. There are several similar products produced by other companies. For example, FreeDOS is a complete, free, MS-DOS compatible operating system. The command interpreter of Windows XP/Vista, *cmd.exe*, maintains most of DOS commands (Wikipedia, 2009).

[4] Well known command *dir* lists the content of a directory; the command *more* displays the output one screen at a time.

[5] *Tcl* (originally from "Tool Command Language") is a scripting language created in the spring of 1988 by John Ousterhout (Wikipedia, 2009).

Example 3: Commands *set* and *expr*[6] of the programming language Tcl

```
set preset1 2
set preset2 3
expr $preset1 + $preset2
```

To search a command history for certain search string, a general search algorithm can read commands from command history and match commands that contain a search string. If a user wants to find only commands that contain single-words equal to search string, he should avoid matching entire command, because the search string can be a substring of the single-word. For example the search string '*set*' is the substring of the variable name '*preset1*'.
Regardless of concrete command language and its lexical structure[7], spaces are inserted between single-words in the command in majority of languages. Hence, in this paper the command is considered as a sequence of single-words separated by spaces. This *abstraction* – "a command as a sequence of single-words separated by spaces, where single-word is a sequence of characters without any spaces" – is a key for the optimization of command history search. So, an *optimization* of the command history search is to find only those commands that contain single-words equal to search string. Instead of matching entire command, the search algorithm parses the command and matches any single-word to find one that is equal to search string. This search can be performed without a need for knowing the semantics of a command language.

## Command history

A command that a user executes at the prompt can be stored as a string into an appropriate data structure. It could be a sequential file. This paper doesn't concern itself with the data structures. It is only important that stored commands are linearly ordered by creation in command history and that they could be read one by one.
Even the commands that have syntax errors should be stored to the command history, because the user can easy recall these commands from command history, edit and then execute.
When the commands are stored in the command history, they can be reused in many ways:

- recalling and executing a previous command;
- recalling and editing a previous command in case of syntax error;

[6] The *set* command is used to write and read variables. The *expr* command evaluates an expression; it treats all of its arguments together as an arithmetic expression (Tcl Developer Xchange, 2008).

[7] "Grammars are capable of describing most, but not all, of the syntax of programming languages. A limited amount of syntax analysis is done by a lexical analyzer." (Aho et al, 1986, 172)

- copying a command from command history to another context (batch script or document);
- a statistical analysis (for example, counting the most frequently used commands in order to be replaced by aliases).

In this part of the paper is an overview of four implementations of the command history in various operating systems (Table 1). The shells in the operating systems DOS and Microsoft Windows XP/Vista store the command history in main memory, while the shell of *Apple Mac OS X* "keeps track of recently entered commands" (Toporek, 2003, 74) in the text file *.tcsh_history* that is located in home directory of the user. The *Bash*[8] shell in Unix also stores the command history in the text file called *.bash_history* (Cameron & Rosenblatt, 1998, 30). To recall previously entered command, the user can use "Up Arrow" key (⇧). When the user issues the command '`doskey /h`' in DOS or Windows XP/Vista operating system, he will see the content of the command history. In Mac OS X and Unix, the command for listing the command history is '`history`'.

Table 1: Implementations of the command history in various operating systems

| Operating system | DOS | Windows XP/Vista | Mac OS X | Unix (with Bash shell) |
|---|---|---|---|---|
| Storage | main memory | main memory | file *.tcsh_history* | file *.bash_history* |
| Recalling previous command | ⇧ | ⇧ | ⇧ | ⇧ |
| Command for the list of commands | `doskey /h` | `doskey /h` | `history` | `history` |

## Command history search

Let $c_1...c_n$, $c_i \in L(G_c)$, $1 \leq i \leq n$ be a sequence of commands in appropriate data structure. The simplest way to find commands that contain search string $x$ is to get commands one by one and put matched commands in output data structure (the pseudo-code of the Algorithm 1):

Algorithm 1: Command history search – matching entire command

*Input: a sequence of commands $c_1...c_n$, $c_i \in L(G_c)$, $1 \leq i \leq n$; search string $x$*
*Output: a sequence of commands $c_x...c_y$, where $c_x$ is from $c_1...c_n$,..., and $c_y$ is from $c_1...c_n$*

[8] *Bash* (short for Bourne Again Shell) is the command interpreter of the operating system Unix (Cameron & Rosenblatt, 1998).

*n = "command count";*
*i = 1;*
*while (i <= n) do*
*get $c_i$ from $c_1...c_n$;*
*if "x is a substring of $c_i$" then put $c_i$ in $c_x...c_y$;*
*i = i + 1;*
*end;*

To search the command history in the operating system Unix, a user can use the software tool called *grep*[9]. Example 4 demonstrates command history search by using the utility *grep*.

Example 4: Command history search for substring '*less*' by using *grep*

Typical content of the text file *.bash_history*:

```
ls | less
ls *.txt
ls *.txt > lstxt.txt
less lstxt.txt
rm lstxt.txt
```

If the user wants to find all the commands that contain single-word '*less*', he (or she) should issue this command:

```
# grep less .bash_history
```

The result of command execution will be:

```
ls | less
less lstxt.txt
```

What if the user wants to search for commands that contain the single-word '*ls*'? If he executes the command '`grep ls .bash_history`', the result will be – entire content of the file *.bash_history*. Proposed optimization of command history search, based on abstraction "a command as a sequence of single-words separated by spaces," solves this problem.

[9] The *grep* utility, originally written for Unix, searches a text file for lines that have a certain text pattern (Peek et al, 1998, 71), formally called *regular expressions* (Robbins, 2000, 2). Regular expressions are not covered in this paper. The *find* command in DOS and *findstr* in Windows XP/Vista are similar to *grep*. There are also *grep* executable files for DOS and Windows XP/Vista operating systems.

## Optimization of command history search

A better way to search commands that match the string *x* is to get commands one by one, parse the command string to divide it up into single-words, and then match any single-word to find one that is equal to search string *x* (Algorithm 2):

Algorithm 2: Command history search – dividing command up into single-words and matching any single-word

*Input: a sequence of commands* $c_1...c_n$, $c_i \in L(G_c)$, $1 \leq i \leq n$; *search string x*
*Output: a sequence of commands* $c_x...c_y$, *where* $c_x$ *is from* $c_1...c_n$,..., *and* $c_y$ *is from* $c_1...c_n$

```
n = "command count";
i = 1;
while (i <= n) do
   get c_i from c_1...c_n;
   founded = 0;
   for each "single-word in c_i" do
      if "x is equal to the single-word" then founded = 1;
   end;
   if (founded = 1) then put c_i in c_x...c_y;
   i = i + 1;
end;
```

In Unix, a user can perform command history search by parsing a command into single-words by using programming language *awk*[10] (Example 5). Each input line (that is a command) of the command history, *awk* will divide into fields (single-words) by "white" space (spaces or tabs); by default, a space is a field separator. "Fields are referred to by the variables *$1*, *$2*, …, *$n*" (Robbins, 2000, 25).

Example 5: Command history search for single-word '*ls*' by using *awk*

The content of Unix file *.bash_history* is shown in Example 4.

The *awk* command, that finds single-word '*ls*' in any command in the file *.bash_history*, is specified on the command line:

[10] Programming language *awk* is designed for processing text-based data, either in files or data streams. It was created at Bell Labs in 1977. The name *awk* is derived from the family names of its authors — Alfred Aho, Peter Weinberger, and Brian Kernighan (Wikipedia, 2009).

```
# awk '{ for (i = 1; i <= NF; i ++) { if ($i == "ls") print
$0 } }' .bash_history
```

Previous command is wrapped-around, because it's too long to be displayed in one row.
This *awk* command corresponds to the statement '*for each "single-word in $c_i$" do (...)*' of the Algorithm 2. A built-in variable *NF* in the *for*-loop stores the number of fields in current line (Robbins, 2000, 29) or single-words in current command in case of the command language. The *$i* variable represents *i*th field[11] in current line (ibid) or single-word in current command. The *print* command (Mateljan et al, 2007, 449) prints out the commands from command history that contain a single-word equal to search string:

```
ls | less
ls *.txt
ls *.txt > lstxt.txt
```

Previously *awk* command can be written inside a script to avoid retyping (Example 6).

Example 6 – A batch script for command history search in Windows XP/Vista

There are *awk* executables for the operating systems DOS and Windows XP/Vista. So, the output of the command '`doskey /h`' can be piped to the *awk* command. First line of the batch script, called e.g. *chs.bat*, can be the '`@echo off`' command; it turns off the command-echoing.
The *type* command shows the content of batch script *chs.bat* that is located in the root directory of the disk *C*:

```
C:\> type chs.bat
@echo off
doskey /h | awk "{ for (i = 1; i <= NF; i++) { if ($i ==
\"%1\") print $0 }}"
```

Although the second command in the script is wrapped-around in this paper (or on the screen), it is one line in the file *chs.bat*.

[11] As mentioned, *awk* is a language for processing files of text (Robbins, 2000, 23). "A file is treated as a sequence of records, and by default each line is a record. Each line is broken up into a sequence of fields, so we can think of the first word in a line as the first field, the second word as the second field, and so on. An *awk* program is of a sequence of pattern-action statements; *awk* reads the input a line at a time" (Wikipedia, 2009).

The script *chs.bat* has one argument ('`%1`') – a search string. The syntax of command *chs* is:

*chs search_string*

To find the *more* command in the command history, the user simply types '`chs more`' at the prompt.

## Conclusion

An optimization of command history search based on abstraction "a command as a sequence of single-words separated by spaces," allows a user to find only those commands that contain single-words that exactly match the search string, regardless of concrete command language.

## References

Abiteboul, Serge; Hull, Richard; Vianu, Victor. Foundation of Databases. Reading : Addison-Wesley, 1995

Aho, Alfred; Sethi, Ravi; Ullman, Jeffrey. Compilers, Principles, Techniques, and Tools. Reading : Addison-Wesley, 1986

Newham, Cameron; Rosenblatt Bill. Learning the Bash Shell. Sebastopol : O'Reilly, 1998

Dovedan, Zdravko. Formalni jezici: sintaksna analiza. Zagreb : Zavod za informacijske studije Odsjeka za informacijske znanosti Filozofskog fakulteta, 2003

Klint, Paul. A Study in String Processing Languages. Berlin Heidelberg : Springer Verlag, 1985.

Mateljan, Vladimir; Požgaj, Željka; Peter Krunoslav. Značaj skriptnih jezika za administraciju operacijskih sustava. // INFuture2007: Digital Information and Heritage / Zagreb : Odsjek za informacijske znanosti Filozofskog fakulteta, 2007, 445-456

Ousterhout, John. Tcl and the Tk Toolkit. Reading : Addison-Wesley, 1994

Peek, Jerry; Tondino, Grace; Strang, John. Learning the UNIX Operating System, 4th Edition. Sebastopol : O'Reilly, 1998

Robbins, Arnold. sed & awk Pocket Reference. Sebastopol : O'Reilly, 2000

Tcl Developer Xchange. Language. 20 October 2008. http://www.tcl.tk/about/language.html (20 August 2009)

Toporek, Chuck. Mac OS X Pocket Guide, 2nd Edition. Sebastopol : O'Reilly, 2003

Wikipedia. awk. 20 August 2009. http://en.wikipedia.org/wiki/AWK (20 August 2009)

Wikipedia. MS-DOS. 5 August 2009. http://en.wikipedia.org/wiki/MS-DOS (20 August 2009)

Wikipedia. Tcl. 13 August 2009. http://en.wikipedia.org/wiki/Tcl (20 August 2009)

Original scientific paper

# Automatic Diacritics Restoration in Croatian Texts

Nikola Šantić
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
nikola.santic@fer.hr

Jan Šnajder
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
jan.snajder@fer.hr

Bojana Dalbelo Bašić
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, 10000 Zagreb, Croatia
bojana.dalbelo@fer.hr

**Summary**

*The absence of diacritics in digitally encoded text is a common problem for languages whose writing systems are not covered by the standard ASCII character set. It is a deterioration of language in its own right, but also a serious impediment to automated text processing and information retrieval. Restoration of diacritics, if performed manually, is a tedious and time-consuming process. In this paper we describe a robust system for automatic diacritics restoration in Croatian texts. The system combines dictionary look-up and statistical language modelling. Diacritics restoration is evaluated on a corpus of newspaper articles and discussion forum posts. Our experiments show that high levels of accuracy can be achieved with fairly simple and computationally inexpensive methods.*

**Key words:** natural language processing, diacritics restoration, statistical language model, Croatian language

## 1. Introduction

Verification and correction of spelling errors is probably one of the most used applications of natural language processing. The most common types of errors are orthographic and typing ones. In addition to these, there is another category of spellchecking, which, although not characteristic for English, is relevant to most other European languages – restoration of diacritics. Diacritics can be

found in, amongst others, Spanish, Portuguese, French, German, and most of the Scandinavian and Slavic languages. Automated restoration of diacritics is useful not only for the restoration of legacy texts that were typeset without diacritics, but also for ever-larger amounts of contemporary content in which the diacritics are absent. The main reason for this is the lack of an accepted encoding standard, so the users find it simpler to leave out the diacritics while typing. This phenomenon is especially common in casual forms of electronic communications such as e-mail, discussion forum posts, and chats. In fact, most texts found on Internet blogs and discussion forums are in "mixed" form, partly with diacritics and partly without. Another common cause of diacritics' absence is conversion to formats of non-compatible encodings, for instance, when text is extracted from a PDF document. Although a minor problem for a human reader, the absence of diacritics is a major setback for automated text processing and information retrieval.

The problem of diacritics restoration, although specific to each language, has given rise to two basic approaches: *word-based* and *character-based* restoration (Mihalcea, 2002). Word-based approaches are commonly knowledge-intensive systems that rely on dictionaries and statistical language models, which makes these systems language dependant. They require a large source of grammatically correct text in order to build a useful model, and need more processing time because of the pre-processing (e.g., tokenization, tagging, etc.). On the other hand, character-based systems rely on language-independent algorithms that use statistical information gained from the training data. They are much simpler, faster, easier to implement, and do not require any language-specific resources. However, in languages where the change of diacritics has a grammatical or semantic role, word-based systems are much more reliable (Tufiş, 2007). Only in languages where the diacritics can be restored without examining the context can the character-based systems be expected to yield high accuracy. Therefore, when choosing an approach, different aspects should be considered: the role of diacritics in the language, availability of adequate training data, required processing speed, and users' requests and needs.

In this paper, we describe and evaluate a word-based diacritics restoration system designed for Croatian language. The system is based on word recognition and selection using a dictionary and a statistical language model. To the best of our knowledge, this is the first work that directly addresses the issue of diacritics restoration for Croatian language.

The rest of the paper is structured as follows. A brief overview of related work is given in the next section. Section 3 discusses the peculiarities of diacritics restoration for Croatian language. In Section 4 we describe our diacritics restoration system, while in Section 5 we present the experimental evaluation. Section 6 concludes the paper.

## 2. Related work

Due to orthographic differences among languages, it is impossible to define a universal solution to diacritics restoration. Still, some generic ideas can be adjusted or refined to fulfil the requirements of a particular language.

Spriet and El-Bèze (1997) use part-of-speech tagging to restore accents in French. Their method tags each word with its type and its relation to other words, based on the context and statistical n-gram information. Unknown words are ignored, since the majority of unknown words need not be restored anyway. They evaluate the method on a 19,000 word corpus, obtaining a rather satisfactory accuracy of 99.31%.

Pauw et al. (2007) have shown that for resource-scarce languages, character-based approaches are more successful than word-based approaches. Using machine learning methods, they managed to restore diacritics of various African languages with an accuracy of up to 70%.

Mihalcea (2002) presents a method based on a genetic algorithm applied on the letter level. In the restoration of a Romanian electronic dictionary, she reports letter-level accuracies of over 99%.

DIAC$^+$ (Tufiş, 2007) is an advanced diacritics restoration tool – originally developed for Romanian language – that uses both word and character based approaches. DIAC$^+$ uses a tagging system similar to that of El-Bèze et al. (1994) and three dictionaries: a dictionary of words with diacritics, a dictionary of words with diacritics stripped off, and a list of words currently being processed but not present in either of the other two dictionaries. Restoration candidates are first chosen from one of the three dictionaries and then morphosyntactically tagged. In case of ambiguity, either the user is queried or the restoration proceeds automatically according to chosen probabilistic parameters. For unknown words, a character-based n-gram model is used. Evaluated on a 118,000 words corpus, DIAC$^+$ achieves accuracy of almost 99% on pre-tagged text and 97% when the text is untagged.

The only published work on diacritics restoration related to Croatian language is (Scheyt et al., 1998). The method described there was developed as a part of the *Serbo-Croatian dictation and broadcast news speech recogniser*. Diacritics restoration was required since their absence from most Internet sites prevented further data processing. The proposed method marks as correct all words found in the dictionary, ignoring possible ambiguity. Unrecognised words are assigned to the nearest neighbour based on string similarity. In the last step, a letter trigram model is generated and used to score likelihood of the different possible character sequences for the remaining words. Accuracy of 95% is reported, a fairly high result considering the simplicity of the procedure.

All the above mentioned examples show that good results can be achieved even with the most rudimentary approaches, but also that the properties of the language still play a major role.

## 3. Diacritics in Croatian

There are five diacritical characters in Croatian: *č*, *ć*, *š*, *ž* and *đ* (a diacritic is also contained in the digraph *dž*). As most other diacritical characters, none of these are present in the standard ASCII character set, the first widely used encoding standard. First code page expansions to start including diacritics were developed during the 1980s – the ISO 8859 family. Croatian language was supported there within both the Central European language family (Latin-2 encoding) and the Southeast European family (Latin-10 encoding). Different operating systems also developed their own encoding standards, and the Croatian diacritical characters were included both in windows-1250 and in Macintosh Southeast code pages. Unfortunately, all the available encodings were mutually incompatible. First unifying standards came with the advent of Internet. Today, the two most prevalent are UCS and Unicode. Both use up to 32 bits per character, which allows the encoding of all characters of all known languages. A lot of effort has been made to synchronize and translate one standard into the other. Despite this, a unified standard still does not exist – for example, the three most popular Croatian news portals still use different encodings: UTF-8, windows-1250, and ISO-8859-2.

### Substitution of diacritics

Although the above-mentioned encoding standards represent an improvement from the times when diacritics were not supported at all, many users still carry a habit of omitting the so-called "Croatian characters", and substituting them with conventionalised variants. Table 1 contains some of the most common substitutes for each diacritical character.

Table 1: Most common diacritic substitutions

| Diacritical character | Substitute characters |
|---|---|
| č, ć | c, cc, ch, cx, cy |
| š | s, ss, sh, sx, sy |
| ž | z, zz, zh, zx, zy |
| đ | d, dj, dy |

None of the above-mentioned substitutions are flawless. The most frequent substitution scheme is the omission of diacritics (*č* and *ć* become *c*, *š* becomes *s*, *ž* becomes *z*, and *đ* becomes *d*). This scheme often causes ambiguity, e.g., between *kuca* (knocks) and *kuća* (house), *kos* (blackbird) and *koš* (basket), *zao* (evil) and *žao* (sorry), *voda* (water) and *vođa* (leader). Fortunately, most words have only one or two valid diacritical forms, while three valid diacritical forms are extremely rare, e.g., *obuci* (wear), *obući* (train), and *obući* (to footwear). Switching to a substitution scheme that adds an extra character eliminates this ambiguity, but it also makes words less readable, e.g., *rječca* (small word) becomes *rjeccca*, *čišći* (cleaner) becomes *cxisxcxi,* etc.

In what follows, words that contain a potential substitution for a diacritic (characters *c, s, z,* or *d*) will be referred to as *C-words*. C-words are words that are to be considered for restoration (this is because all substitutions schemes use one of these four letters as the substitute character). Words that contain a diacritical character will be referred to as *D-words*. Note that a word can be both a C-word and a D-word, e.g., *cvijeće* (flowers). A *C-word variant* is obtained by applying (inverse) substitutions on a given C-word; a complete set of C-word variants of a given C-word is obtained by permuting all possible substitutions. For example, C-word variants of *staza* are *staza, staža, štaza,* and *štaža,* of which only the first two are valid words.

**Statistical analysis**

Difficulty in selecting a correct word form depends greatly on the properties of the language itself. Table 2 shows statistical data for Croatian language gathered from three different samples: newspaper articles with correct diacritics (*valid*), discussion forum posts with both diacritics and substitutions (*mixed*), and newspaper articles and discussion forum posts from which the diacritics were removed (*removed*). The evaluated texts were of various sizes, ranging from 10,000 to 100,000 word tokens. The C-word and D-word ratios are calculated on the word level, while for substitute characters and diacritics ratios this is done on the letter level.

Table 2: Statistical analysis of diacritics on three samples

| | **Samples** | | |
|---|---|---|---|
| | **Valid** | **Mixed** | **Removed** |
| **C-word ratio** | 45.7% | 48.5% | 53.8% |
| **D-word ratio** | 16.3% | 10.1% | – |
| **Substitute characters ratio** | 10.5% | 12.2% | 14.1% |
| **Diacritics ratio** | 3.2% | 1.3% | – |

The C-word ratio on valid text sample suggests that more than half the words can be marked off as correct without any further processing. Also, the D-word ratio on the same sample suggests that only every sixth word actually contains a diacritic, which means that a text in which the diacritics are absent is already more than 80% correct. As expected, compared to the diacritically valid text, in forum texts (*mixed*) the proportion of C-words is higher whereas the proportion of D-words is lower.

Another interesting parameter of a C-word is the number of its valid C-word variants. C-word variants are considered valid if they are present in the dictionary. Based on a sample of texts containing up to 100,000 words with diacritics removed, the following results were obtained: 88.6% of the C-words have a single valid variant, 7.4% have two, and only 0.1% have three valid variants. The rest (3.9%) are words not present in the dictionary, either misspelled or un-

common. These results imply that most of the C-words can be unambiguously restored using a wide-coverage dictionary. Still, for the process to be fully automated, a method to restore the ambiguous cases is required.

## 4. System description

The diacritics restoration system described in this work uses a dictionary backed up by a statistical language model. The restoration procedure is divided into three successive steps: tokenization, candidates generation, and the selection of the correct word form. The architecture of the system is shown in Figure 1.

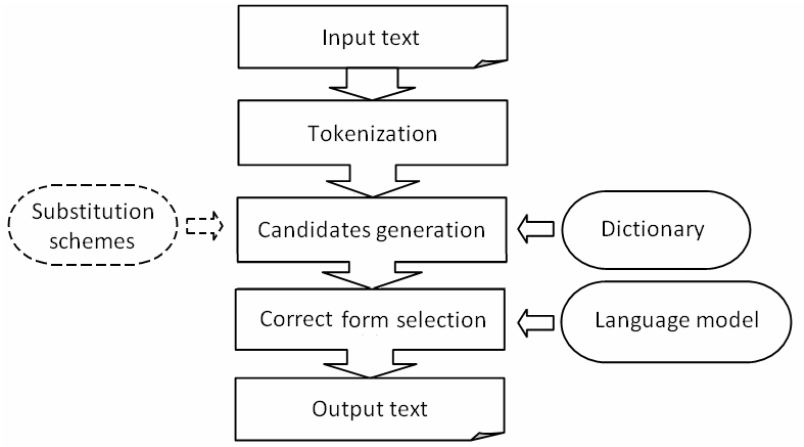



Figure 1: Architecture of the diacritics restoration system

**Tokenization**

Diacritics restoration in this system is word-based, meaning the lines of input text first have to be tokenised into word tokens. The line is split on whitespace and punctuation characters. Word-based diacritics restoration relies on a context of each word, i.e., a sequence of words that surround it. Our system uses a left and a right context – words that precede and follow a given word – with an adjustable size of the context window. The system processes the text one line at a time, even if the lines contain multiple sentences; preliminary evaluation has shown that, even if the context extends beyond the sentence boundaries, the results do not deteriorate in any way.

**Candidates generation**

In this step all C-word variants of a word currently being restored are generated. If the word is not a C-word, it is marked off as correct; as shown in the previous section, this eliminates more than half of the words. For the remaining words all their C-word variants are generated. The number of variants grows exponentially with the number of substitute characters. Since every character can be

substituted by either two or three characters ($s$ : {*s,š*}; $c$ : {*c,č,ć*}), the base of the exponential function takes on values between 2 and 3. In practice, a C-word has on average 1.1 substitution characters in valid texts and 1.2 in texts with removed diacritics. These numbers are small enough to avoid combinatorial explosion.

The input text is processed from left to right, token by token, so the left context of the word will have been processed by the time we create the C-word's variants. However, the right context will not be processed yet, so the C-word variants of each token from the right context need to be created. Ultimately, a Cartesian product of generated variants sets is used; its size increases exponentially with the size of the right context.

Most of the C-word variants generated in this step are not valid words, so they have to be checked against a dictionary. The dictionary used in this system is derived from an automatically acquired inflectional lexicon totaling over 3.5 million words (Šnajder et al. 2008). Since all the words looked up in the dictionary will be C-word variants, the dictionary needs to contain only C-words and D-words; this reduces the size of the dictionary to 750.000 entries.

**Selection of the correct word form**

As shown previously, using dictionary alone will correctly and unambiguously restore the majority of C-words (more than 88% on our sample). We consider this performance as the baseline. The remaining 12% of the C-words have ambiguous variants, e.g., *suma* being either *suma* (a sum) or šuma (a forest). In such cases, the word's context is examined using a language model.

The language model is a statistical model of word sequences (Jurafsky, 2000). It contains data on probabilities of specific word combinations to appear in text. In our case, the input to the language model is a word sequence $w_1^n$ consisting of the C-word and its left and right contexts. The output of the language model is the probability $P(w_1^n)$ of the given sequence. The probabilities of different sequences are compared and the most likely one is chosen as the correct one. For practical reasons (memory consumption and training data availability), in a language model a sequence probability $P(w_1^n)$ is approximated by n-gram probabilities. In this work, we chose to use a bigram language model, thus the probability of word sequence $P(w_1^n)$ is approximated by a product of conditional probabilities:

$$P(w_1^n) = \prod_{k=1}^{n} P(w_k \mid w_1^{k-1}) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1}) \ .$$

In other words, the probability of a word in a sequence is calculated based only on a single word preceding it. It has been shown that this simplification still produces satisfactory results.

The main drawback of the above-described model is in the limited size of the training corpus: no matter how large the corpus, there will always exist bigrams

that are not present in the training corpus but perfectly acceptable in the language. Such bigrams would be assigned a zero probability, and consequently the whole word sequence would be discarded. This problem is commonly addressed by *smoothing*, the technique of assigning non-zero probabilities to "zero probability n-grams". The system described here uses the Witten-Bell discounting method (Witten & Bell, 1991), treating the zero-frequency bigrams as events that did not yet occur. Their probability can be calculated by counting the number of times a new bigram appears in the corpus. This number is, in turn, equal to the number of bigram types – different bigrams that appear in the corpus. Thus, if bigram $w_{n-1}w_n$ is not contained in the corpus, in order to calculate the probability $P(w_{n-1}w_n)$, we use the number of bigram types starting with $w_{n-1}$. Let $T$ be this number, $Z$ the number of zero-probability bigrams, and $N$ the total number of bigrams starting with $w_{n-1}$. The smoothed probability $P^*$ of word $w_i$ conditioned on a preceding word $w_{i-1}$ equals to:

$$P^*(w_i \mid w_{i-1}) = \frac{T(w_{i-1})}{Z(w_{i-1})(N(w_{i-1}) + T(w_{i-1}))}.$$

This extra probability must be discounted from the probability of all the seen bigrams, using the following equation for the total probability mass where $C$ is the total count of the given word sequence:

$$\sum_{i:C(w_x w_i)>0} P^*(w_i \mid w_x) = \frac{C(w_x w_i)}{C(w_x) + T(w_x)}.$$

Smoothing is a relatively simple concept, but one that can significantly improve the results, especially when training corpora is small.

## 5. Evaluation

For the purposes of evaluation, two different corpora were used: a ca. 100,000 word corpus of newspaper articles collected from the Internet site of *Glas Slavonije*[1] and a ca. 30,000 word corpus of discussion forum posts collected from the *forum.hr* web site. Diacritical characters were removed from the newspaper corpus and replaced with substitute characters. The substitution scheme used was the omission of diacritics because this scheme causes the most ambiguity. The forum corpus was not altered. These two corpora were used to evaluate different system configurations by comparing the original corpus to the one automatically restored (for newspaper corpus) or comparing manually and automatically restored text (for forum corpus). The context window extended one word to the left and one word to the right; in our preliminary experiments, this showed to yield the best results. The accuracy of diacritics restoration on the two corpora is shown in Table 3; performance improvement over the dictionary baseline is shown in parentheses.

[1] http://www.glas-slavonije.hr

Table 3: Accuracy of diacritics restoration on different corpora

| | Newspaper articles | Forum posts |
|---|---|---|
| Unrestored | 80.72% | 92.00% |
| Dictionary only (baseline) | 97.07% | 97.95% |
| Dictionary + Language model (unsmoothed) | 97.65% (+ 0.6%) | 98.38% (+0.4%) |
| Dictionary + Language model (WB smoothing) | 98.81% (+1.8%) | 99.35% (+1.4%) |

As shown previously, unrestored texts are already over 80% correct, and using only a dictionary-lookup restores the majority of unambiguous words. The improvement over the dictionary baseline is achieved using the bigram language model built from a 2.5 million word corpus of newspaper articles and literary works. Using the same model with smoothing results in restoration accuracy of up to 99%. In comparison to diacritics restoration systems for other languages, this is a very satisfying accuracy, especially considering the computational simplicity of the method.

The incorrectly restored words are of various types, but in each of the corpora a different type prevails. In the newspaper corpus, most errors are made on location and person named entities. This problem cannot be solved using the dictionary nor the language model due to limited corpus coverage – one possible solution would be for the user to define the corresponding substitutions explicitly. Most mistakes made in the forum corpus are misspellings not recognizable by the dictionary. This could be remedied by adding common spelling mistakes to the dictionary or by approximate matching. This can, however, increase the number of errors on diacritically valid words.

Although the language model does not contribute to the performance as much as the dictionary does, it is nevertheless required to achieve peak performance. This is evident by examining the most frequent restoration errors in the corpus (Table 4). The use of the language model reduces the frequency of these errors by more than 90%. This is a substantial improvement, especially if the text is meant for further automatic processing.

Table 4: Most frequent restoration errors in the newspaper articles corpora

| Ambiguous C-word variants | Error count | |
|---|---|---|
| | Dictionary | Dictionary + Language model (smoothed) |
| posto – pošto | 74 | 6 |
| gdje – gđe | 69 | 2 |
| nas – naš | 47 | 3 |
| grada – građa | 44 | 2 |
| posao – pošao | 30 | 1 |

## 6. Conclusion

Restoration of diacritics is relevant for most European languages. It is especially important in view of the fact that – despite the efforts to define a unifying encoding standard – a significant amount of diacritically invalid text is still being generated on a daily basis. Such texts present a problem for automated text processing and information retrieval.

In this work we have presented a robust word-based system for diacritics restoration in Croatian texts. The system relies on a dictionary and a bigram language model, it does not require any pre-processing, and is computationally inexpensive. Evaluation shows that good results can even be achieved using only dictionary-lookup, while the use of the language model increases restoration accuracy to almost 99%. The described system will soon be available on-line.[2]

For future work, the performance of the system could be further improved by using a user-defined list of unknown words and dynamic loading of corpus-specific language models.

## Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia and under the Grant 036-1300646-1986.

## References

Bèze, M.; Mérialdo, B.; Rozeron, B.; Serouault, A., M. Accentuation automatique de texte par des méthodes probabilistes. // *Technique et sciances informatiques*. 13 (1994), 6; 797–815

De Pauw , G.; Wagacha, P.W.; de Schryver G. Automatic diacritic restoration for resource-scarce languages. // *Lecture Notes in Computer Science*. 4629 (2007); 170–179

Jurafsky, D.; Martin, J.H.; Kehler, A.; Vander Linden, K.; Ward, N. An introduction to natural language processing, computational linguistics, and speech recognition. MIT Press, 2000

Mihalcea, R.F. Diacritics restoration: Learning from letters versus learning from words. // *Lecture notes in computer science*, 2276 (2002); 96–113

Scheytt, P.; Geutner, P.; Waibel, A. Serbo-Croatian LVCSR on the dictation and broadcast news domain. // *IEEE international conference on acoustics, speech and signal processing*, *Volume 2*. 1998, 897-900

Šnajder, J.; Dalbelo-Bašić, B.; Tadić, M.. Automatic acquisition of inflectional lexica for morphological normalisation. // *Information Processing and Management*, 44 (2008), 5; 1720–1731

Tufiş, D.; Ceauşu, A. DIAC$^{+}$: A professional diacritics recovering system. // *6$^{th}$ language resources and evaluation conference*. ELRA, 2007, 167–174

Witten, I. H.; Bell, T. C. Estimating the probabilities of novel events in adaptive text compression. // *IEEE transactions on information theory*. 37 (1991), 4; 1085–1094

[2] http://ktlab.fer.hr/diacro

Original scientific paper

# Evaluation of the Statistical Machine Translation Service for Croatian-English

Marija Brkić
Department of Informatics, University of Rijeka
Omladinska 14, 51000 Rijeka, Croatia
mbrkic@uniri.hr

Tomislav Vičić
Freelance teacher of economics and translator
Zagreb, Croatia
ssimonsays@gmail.com

Sanja Seljan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
sanja.seljan@ffzg.hr

**Summary**

*Much thought has been given in an endeavour to formalize the translation process. As a result, various approaches to MT (machine translation) were taken. With the exception of statistical translation, all approaches require co-operation between language and computer science experts. Most of the models use various hybrid approaches. Statistical translation approach is completely language independent if we disregard the fact that it requires huge parallel corpus that needs to be split into sentences and words. This paper compares and discusses state-of-the-art statistical machine translation (SMT) models and evaluation methods. Results of statistically-based Google Translate tool for Croatian-English translations are presented and multilevel analysis is given. Three different types of texts are manually evaluated and results are analysed by the χ2-test.*

**Key words:** SMT (statistical machine translation), online, Google Translate, MT, Croatian-English, manual evaluation, fluency, adequacy, χ2-test

## Introduction

The translation process is conducted differently by different translators and results are, therefore, not uniform (MT Marathon, 2008). As Knight (2003) points

out, the non-existence of right answers in translation does not imply the non-existence of wrong answers.
MT was first conceived as a technology that significantly speeds up the translation process and offers human-like quality translations (Valderrábanos, 2003). Nowadays, it is seen as a tool of limited use. Current computational models of MT can address a number of non-literary translation tasks, like tasks for which a rough translation is adequate, tasks where a human post-editor is needed and tasks limited to sub-domains in which fully automatic high quality translation is achievable (FAHQT) (Jurafsky & Martin, 2009).
The basic idea behind the development of MT (compared to the human translation) is to find a way for busting up speed while reducing the cost of the translation process (i.e. removing human component as much as possible) (Awatef, 2005). Further development focuses itself on the precision and overall quality of the output.
So far, the MT has been directly associated with (and mostly restricted to) the translation of the written language. This is probably due to the fact that most of contemporary communication (legal, commercial, Internet and so forth) is in written form and still too often on paper.
There are various approaches to MT, such as word-for-word translation, syntactic transfer, interlingual approaches, controlled language, example-based translation, and SMT (MT Marathon, 2008). SMT has low development cost and it is portable across languages (Valderrábanos, 2003). The only requirement SMT imposes is a large parallel corpus.
The paper explores the development of MT. A particular attention is paid to the Google Translate system, which exemplifies SMT. The system is tested for Croatian-English language pair. MT systems need to be evaluated in order to be ranked. For that purpose, different evaluation methods are introduced and the results of a conducted manual evaluation method are given and discussed.

## MT

Although most of the MT approaches integrate different methods (e.g. integration of statistical MT and syntactic transfer, or example-based MT with rule-based method), basic approaches in MT are, according to Hutchins[1] the following: "syntactic transfer", "example-based" and "statistical systems".

### Syntactic transfer

Syntactic transfer approach applies linguistic rules to some extent, analyzing source text and creating translated text accordingly, involving some variety of intermediary linguistic representation, with morphological, syntactic and semantic analysis (Lavie, 2006). Since the 1980s, many new operational MT sys-

[1] Hutchins, J.: *Machine Translation: past, present, future* (Ellis Horwood, UK, 1986)

tems appeared and included this approach: the French multilingual system TITUS; the Chinese-English CULT system; the Spanish-English SPANAM; the Russian-English system Systran which was adopted by the US Air Force and the European Community; the System of Logos Corporation. (Awatef, 2005) In Europe, the Commission of the European Communities (CEC) supported a lot of work on the English-French version of the Systran. In Germany it was SUSY (Saarbrucker Ubersetzungssystem), the French-German System (ASCOF) and (SEMSYN) for the translation of Japanese scientific articles into German. A more ambitious and reputable system developed in this era is the EUROTRA project of the European Communities, which aimed at developing multilingual transfer system for translating among all the Community languages. In the 1980s, according to Hutchins (1992), Japan maintained the greatest commercial activity where most computer companies developed software for computer-aided translation mainly for the Japanese-English market. According to WTEC Hyper Librarian (1994), MT in Japan is viewed as an "important strategic technology that is expected to lay a key role in Japan's increasing participation in the world economy". The most sophisticated commercially available system was METAL, a German-English system, which originated from the research at the University of Texas at Austin and supported by Siemens, which obtained commercial rights for marketing it (Lehmann 2000: 162).

**Example-based translation**

Example-based MT was first suggested by Nagao Makoto in 1984[2]. He suggested the method which may be called *MT by example-guided inference* or MT by the analogy principle. One of the strong reasons for this approach has been that the detailed analysis of a source language sentence is of no use for the translation between languages that have completely different structure (for example, English and Japanese). In this approach, the translation unit is a block of words. This is accomplished by storing *varieties of example sentences in the dictionary* and deploying a mechanism for finding analogical example sentences.

The process of mechanical translation by analogy is time-consuming in its primary structure. Therefore, the process is divided into substages and the system

[2] "Problems inherent in current MT systems are shown to be inherently inconsistent. The present paper defines a model based on a series of human language processing and in particular the use of analogical thinking. Machine translation systems developed so far have a kind of inherent contradiction in themselves. The more detailed a system has become by the additional improvements, the clearer the limitation and the boundary will be for the translation ability. To break through this difficulty we have to think about the mechanism of human translation, and have to build a model based on the fundamental function of language processing in the human brain. The following is an attempt to do this based on the ability of analogy finding in human beings." *in ARTIFICIAL AND HUMAN INTELLIGENCE (A. Elithorn and R. Banerji, editors). Elsevier Science Publishers. B.V., NATO, 1984*

is fed with all the information available in the initial system construction. The learning comes in only during the augmentation stage of the system, which mainly refers to the increase of example sentences and the improvement of the thesaurus (Nagao, 1984). Examples of this approach are translation memories, which are often integrated with language-dependant approach.

**Statistically-based translation**

Further development in MT took place in the 1990s as computers became more powerful and storage capacities much larger and cheaper. The new development shifts from syntactic transfer to what has been called "statistical approaches" with provenance from the "corpus linguistics". Statistical translation systems do not depend on underlying grammatical rules any longer. Statistically-based MT systems rely on statistical models whose parameters are derived from bilingual corpus.

Put very simply, as Farah (2003) put it in an article for the *New York Times* (reprinted in the *International Herald Tribune*), traditional MT relied heavily on bilingual programmers entering the vast wealth of information, needed by the computer, in the lexicon and syntax. A team from IBM in the 1990s tried to make the computer learn the second language by feeding a computer with English text and its translation in a different language, and then analyzing it statistically. The example given by Farah (2003) is revealing:

"Compare two simple phrases in Arabic: "rajl kabir" and "rajl tawil. If a computer knows that the first phrase means "big man" and the second means "tall man," the machine can compare the two and deduce that *rajl* means "man," while *kabir* and *tawil* mean "big" and "tall," respectively." Phrases like these, called N-grams (with "N" representing the number of terms in a given phrase), are the basic building blocks of SMT.

Mackin (2003), in an article interestingly entitled "Romancing the Rosetta Stone," reports on work on translation using statistical approaches. Mackin quotes the computer scientist Franz Joseph Och boasting: "Give me enough parallel data, and you can have a translation system in hours." The new approach for translation uses huge volumes of "matched bilingual texts" which are the encoded equivalents. Och (Makin 2003) asserts that the new approach uses statistical models to find "the *most likely* translation for a given input." The new approach ignores explicit grammatical rules and traditional dictionary lists of the lexicon in order to have the computer itself match up patterns between original texts and translations. Och's work (Makin 2003) is an improvement of the earlier work on the statistical approach that started back in the late 1980s and early 1990s by Peter F. Brown and his colleagues at IBM's Watson Research Center.

Statistical approach to MT tackles the MT problem by finding the maximum likelihood solution (Watanabe & Sumita, 2002). According to Wang and Waibel (1997), SMT systems deal with the following problems:

- the modelling problem (in order to create language and translation models, with problems involving idioms, compounds, morphology and different word order),
- the learning problem (in order to estimate parameters from bilingual corpora), and
- the decoding problem (which essentially comes down to finding an efficient way of searching for a target language sentence).

SMT systems produce a general model of the translation process. Specific rules are acquired automatically from bilingual and monolingual text corpora. Although all of these systems share the same underlying principle, they differ in the structures and sources of their translation models.

In a *word-based approach*, words are treated like tokens, independently from other words. This poor handling of morphology is one of the major drawbacks of this approach. A word-based system may recognize one form of a word, but not the other form of the same word (MT Marathon, 2008). This is particularly apparent with morphologically rich languages, as shall be seen in our study. IBM models 1-5 fall into this category (Koehn, Och, & Marcu, 2003).

Nowadays, many systems implement *phrase-based models*. What differentiates them from word-based models is a lexicon, which is not single-word-based, but phrase-based (Och & Ney, 2004). In addition, phrase length should not exceed three words (Koehn, et al., 2003). Phrase-based models translate small word sequences at a time and do not use explicit syntactic or morphological information (MT Marathon, 2008). Moreover, as Koehn, et al. (2003) report, imposing syntactic restrictions on phrases does not lead to better system performance. The number of useful phrases grows with the size of the training corpora. Log-linear models are variations to a standard model. However, since phrase-based models cannot model grammaticality and long-distance dependencies, they are not suitable for large-scale restructuring of sentences. Furthermore, they cannot generalize.

*Syntax-based models* can be classified according to the underlying syntactic formalism. Representatives of this approach, tree-based models, have proved to have performance comparable to phrase-based models (MT Marathon, 2008).

Bearing in mind that SMT is language-independent and that existent language resources are sparse, moderate results should be expected. According to Sepesy Maucec and Kacic (2007), a hybrid approach, which combines SMT with rule-based MT, would presumably give much better results.

## Evaluation

MT evaluation is not a straightforward task. Different translators translate the very same sentence differently. Evaluation methods can be manual or automatic. Nevertheless, both categories are extremely subjective (Jurafsky & Martin, 2009).

The correlation between two metrics is usually computed using the Pearson correlation coefficient in (1), whereas sample means and variances are expressed in (2) and (3), respectively (MT Marathon, 2008).

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}. \quad (1)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (2)$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2. \quad (3)$$

**Manual evaluation**

Unfortunately, bilingual evaluators, which are best suited for manual evaluation task, are not always available. If that is the case, monolingual target language speaking evaluators are given reference translations and employed for the task. In this case study, two criteria are taken into consideration:

- *fluency,* which refers to grammaticality and word choices (MT Marathon, 2008); according to Jurafsky and Martin (2009), there are three aspects of *fluency – clarity*, *naturalness* and *style*, and
- *adequacy*, which, on the other hand, questions whether any part of a message is lost, added, or distorted; Jurafsky and Martin (2009) group *adequacy* and *informativeness* into another dimension – *fidelity*.

In manual evaluation task, evaluators are asked to score output on a 1-5 scale according to both criteria. It is advisable that evaluators read the output prior to reading the reference translation, because human mind tends to fill in the missing information if reference translation is read first or evaluators are acquainted with the domain. Judgements of *fluency* and *adequacy* are usually related, which either points to the difficulty in distinguishing the two criteria or just to the fact that ungrammatical sentences and wrong word choices carry less meaning (MT Marathon, 2008).

Besides the described procedure for measuring *fluency* and *adequacy*, *fluency* can also be measured through the time needed for reading the translation (Jurafsky & Martin, 2009) or through cloze test (Taylor, 1953, 1957 in Jurafsky & Marin, 2009).

Furthermore, described dimensions can be measured through the edit cost of post-editing the MT output into a satisfying translation. This can be done on word-level, time-level or keystrokes-level (Jurafsky & Martin, 2009).

Hajič, Homola, and Kuboň (2003) present a way of exploiting TM (translation

memory) tools for MT manual evaluation. A TM is created by aligning source text and corresponding MT output. The source text is then translated by a human translator, and with the aid of the newly-built TM. Finally, MT system is used to determine the percentage of similarity between the MT output and the human translation of the same sentence (reference translation), which is stored in the TM.
Evaluation procedure is of crucial importance in comparing different translation models. Manual evaluation methods are too expensive and time-consuming (Papineni et al., 2001). Hence, automatic evaluation methods are needed.

**Automatic evaluation**
All automatic evaluation metrics use one or more reference translations. These reference translations are used for comparison with MT output or candidate translations (MT Marathon, 2008). Automatic method is considered to be better if it has higher degree of correlation with human judgements. There are a number of automatic methods, such as Bilingual Evaluation Understudy (BLEU), NIST, TER, Precision and Recall, and METEOR. Although they differ in the way they measure similarity, they all rank better the candidate translation which is closer to human translation (Jurafsky & Martin, 2009).

## Experimental study

### Google Translate Service

The tool Google Translate is chosen in this case study for two basic reasons: it is statistically-based and only of a kind that offers Croatian as one of the languages in the translation pairs. Furthermore, Google developed its own statistical software for translation. According to Och (now head of Google MT department), a solid base for the development of a usable SMT system for a new language pair from scratch, would consist in having a bilingual text corpus (or parallel collection) of more than a million words and two monolingual corpora of each more than a billion words. Statistical models built from this data would then used for translating between those languages.
Google acquired the initial amount of linguistic data from United Nations' documents, which are available in six official UN languages (Arabic, Chinese, English, French, Russian and Spanish). To quote Google: "Our system takes a different approach: we feed the computer billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. We then apply statistical learning techniques to build a translation model. We've achieved very good results in research evaluations."[3]
This service now (2009) offers following languages for bidirectional translation

[3] http://www.google.com/intl/en/help/faq_translation.html#statmt

(alphabetically): Arabic, Bulgarian, Catalan, Chinese (Simplified), Chinese (Traditional), *Croatian*, Czech, Danish, Dutch, Filipino, Finnish, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Latvian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian and Vietnamese.

**Croatian-English Language Pair**

The translation process is hindered by the fact that languages involved often differ culturally, stylistically, syntactically, and lexically. These differences are called translation divergences and they can be, according to (Jurafsky & Martin, 2009):

- systematic,
- idiosyncratic, and
- lexical.

While *systematic* differences can be modelled in a general way, idiosyncratic and lexical differences must be dealt one by one (Jurafsky & Martin, 2009). Croatian language is essentially very different from English.

Croatian, in comparison to English, has relatively free word order, does not have articles and uses fewer pronouns. Languages which omit pronouns are called pro-drop, referentially sparse or cold languages because they require the hearer to do more inferential work to recover antecedents. Translating from pro-drop languages into non-pro-drop languages is exhaustive because each zero has to be identified and anaphor recovered.

*Idiosyncratic* differences also have to be tackled for the translation process to succeed. For example, 'existential *there*' is the name of an English idiosyncratic construction used to introduce a new scene (Jurafsky & Martin, 2009). Croatian does not have a similar construction.

Finally, there are *lexical divergences* which further complicate the translation process. Besides difficulties in disambiguating homonymous and polysemous expressions, divergences can also be grammatical. For example, part-of-speech (POS) tags between source words and corresponding target words do not have to overlap. Another divergence is that Croatian marks gender, number and case on adjectives, while English does not. Nevertheless, one of the languages may have a lexical gap (the meaning of a word or phrase cannot be conveyed in another language because there is no corresponding word or phrase) (Jurafsky & Martin, 2009).

**Examples and translations**

Croatian-English translation is done on three different types of texts:

- text on corpus linguistics, annotation and research methods,
- text on small, medium and large enterprises and Government's plan for reform, and

- text on purchasing washing machine.

The Croatian texts and the reference texts, i. e. English translations are taken from the Internet and used without any modifications.

**Comparison and analysis**

The task in these examples was to compare human and MTs from Croatian to English, using Google Translate service. Source texts on Croatian and reference translations on English, taken from the Internet[4], had no restrictions for use and have not been modified in any way.

The comparison and analysis of translations has been done on lexical, morphological, syntactic and semantic level. The usage of punctuation marks has also been analysed.

On the *lexical level* (i.e. wrong translation / misuse of words), the lack of translation indicates that the system does not "recognize" single words, even repeatedly used, although these words are internationalisms (e.g. *leksičko, inherentno, kontekstno*). These untranslated units are called zerotones according to Sepesy Maucec and Kacic in (2007).

The usage of "not appropriate" words in the translation (i.e. synonyms or words that do not warp the meaning) does not significantly affect intelligibility (e.g. *rewriting* instead of *processing*, *certain* instead of *determined by* or *stipulated by*), since the rest of the translation provides an understandable message. There is also an issue with personal pronouns (*he* instead of *it)* or expressions (*great body* instead of *large corpora*). On the other hand, if something is not translated and cannot be "deducted" from the similarity in expression (for example, in a language the user does not speak at all), it can make the message undecodable, although a partial translation is available.

As for *syntax*, the word order in Croatian is relatively free, and in English is basically determined with the rule SVO. This order (along with formal structure) is also common in "bureaucratic languages". Therefore, it should not come as a surprise that the best results were achieved in texts 2 and 3, since Google Translate obtained its basic language corpora from the official documents of UN and EU. On the other hand, most mistakes are found in text 1, written in more of a "scientific language" (longer and somehow more complicated sentences).

*Morphological* analysis shows results similar to those of syntactic analysis. The most notable mistake is the frequent misuse of singular/plural. Some mistakes are due to the "odd" use of expressions (e.g. "environment friendly" should be used in general, but MT translated the original almost literally ("not burden the environment"). Another issue is the usage of cases in Croatian, which explains the lack of translation for "already translated" words.

[4] Respectively: (1.a) http://ling.unizd.hr/znanost/projekti/index.hr.html, (2.a) http://www.mingorp.hr/default.aspx?id=8 and (3.a) http://products.gorenje.si

On the *semantic* level (i.e. preservation of original message), Google MT shows some "effort", although in some cases the user has a lot of inferential work. It is also obvious that statistical MT lacks in taking context into account, which could significantly affect original message. The usage of *punctuation* marks is mainly taken over from the original text.

**Manual evaluation**

We employed manual evaluation method in order to obtain results which could later be used in evaluating automatic methods and determining their correlation with human judgements.

Six evaluators were kindly asked to score 21 machine-translated sentences according to a scale given in Table 1, and with regard to corresponding reference translations.

Table 1: *Fluency* and *adequacy* scale

| | Fluency | Adequacy |
|---|---|---|
| 1 | incomprehensible | none |
| 2 | disfluent English | little meaning |
| 3 | non-native English | much meaning |
| 4 | good English | most meaning |
| 5 | flawless English | all meaning |

Source: MT Marathon, 2008

The results are as follows. The average *fluency* judgement per judge ranges from 2.14 to 3.57, while the average *adequacy* judgement per judge ranges from 2.71 to 3.67. The average of a set of judgements is calculated according to the formula in (2). The averages are 2.98 for *fluency* and 3.36 for *adequacy*. The standard deviation of experimental data is calculated using the formula in (4), where $n$ stands for the number of different values and $n_i$ for the total frequency of each value. The standard deviation per question ranges from 0.52 to 1.03 for *fluency* and from 0.41 to 1.05 for *adequacy*, while standard deviation per judge according to the *fluency* criterion ranges from 0.60 to 1.06, and according to the *adequacy* criterion from 0.60 to 1.32.

$$\sigma = \frac{1}{n}\sqrt{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}\,. \tag{4}$$

We used the $\chi^2$-test to determine whether there is a difference in the distribution of grade 3 among different evaluators for *fluency* and *adequacy* separately. We had to pool categories in the above mentioned way. Otherwise, one or more of the expected frequencies would fall below five, which would invalidate the chi-square test results. The $\chi^2$ formula is given in (5), where $O$ stands for observed

frequencies and *E* for expected frequencies (6). When $\chi^2$ is used as a test of association, the expected frequencies are calculated directly from the observed frequencies by assuming independence between the categories. We applied the test to the data in tables 2 and 3.

$$\chi^2 = \sum \frac{(O-E)^2}{E}. \tag{5}$$

$$E = \frac{rowTotal \times columnTotal}{overallTotal}. \tag{6}$$

Table 2: Score frequencies according to *fluency* criteria

| Fluency | Eval1 | Eval2 | Eval3 | Eval4 | Eval5 | Eval6 | Total |
|---|---|---|---|---|---|---|---|
| score 3 | 5 | 10 | 6 | 9 | 3 | 10 | 43 |
| other scores | 16 | 11 | 15 | 12 | 18 | 11 | 83 |
| Total | 21 | 21 | 21 | 21 | 21 | 21 | 126 |

Table 3: Score frequencies according to *adequacy* criteria

| Fluency | Eval1 | Eval2 | Eval3 | Eval4 | Eval5 | Eval6 | Total |
|---|---|---|---|---|---|---|---|
| score 3 | 5 | 10 | 7 | 11 | 5 | 10 | 48 |
| other scores | 16 | 11 | 14 | 10 | 16 | 11 | 78 |
| Total | 21 | 21 | 21 | 21 | 21 | 21 | 126 |

The number of the degrees of freedom is 5 ((*rowTotal* - 1) x (*columnTotal* - 1)). The table value for the $\chi^2$ with 5 degrees of freedom at the 5 per cent significance level is 11.070. We obtained $\chi^2$ values for *fluency* and *adequacy*, 9.073 and 7.269 respectively. Since these values are smaller than the appropriate table value, we can conclude that the evaluators do not significantly differ in assigning score 3, neither for *fluency*, nor for *adequacy*. The same counts at the 1 per cent significance level because the appropriate table value is 9.236.

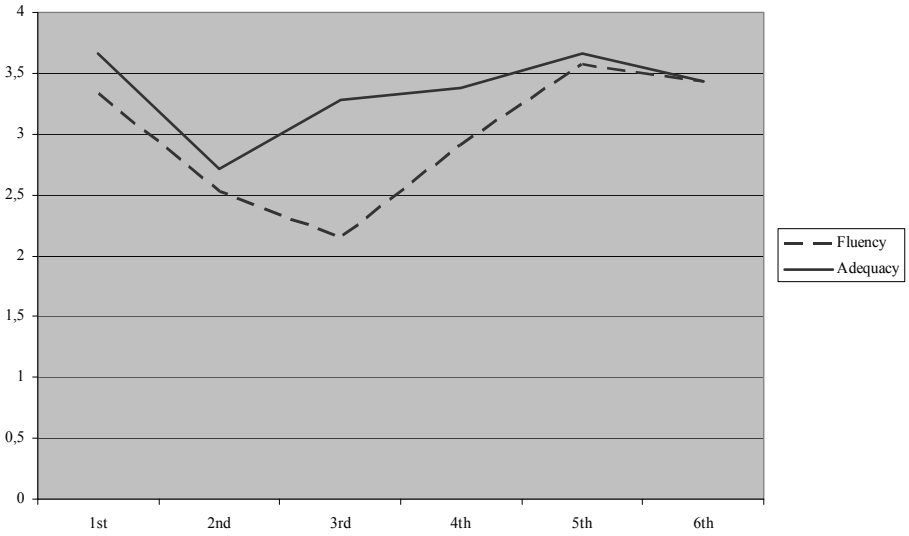
Table 4: Association of two criteria with regard to different evaluators

| | Eval1 | Eval2 | Eval3 | Eval4 | Eval5 | Eval6 |
|---|---|---|---|---|---|---|
| $\chi^2$ | 7.118 | 0.067 | 4.571 | 0.077 | 13 | 0.048 |

We performed the same test to see whether there is a significant difference in assigning scores for the two criteria per each evaluator applying the same pooling strategy. The results are shown in table 4. Since the table value with one degree of freedom at the 5 per cent significance level is 3.841, we may conclude that there is a significant difference in assigning *fluency* and *adequacy* scores for the first, third and fifth evaluator, while there is almost no difference for the remaining evaluators.

In general terms, *adequacy* scored slightly better than *fluency*, as evident in chart 1. Histograms of *adequacy* judgements show that different human evaluators use the scale 1-5 differently. Histograms of *fluency* judgements point to the same phenomenon.

Chart 1: *Fluency* and *adequacy* average judgements




Results of the language independent statistically-based MT service could be improved by the integration of the language-dependant module, which already exists for a number of languages. Human intervention in the post-editing step could certainly improve the output, although even the raw output, taken *cum grano salis,* could be useful and even usable for the basic information transfer and personal use.

## Conclusion

In this case study the statistically-based MT service has been evaluated on the Croatian-English language pair. The results of the $\chi^2$ test show that different evaluators do not significantly differ in assigning score 3, neither for *fluency*, nor for *adequacy*. Furthermore, the same test points that half of the evaluators find *fluency* and *adequacy* criteria to be closely related as far as grade 3 is concerned, while the other half of them can better distinguish between these criteria, and, therefore, rates them differently. In order to perform the $\chi^2$ test, the pooling strategy had to be applied. This highlighted the need for the greater number of evaluators, and, accordingly, higher frequencies.

Since user expectations are of considerable importance (including education, intelligence, culture), it should be pointed out that one should be aware of MT

limitations and possibilities, even though SMT service could be improved by the integration of language-dependant module or by introducing the post-editing step.

## References


Bhagat, Rahul; Ravichandran, Deepak. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. // *Proceedings of Association for Computational Linguistics (ACL).* OH, Columbus, 2008, 674-682

Brants, Thorsten; Popat, Ashok; Xu, Peng; Och, Franz, Dean, Jeffrey. Large Language Models in Machine Translation. // *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing*. Czech Republic, Prague, 2007

Direct mailing with Mr. Josh Estelle, Google Translate, Google Inc.

Dovedan, Zdravko; Seljan, Sanja; Vučković, Kristina. Machine Translation as Help in the Communication Process. // *Informatologia*, vol. 4 (2002), 35, 283-291

Google Translate. http://translate.google.com/# (12.08.2009)

Hajič, Jan; Homola, Petr; Kuboň, Vladislav. A simple multilingual machine translation system. // *Proceedings of the MT Summit IX*. New Orleans, Louisiana, 2003

Hutchins, John; Somers, Harold. An Introduction to Machine Translation. UK, London : Academic Press, 1992

Jayaraman, Shyamsundar; Lavie, Alon. Multi-Engine Machine Translation Guided by Explicit Word Matching. // *10th Conference of the European Association for Machine Translation (EAMT)*, Hungary, Budapest, Hungary, 2005, 143-152

Jurafsky, Daniel; Martin, James H. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. New Jersey : Pearson education, 2009

Knight, Kevin. Teaching Statistical Machine Translation. // *Proceedings of the MT Summit IX Workshop on Teaching Translation Technologies and Tools.* New Orleans, 2003, 17-19

Koehn, Philip; Och, Franz J.; Marcu, Daniel. Statistical Phrase-Based Translation. // *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1.* New York, Morristown : Association for Computational Linguistics, 2003, 48-54

Lin, Dekang, Wu, Xiaoyun. Phrase Clustering for Discriminative Learning. // *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Singapore, 2009, 1030-1038

Macherey, Wolfgang; Och, Franz J. An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. // *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Czech Republic, Prague, 2007, 986-995.

Mohri, Mehryar. Statistical Natural Language Processing. In: M. Lothaire (Ed.), *Applied Combinatorics on Words*. Cambridge: Cambridge University Press, 2005

Nagao Magao, A framework of a mechanical translation between Japanese and English by analogy principle. // *Proceedings of the international NATO symposium on Artificial and human intelligence*. New York : Elsevier North-Holland, Inc., 1984, 173-180

Och, Franz J.; Ney, Hermann. The Alignment Template Approach to Statistical Machine Translation. // *Computational Linguistics.* 30 (2004), 4; 417-449

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing. BLEU: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), 2001

Pasca, Marius; Dienes, Peter. Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web. // *Proceedings of the 2nd International Joint Conference on Natural Language* Processing (IJCNLP-2005). Korea, Jeju Island, 2005, 119-130

Second Machine Translation Marathon, Lecture Notes, Germany : Berlin, 2008

Sepesy Maucec, Mirjam; Kacic, Zdravko. Statistical Machine Translation from Slovenian to English. // *Journal of Computing and Information Technology*. 15 (2007), 1; 47-59

Valderrábanos, Antonio S.; Esteban, José; Iraola, Luis. TransType2 - A New Paradigm for Translation Automation. // *Proceedings of the MT Summit IX*. New Orleans, 2003, 498-501

Wang, Ye-Yi; Waibel, Alex. Decoding Algorithm in Statistical Machine Translation. // *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.* New York, Morristown : Association for Computational Linguistics, 1997, 366-372

Watanabe, Taro; Sumita, Eiichiro. Bidirectional Decoding for Statistical Machine Translation. // *Proceedings of the 19th international conference on Computational linguistics – Volume 1*. Taipei, Taiwan, 2002, 1-7

Zollmann, Andreas; Venugopal, Ashish; Och, Franz; Ponte, Jay. A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT. // *Proceedings of 22nd International Conference on Computational Linguistics Coling*, Manchester 2008, 1145–1152

Zughoul, Muhammad R.; Abu-Alshaar, Awatef M. English/Arabic/English Machine Translation: A Historical Perspective. // *Journal des traducteurs / Meta: Translators' Journal*. 50 (2005), 3; 1022-1041

# Statistical Language Models for Croatian Weather-domain Corpus

Lucia Načinović
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
lnacinovic@inf.uniri.hr

Sanda Martinčić-Ipšić
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
smarti@inf.uniri.hr

Ivo Ipšić
Department of Informatics, University of Rijeka
Omladinska 14, Rijeka, Croatia
ivoi@inf.uniri.hr

**Summary**

*Statistical language modelling estimates the regularities in natural languages. Language models are used in speech recognition, machine translation and other applications for speech and language technologies. In this paper we will present a procedure for language models building for the Croatian weather-domain corpus. Different types of n-gram statistic language models and smoothing methods for language modelling are presented. Those models are compared in terms of their estimated perplexity.*

**Key words:** statistical language modelling, n-gram, smoothing methods, Croatian weather-domain corpus

## Introduction

Language models are employed in many tasks including speech recognition, optical character recognition, handwriting recognition, machine translation, and spelling correction (Chen & Goodman, 1998). Speech recognition is concerned with converting an acoustic signal into a sequence of words. Through language modelling, the speech signal is being statistically modelled. Language model of a speech estimates probability Pr(W) for all possible word strings $W=(w_1, w_2,\ldots w_i)$. (Chou & Juang, 2003) Before language models can be estimated, text corpora must be appropriately processed. As can be seen in Figure 1, language models for automatic speech recognition are usually estimated from manual

transcriptions of speech signals and from normalized text corpora. Furthermore, text preparation includes locating appropriate sources of text data and audio transcriptions and processing them in homogeneous manner. (Chou & Juang, 2003)

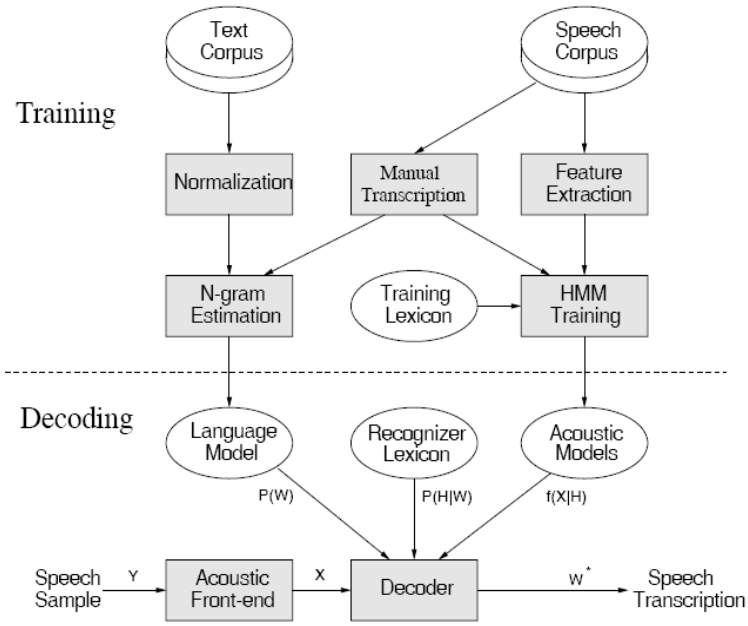



Figure 1: System diagram of a generic speech recognizer based on statistical models, including training and decoding processes and the main knowledge sources (Chou & Juang, 2003).

In this work, we carry out a comparison of the most widely-used smoothing techniques built on Croatian weather-domain corpus using n-grams of various order, and show how these factors affect the relative performance of smoothing techniques which is measured through the estimated perplexities of models.
This paper is organized as follows. The next section gives general information on language models followed by a section on concept of a language model perplexity which is used in this work as a metrics for language models comparison. Then we give information on smoothing and smoothing techniques that we used in our research. Afterwards we describe used text corpus and the implementation of smoothing techniques that we used and we give the obtained results. We end the paper with the conclusion.

## Language models

Language models estimate the probabilities of word sequences which are usually derived from large collections of text material (Manning & Schütze, 1999).

The models of word sequences we will consider in this work are probabilistic models - ways to assign probabilities to strings of words, whether for computing the probability of an entire sentence or for giving a probabilistic prediction of what the next word will be in a sequence.
The most widely-used language models are n-gram language models. The central goal of the most commonly used - trigram models, is to determine the probability of a word given the previous two words:

$$p(w_i \mid w_{i-2} w_{i-1})$$

The simplest way to approximate this probability is to compute

$$pML(w_i \mid w_{i-2} w_{i-1}) = \frac{c(w_{i-2} w_{i-1} w_i)}{c(w_{i-2} w_{i-1})}$$

i.e. the number of times the word sequence $w_{i-2}w_{i-1}w_i$ occurs in some corpus of training data divided by the number of times the word sequence $w_{i-2}w_{i-1}$ occurs. This value is called the maximum likelihood (ML) estimate.

**Language model perplexity**
The most common metric for evaluating a language model is the probability that the model assigns to test data, or the derivative measures of cross-entropy and perplexity. The cross-entropy $H_p(T)$ of a model *p(T)* on data T is defined as

$$H_p(T) = -\frac{1}{W_T} \log_2 p(T)$$

where $W_T$ is the length of the text *T* measured in words. This value can be interpreted as the average number of bits needed to encode each of the $W_T$ words in the test data using the compression algorithm associated with model *p(T)*. The perplexity $PP_p(T)$ of a model *p* is the reciprocal of the average probability assigned by the model to each word in the test set *T*, and is related to cross-entropy by the equation

$$PP_p(T) = 2^{H_p(T)}$$

Clearly, lower cross-entropies and perplexities are better. (Chen & Goodman, 1998)

**Smoothing**
One major problem with standard N-gram models is that they must be trained from some corpus, and because any particular training corpus is finite, some perfectly acceptable N-grams are bound to be missing from it. (Jurafsky & Martin, 2000). To give an example from the domain of speech recognition, if

the correct transcription of an utterance contains a trigram $w_{i-2}w_{i-1}w_i$ that has never occurred in the training data, we will have $pML(w_i|w_{i-2}w_{i-1})=0$ which will preclude a typical speech recognizer from selecting the correct transcription, regardless of how unambiguous the acoustic signal is.
Smoothing is used to address this problem. The term smoothing describes techniques for adjusting the maximum likelihood estimate of probabilities to produce more accurate probabilities. These techniques adjust low probabilities such as zero probabilities upward, and high probabilities downward. Not only do smoothing methods generally prevent zero probabilities, but they also attempt to improve the accuracy of the model as a whole. (Chen & Goodman, 1998)

*Smoothing techniques used in our research*
In our research, we used four different smoothing techniques including additive smoothing, absolute discounting, Witten-Bell smoothing technique and Kneser-Ney discounting. General information on each of those smoothing techniques is given bellow.

*Additive smoothing*
Additive smoothing is one of the simplest types of smoothing. To avoid zero probabilities, we pretend that each n-gram occurs slightly more often than it actually does: we add a factor δ (0< δ ≤1) to every count. Thus, we set

$$p_{add}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\delta + c(w_{i-n+1}^{i})}{\delta \mid V \mid + \sum_{w_i} c(w_{i-n+1}^{i})}$$

where V is the vocabulary, or set of all words considered and *c* is the number of occurrences. Lidstone and Jeffreys advocate taking δ=1 (Chen & Goodman, 1998). We used three different values of δ parameter (0.1, .05 and 1) in our research. More information on how the change of those values affected language models is given in section *Results* bellow. Although additive smoothing does not perform well and is not commonly used, it makes a basis for other smoothing techniques.

*Absolute discounting*
In absolute discounting techniques, the linear interpolation algorithm is used. When there is little data for directly estimating an *n-gram* probability useful information can be provided by the corresponding *(n-1)-gram*. A simple method for combining the information from lower-order n-gram models in estimating higher-order probabilities is linear interpolation, and a general class of interpolated models is described by Jelinek and Mercer (1980) (Chen & Goodman, 1998):

$$p_{\text{int}erp}(w_i \mid w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+11}^{i-1}} p_{ML}(w \mid w_{i-n+1}^{i-1}) + (1-\lambda_{w_{i-n+1}^{i-1}}) p_{\text{int}erp}(w_i \mid w_{i-n+2}^{i-1})$$

The nth-order smoothed model is defined recursively as a linear interpolation between the nth-order maximum likelihood model and the (n-1)-th-order smoothed model. Given fixed pML, it is possible to search efficiently for the

$$\lambda_{w_{i-n+1}^{i-1}}$$

that maximizes the probability of some data using the Baum–Welch algorithm (Chou & Juang, 2003).

In absolute discounting smoothing instead of multiplying the higher-order maximum-likelihood distribution by a factor

$$\lambda_{w_{i-n+1}^{i-1}}$$

the higher-order distribution is created by subtracting a fixed discount D from each non-zero count:

$$p_{abs}(w_i \mid w_{i-n+1}^{i-n}) = \frac{\max\left\{c(w_{i-n+1}^{i}) - D, 0\right\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + (1-\lambda_{w_{i-n+1}^{i-1}}) p_{abs}(w_i \mid w_{i-n+2}^{i-1})$$

To make this distribution sum to 1, we take:

$$1-\lambda_{w_{i-n+1}^{i-1}} = \frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_1 + (w_{i-n+1}^{i-1}\bullet)$$

Ney et al. (1994) suggest setting D as follows:

$$D = \frac{n_1}{n_1 + 2n_2}$$

where $n_1$ and $n_2$ are the total number of n-grams with exactly one and two counts in the training data. According to that formula we came to the value of D=0.06 in our research. Besides this particular value, we also experimented with three more values: 0.3, 0.5 and 1. Information on how changing those values affected language models is given in section *Results*.

*Witten-Bell smoothing*

The nth-order smoothed model is defined recursively as a linear interpolation between the nth-order maximum likelihood model and the (n-1)-th-order smoothed model. To compute the parameters

$$\lambda_{w_{i-n+1}^{i-1}}$$

for Witten–Bell smoothing, we will need to use the number of unique words that follow the history

$$w_{i-n+1}^{i-1}$$

We will write this value as follows:

$$N_1 + (w_{i-n+1}^{i-1}\bullet) = |\{w_i : c(w_{i-n+1}^{i-1} w_i) > 0\}|$$

We assign the parameters

$$\lambda_{w_{i-n+1}^{i-1}}$$

for Witten–Bell smoothing such that

$$1 - \lambda_{w_{i-n+1}^{i-1}} = \frac{N_1 + (w_{i-n+1}^{i-1}\bullet)}{N_1(w_{i-n+1}^{i-1}\bullet) + \sum_{w_1} c(w_{i-n+1}^{i})}$$

*Kneser-Ney smoothing*

Kneser and Ney (1995) have introduced an extension of absolute discounting where the lower-order distribution that one combines with a higher-order distribution is built in a novel manner. In previous algorithms, the lower-order distribution is generally taken to be a smoothed version of the lower-order maximum likelihood distribution. However, a lower-order distribution is a significant factor in the combined model only when few or no counts are present in the higher-order distribution. Consequently, they should be optimized to perform well in these situations.

According to Kneser-Ney smoothing, the lower-order distribution such that the marginals of the higher-order smoothed distribution match the marginals of the training data are being selected. For example, for a bigram model we would like to select a smoothed distribution $p_{KN}$ that satisfies the following constraint on unigram marginals for all $w_i$:

$$\sum_{w_{i-1}} p_{KN}(w_i - 1 w_i) = \frac{c(w_i)}{\sum_{w_i} c(w_i)}$$

The left-hand side of this equation is the unigram marginal for $w_i$ of the smoothed bigram distribution $p_{KN}$, and the right-hand side is the unigram frequency of $w_i$ found in the training data. (Chen & Goodman, 1998)

## Smoothing implementations

In this section we describe our smoothing techniques implementations. 2-gram, 3-gram and 4-gram language models were built. On each of these models, we applied four different smoothing techniques – additive smoothing, Witten-Bell smoothing, absolute discounting and Kneser-Ney smoothing.

Language models were built from the Croatian weather-domain corpus (Martinčić-Ipšić, 2007). Corpus contains 290 480 words, 2 398 1-grams, 18 694 2-grams, 23 021 3-grams and 29 736 4-grams.

Major part of the corpus was developed in the period from 2002 until 2005 by recording radio weather forecasts and some parts were added later. It includes the vocabulary related to weather, bio and shipping forecast, river water levels and weather reports.

We divided corpus into ten parts. We used nine parts as train data for building language models and one part as test data for evaluating those models in terms of their estimated perplexities.

Different language models were built and tested with SRILM language modelling toolkit. (Stolcke, 2002)

## Results

After building 2-gram, 3-gram and 4-gram language models and applying different smoothing techniques on those models, the perplexities of models were estimated. Those perplexities are given in Table 1.

As mentioned before, it is usually considered that models with lower perplexities are better. According to that, additive smoothing gave the worst results. The perplexities of models after applying additive smoothing were even higher than those of the models built without implementing smoothing techniques. By increasing the parameter $\delta$ in additive smoothing, the perplexities of the built language models increased as well.

Absolute discounting gave the best results with the parameter D=0.3 and the worst with the parameter D=1. With the parameter 0.3, perplexities of 2-gram, 3-gram and 4-gram models were lower than the perplexities of those models without smoothing. However, it gave poor results with the parameter D=1. According to the obtained perplexities, we can also come to the conclusion that absolute discounting gives better results on higher-order n-grams such as 4-grams.

Witten-Bell smoothing gave good results on 2-gram, 3-gram and 4-gram models. The perplexities of the models after implementing the smoothing are lower than the perplexities of those models without smoothing implementation.

The best results gave the implementation of Kneser-Ney smoothing. The perplexities of the models after implementing that smoothing technique are lower than all other perplexities.

The presented results were expected because the used text covered only the weather domain vocabulary.

Table 1: The perplexities of tested language models

| | Without smoothing | Additive smoothing | | | Absolute discounting | | | | Witten-Bell | Kneser-Ney |
|---|---|---|---|---|---|---|---|---|---|---|
| | | δ parameter | | | D parameter | | | | | |
| | | 0,1 | 0,5 | 1 | 0,06 | 0,3 | 0,5 | 1 | | |
| 2-gram | 19,87 | 28,8 | 51,6 | 73,5 | 20,39 | 19,61 | 19,64 | 21,6 | 19,75 | 18,96 |
| 3-gram | 8,45 | 30,04 | 86,9 | 144,2 | 8,55 | 8,17 | 8,22 | 9,30 | 8,25 | 7,63 |
| 4-gram | 6,04 | 42,9 | 142,6 | 239,87 | 5,93 | 5,64 | 5,71 | 6,76 | 5,76 | 5,24 |

## Conclusion

In this paper we described our research on different smoothing techniques which were applied to language models built from the Croatian weather-domain corpus. 2-gram, 3-gram and 4-gram language models were built. On each of these models, we applied four different smoothing techniques – additive smoothing, Witten-Bell smoothing, absolute discounting and Kneser-Ney smoothing. After we built the models, we estimated and compared perplexities of those models. We came to the conclusion that Kneser-Ney smoothing technique gives the best results.

Further we will prepare the more balanced corpus of Croatian text and thus build more complete language model.

## References

Chen, Stanley F.; Goodman, Joshua. An empirical study of smoothing techniques for language modelling. Cambridge, MA: Computer Science Group, Harvard University, 1998

Chou, Wu; Juang, Biing-Hwang. Pattern recognition in speech and language processing. CRC Press, 2003

Jelinek, Frederick. Statistical Methods for Speech Recognition. Cambridge, MA: The MIT Press, 1998

Jurafsky, Daniel; Martin, James H. Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River, New Jersey: Prentice Hall, 2000

Manning, Christopher D.; Schütze, Hinrich. Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press, 1999

Martinčić-Ipšić, Sanda. Raspoznavanje i sinteza hrvatskoga govora konteksno ovisnim skrivenim Markovljevim modelima, doktorska disertacija. Zagreb, FER, 2007

Milharčič, Grega; Žibert, Janez; Mihelič, France. Statistical Language Modeling of SiBN Broadcast News Text Corpus.//Proceedings of 5th Slovenian and 1st international Language Technologies Conference 2006/Erjavec, T.; Žganec Gros, J. (ed.). Ljubljana, Jožef Stefan Institute, 2006

Stolcke, Andreas. SRILM – An Extensible Language Modeling Toolkit.//Proceedings Intl. Conf. on Spoken Language Processing. Denver, 2002, vol.2, pp. 901-904

Original scientific paper

# Evaluation of Electronic Translation Tools Through Quality Parameters

Vlasta Kučiš
Department for Translation Studies
Faculty of Arts, University of Maribor
Koroška cesta 160, 2000 Maribor, Slovenia
asta.kucis@siol.net, vlasta.kucis@uni-mb.si

Sanja Seljan
Department for Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
sanja.seljan@ffzg.hr

Ksenija Klasnić
Department of Sociology
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
kklasnic@ffzg.hr

**Summary**

*In the paper, the difference of translation quality of texts obtained through traditional reference works and online electronic translation tools (corpus and multilingual terminology database) will be measured in three main categories: lexical, orthographic and punctuation; syntactically and stylistically using paired samples t-test. The translation was made with the support of electronic translation tools, using the example of a Slovenian bilingual corpus called Evrokorpus and the multilingual terminology database Evroterm. In the paper, the hypothesis that modern online translation tools contribute to the quality and consistency of expert translations, as well as to the acquisition of new competitive skills and knowledge is to be examined.*

**Key words**: translation quality, consistency, translation tools, mistake categorisation, t-test

## Introduction

The translation profession underwent a metamorphosis at the turn of the last century, embracing new information and communication knowledge and skills, as well as adopting the usage of modern multilingual technologies, according to

Seljan (2007); Vintar (2008); Željko (2004). The acceptance, implementation and application of translation technologies, as well as the exploitation of their potential by translators during the translation process aim to enhance productivity, competitiveness and the quality of the work. They should therefore be perceived as an integral part of a translator's reality necessitated by globalization and the need for swift information flow.

Over the last ten years, the European Union has been intensively preoccupied with the inherent problems of a multilingual environment, which is a demanding and ambitious project. EU translations have to be unambiguous and terminologically consistent. Such unambiguousness can only be achieved through the consistent and synchronized use of terminology databases and other translation tools.

The GILT sector (Globalization, Internationalization, Localization and Translation) has been facing an increase in translation demands. Due to EU enlargement and the use of the English language as lingua franca on one side, and the growing interest for the protection of national cultures and identities on the other side, the development of multilingual services plays a key role in written communication.

Technical innovations, research and quality management aim to compensate for the lack of translators and increased demands within a time constraint. Several key drivers, such as multilingualism and language technology, market changes (commercial translations, web products, localization) and the Internet (e-books, language barriers in communication, multilingual services, web translations, newsletters) have caused considerable changes in the translation process, relating also to expectations in terms of quality, time and consistency.

The importance of translation practices using ICT does not only witness individual experiences, but also examples in large national translation companies, as presented in (Ørsted, 2001), where assessment procedures aim to evaluate the working environment of translators and support services in IT departments, becoming a corporate issue.

Starting from individual education and practice, up to integrated document workflow, translation quality has been a matter of numerous business applications and workflow document changes.

In the paper, the differences in the translation quality among two groups were analyzed and statistically evaluated. The translation quality of texts obtained through traditional reference works and online electronic translation tools (corpus and multilingual terminology database) will be measured through three main categories: lexical, orthographic and punctuation, syntactically and stylistically using t-test.

This paper analyzes the quality and consistency of translations made with the support of electronic translation tools, using the example of a Slovenian bilingual corpus called *Evrokorpus* and the multilingual terminology database *Evroterm,* which are available at http://www.evrokorpus.gov.si and http://www.

evroterm.gov.si respectively. In the paper, the hypothesis that modern online translation tools contribute to the quality and consistency of expert translations, as well as to the acquisition of new competitive skills and knowledge will be examined.

## Related work

Quality assurance is also one of the key issues of the language policy of the European Commission's Directorate-General for Translation (DGT). Documents are mostly translated and revised in-house, demanding the quality standards that apply, according to Farkas, to completeness, terminology, clarity, compliance with linguistic and idiomatic requirements of EU legislation, while revisers consider the text from several points of view including meaning, content, language, style, form and editing. Therefore, the DGT is encouraging the use of translation tools through education, in-house open access and document workflow. To ensure a high quality standard, translators are required to use translation tools, memories and databases. Terminological resources and related databases generally include the translation database of the Ministry of Justice, Eurlex or the CELEX database of legal texts, IATE (Inter-Agency Terminology Exchange) and EURAMIS (European Advanced Multilingual Information System).

According to Hemera and Elekes (2008), apart from the growing need for translations within a very short time period, the Central and Eastern European translation markets have faced problems in the translation business in terms of different expectations when it comes to technical aspects, prices and quality levels. While the U.S. and Western European markets had enough time to learn through educational phases, to experiment with business models and to learn business ethics, CEE countries had to learn very fast and under more difficult circumstances, with no time to experiment, but having to meet high and sophisticated quality standards that have become an indispensable issue in information and communication technology, adequate project management and business flexibility.

According to Waddington (2006), there are no standards in the evaluation of translation quality. Often, we judge whether a translation is more or less appropriate. Contrary to right or wrong answers, it is possible to develop non-binary categories that relate to the degree of acceptability, ranging from the least to the most acceptable translation (1 to 5). Like Waddington, Sager (1989) lists five different types of errors: inversion of meaning, omission, addition, deviation, modification, but also linguistic, semantic and pragmatic effects. Another classification relates to the communicative function, evaluating the degree to which it affects communication in the target language. When comparing source and target tests of several software products in order to determine the translation quality, Gerasimov (2007) includes the following errors: inconsistency, inadequately translated terms, omission, identical source and target segments, punc-

tuation, capitalization, number/value formatting errors, incorrect untranslatables and tags.
As this research was conducted on students' assignments, the evaluation was performed through a points system in which every mistake carried one point. Mistakes were classified in three categories: lexical, orthographic and syntactic/stylistic. This kind of text processing was used for easier data processing and an easy-to-survey mistake evaluation.

## Goals and operationality

The pilot project was made at the Department for Translation Studies at the Faculty of Arts, University of Maribor. A random sample of 51 students (N=51) from all four years of study was taken. For this purpose, the same group of students translated two texts of similar length from the same domain, differing in the type of tools used.
The students translated two texts from German into Slovene:

- Group A: Text 1 representing part of the *acquis communaitaire*,
- Group B: Text 2 about intercultural communication in the EU

The students were given 45 minutes to translate both texts, which had approximately the same length and were equally as difficult to translate. The first text was 159 words long, the other 140 words. In both experiments, the translation was made from German into Slovene. Both translations were evaluated by a professional bilingual translator, with both German and Slovene as mother tongues and a degree from the Department of German Language.
The students translated the first text (group A) with the help of German-Slovene/Slovene-German electronic dictionaries Debenjak (2003) installed on the computer and a Duden dictionary http://www.duden.de, while also using Goggle and Yahoo search engines. The use of online dictionaries and search engines was provided with the belief that translators without special education are able to use the mentioned tools.
For the translation of the second text (group B), more specialized translation tools were available:

- a Slovenian bilingual corpus called *Evrokorpus* http://www.evrokorpus.gov.si
- the multilingual terminology database *Evroterm* http://www.evroterm.gov.si
- a terminology base integrated into the SDL Trados translation program, with prior 15-minute training (all students were familiar with Trados from the course "Computer-Aided Translation")

Expert evaluation of the translations of both texts was done for each student, with mistakes in the translations measured in three main categories:

- lexical mistakes,
- spelling and punctuation mistakes, and

- syntactic and stylistic mistakes

The basic goal of the research was to determine the differences in translation between both texts with regard to the introduction of additional interactive, computer-aided tools in the translation process. The mentioned research aimed to examine the hypothesis of whether computer-aided translation tools and resources improve the quality and consistency of translation.

As part of this research, the following theses were tested:

1. Differences in average results between translations are to be statistically significant considering lexical mistakes.
2. Differences in average results between translations are to be statistically significant considering spelling/punctuation mistakes.
3. Differences in average results between translations are to be statistically significant considering syntactic/stylistic mistakes.

## Sample

The research was done on a sample (N=51) of students from all four years of study. This was a non-probability convenient sample, i.e. one that encompasses a group of individuals available in a certain situation.

There are some methodology issues arising from this sample. First of all, such samples are not representative because they do not encompass the part of the student population interested in attending classes. That is why the interpretation and conclusions arising from this research cannot be generalized against the complete student population. But, the purpose of the research itself is precisely to check whether interactive tools have any influence on the quality of translation.

Moreover, an appropriate sample is the optimal choice because it encompasses a smaller part of the population that can be regarded as being defined by a mutual characteristic (in this case, all respondents work with foreign languages and study translation at university level), which makes it homogenous. With a larger number of respondents, differences would arise only among students of lower and those of higher years of study. It would be expected that translation ability increases with the progress in the years of study due to more experience and practical work in translation.

However, in this research, because of the size of the sample, differences in average results between students of certain years of study will not be taken into account. Another advantage of this sample is the fact that it is economic and easily realized. It is worth repeating that, regardless of the fact that this is a homogenous sample, generalization against the complete student population would not be justified, because the sample is not representative.

Still, it is possible to make certain conclusions regarding the quality and consistency of translation based on statistical processing using t-tests. If the hypotheses prove to be correct, there is justification for the introduction and use of

interactive translation tools that contribute more to quality, speed and consistency in the translation process.

## Results

### Comparison of total mistakes

Generally speaking, all respondents (N=51) translated two texts from the same domain that were equally as difficult, of similar length, and were translated under similar conditions. When comparing the total number of lexical, spelling, punctuation, syntactic and stylistic mistakes the students made in both texts, we can see that in the first translation there was a total of 958 mistakes, in the second a total of 571 (Table 1). Average number of mistakes in the first translation was 18.78 which decreased in the second translation on the average of 11.20 mistakes. The coefficient of variability presented in Table 1. represents the ratio of the standard deviation of a variable relative to its mean and it measures the degree of variation in each variable. It can be seen than there is a slightly less variability of mistakes in the second translation.

Table 1. Total number of mistakes and paired samples statistics

| | Total No. of mistakes | N | Average result | Standard deviation | Coefficient of variability |
|---|---|---|---|---|---|
| Group A | 958 | 51 | 18.78 | 6.100 | 32.48 |
| Group B | 571 | 51 | 11.20 | 3.742 | 33.41 |

Seeing how this is the same sample of respondents in both tests with changed conditions, to test the statistical significance of the difference between the arithmetic means of the samples we used t-test for dependent samples which is a standard parametric test used to test the significance of the change in the average result after the controlled change of conditions. The t-test is based on the comparison of the calculated t-value with the theoretical t-value from the table of critical t-values with respect to different number of degrees of freedom and different risk levels. The calculation of the observed t-value was done using the formula in which the t-value is expressed as the ratio of the difference of arithmetic means and the standard error of difference between means.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

This method is often called a correlated t-test because the Pearson's coefficient of correlation between two measurements is used in the computing of the standard error of difference between means.

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2} - 2r_{1,2}s_{\bar{X}_1}s_{\bar{X}_2}}$$

So, to test weather there is a statistically significant change in the average number of mistakes after the repeated testing introduced new parameters and we tracked their influence on the quality of translation, the data were introduced in the formula presented above and the corresponding t-value and border p-value were calculated. The statistical testing was performed two-sided, at risk level $\alpha$=0.05 and degrees of freedom df=50. Seeing how the border p-value (which represents the probability of the type I error: the rejection of the null hypothesis that is correct) is less than 0.001, we can conclude that the average number of all mistakes has statistically significantly decreased after the introduction of electronic translation tools suggesting the need for adequate education and use of translation tools.

Table 2. Paired samples t-test of statistically significant difference between average number of mistakes

| t | p | df |
|---|---|---|
| 10.553 | <.001 | 50 |

*Lexical mistakes*

As has already been pointed out, the same sample of students translated the first text with the help of dictionaries and web search engines, and the second text with the help of web sources, and *Evroterm* and *Evrokorpus*.

The students (N=51) made a total of 479 mistakes in the first text and 302 in the second text. The average result is shown in Table 3. The coefficient of variability is higher in group B, suggesting the bigger variations when using translation tools.

Table 3. Number of lexical mistakes and paired samples statistics

| | No. of lexical mistakes | N | Average result | Standard deviation | Coefficient of variability |
|---|---|---|---|---|---|
| Group A | 479 | 51 | 9.39 | 3.567 | 37.98 |
| Group B | 302 | 51 | 5.92 | 2.489 | 42.04 |

As presented in Table 3, the averages of samples differ, and the t-test has determined (t=7.175) that there is a statistically significant difference at the level p<0.001 (Table 4). Therefore, the first hypothesis can be accepted. This means that the comparison of the two translations can lead to the conclusion that interactive tools significantly contributed to the quality of translation, at least when it comes to lexical mistakes, where the number of lexical mistakes was significantly lower using additional interactive tools.

Table 4. Paired samples t-test of statistically significant difference between average number of lexical mistakes

| t | p | df |
|---|---|---|
| 7.175 | <.001 | 50 |

*Spelling and punctuation mistakes*

In the same way as in the case of lexical mistakes, spelling mistakes in both translations were analyzed. In total, the number of mistakes students made amounted to 243 in the first text and 131 in the second text (Table 5.) The coefficient variability is considerably bigger in group B.

Table 5. Number of spelling and punctuation mistakes and paired samples statistics

| | No. of spelling and punct. mist. | N | Average result | Standard deviation | Coefficient of variability |
|---|---|---|---|---|---|
| Group A | 479 | 51 | 4.76 | 2.566 | 53.566 |
| Group B | 302 | 51 | 2.57 | 1.814 | 70.583 |

The t-test determined that in this case there is also a statistically significant difference between the average number of spelling and punctuation mistakes in two translations (t=5.887). We can conclude that the second hypothesis is accepted as well, i.e. that the use of additional translation tools significantly decreased the number of spelling mistakes ($p<0.001$) (Table 6).

Table 6. Paired samples t-test of statistically significant difference between average number of spelling and punctuation mistakes

| t | p | df |
|---|---|---|
| 5.887 | <.001 | 50 |

*Syntactic and stylistic mistakes*

In the same way, we compared syntactic and stylistic mistakes in both translations. The total number of mistakes the students made amounted to 236 in the first text and 138 in the second (Table 7). The coefficient of variability is considerably bigger in group B.

Table 7. Number of syntactic and stylistic mistakes and paired samples statistics

| | No. of syntactic and styl. mist. | N | Average result | Standard deviation | Coefficient of variability |
|---|---|---|---|---|---|
| Group A | 236 | 51 | 4.63 | 2.425 | 52.375 |
| Group B | 138 | 51 | 2.71 | 1.701 | 62.76 |

The t-test determined that in this case there is also a statistically significant difference in the average number of syntactic and stylistic mistakes between the two translations. We can conclude that the third hypothesis is accepted as well, i.e. that the use of electronic translation tools has, on average, significantly decreased the number of syntactic and stylistic mistakes ($t=4.43$) with $p<0.001$ (Table 8).

Table 8. T-test of statistically significant differences of syntactic and stylistic mistakes

| t-test | p | df |
|---|---|---|
| 4.43 | <.001 | 50 |

## Interpretation of results

Analyzing the quality of translation and type of mistakes (lexical, spelling and punctuation, syntactic and stylistic), the general conclusion is that the introduction of additional computer-aided translation tools significantly influences the quality and consistency of translation.

Taking into account conditions for translation, time and identical text types, it can be concluded that the use of electronic tools was of significant help to students regarding the quality of their translation, although we cannot make conclusions against the entire population of students of the same departments. In the case of such an analysis, other variables would be important, such as the year of study, success, (lack of) motivation, etc.

T-tests, resulting with t-values 7.175, 5.887 and 4.43 respectively, have all shown statistically significant differences at the level of probability lesser than 0.001 and indicated the acceptance of hypothesis 1, 2 and 3 claiming that translation tools improve the quality of translation at lexical, spelling and punctuation and also syntactic and stylistic level.

In any case, the same sample of students showed significantly better results when using an online corpus and terminology databases. It is important to mention that the introduction of additional electronic tools in translation has, on average, decreased the number of mistakes in all analyzed categories. This means that additional online tools contribute to the quality and consistency of translation on all of the most important levels.

Table 9: Percentage of translation improvements

| Mistakes | Group A | Group B | Improvement in % |
|---|---|---|---|
| Lexical | 479 | 302 | 22.66 |
| Spelling | 243 | 131 | 29.96 |
| Syntactic / stylistic | 236 | 138 | 26.20 |
| TOTAL | 958 | 571 | 25.31 |

## Conclusion

The increasing demand for simultaneous translation and integrated solutions also suggests high quality translations. Adequate education and the use of ICT, i.e. computer-assisted translation tools and their integration into document workflow, could help in the translation process during preparation, translation and revision.

The use of additional translation tools (online terminology base, created terminology base and online corpus) significantly influenced the quality and consistency of translation in general (25.31%), but also on all levels (lexical, spelling and punctuation, syntactic and semantic) ranging from 22.66 – 29.96%. The hypothesis that modern electronic translation tools contribute to the quality and consistency of translation has been accepted with the probability of a type I error being lower than 0.1%. The differences among the results on the three mentioned levels are statistically significant at the level $p<0.001$.

With high expectations regarding the translation quality, time constraints and demand for increased productivity, translators are faced with new challenges in education and in business. The use of translation tools certainly improves the quality of professional translations, but has become a corporate issue, asking for horizontal and vertical integration.

## Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grant No. 130-1300646-0909.

## References

Bonet, Josep. Present and future of machine translation in the Directorate General for Translation of the European Commission. Barcelona, Spain. June 19-21, 2006. http://www.elda.org/tcstar-workshop_2006/pdfs/keynotes/tcstar06_bonet.pdf

BRUNDAGE Jennifer. Translation Quality for Professionals, 3.1.2000. http://www.istworld.org/ProjectDetails...

Debenjak,D. *Veliki slovensko-nemški slovar*, DZS and*Veliki nemško-slovenski slovar*, DZS, 2003.

DGT of the EC. *Translation Tools and Workflow*, 2009. http://ec.europa.eu/dgs/translation/bookshelf/tools_and_workflow_en.pdf

Farkas, Ágnes. Translation at the European Commission: quality criteria and quality assurance. http://www.translationconference.com/2006_eloadasok/FarkasEN.pdf

Gerasimov, Andrei. Review of Translation Quality Assurance Software. 2007. http://www.translatorscafe.com/cafe/article71.htm

Groenewald, H. J.; Fourie, Wildrich. Introducing the Autshumato Integrated Translation Environment. EAMT, 2009. http://www.mt-archive.info/EAMT-2009-Groenewald.pdf

Hager, Astrid. The translation market in ten years' time - a forecast. 6/ 2008, pp. 14. http://www.tekom.de/upload/alg/tcworld_608.pdf

Hemera, Annette; Elekes, György. The Central and Eastern European translation market. Multilingual Computing Inc., March 2008. http://www.leg.eu/images/File/MultiLingual_march.pdf

Joann, Drugan. Multilingual document management and workflow in the European institutions. http://www.leeds.ac.uk/cts/research/publications/leeds-cts-2004-11-drugan.pdf

Muzii, Luigi. Quality Assessment and Economic Sustainability of Translation. ITI conference London, 2009. http://www.openstarts.units.it/dspace/bitstream/10077/2891/1/ritt9_05 muzii.pdf

Ørsted, Jeannette. Quality and Efficiency: Incompatible Elements in Translation Practice? Meta : journal des traducteurs / Meta: Translators' Journal, vol. 46, n° 2, 2001, p. 438-447. http://id.erudit.org/iderudit/003766ar

Seljan, Sanja; Gaspar, Angelina. Primjena prevoditeljskih alata u EU i potreba za hrvatskim tehnologijama: Translation Tools in EU and need for Croatian Language Resources. Jezicna politika i jezicna stvarnost / Language Policy and Language Reality. Zagreb: HDPL, 2009. pp. 617-625.

Seljan, Sanja; Gašpar, Angelina; Pavuna, Damir. Sentence Alignment as the Basis For Translation Memory Database // The Future of Information Sciences: INFuture 2007 - Digital Information and Heritage / Seljan, Sanja ; Stančić, Hrvoje (ur.). Zagreb : Odsjek za Informacijske znanosti, Filozofski fakultet Zagreb, 2007. Str. 299-311.

Seljan, Sanja; Pavuna, Damir. *Why Machine-Assisted Translation (MAT) Tools for Croatian?* // Proceeding of 28th International Information Technology Interfaces Conference - ITI 2006. pp. 469-475

Vintar, Špela. Corpora in translation: a Slovene perspective. *The journal of specialised translation*, 2008. http://www.jotrans.org/issue 10/art vintar.pdf.

Vintar, Špela. Localization, globalisation or Slovenization? *Mostovi,* 2004, 38, pp. 74-82.

Željko, Miran. Evroterm and Evrokorpus – a terminology database and a corpus of translations. In: Humar, Marjeta (ed.) *Terminology at the Time of Globalization*, ZRC Publishing, 2004, pp. 139-149.

Waddington, Christopher. Measuring the effect of errors on translation quality. Lebende Sprachen. Volume 51, Issue 2, pp. 67-71, 2006. http://www.reference-global.com/doi/abs/10.1515/LES.2006.67

# Using Translation Memory to Speed up Translation Process

Marija Brkić
Department of Informatics, University of Rijeka
Omladinska 14, 51000 Rijeka, Croatia
mbrkic@uniri.hr

Sanja Seljan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
sanja.seljan@ffzg.hr

Božena Bašić Mikulić
Primary school "Turnić"
Franje Čandeka 20, 51000 Rijeka, Croatia
bozena.basic@gmail.com

**Summary**

*Translation process is one aspect of human creativity. Due to globalization, EU accession negotiations, and the need for information exchange, the amount of translation work increases on a daily basis. The translation process is hindered by the fact that the languages involved differ culturally, stylistically, syntactically and lexically. This paper explores the benefits and limitations of TMs (translation memories). TMs are not used for replacing humans in the translation process, but rather for enhancing the human translation process. In this paper, a detailed analysis of Atril's Déjà Vu X system is presented, along with its time-saving implications, which are based on the reuse of previously stored segments. Excerpts from three different digital camera user manuals are translated from English into Croatian. Evaluation is performed by measuring the time difference between human and TM-based translation speeds in preparation, translation, and revision phases, and with regard to six different parameters.*

**Key words:** Translation Memory (TM), Déjà Vu, Computer-Assisted Translation (CAT), language pair, translation unit (TU), translation speed

## Introduction

The EU relies on the principles of open access to documents, multilingualism and democracy. Therefore, the EU legislation needs to be translated into each of

the official languages. On the other hand, the legislation of each particular member state needs to be translated into one of the EU’s official languages (Seljan & Pavuna, 2006b). To sum up, translation demands in the EU surpass human capacities. There are 23 official languages with 23 x 22 = 506 language pairs. A huge number of pages need to be translated on a daily basis. Short deadlines, demands for consistency and data-sharing, and insufficient number of translators further impede the translation process, particularly for newly admitted countries (Seljan & Pavuna, 2006a). Nowadays, fortunately, translators have fully automatic MT systems and CAT tools at their disposal (Valderrábanos, 2003). Moreover, the usage of such tools has been recommended by the Directorate General for Translation of the European Commission (DGT), which is European Commission’s in-house translation department (Seljan & Pavuna, 2006a). Depending on their needs, translators can opt for MT or CAT tools.

MT was first conceived as a technology that significantly speeds up the translation process and offers human-like quality translations. Soon, it became clear that such a goal is far-fetched (Valderrábanos, 2003), which led to the development of TM technology. Current computational models of MT are limited to tasks for which rough translations are adequate, tasks where human post-editors are used and tasks limited to sub-domains in which fully automatic high quality translations are achievable (FAHQT) (Jurafsky & Martin, 2009). TMs, on the other hand, exploit machine memory for storing translated segments in order to reuse them in future translations. Their usability, therefore, increases with the size of the stored data.

As Croatia’s EU accession negotiations are underway, it is high time for the development of Croatian language tools and resources. This paper explores the benefits of using TMs, in particular Atril’s Déjà Vu X (DVX) system, and presents the results of a study in which English and Croatian are source and target languages, respectively.

## Translation memory

TM technology is based on the notion of reuse of previously translated segments. It is usually integrated into a system which has a terminology management module and a lexicon (Valderrábanos, 2003). This technology does not aim at replacing humans, but rather at enhancing the human translation process. A TM can be defined as a database which stores corresponding source and target language translations, called translation units (TUs).

### Approaches

There are two TM implementation approaches. Despite the differences in implementation, TMs are designed with the common purpose of storing previously translated material in an organized way, in order to present it to the user in future translations (Gow, 2003).

*Sentence-based approach*

A sentence-based approach divides source and target language texts into corresponding TUs, which can be sentences, titles, subtitles or list entries. TUs are stored in a database and retrieved in future translations in cases of identical or similar TUs in new source texts. Sentence units are easy to identify if they start with capital letters and end with full stops (Gow, 2003). However, abbreviations or full stops which are not at the end of sentences pose problems. These problems can be solved by defining new sentence delimitation rules (Déjà Vu, 2009). The main benefit of the sentence-based approach, compared to a character-string-in-bitext-based approach, is that exact matches are more likely to be relevant because sentence-based TMs represent an extreme form of high precision, low recall search (Simard & Langlais, 2000). Fuzzy matching algorithms, on the other hand, are based on statistical models of similarity. Since these models are only loose approximations, the matching algorithms sometimes create useless matches, known as 'noise', or fail to generate matches, the phenomenon known as 'silence' (Bowker, 2002 in Gow, 2003).

*Character-string-in-bitext(CSB)-based approach*

A CSB-based approach involves storing of source texts and corresponding translations in a database. The resulting texts are called bitexts. Bitexts can be used for preparatory background reading. In this approach, identical character strings of any length are recognized and reused (Gow, 2003). Working with sentence segments, instead of entire sentences, has its advantages (Simard & Langlais, 2000). It enables identification of identical sentence segments or even several consecutive identical sentences at once (Macklovitch & Russell, 2000 in Gow, 2003). 'Noise' phenomenon is still present, but this time as a result of finding unreliably small matches. 'Silence', on the other hand, occurs because there is no support for fuzzy matching. One of the disadvantages of this approach is that internal repetitions have to be recycled exclusively through terminology databases, because only entire translations are added to databases (Gow, 2003).

**Advantages**

Two major advantages of TM technology are consistency and speed. Consistency is of crucial importance in non-literary texts, for example software and hardware manuals (Valderrábanos, 2003), or business, legal, scientific and technical texts (Gow, 2003). These texts are highly repetitive. The longer they are, the more likely they are to contain repetitive content (Austermühl, 2001 in Gow, 2003). Repetition can occur, not only internally, but also across several texts in the same domain (Gow, 2003). Moreover, software documentation, besides being highly repetitive, is subject to frequent version updates. It is thus an ideal candidate for exploiting TM benefits (Bruckner, 2001). Consistency is es-

pecially important in cases where several translators work on the same project and share the same TM on the network (O'Brien, 1998 in Gow, 2003).
Speed is important, regardless of the domain, because globalization has brought forward endless translation demands (Valderrábanos, 2003). Using TMs in the translation process implies cost reduction. For example, the translation process can be started as soon as the first draft of the document to be translated is obtained. Furthermore, translation vendors can lower prices and thus earn more contracts (Gordon, 1997 in Gow, 2003). On the other hand, freelance translators can save up their valuable time or increase their earnings by increasing the translation speed (Gordon, 1996 in Gow, 2003).
In addition, using TMs preserves original page layouts in translated documents, because formatting information is hidden in embedded codes (Seljan & Pavuna, 2006a).
Finally, TMs are usually integrated into systems which have tools for building dictionaries (Webb, 1998) and reporting detailed statistics on internal and external repetition and word counts. These tools can help project managers in scheduling localization products (Esselink, 2000).

**Disadvantages**
Although TMs bring a lot of advantages, there are also some limitations of their usage.
First of all, using TMs implies initial decrease in productivity because translators need to master the environment (Webb, 1998). The odds of finding quality matches increase with the size of TMs (Gow, 2003). Moreover, the beneficiary effects of using TMs are felt only on repetitive texts (Valderrábanos, 2003). Therefore, cost-effects of investing additional time into importing existent translations through the process of alignment should be calculated (Seljan & Pavuna, 2006a).
Although, according to Esselink (2000), TMs indisputably save time, regular database maintenance is time-consuming (Austermühle, 2001 in Gow, 2003).
Furthermore, source texts need to be in digital form (Gow, 2003) and suitable file formats because not all formats are supported since TM systems require filters to preserve formatting. As an effect, they are usually bundled only with filters for most commonly used formats (Esselink, 2000).
TMs affect quality of the translation because, by using them, translators tend to avoid using anaphoric or cataphoric references in order to make segments more 'universal'.
Seljan and Pavuna (2006a) add lack of language knowledge and context insensitivity to the list of drawbacks and point out additional software, maintenance and education costs.
There are also other concerns with regard to using TMs. For example, it is questionable whether translators should be paid differently for identical and

fuzzy matches recovered from TMs. Furthermore, the ownership of final TMs is also unclear.

## Déjà Vu X

DVX is a very powerful and adaptive CAT system which integrates several CAT tools – lexicon, terminology database, TM, alignment module, etc. The first version of this system appeared in 1993. DVX is an example of a hybrid approach. Matches are ranked into the following categories: perfect/exact match, fuzzy match, guaranteed match (there is overlapping of neighbouring TUs as well) and assembled from portions match. It features auto-search, assembling, propagation and pre-translation. Besides, it gives a detailed statistical analysis of source and target language texts. Terminology database is the only database that needs to be manually filled and it allows linguistic enrichment of inserted terms. A list of lexicon entries is automatically built. However, entries which are to be included need to be manually translated. DVX combines modern TM technology with example-based machine translation (EBMT). EBMT implies translation by analogy and enables combining several segments into one translation segment (Déjà Vu, 2009).

According to a survey, which included 699 translators from 50 different countries, DVX was the second TM on the list of popularity, meaning that 61% of translators were acquainted with the system. On the usage list, it was the fourth. To be precise, there were 23% of translators using the system (Laugodaki, 2006). The statistics show that the system is preferred by translators with higher level of information literacy. DVX scored better than competitors' systems in functionality, efficiency, speed, reliability, price, and usability. According to the survey, it also had better customer support.

## Experimental study

The feature to be examined in this study was the speed of the translation process, with the goal of measuring the time difference between human and TM-based translation speeds (Bruckner, 2001). The following parameters were taken into account:

- Measure (minutes),
- Evaluation procedure (comparing times needed by each translator to deliver translation),
- Score (time needed for the translation process),
- Metric (faster / slower),
- Languages (English – source language, Croatian – target language),
- Text type (hardware documentation), and
- System used (Déjà Vu).

The study was conducted by two expert translators, who translated three excerpts from Kodak's, Nokia's and Canon's digital camera user manuals, re-

spectively. Each excerpt was two standard pages long and contained information on battery and relating equipment care and maintenance. Therefore, each excerpt contained terms and phrases from the same domain. The excerpts were translated from English into Croatian.
The experimental study was conducted in the following phases:

1. Text selection phase
2. Preparation phase (lexical analysis of the source language texts and their technical registers)
3. Translation phase (setting up environment, translating, building up lexicon, filling terminology database)
4. Revision phase (post-editing)

**Preparation phase**
For all the excerpts, the preparation phase was performed jointly by the two translators. The lexical analysis of the source language texts (English) lasted 45, 10 and 18 minutes, respectively.

**Translation phase**
Prior to the translation process, the translator using the TM spent 5 minutes setting up the environment. The translator had some previous experience with DVX translation memory. After the text was inserted into DVX, the system calculated that the internal repetition was 6 per cent.
The translation speed in the first translation (Kodak) was identical for both translators (37 minutes).
After the first text was translated, the TM translator had to manually build up the system's lexicon, which lasted 15 minutes and included 68 entries. The inserted entries were mostly nouns in nominative singular and plural, and masculine adjectives in singular. Filling the terminology database lasted 10 minutes. Only 18 phrases were added due to Croatian rich morphological system.
The second excerpt (Nokia) first underwent the pre-translation process. Besides the exact matches, fuzzy matches and parts assembled from portions were also allowed. After processing 69 source language sentences, the system found 6 sentences with 1 or more fuzzy matches and 31 sentences assembled from portions, some of which are presented in Figure 1.
The traditional translation process for the second text lasted 31 minutes, while, with the help of the TM, it lasted 30 minutes. After the second text translation, 111 new entries were added to the lexicon and 13 new phrases were added to the terminology database. These processes lasted 15 and 10 minutes, respectively.
The third excerpt (Canon) also underwent the process of pre-translation prior to the translation. The system processed 41 source sentences and found 1 sentence with a unique exact match and 32 sentences assembled from portions. The system found substantially more matches than it had in the second text (Figure 2).

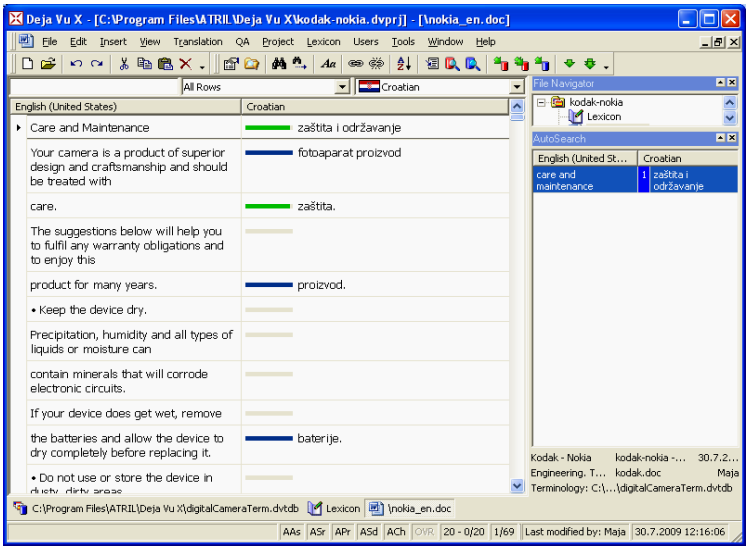



Figure 1: Matches found by DVX when the second source text was inserted

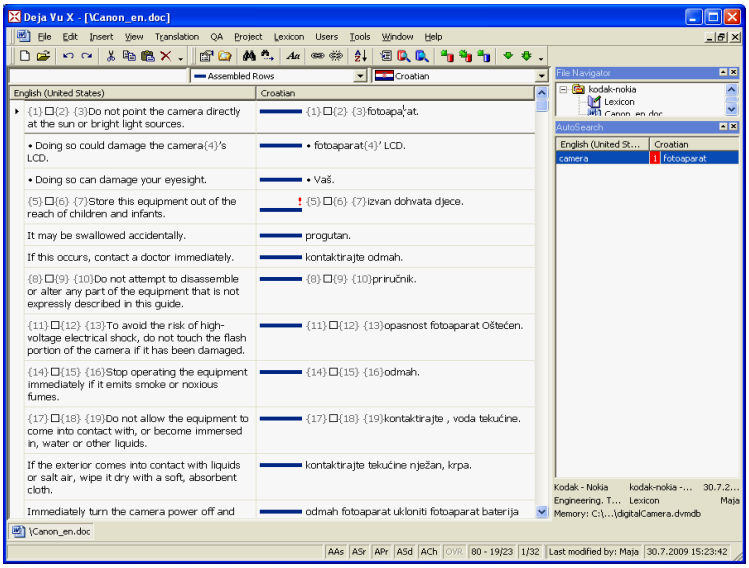



Figure 2: Matches found by DVX after the third source text was inserted

The translation phase for the third text lasted 29 minutes for the traditional translation and 23 minutes with the TM. It took 5 minutes to add 45 new entries to the lexicon, and 3 minutes to add 7 new phrases to the terminology database.

**Revision Phase**

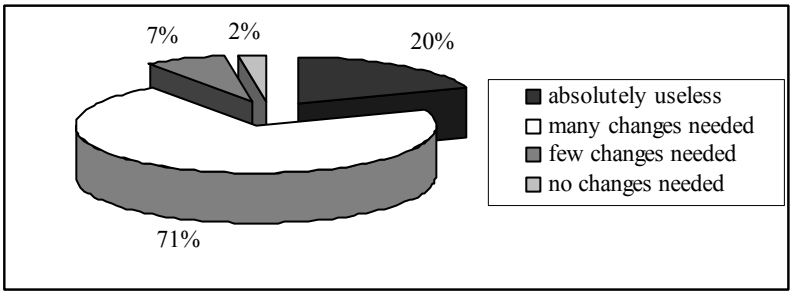
The revision phase for the three translations in the traditional translation process lasted 5, 7, and 5 minutes, respectively, while in the TM translation process it lasted 5, 8, and 5 minutes, respectively.

**Evaluation**

According to the automation level, evaluation methods can be divided into automatic and manual. Although time-consuming, manual methods better suit real-life system application. They can be implemented in two different ways. One way is that a translator scores each segment on a 1-4 scale from "absolutely useless" to "no changes needed". The other way includes modifying each resulting segment into acceptable translation and counting post-editing steps needed (Hodász, 2006).

The results of a 1-4 scale test performed on the third translation are presented in Chart 1. Since the TM is still under development, these are only initial results. With the growth of the TM, the increase in the percentage of segments classified as 'few changes needed' or 'no changes needed' can be expected.

Chart 1: Effectiveness of TM




**Discussion**

The search process gives the highest priority to the TM matches and the lowest to the lexicon matches, with the terminology database matches in between. Nevertheless, the user is supplied with all the matches in a separate window, which enables them to choose the most appropriate match for the given context.

Most of the problems with DVX regard morphology and word order. A word extracted from the lexicon which is a masculine adjective needs to be changed into feminine or neuter in order to match the context syntactically. The same is valid for verbs, since the lexicon contains their infinitive form.

Furthermore, words are extracted in order of appearance and they often need to be rearranged because of the syntactic rules of the target language.

There are also capitalization issues which need to be resolved. For example, if there is a word which was the first word in a previously stored segment, it remains capitalized regardless of its position in subsequent occurrences.

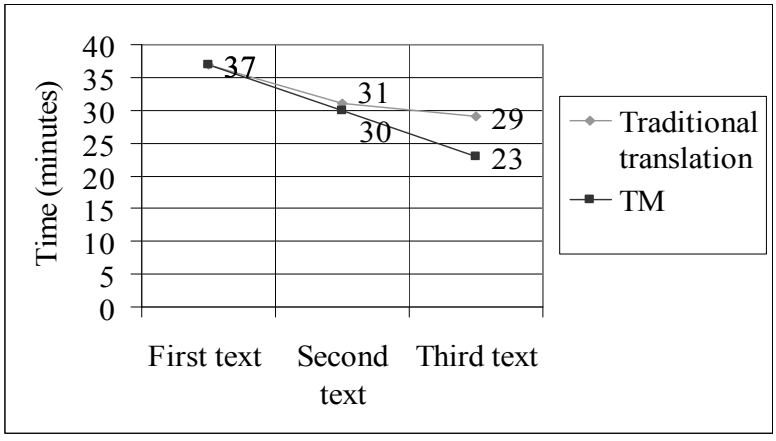
Additionally, punctuation is also retained if the word found in one of the resources is followed by a punctuation mark.
However, one of the main advantages of using the TM is the user interface, which enables the translator to see parallel sentences or units in the same row. That eases the translation process because the translator does not have to spend a lot of time inserting or deleting units of the source text and scanning through both texts.

*Speed of translation process*
Since the main advantage of any TM system should be speeding up the translation process, here follow the results of this case study. Table 1 presents time (expressed in minutes) needed for each phase (preparation phase has been omitted because it was done jointly by the two translators).

Table 1: Time spent for each phase

| | | Translation phase (minutes) | Lexicon and terminology database filling (TM) (minutes) | Revision phase (minutes) | Total (minutes) |
|---|---|---|---|---|---|
| 1st text | Traditional | 37 | / | 5 | 42 |
| | TM | 37 | 25 | 7 | 69 |
| 2nd text | Traditional | 31 | / | 5 | 36 |
| | TM | 30 | 25 | 8 | 63 |
| 3rd text | Traditional | 29 | / | 5 | 34 |
| | TM | 23 | 8 | 5 | 36 |

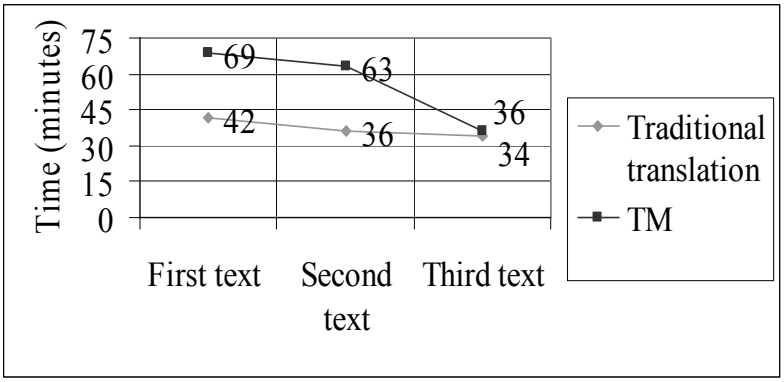
Chart 2: Time spent in translation process




It is evident that the TM translator becomes faster than the traditional translator in the second, and even more, in the third translation. Since the TM was empty prior to the first translation, the translation phase of the first excerpt was of the same length for both translators. As DVX grows in size, the TM translator be-

comes increasingly faster than the traditional one (Chart 2). The difference in speed is the most obvious in the third translation, where the TM translator is 6 minutes faster.
Taking into account the time which the TM translator needs to spend in the process of filling the lexicon and the terminology database, it is evident that using TM systems can be somewhat time-consuming in the beginning (Chart 3).

Chart 3: Total time for traditional translation and TM




Nevertheless, it is quite plausible that the time which the TM translator spends in filling the lexicon and terminology database gradually decreases because both databases have smaller number of entries to be added in, while the time needed for human translation remains the same. Therefore, the time difference for the TM translator and the traditional translator becomes less significant.

## Conclusion

This paper presents a detailed analysis of the benefits of the Atril's Déjà Vu X translation memory system, with its time-saving implications based on the reuse of previously stored translation units.
In this case study, segments from three different digital camera user manuals are translated, with the aim of presenting how TMs ensure consistency in non-literary texts. After measuring the time difference between human and TM-based translation speeds and taking into account 6 parameters, those being measure, evaluation procedure, score, metric, languages and text type, it can be concluded that TMs speed up the translation process, especially in later phases, i.e. when texts show certain level of local or global repetition. Nevertheless, even though TMs unquestionably save time, regular lexicon and terminology database maintenance is still time-consuming. However, this task does not take as much of translator's time as databases grow in size. On the other hand, lack of knowledge asks for intensive work in the revision phase. Even so, TMs represent a valuable resource, especially when several translators work in the same

domain, and aim to produce fast, consistent and professional-quality translations.

## Acknowledgments

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grants No.130-1300646-0909 and No.009-0361935-0852.

## References

Bruckner, Christine; Plitt, Mirko. Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input. // *Proceedings of the VIII MT Summit*. Spain, Santiago de Compostela, 2001

Déjà Vu – Translation memory and productivity system. http://www.atril.com/ (July 25, 2009)

Esselink, Bert. A Practical Guide to Localization. Amsterdam/Philadelphia : John Benjamins Publishing Company, 2000.

Gow, Francie. Metrics for Evaluating Translation Memory Software. M.A. thesis. University of Ottawa, Canada, 2003. http://www.chandos.ca/Metrics_for_Evaluating_Translation_Memory_Software.pdf (August 3, 2009)

Hodász, Gábor. Evaluation Methods of a Linguistically Enriched Translation Memory System. // *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Italy, Genoa, 2006, 2044-2047

Jurafsky, Daniel; Martin, James H. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. New Jersey : Pearson education, 2009

Lagoudaki, Elina. Translation Memories Survey 2006: Users' perceptions around TM use // *Proceedings of the ASLIB International Conference Translating & the Computer 28*. UK, London, 2006.

Seljan, Sanja; Pavuna, Damir. Translation Memory Database in the Translation Process. // *Proceedings of the 17th International Conference on Information and Intelligent Systems IIS 2006* / Aurer, B.; Bača, M. (ed.). Croatia, Varaždin : FOI, 2006a, 327-332

Seljan, Sanja; Pavuna, Damir. Why Machine-Assisted Translation (MAT) Tools for Croatian? // *Proceedings of the 28th Conference on Information Technology Interfaces – ITI 2006*. Croatia, Cavtat, 2006b, 469-474

Simard, Michel; Langlais, Philippe. Sub-sentential Exploitation of Translation Memories. // *LREC 2000 Second International Conference on Language Resources and Evaluation.* Greece, Athens, 2000

Valderrábanos, Antonio S.; Esteban, José; Iraola, Luis. TransType2 - A New Paradigm for Translation Automation. // *Proceedings of the MT Summit IX*. New Orleans, Louisiana, 2003, 498-501

Webb, Lynn E. Advantages and Disadvantages of Translation Memory : A Cost/Benefit Analysis. M.A. thesis. Monterey Institute of International Studies, Monterey, California, 1998. http://www.tradulex.org/Bibliography/Webb.htm (October 31, 1998)

# Transcription and Transliteration in a Computer Data Processing

Greta Šimičević
Library, Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
gsimicev@ffzg.hr

Ana Marija Boljanović
Croatian Standards Institute
Ulica grada Vukovara 78, 10000 Zagreb, Croatia
ana.marija.boljanovic@hzn.hr

**Summary**

*This paper shows different methods of transliteration of Cyrillic characters into Latin characters in bibliographic databases. The differences between transliteration and transcription have been presented. The paper draws attention to the problems that Latin databases are faced in the process of transliteration Cyrillic characters into Latin. Several examples of bibliographic databases, which apply different rules and standards – have been searched (the British Library, the Library of Congress, libraries in Croatia etc.). Searches were made according to certain criteria and the results have been presented. As the result of research there is a rate of application of certain existing rules and standards, which indicates that there is a need for standardization in this area and implementation of a unified system for transliteration at the international level.*

**Key words:** conversion, transliteration, transcription, characters, digraphs, diphthongs, diacritical marks, standards, database

## Introduction

In the course of the second half of 20th century computer catalogues, i.e. bibliographic databases replaced library catalogues on cards. Digitalization of library operations changed the way libraries function, which was most evident in the area of library material processing, i.e. data processing that it involves, its search possibilities and provision of user access to this data. During the said period little attention was paid to indexing and standardization, or to index languages and general data entry standards, as their purpose was barely understood. Databases were continually filled with data and became larger, clumsier and increasingly disorganized, resulting in maintenance problems, problems in data management and search problems. These increased even further once data-

bases became accessible to a large community via Internet. As the number of stored data continually increases, it has become essential to manage information systems in a precise manner, i.e. to select appropriate information language for storing and searching.

Amongst various other problems, bibliographic databases also faced problems of transcription and transliteration. Given that the subject of transfer of different scripts from one into another is both very broad and very demanding, in this paper we specifically cover the transfer of Russian Cyrillic characters into Latin, stipulating individual examples in practice of global bibliographic databases, with a particular overview of Croatian practice. We will try to specify problems that we discovered, possible methods for their solution through standardization, and certain discrepancies from international standards, as well as to explain the reasons for such practices and divergences.

## Transcription and transliteration: possible procedures for transfer of one script into another

Transcription is a transfer of pronunciations and phonemes of one language into graphical system for phonetic recording of phonemes of another language, i.e. pronunciation of words in one language adapted to pronunciation in another language and to this other language's vocalization. Transcription respects phonetic characteristics of different languages and national variants, and need not necessarily involve transfer of one script into another, but may concern graphical transcript of words from one language into another even in cases when both language systems use the same script.[1] The transcription process is connected to a significantly narrower space than the global one, for it is often limited by a language system; in other words, by the rules (orthography) of the specific language system within which the process of transcription is being carried out. The most frequent differences between systems lie in the diverse phonetisation of certain graphemes that we transfer. For example: surname of the Russian author *Цветаева* looks in transcription of different language systems as follows: *Tsvetaeva* (Eng.), *Zwetajewa* (Ger.), *Cvetaeva* (Ita.), *Cvjetajeva* (Cro.), *Tswetaewa* (Pol.). The differences appear due to existence, or else lack of, specific graphemes and phonemes in different systems. For example, Latin Slavic languages have the diacritical characters *č, ž, š*, which other Latin language systems do not have so, they transfer Cyrillic characters for this phonetisation: *ч, ж, ш* into *ch, zh, sh*…etc. as well as Latin versions of these phonemes. Concurrently, Slavic language systems will transfer a diacritical character from another Slavic language as a diacritical character, for they both contain relevant

[1] Badurina, Lada, Ivan Makarović i Krešimir Mićanović. *Hrvatski pravopis.* Zagreb: Matica hrvatska, 2007., str. 221.

graphemes and phonemes as such. Similarly, German umlaut characters, such as for example *ü*, are transferred into Croatian language in a similar way, by applying phonetisation of the Croatian language system into *ue*, for umlaut characters as such do not exist in Croatian language system.

Russian surname *Щедрин* is another example, and in Croatian and Czech it is transferred as *Ščedrin*, in Polish it is *Szczedrin*, in English *Shchedrin*, in French *Chtchedrine*, in Dutch *Sjtsjedrin* and, in German *Schtschedrin*. This example shows that sometimes as much as seven letters of Latin alphabet are needed for the Cyrillic character *щ*, which makes it difficult to establish international catalogues or lists. This example also shows the issues concerning all four Russian Cyrillic diphthongs *я, ю, ё, щ* for which there are no graphemes in Latin alphabet. It is necessary to mention that a similar problem occurs in transfer of Latin diagraphs in other Latin language systems, or else into Cyrillic; such is the case for example with graphemes *dž, lj, nj, sch, ch*, as well as already mentioned diacritical characters and other special characters of Slavic and non-Slavic language systems that are not commonly accepted. Examples of such special characters in Russian Cyrillic are characters with strong or soft phonetisation, such as *э, ы, ь* and characters that existed throughout history of language; hereby noting that many Latin and other Cyrillic scripts abound with similar examples. There are recommendations on the global level that we may, but also need not accept. One of recommendations, for example, is to transfer the words from one language system into another in the same way that the language community of the former would transfer their words. However, this is primarily valid for idiomatic scripts.

From the above-mentioned it would arise that there would be as many transcription rules in the world as there are languages, so such transfer process from one script into another could not bring about uniformity on the global level.

Transliteration, on the other hand, is a transfer of characters (graphemes) of one script into characters of another script (e.g. from Glagolitic into Latin, from Cyrillic into Latin, etc.). This should occur almost automatically and both ways, so that the regress into original text should be possible. But, clearly – with 25 or 26 globally accepted characters in Latin script it is not possible to transfer 40 or 50 Cyrillic characters without occasional recourse to combinations of the usual Latin graphemes for the special characters.[2] The same symbols should not be used in transliteration of different characters in any language and, using two or more characters for one character is only acceptable when there is no better solution. As a possibility, transliteration has, on the global level, proven to be a much better procedure than transcription concerning harmonization of data entry into databases from other scripts into Latin, which brought about attempts at creating various international rules for transliteration.

[2] British standard BS 2979:1958. Transliteration of Cyrillic and Greek characters, BSI 1958.

Even though the definition of the procedure in itself should guarantee a simple and unambiguous solution of the problem of transfer in different scripts from one into another – given that the procedure itself should not be bound to any rules of various language traditions – this is not the case and in this procedure the issues of diversity in the use of graphemes and phonemes in language and script traditions becomes quite obvious.

Researching through various international bibliographic databases we found out numerous "inconsistencies" in application of different transliteration rules that were agreed on the global level. Examples of such "inconsistencies" in application of international transliteration rules are many but they generally boil down to accepted procedures linked to existence of large global language groups and their language or script traditions, which resulted in emergence of variants of international rules at the level of large language groups. This realization brings us inevitably to the application of transcription process within transliteration and, to the entire, already mentioned, body of issues that such practice brings about during transfer from one script to another, as well as to the hybridism of transcription and transliteration procedures. In transfer of Russian Cyrillic into Latin this is particularly manifest in the transfer of diphthongs *я, ю, ё, щ* and diacritical characters existent only in Slavic languages *ч, ж, ш* into Latin script of non-Slavic languages. There is also the problem of transfer of Russian graphemes such as *ц, х*, which are, for example, under a strong transcription tradition within databases of English speaking areas transferred as *ts, kh*, within German as *z, h/ch* (depending on the position of the character within a word)..., etc. Hence, we have – for example – transliteration procedure for one language group and transliteration procedure for another language group (e.g. transliteration procedure from Cyrillic into Latin for Slavic languages and transliteration procedure for non-Slavic languages).[3]

## Standardization and other systems in the area of transliteration

In order to facilitate and improve communication and data and information exchange standards for transliteration of all international scripts into Latin script were developed by International Organization for Standardization (ISO) (see Table 1).

The fact that it has 162 member states speaks best about this international organization's significance. Experts from its member states contribute to the development of standards and standardization work and published standards are mostly adopted by the member states. As for international standard ISO 9 for transliteration of Cyrillic characters into Latin characters even back in 1954 the first issue of this standard was published. From the very beginning ISO 9 had the status of recommendation, established a provision within the text itself that

[3] As the result of research of the foreign bibliographic databases

the international standard may, in transliteration from Cyrillic into non-Slavic language, be amended or replaced with the national system accepted as usual practice. Nowadays, the third edition issued in 1995 is widely adopted by most European countries but with certain national modifications (e.g. Denmark, Deutschland, France, Italy, Poland, Russian Federation, Serbia, Sweden, Turkey, United Kingdom).

Table 1. List of valid international standards for transliteration

| Document identifier | Title (English) | Publication date |
|---|---|---|
| ISO 9 | Information and documentation – Transliteration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages | 1995-02-00 |
| ISO 233 | Documentation; Transliteration of Arabic characters into Latin characters | 1984-12-00 |
| ISO 233-2 | Information and documentation; transliteration of Arabic characters into Latin characters; part 2: Arabic language; simplified transliteration | 1993-08-00 |
| ISO 233-3 | Information and documentation – Transliteration of Arabic characters into Latin characters – Part 3: Persian language – Simplified transliteration | 1999-01-00 |
| ISO 259 | Documentation; Transliteration of Hebrew characters into Latin characters | 1984-10-00 |
| ISO 259-2 | Information and documentation – Transliteration of Hebrew characters into Latin characters – Part 2: Simplified transliteration / Note: Corrected and reprinted in 1995-07 | 1994-12-00 |
| ISO 843 | Information and documentation – Conversion of Greek characters into Latin characters / Note: Corrected and reprinted in 1999-05 | 1997-01-00 |
| ISO 3602 | Documentation; romanization of Japanese (kana script) | 1989-09-00 |
| ISO 7098 | Information and documentation; romanization of Chinese | 1991-12-00 |
| ISO 9984 | Information and documentation – Transliteration of Georgian characters into Latin characters | 1996-12-00 |
| ISO 9985 | Information and documentation – Transliteration of Armenian characters into Latin characters | 1996-12-00 |
| ISO 11940 | Information and documentation – Transliteration of Thai | 1998-06-00 |
| ISO 11940-2 | Information and documentation – Transliteration of Thai characters into Latin characters – Part 2: Simplified transcription of Thai language | 2007-05-00 |
| ISO/TR 11941 | Information and documentation – Transliteration of Korean script into Latin characters | 1996-12-00 |
| ISO 15919 | Information and documentation – Transliteration of Devanagari and related Indic scripts into Latin characters | 2001-10-00 |

Source: Perinorm International Database, British Standards Institute, 2009

Croatia had adopted almost all international standards for transliteration without any modifications. Even though ISO 9:1995 is adopted as the national standard in the Republic of Croatia, in practice of data entry into bibliographical data-

bases this standard is not applied consistently; entry practice is closer to ISO R 9:1968. Furthermore, the table which was given in ISO/R 9:1968 representing international system is in fact extended transliteration system for Serbian Cyrillic into Croatian Latin (see Table 2).

Table 2 Transliteration of the modern Russian alphabet (extracted from table ISO/R9)



TABLE 1. — Transliteration of the modern Russian alphabet

| Russian | | | | | Transliteration | Examples | | |
|---|---|---|---|---|---|---|---|---|
| Letter numbers | printed | | written | | | | | |
| 1 | а | А | *а* | *А* | a | адрес | — | adres |
| 2 | б | Б | *б* | *Б* | b | баба | — | baba |
| 3 | в | В | *в* | *В* | v | вы | — | vy |
| 4 | г | Г | *г* | *Г* | g | голова | — | golova |
| 5 | д | Д | *д, д* | *Д* | d | да | — | da |
| 6 [1]) | е (ё) | Е (Ё) | *е (ё)* | *Е (Ё)* | e (ë) | ещё | — | eščë |
| 7 [2]) | ж | Ж | *ж* | *Ж* | ž | журнал | — | žurnal |
| 8 | з | З | *з, з* | *З* | z | звезда | — | zvezda |
| 9 | и | И | *и* | *И* | i | книга | — | kniga |
| 10 [2]) | й | Й | *й* | *Й* | j | первый | — | pervyj |
| 11 | к | К | *к* | *К* | k | как | — | kak |
| 12 | л | Л | *л* | *Л* | l | липа | — | lipa |
| 13 | м | М | *м* | *М* | m | муж | — | muž |
| 14 | н | Н | *н* | *Н* | n | нижний | — | nižnij |
| 15 | о | О | *о* | *О* | o | общество | — | obščestvo |
| 16 | п | П | *п* | *П* | p | пара | — | para |
| 17 | р | Р | *р* | *Р* | r | рыба | — | ryba |
| 18 | с | С | *с* | *С* | s | сестра | — | sestra |
| 19 | т | Т | *т, т, т* | *Т* | t | товарищ | — | tovariš |
| 20 | у | У | *у* | *У* | u | утро | — | utro |

4

Source: ISO R/9:1968

In addition to standards there are also several global systems, which establish transliteration rules concerning practice of data entry into computer databases. Table 3 presents parallel transliteration systems by several different rules and/or recommendations, based upon Russian Cyrillic as example. Data entry into bibliographical databases in Croatia actually matches best to UN transliteration rules (see Table 3), with the exception of character *ë*, which is in current prac-

tice of Croatian bibliographical databases transliterated into *e*, and not into *ë* as required by the rule, probably due to the simple reason that this character is also often presented in Russian graphics with the grapheme *e*.[4] Such transliteration method can be traced back to the Rulebook and manual for preparation of alphabetical catalogues of Eva Verona from 1986, which was issued prior to the acceptance of international standards at the level of Republic of Croatia (see Table 3).

Table 3. Parallel overview of several transliteration rules and standards (extracted from Transliteration table)

| Cyrillic | Scholarly | ISO/R 9:1968 | GOST 1971 | UN | ISO 9:1995; GOST 2002 | ALA-LC | BGN/PCGN |
|---|---|---|---|---|---|---|---|
| А а | a | a | a | a | a | a | a |
| Б б | b | b | b | b | b | b | b |
| В в | v | v | v | v | v | v | v |
| Г г | g | g | g | g | g | g | g |
| Д д | d | d | d | d | d | d | d |
| Е е | e | e | e | e | e | e | e, ye † |
| Ё ё | ë | ë | yo | ë | ë | ë | ë, yë † |
| Ж ж | ž | ž | zh | ž | ž | zh | zh |
| З з | z | z | z | z | z | z | z |
| И и | i | i | i | i | i | i | i |
| Й й | j | j | j | j | j | ĭ | y |
| К к | k | k | k | k | k | k | k |
| Л л | l | l | l | l | l | l | l |
| М м | m | m | m | m | m | m | m |
| Н н | n | n | n | n | n | n | n |
| О о | o | o | o | o | o | o | o |
| П п | p | p | p | p | p | p | p |
| Р р | r | r | r | r | r | r | r |
| С с | s | s | s | s | s | s | s |
| Т т | t | t | t | t | t | t | t |
| У у | u | u | u | u | u | u | u |
| Ф ф | f | f | f | f | f | f | f |
| Х х | x | ch | x | h | h | kh | kh |
| Ц ц | c | c | cz, c | c | c | t͡s | ts |
| Ч ч | č | č | ch | č | č | ch | ch |
| Ш ш | š | š | sh | š | š | sh | sh |
| Щ щ | šč | šč | shh | šč | ŝ | shch | shch |

[4] As the result of research of the Croatian bibliographic Databases

Table 3 cont.

| Cyrillic | Scholarly | ISO/R 9:1968 | GOST 1971 | UN | ISO 9:1995; GOST 2002 | ALA-LC | BGN/PCGN |
|---|---|---|---|---|---|---|---|
| Ъ ъ | " | " | " | " | " | " | " |
| Ы ы | y | y | y' | y | y | y | y |
| Ь ь | ' | ' | ' | ' | ' | ' | ' |
| Э э | è | ė | eh | è | è | ė | e |
| Ю ю | ju | ju | yu | ju | û | i͡u | yu |
| Я я | ja | ja | ya | ja | â | i͡a | ya |
| Pre-1918 letters | | | | | | | |
| I i | i | i | i, i' ** | ĭ | ì | ī | – |
| Ѳ ѳ | f | ḟ | fh | ḟ | f̀ | ḟ | – |
| Ѣ ѣ | ě | ě | ye | ě | ě | i͡e | – |
| V v | i | ẏ | yh | ẏ | ỳ | ẏ | – |

Source: http://en.wikipedia.org/wiki/Romanization_of_Russian#Transliteration_table

## Computer systems and transliteration

The most important task in data processing is how to store all information contained in a unit of the material in such a way as to make it easily searched and successfully found at the request of a user. Computer systems differentiate various data by distinctiveness of characters. Thus it can happen that one and the same information entered into the computer system via both the transcription process and transliteration process would, in fact, signify two different pieces of information for the computer. If at the same time several different rules are applied for transcription and transliteration, we could from one semantically identical data item create, as far as computer is concerned, a multitude of different data items. This is particularly important for the entry of normative data, and for indexing, which in concrete terms of bibliographical databases represents data on authoring, subject, etc. Subsequent search results will depend exclusively on that, which rules have been applied to enter specific data into the computer system, and which rules have been applied to define the search. Researching through the largest and the most prominent bibliographic databases in the world and in the region we have established that in practice ISO international standard for transliteration from Russian Cyrillic into Latin is never completely and fully implemented. Most frequently this standard represents merely a basis upon which other rules generally linked with the transcription process are built, or which are imposed by certain large language groups and their language traditions. In this way, what may be called national variants of ISO standard emerged, which are then consistently applied in the majority of researched databases. Within narrow national levels such script transfer method functions very well, for it is familiar to this national group's users. But, the problem occurs on the global level, given that databases have now outgrown the national

level. Global user does not have access to the data transferred into Latin script through the transcription process; or even through transliteration process heavily influenced by the transcription tradition, especially if the information on the transfer method is lacking. Hence, a problem has been noted concerning search where transparent information on the rules based upon which transliteration has been implemented within certain database is lacking. As it were, this information does in fact exist, but it is hidden within encoded fields of entry and it is inaccessible for regular user.

There are numerous user oriented Internet pages covering issues of transliteration, and offering one-stop-shop information on various rules for transliteration of all scripts, Russian Cyrillic included. In addition to tabular overview of standards, many also have built-in software for automatic transliteration of Cyrillic characters, as well characters of other scripts into Latin (and vice versa) by different standards.[5]

## Conclusion

Whilst trying to find an answer to the question why the practice on international level does not apply the single transliteration standard that exists and that has been adopted by consensus exactly for the purpose of bridging the recognized problems that arose from publication of bibliographic databases on the global level, we found that the answers are self-evident. One of the first facts is that bibliographic databases were established and were becoming larger and larger way before the problem of transfer of scripts on the global level became recognized. Subsequently adopted international standards became sort of an attack on language traditions of large groups. Additionally, it was very difficult to adopt a standard that would reconcile all language traditions. One of the larger problems is also the need to use a multitude of special characters whose perusal used to be far from simple. If today we tried to apply the unified standard, we would find numerous problems in translating the data that has already been entered. Naturally, all of this would not justify why all databases should not start applying single standard in the future, which would greatly facilitate search for the global user and would make many currently "unavailable" information accessible.

Transliteration process should serve as a unique technical aid in a transfer of characters of one script into characters of another script regardless of linguistic rules and traditions of any existing language system. Accordingly, this process should not produce a system for the original reading and writing, but the system for conversion of written sources in other written form, for its recording, storing and searching in another script, with the possibility of regress into the original

[5] Examples available through following web sites:
http://www.russki-mat.net/trans2.html
http://www.allmend-ru.de/etc/transliteration.html

system of characters. Tendency towards greater unification of standards on the global level continues to exist. Aspirations towards use of transcription process, which makes harmonization process on the wider level impossible, have been overcome and transliteration process, which is even with all transcription interventions on national levels still more homogenous than it would have been possible with transcription, has been completely accepted. This in itself facilitated user-friendly access to information that has been transferred from other scripts into Latin databases.

## References

British standard BS 2979:1958. Transliteration of Cyrillic and Greek characters, BSI 1958

Badurina, Lada, Ivan Makarović i Krešimir Mićanović. *Hrvatski pravopis.* Zagreb: Matica hrvatska, 2007.

Deutches standard DIN 1460:1982. *Conversion of cyrillic alphabets of slavic languages,* DIN 1982.

International standard ISO/R 9:1954. *International system for the transliteration of Cyrillic characters,* ISO 1954.

International standard ISO/R 9:1968. *International system for the transliteration of Slavic Cyrillic characters*, ISO 1968.

Croatian standard HRN ISO 9:1995. *Information and documentation – Transliteration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages,* ISO 1995

Katalog Nacionalna i sveučilišna knjižnice u Zagrebu http://katalog.nsk.hr/cgi- bin/Pwebrecon. cgi?DB=local&PAGE=First (2009-05-14)

Knjižnice iz sustava Cobiss Net http://www.cobiss.net/default-SR.asp (2009-06-08)

Library of Congress Online Catalog http://catalog.loc.gov/ (2009-06-17)

Parallel overview of several transliteration rules and standards, http://en.wikipedia.org/wiki/ Romanization_of_Russian#Transliteration_table (2008-02-12)

Perinorm International Database on CD ROM, British Standards Institute, 2009.

Russian standard GOST 7.79:2000 *System of standards on information, librarianship and publishing. Rules of transliteration of Cyrillic script by Latin alphabet,* GOST 2000.

Russki-mat.net. *Transliteracija russkogo alfavita: Transliteration and transcription using the latin alphabet*. http://www.russki-mat.net/trans.htm (2008-02-12)

Russki-mat.net. *Transliteracija russkogo alfavita: Automatic transliteration of Russian.* http://www.russki-mat.net/trans2.html (2008-02-12)

The European Library http://search.theeuropeanlibrary.org/portal/en/index.html (2009-05-20)

Verona, Eva. Pravilnik i priručnik za izradu abecednih kataloga. 1. dio, Odrednice i redalice. Zagreb : HBD, 1986.

# USING INFORMATION RESOURCES IN RESEARCH, EDUCATION AND PRESENTATION

Original scientific paper

# Digital Information Services of Heritage Institutions – Exploiting Potentials of Web 2.0 Technologies

Lejla Kodrić
Department of Comparative Literature and Librarianship
Faculty of Philosophy in Sarajevo
Franje Račkog 1, Sarajevo, Bosnia and Herzegovina
lejla.kodric@ff.unsa.ba

**Summary**

*As a consequence of redefinition in a broader information and service environment, information services of heritage institutions have been particularly reshaped, encountering potentials offered in the field of information services first by Web 1.0, and later, in a more intensive manner, Web 2.0 technologies. Information services of heritage institutions, due to their immanently communicative and participatory nature, are among the first heritage services that have evidently and more intensively begun using the opportunities offered by the Web 2.0 technology within the field of providing digital information services. Heritage institutions are improving their information services ever more intensively with tools such as podcasts, blogs, wikis, digital video games, Web 2.0 social networks, virtual worlds and other tools to be developed in the future. It is sure that Web 2.0 technologies make possible diverse manifestations of heritage information services in a redefined environment.*

**Key Words:** Heritage Institutions, Digital Information Services, Web 2.0 Technologies

## Towards 2.0 Information Services

As a consequence of redefinition in a broader information and service environment, information services of heritage institutions have been particularly reshaped, encountering potentials offered in the field of information services first by Web 1.0, and later, in a more intensive manner, Web 2.0 technologies. In the last few years, information services of heritage institutions, as well as heritage services in general, have gradually raised the awareness of the so called 2.0 concept, which appears to intervene in a whole range of activities „sensitive“ to technological innovations, thus opening new opportunities for transformed and advanced social practices. Information and service models of heritage institutions, even in the Web 1.0 environment, due to their communicative and par-

ticipatory nature, were a solid base on which the following phase of Web technology development, known as Web 2.0, encountered service manifestations which would become a form of protoservice for other forms of heritage services. Information services of heritage institutions, due to their immanently communicative and participatory nature, are among the first heritage services that have evidently and more intensively begun using the opportunities offered by the Web 2.0 technology within the field of providing digital information services.

## Library Instruction in the Web 2.0 Technology Environment

At the very start of redefining information services in accordance with the so called 2.0 concept, or their move towards 2.0 information services, the modified nature of library/heritage instruction began to manifest itself as one form of heritage information services that would very early "migrate" into the socially richer Web 2.0 environment. The first generation of library instruction is much less dynamic than the current, more interactive, sophisticated and emphatically multimedia library instruction within the Web 2.0 environment. This is achieved in various interactive ways, either by instructing the user through an online quiz, that is a range of answers to a set of posed questions, or by using current tools such as wiki, blogs or virtual "rooms" for real-time message exchange (chat rooms) to create an atmosphere of "classroom learning", thus creating a "space" for active information exchange among a large number of participants and information experts.

In addition to that, an insight into instruction materials which heritage institution users find attractive and useful confirms the increasingly appealing, and thus necessary use of one of the Web 2.0 original technologies – podcast. According to K. De Voe "podcast is a digital audio recording available on-line (...) the word podcasting is derived from the combination of the words iPod and broadcasting."[1] Heritage institutions are still, in most communities, in the experimental phase of implementing podcasting as a Web 2.0 tool in the field of providing heritage information services. There are still no sure indicators of advantages and/or disadvantages of their use in the heritage context. Furthermore, theoretical literature on this subject has not made a significant contribution, majority of published texts on podcast use in the field of library/heritage services being based on technical or implemental aspects, without any significant identification of advantages or potential disadvantages of their use. In other words, there are no necessary generalizations that would be helpful in better

[1] DeVoe, K. 2006. Quoted in: Jowitt, Angela L. Perceptions and Usage of Library Instructional Podcasts by Staff and Students at New Zealand's Universal College of Learning (UCOL). // Reference Services Review. Vol. 36 (2008), No. 3; pp. 312-313. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/2400360309.pdf (28 April 2009)

considered use of this type of tools in the field of library/heritage information services. Gradually awoken interest in podcasts, expressed by a few farsighted authors, has, however, confirmed that podcast as a tool of information and other heritage institution services has become an interesting technological opportunity. K. De Voe (2006) suggests that podcasts are innovative solutions to be by all means taken into consideration by libraries. L. Balas (2005) highlights the fact that podcasts should not be seen as yet another massive whim. Furthermore, A. Brooks-Kirkland (2004) and K. Graham (2005/2006) both hold the opinion that this technology should be studied so that it could be used for the advancement of library instruction programs. E.K. Eash (2006) and J. Janes (2002) both emphasize that podcasts should be adopted not only because they are yet another innovation but also because they support library goals and are good tools for performing particular tasks.[2] Heritage institutions, aware of the fact that by implementing podcasts new communication channels are opened towards new users and that the entry in "the podcast world" is an unstoppable process in which a growing number of business and educational institutions have been taking part, have increasingly explored possibilities of using the tool with the aim of ensuring a more successful service provision. In a variety of heritage institution services, podcasts have thus far proven themselves to be effective tools in providing information services of heritage institutions, frequently instruction ones. In addition to that, users are increasingly interested in enriching podcast instruction and information services with visual components added to the form of information service or information training of visual learners or the types of training that views visualization of a particular unit as a more successful way of learning it.

In an attempt to successfully provide instruction and information services, heritage institutions have continuously adopted and tested various tools. In the last few years, the so called *game technology*, for a long time used in libraries in a more traditional manner, has been actualized. Current models of using the so called serious or peer designed games in the field of library instruction, or library information services in general, are part of more general interest of various types of institutions, ranging from business to educational ones, in possibilities offered by digital video games in the field of training or promotion. In addition to that, the sustained interest in game technology in the field of providing heritage information services is a response to the fact that man has always been *homo ludens*, games certainly being "a lifelong human habit".[3] Re-

[2] See: Jowitt, Angela L. Perceptions and Usage of Library Instructional Podcasts, p. 315.

[3] Cross, Carl. Making Games Seriously: Creating a Peer Designed Video Game for Use in Library Promotion and Instruction. // Library Review. Vol. 58 (2009), No. 3; p. 215. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/0350580305.pdf (6 May 2009)

search has also shown that, especially in highly developed societies, "playing videogames has become equally habitual as watching television, movies or reading books."[4] The complete incorporation of videogames in educational or customer service processes within heritage institutions should not come as a surprise, taking into account that "early videogames come from graduate laboratories of prestigious American universities, their primary nature being educational."[5] Although studies of the use of traditional games in the field of providing library/heritage services are rare or insufficient, heritage institutions are certainly those having a developed tradition of providing part of their services through game technologies. Taking other types of institutions as a model, heritage institutions use videogames in order to provide important services such as information as well as recreation ones. In fact, due to the development of digital technologies or more advanced possibilities of videogames in digital environment, heritage institutions have a great opportunity to serve and teach the user through a synergy of information and recreation tasks in an entertaining and relaxing way. Information services of heritage institutions can gain a great deal of advantage from videogames that incorporate certain objects or learning units in the game environment.

## 2.0 Reader's Advisory Services

2.0 reader's advisory services, as part of reference services, together with other related services of user education such as instruction service, have undergone a significant change within the current information environment. 2.0 reader's advisory services, as part of 2.0 library services or 2.0 information services have very interesting manifestations in the current environment. Their redefinition is a consequence of reader's advisory service redefinition that has appeared in the commercial, non-heritage environments, thus evidently becoming competition to heritage services of this kind. LibraryThing[6], Shelfari[7], Goodreads[8] and

[4] See: BBC New Media Research. Gamers in the UK: Digital Play, Digital Lifestyles. http://open.bbc.co.uk/newmediaresearch/files/BBC_UK_Games_Research_2005.pdf (7 May 2009)

[5] Cross, Carl. Making Games Seriously, p. 216.

[6] LibraryThing is a web application of online social cataloguing used by book lovers to organize personal book collections. LibraryThing helps in creating catalogues of personal collections modelled on library catalogues. Since cataloguing takes place online, it is a joint cataloguing at the same time. LibraryThing connects people according to books they read and book reviews they share. In fact, LibraryThing enables unit tagging as well as contributions in the form of their broader reviews and evaluations. Viewing collections of other users of the tool is also possible, based on similarities of archived materials or books tagged by equal descriptors. In that way a contribution is made to the entire community of LibraryThing users. See: http://www.librarything.com (10 May 2009). The appearance of web applications for online social cataloguing such as LibraryThing has influenced the library/heritage community. Although LibraryThing as a web tool has been available since 2005, the early examples of evident connection of library catalogues and LibraryThing have been present since 2007 in the form of the project called LibraryThing for Libraries (LTFL). LTFL is based on using the data provided by personal contributions of

Literature Maps[9] offer services which are traditionally in the domain of heritage institutions, in ways relevant to and highly convenient for present day users. Web 2.0 technologies open up new possibilities in the field of reader's advisory services. Blogs and wikis are becoming spaces for providing such services. At the same time, the library catalog, traditionally closed for direct reader's advisory services, incorporating originally non-heritage tools, that is the advanced options of Web 2.0 social networks such as LibraryThing, is becoming a space for expressing user's opinions and advising readers. Abundant possibilities of the advanced reader's advisory services in the Web 2.0 environment have been confirmed by pioneer projects carried out by a large number of libraries in the field.[10] Web 2.0 have evidently contributed to highly important reader's advisory services by advancing conversation on library materials, currently held not only between information experts and users, but also among users themselves. Sources of "recommending" materials have also been upgraded and generated by various points of view, equally open to everyone. The greatest change, however, is that it has been incorporated into the library catalog, up until now exclusively accessible to information experts.

## Heritage Information Services within Web 2.0 Social Network Spaces

Since heritage institutions, in the long tradition of their existence, have always been "local gathering points" or spaces for communication and specific sort of conversation, it is not surprising that exceptionally up-to-date social networks, developed as social tools within Web 2.0 technologies, are also promising in the context of activities of such institutions. Due to their potentials of informing and connecting members within a community, heritage institutions have always

LibraryThing users. They are then used to enrich the library catalogue. LTFL is a comparatively simple way of implementing the new generation of library catalogues or the redefined WebPAC through using the user generated folksonomy in WebPAC. LTFL is at the same time a new possibility for using WebPAC that transcends exclusively traditional access points such as the author's name, title, subject and key words. The library catalogue is also enriched with annotations on library items in ways that did not exist before. See: The catalogues of The Libraries at the Claremont University Consortium Libraries that were changed using this tool and the catalogues of the first American academic libraries that have implemented LTFL. http://libraries.claremont.edu/ (10 May 2009)

[7] See: Shelfari. http://www.shelfari.com/ (10 May 2009)

[8] See: Goodreads. http://www.goodreads.com/ (10 May 2009)

[9] See: Literature Maps. http://www.literature-map.com/ (10 May 2009)

[10] See: Efforts to advance reader's advisory services in the following libraries, for instance: Ann Arbour District Library Catalog. http://www.aadl.org/catalog (12 May 2009); Carnegie Library of Pittsburgh – Teen: Updates and Recommendations from the Teen Staff here at the Main Library in Oakland. http://clpteens.blogspot.com/ (12 May 2009); Danbury Library. http://www.danbury library.org/ (12 May 2009)

functioned as multilayered, multifunctional and, above all, relevant social institutions or important social networks. The appearance of tools such as Web 2.0 will, therefore, engender another abundant possibility for the continued application of fundamental and lasting principles of heritage institutions within the redefined heritage and information service system. Heritage institutions will thus "oblige" Web 2.0 social networks by appearing in their virtual space in order to confirm their inclination to the idea of the lasting accomplishment of their mission directed towards users by different means, including penetration into the nowadays more and more interesting Web 2.0 social networks. The appearance of heritage institutions in less formal environments such as Web 2.0 social networks is a consequence of developing awareness of the importance of the so called *push* principle, to which social institutions resort in order to "impose" themselves on users in the environments in which they usually live and work. Web 2.0 social networks are also characterized by hybridization of various Web 2.0 tools, which many users find attractive. Those include real-time message exchange, enriched with multimedia components, blogging, tagging etc., which contribute to their overall popularity. Web 2.0 social networks, furthermore, allow not only multimedia enriched real-time message exchange but also a dynamic sharing and exchange of information sources among social network members in the electronic environment. Web 2.0 social networks are becoming a cultural phenomenon of today, often reserved for entertainment as well as business and educational environments, implementing relationships that virtually transcend geographical, gender, age, racial, economic and cultural boundaries. As there are numerous features heritage institutions and Web 2.0 social networks share in common (e.g. both are communication spaces), many opportunities are being opened for provision of certain heritage services within virtual spaces of Web 2.0 social networks, whose membership constantly grows. The omnipresence of Web 2.0 social networks has resulted in a large number of articles on their appearance, importance and characteristics, published in the last few years. The heritage officer community, and particularly the librarian, academic and expert ones, have written about consequences of the frequent presence of heritage institutions within Web 2.0 social networks. Certain relatively "conservative" doubts on the part of information experts are still occasionally encountered about the possibility of successful provision as well as justifiability of offering information services of heritage institutions within frequently informal and entertainment environment as certain Web 2.0 social networks are. However, the presence of heritage institutions in Web 2.0 social networks, primarily in the form of information services, is the reality of a large number of heritage institutions and a position increasingly expected and assumed by users of both heritage institutions and Web 2.0 social networks. The suspicion caused by the fear of losing the "professional" atmosphere in providing heritage information services in an informal environment is accompanied by another, greater apprehension. Namely, while providing heritage information services

within Web 2.0 social networks, a heritage information service leaves the home ground, becoming part of a completely different network. A heritage institution, or more exactly its information service, is just another helping hand lent to the user or their so called *friends*, speaking the matalanguage of Web 2.0 social networks. It is, however, up to heritage institutions to wake up to the reality of the revised information and service environment in which they are no longer the only game in town but just one participating element standing at the user's service in the environment in which the user participates in a number of social networks. In fact, it is the user around whom the entire information and service network is built up, the most up-to-date and useful services being those in his immediate surroundings. If this means that heritage institutions, in their struggle over the user's interest, for their social relevance, and finally, their survival, need to step out of their original contexts and temporarily move into the customer's context, than it becomes the reality of user service today which heritage institutions can no longer ignore. The importance of heritage institutions will not diminish if their services are offered at a "trivial" place such as MySpace or Facebook. Numerous prominent heritage institutions, having recognized the importance of active participation in networks users see as useful, attractive and entertaining, have established their presence on Facebook.
In order to keep their proximity to networks in which users take interest, heritage institutions have tested their information and service activities within the so called virtual worlds such as Second Life.[11] With growing interest in virtual worlds and the fact that an increasing number of prominent business and educational institutions[12] around the globe conduct some of their activities on Second Life, heritage institutions are establishing their Second Life presence, using this network space to test possibilities for providing some of their services, primarily information ones. In fact, information services are among the first services offered as a consequence of the presence of heritage and information institutions within virtual worlds such as Second Life, for as J. Jane suggests "give users and libraries tools they can communicate with, and the tools will soon be used for reference transactions." [13] In addition to that, in the redefined information and service environment, the user needs to be served in the space in which he usually is, as well as at the point of need, which is contrary to passive waiting for his arrival in either physical or virtual spaces of heritage institutions. Despite certain dilemmas that need to be resolved, the promotion of the so

[11] See: Second Life. http://secondlife.com/ (19 May 2009)

[12] Among the institutions and companies are the prominent Harvard University, IBM, Reuters, Sun and many others.

[13] Janes, Joseph. 2008. Quoted in: Godfrey, Krista. New World for Virtual Reference. // Library Hi Tech. Vol. 26 (2008), No. 4; p. 525. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/2380260403.pdf (19 May 2009)

called anonymity culture and being "at user's fingertips" are justified by experiments with this sort of services in virtual worlds.

## Adjusting Heritage Information Services to 2.0 Information Service Environment

One relatively simple and fast way of adjusting information services of heritage institutions to the broader 2.0 information and service environment is the provision of certain heritage services along with using the possibilities of Web 2.0 tools such as blogs, wikis or RSSs. Since blogs and wikis enable exceptional user's participation and RSSs a complete information personalization, it does not surprise that these tools enabled new manifestations in the field of information services of heritage institutions. The new Web 2.0 tools are effecting evident changes within both internal communication processes among heritage institution employees and external communication processes with user communities. The importance of the blog, as a new, more participatory and truly collaborative communication tool was relatively early recognized by many information expert communities, especially librarian ones, particularly in the part dealing with the design of personal blogs[14] as communication channels for joint discussion of the influence and significance of the Web 2.0 tools increasingly present in the practice of heritage institutions. A more essential exploitation of blog potentials followed soon in the form of *institutional blogs* as tools for providing certain heritage institution services, often information ones. A great number of heritage institutions have recently begun to use blogs actively as spaces for providing their information services.[15] Reports of blog use in the field of heritage institution service provision suggest that this technology promotes internal communication among heritage institution employees and employee-customer communication as well as inter-user communication. This cooperative environment of communication carried out in numerous available ways is a principle

[14] See: For instance, some of the best known blogs for exchanging opinions and experiences of information experts related to blogs and other Web 2.0 tools that enable us to speak about the 2.0 library/heritage institution: Information Wants to be Free.
http://meredith.wolfwater.com/wordpress/index.php (19 May 2009); The Shifted Librarian.
http://www.theshiftedlibrarian.com/ (19 May 2009); Phil Bradley's Web 2.0 Blog.
http://philbradley.typepad.com/i_want_to/ (19 May 2009); Are You 2.0 Yet?
http://briangray.alablog.org/blog (19 May 2009);
Library Crunch. http://librarycrunch.com/ (23 May 2009)

[15] See only some of the successful examples: Madison–Jefferson County Public Library.
http://mjcpl.org/ (23 May 2009); Ohio University Library Business Blog.
http://www.library.ohiou.edu/subjects/businessblog/ (23 May 2009);
Kansas State University Library Blogs. http://ksulib.typepad.com/ (19 May 2009); University of Bath Library Subject Blogs.
http://www.bath.ac.uk/library/subjects/blogs.html (23 May 2009);
University of Canterbury Library, University's College of Business and Economics, The Economics Library Blog. http://blogs.libr.canterbury.ac.nz/econ.php (23 May 2009)

promoted by Web 2.0 technologies and blogs. Internal blogs of heritage institutions are becoming "knowledge management tools par excellence" or "repositories of institutional knowledge that incorporate information that would otherwise remain unrecorded. Since all participants contribute with their knowledge and experience, an internal blog is shaped by collective memory of the employee community and the recording of its good practice."[16] With equal importance, the blog is appearing as a tool of external communication with users and among users themselves or as a tool of successful provision of the upgraded 2.0 information service, which is not surprising given that "blogging is another word for conversation"[17], information service being an evident communicative and participatory act. 2.0 information service can manifest itself in the blog space as one of the most popular Web 2.0 tools as well as through an asynchronic message exchange among a number of information experts and users, where the so called referential blog becomes part of a "Question & Answer" community, whose records are then permanently stored and thus made accessible to public as information sources for other users. Apart from that, blogs frequently become advanced solutions for services such as current awareness services, for newly arrived library items, new articles in databases etc., since email user notifications are becoming less popular due to frequent inbox overload. Marketing opportunities in blog space are immense, while referring to complete texts as results of the process of responding to blog users' information requests are becoming the reality of 2.0 information service.

Library wiki, in a similar way, enables strong integration of information experts and users in providing information services, moving "virtual group instruction room" to the online environment. Users and information experts create "the world of questions and answers" within wikis, while records of these transactions ensure help not only to current users/transaction participants, but also to prospective users to whom the record of reference transactions will become an additional or initial source as well as the final, satisfactory information and reference source.[18]

Since blogs and wikis are pull, not push, technologies, the use of RSS in the blog environment and wikis makes it possible for 2.0 information service to become truly personalized and user directed, the one dominated by the push, not only the pull principle in service provision.

---

[16] McIntyre, Alison; Nicolle, Janette. Biblioblogging: Blogs for Library Communication. // The Electronic Library. Vol. 26 (2008), No. 5; p. 685. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/2630260506.pdf (24 May 2009)

[17] McIntyre, Alison; Nicolle, Janette, Biblioblogging, p. 687.

[18] A whole range of wikis designed for internal communication among librarians is known today. Meredith Farkas is a prominent one. See: Library Success: A Best Practices Wiki. http://www.libsuccess.org/index.php?title=Main_Page (24 May 2009)

## Conclusion

Heritage institutions will certainly advance their information services using tools such as podcasts, digital videogames, Web 2.0 social networks, virtual worlds and other tools to be developed in the future. Such developments become understandable when it is born in mind that heritage institution information services are necessarily influenced by processes of redefinition in a wider information and service environment. Wanting to remain “attractive” and up-to-date for user communities, they have a social duty, or the responsibility of continued notification and adoption of new, more convenient tools of information and communication. The appearance of new information and communication tools and their timely and rightly directed use in the context of heritage institutions effects redefinition within heritage institution information service models. The provision of heritage information services needs to be analyzed within the Web 2.0 technology environment, which enabled various manifestations of information services in the redefined environment.

## References

BBC New Media Research. Gamers in the UK: Digital Play, Digital Lifestyles. http://open.bbc.co.uk/newmediaresearch/files/BBC_UK_Games_Research_2005.pdf (24 May 2009)

Cross, Carl. Making Games Seriously: Creating a Peer Designed Video Game for Use in Library Promotion and Instruction. // Library Review. Vol. 58 (2009), No. 3; pp. 215-227. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/0350580305.pdf (6 May 2009)

Godfrey, Krista. New World for Virtual Reference. // Library Hi Tech. Vol. 26 (2008), No. 4; pp. 525-539. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/2380260403.pdf (19 May 2009)

Jowitt, Angela L. Perceptions and Usage of Library Instructional Podcasts by Staff and Students at New Zealand’s Universal College of Learning (UCOL). // Reference Services Review. Vol. 36 (2008), No. 3; pp. 312-313. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/2400360309.pdf (28 April 2009)

McIntyre, Alison; Nicolle, Janette. Biblioblogging: Blogs for Library Communication. // The Electronic Library. Vol. 26 (2008), No. 5; pp. 683-694. https://proxy.knjiznice.ffzg.hr/proxy/nphproxy.cgi/000100A/http/www.emeraldinsight.com/Insight/viewPDF.jsp=3fcontentType=3dArticle=26Filename=3dhtml/Output/Published/EmeraldFullTextArticle/Pdf/2630260506.pdf (24 May 2009)

Original scientific paper

# Information Literacy in the Academic Context: Global Trends and Local Issues

Sonja Špiranec
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sspiran@ffzg.hr

Tibor Toth
Croatian Information and Documentation Society
Zagreb, Croatia
tibor_toth@hotmail.com

Mihaela Banek Zorica
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mbanek@ffzg.hr

**Summary**
*Information literacy is generally perceived as the* spiritus muovens *of learning processes, reflecting the premise of information as the basic building block of education. This idea gained relevance and new facets with the proliferation of the Web 2.0, which has brought about new, speculative and concerning issues.*
*In the first part of the paper the authors will discuss the importance of information literacy in the higher education sector. They will particularly discuss and reexamine the question what it has meant to be information literate in the print era, the digital era and in the context of the Web 2.0. Taking conceptual shifts between those different information ages as a point of departure, the particular cluster of competencies needed today to support educational processes in the higher education sector will be identified. After commenting global issues the authors will present preliminary results from a local (national) survey about the inclusion and integration of information literacy elements into Croatian new higher education (i.e. Bologna) curricula.*

**Key words:** Information literacy, higher education, information behaviour, Web 2.0, Bologna programs, Croatia

## Introduction

The information literacy (IL) movement has grown dramatically over the past quarter century. The rationale for its positive perception is located within the concept of the information age that places high value on efficient and effective acquiring and use of information (Badke, 2008). According to Lloyd, information literacy has been derived from librarians’ discourses of empowering and facilitating lifelong learning skills and through their discursive practices, which focus on the ‘end users’ developing a proficiency with information, through the seeking, interrogation and evaluation of information sources and the appropriate and ethical use of information (2005, 82). In the higher education sector, the emphasis is on acquiring, developing and demonstrating individual skills and competency which will support independent lifelong learning, critical thinking and problem solving.

Despite the potential positive impact of information literacy on learning, our perceptions of the student population and their learning performance are determined by plagiarism, horizontal information seeking, the lack of the habit of evaluating information and the cut-paste syndrome. According to numerous authors, the described anomalies are a result of information literacy inadequacy in higher education, which is leaving university graduates devoid of the very skills they require to function well within the information workplace (Maughan, 2001; Cheuck, 2002). At the root of the problem is the fact that information literacy is rarely addressed as an educational objective and therefore is not systematically covered in academic program curricula.

## Correlative dimensions of information and learning environments

The process of learning begins with dealing with information; therefore information environments with libraries as their traditional proponent and learning environments created by educational institutions have always been connected and interrelated. This was certainly the case in print-based environment, when learning processes relied on print resources and the capacities of using libraries. With the transition towards electronic and hybrid learning environments the structure of capacities that have the potential to support learning processes has changed. Several new clusters of skills and competencies have emerged with the assumed potential to influence learning processes in diverse environments; Information literacy (IL) is one of them. The uniqueness of this particular literacy refers to its relevancy in analogue, electronic and hybrid environments. Current IL frameworks cover a wide spectrum of capacities such as the ability to access, evaluate, and apply information effectively to situations requiring decision making, problem solving, or the acquisition of knowledge. IL therefore refers to a set of abilities enabling individuals to "recognize when information is needed and the capacity to locate, evaluate, and use effectively the needed information"' (ALA, 1989).

Pervasive computing and the integration of ICT into learning processes have further emphasized the importance of IL as a precondition for learning success in new environments. New pedagogical paradigms are based on the premise of constant interactions with the present complex information environment, where the learner constructs knowledge rather than passively receives it. People who are learning and working in new virtual learning settings have to be independent and self-sufficient learners and users, but are faced with abundant information and unfiltered, unorganized information floods. Hence, the ability to meaningfully interact with a wealth of information is deemed more important than ever. A common response of educational systems to those new conditions in education was and still is a focus on ICT and digital literacies and the effort to integrate those into academic curricula. No doubt, these are legitimate, but not sufficient efforts. Most public policy discussion of education have centred on technologies—tools and their affordances. The computer is discussed as a magic black box with the potential to create a learning revolution. Nevertheless, computer and IT skills allow individuals to use computers, software applications, databases and various other technologies, while IL, as a broader concept, focuses on social applications of information skills and embrace questions of critical evaluation and selection of information or issues of efficient and ethical information use. Having in mind the cluster of skills, competencies and habits IL includes, its importance in educational processes can hardly be denied. The question remains whether IL can be one of the answers for current actual questions, issues and challenges higher education has to face.

## The contemporary higher education context: problems and issues

### "I google, therefore I am"

This quote could be attributed to any member of the *Google generation*, the *Net generation*, *Napster generation* or to the typical *digital native*. All this popular phrases are used to describe young people who intensively use technologies, services and tools that affect their information seeking behaviours, communication styles and habits. This group of users, consisting mostly of students and pupils who have grown up in an online world with little or no recollection of life before the web, for whom technology is a way of life, also represents the very audience that both quickly adopts and frequently uses Web 2.0 services. Their identification as a specific user group which is characterized by specific information behaviour is not just a matter of technology and using new tools and services; they can be differentiated by a particular state of mind that involves attitudes, emotions and preferences, thinking and learning styles. The described assumptions have been identified and discussed in a range of studies (OCLC, 2004; UCL CIBER group etc.). One of the most prominent surveys conducted recently elicited very important findings that argument the need for information literacy, such as:

- search engines fit students' life styles better than physical or online libraries (students begin their information search with search engines)
- students perform horizontal information seeking, which could be described as a form of skimming activity, where people view just one or two pages from an academic site and then "bounce" out, perhaps never to return
- the Net generation has high expectations of ICTs and has zero tolerance for delay
- they prefer visual information over text
- they do not respect intellectual property
- young people have a poor understanding of their information needs and thus find it difficult to develop effective search strategies
- the speed of young people's web searching means that little time is spent in evaluating information (UCL CIBER group, 2008).

The consequences of described transformations in the information universe and the resulting changes in information behaviour of young people are particularly present in the higher education sector. The majority of university teachers will confirm that students are reading less, referencing less, and writing with less clarity, or, to express it in the words of T. Brabazoon: "Clicking replaces thinking" (2007).

The referenced studies show that the hallmarks of university education, like understanding and application of good practice in constructing searches, establishing the validity of sources and, by extension, attributing them when appropriate, are endangered, and the consequences of these occurrences require attention and concrete actions.

**The Web 2.0 in education: potentials and risks**

The described issues have gained more relevance and require more attention in Web 2.0 environments. Educational institutions are beginning to inject Web 2.0 services and tools into classrooms in order to construct active learning environments where knowledge is allowed to shared, used and reused. Such trends have resulted from new perceptions of learning, which is considered as conversation and sharing and is characterized through open environments constructed with social software such as blogs, wikis, podcasts etc. Yet, despite the educational potentials of the Web 2.0, one should not ignore the large number of doubtful or dangerous implications we are starting to see. New technologies make it possible for average consumers to generate and use, archive, annotate and recirculate content in powerful new ways. Such new spectrum of user activities generates a new knowledge culture and the concept of collective intelligence as its central plank. Like-minded individuals gather online to embrace common enterprises, which often involve access and processing information. According to Levy, who has first coined the term collective intelligence, in such a world "everyone

knows something, nobody knows everything, and what any one person knows can be tapped by the group as a whole" (Jenkins, 2006, 39). The most representative marker of collective intelligence is Wikipedia. The original idea with Wikipedia was that everyone could write, but everyone could also correct and rewrite: the massive amount of readers would eventually make sure that an article on every topic would "converge" to the truth. In such a new knowledge culture, students must acquire deeper skills at assessing the reliability of information, which may come from multiple sources, some of which are governed by traditional gatekeepers, others of which must be crosschecked and scrutinized (Jenkins, 42). Students as well as educators have to be aware that learning and working within such environments involves a large number of errors. Misinformation emerges, is worked over, refined or dismissed before a new consensus emerges. Web 2.0 with its collaborative model of knowledge production and mash-up philosophy obviously has brought an end to the stability of information context by creating flat and fluid information spaces. There is enough evidence that the interlinking of learning with these new information spaces requires specific competencies, such as the analysis and identification of the context of generation of information and permanent practice of determining the authority, authenticity and accuracy of encountered information. Students must be taught to read sources from a critical perspective.

Privacy violation is a further danger the students face in their online activities, specifically when using Web 2.0 services and tools, but are rarely aware of. Moreover, students trust information on the web to easily, when searching for some information on the web they tend to accept what they have found as true information, often without looking at other sources and hence having no justification to accept the information at face value. Schools and universities have more and more problems with students who prepare essays by using material from websites or blogs just by copying pieces of information that look relevant and paste them together, without sometimes even understanding them, let alone citing them. Nevertheless, the copy-paste syndrome has not just consequences in the sense of plagiarism. As T. Brabazoon emphasizes, copy-paste, SMS, blogging and twittering undermines the capacity of "reading with understanding". To put it differently: students who keep reading only small junks of information and who compose essays by mainly copying never learn to read larger segments of complicated text. Thus, Web 2.0 may well be one of the reasons why "high quality literacy" seems to be on the decline (Maurer, 2009).

## IL as the corrective of anomalies in educational processes

The problems of educational processes, however modern and technology-enhanced contemporary education may be, are a result of following massive assumptions: students somehow intuitively understand the research process, can take notes, compare arguments, evaluate information resources and organize them, regulate their own learning and do all this in an ethical manner. Although

the possibilities of the web stimulate inadequate information handling and behaviour, the problem is not Google. “The concern is that teachers and librarians are not being given a chance to instruct the literacies required to transform Google from a leisure application and into a starting point for a critical and reflexive research process" (Brabazoon, 2007, 145).

An analysis of the core key words that describe the main problems higher education is currently facing (triggering the research process, locating high quality information, accessing and evaluating information, organizing it, plagiarism etc.) shows that those can be directly mapped to key concepts that define information literacy. This does not mean that IL is the ultimate panacea for solving problems occurring in academic learning environments, but it certainly is a valid strategy and logical means for dealing with existing anomalies. This interrelation is explicitly expressed in Bent et al, who claim that IL can be thought of as “an individual’s attitude to their learning and research such that they are explicitly thinking about how they use, manage, synthesize and create information, in a wise and ethical manner, to the benefit of society, as part of their learning life. In this view, IL is central to learning and research and is about changing people’s learning attitudes and habits so that they understand how information fits into their learning lives” (Bent et al, 2007, 84).

IL goes beyond surface and technical skills and deals with conceptual insights, the construction of strategies, with assessment and sense-making, the formation of information and learning habits, with the ability of distinguishing between fact and fiction, fact and opinion, with providing arguments and collecting evidence. This literacy allows students not only to handle a search engine but provides the interpretative capacities to handle the results (Brabazoon, 2007).

As studies have shown, information behaviour patterns of students display a variety of plagiaristic activity from poor paraphrasing, plagiarism consisting of pasting together quotes from different sources to complete copying of unacknowledged work (Nadelson, 2007). Even where there is no or little evidence of plagiarism, teachers complain about incompetent referencing of sources. Intellectual property rights issues are certainly raised by technology and digitization because it is extremely easy to reproduce and distribute. Anyway, not the technology per se should be blamed for progress of plagiarism incidences, usually it is the student who does not know that what he is doing counts as plagiarism or he does know but lacks the skills to do anything about it (or thinks that it is acceptable practice). One of the goals of IL is the ethical use of information, which is usually achieved through educating students about plagiarism, making them aware of what constitutes plagiarism, how to reference properly, together with knowledge of the penalties for plagiarism.

The capacities a student can built within IL courses are paramount for higher education, nevertheless, it is important to bear in mind that IL is always a reflection of the current information universe and has to change parallel with the information universe. As the impact of the Web 2.0 on information environ-

ments is huge, it reflects on central conceptions of IL as well. IL programs should raise among students the awareness that information and knowledge are socially produced and distributed, and that they can therefore be effectively accessed through social relationships as well (Lloyd, 2006). Therefore, IL should also focus on social skills, on ways of interacting within a larger community, working within social networks, compound knowledge within a collective intelligence but also discern high quality information from diverse pools of collective intelligence.

## Elements of information literacy in HE curricula: insights from Croatia

Due to the described correlations between IL and modern education, its congruency with contemporary educational goals (promoting critical thinking skills and developing the capacity for lifelong learning) and its potential to respond to different issues arising in new learning environments, one could expect that IL is recognized as central to the mission of higher education, and will have its expression in academic curricula.

Moreover, the integration of IL into higher education curricula is one of its main determinants and the majority of authors claim that IL cannot be realized outside curricula. In contemporary higher education systems acknowledged content that is officially endorsed by the academy has credit bearing status. Credit offerings command the attention of students, faculty, and administrators and serve as the key indicator of what an institution considers essential in the education of its students (Badke, 2008). Although there are several models of offering courses relating to information literacy, real impact of IL is to be expected if its part of the curricula and if it is a credit bearing subject. The main drivers of IL initiatives are libraries, who share the responsibility of creating and offering IL programs with teachers. Despite progress made by academic libraries in advancing their instructional activities, their teaching role continues to be predominantly restricted to limited classroom engagements. The vast majority of librarian time is spent doing one or two hour sessions at the invitation of subject faculty or providing basic generic instruction to incoming freshmen (but even this limited approach is not the rule). Few professionals in the field would argue that such minimal exposure to information literacy instruction can fulfil the goals IL.

### Survey and preliminary results

In order to determine the actual state of IL within the Croatian higher education sector, the level of inclusion and integration of information literacy elements into new higher education curricula (i.e. Bologna programs) has been surveyed.[1]

[1] The survey was initiated by the Croatian Information and Documentation Society (HIDD: Hrvatsko informacijsko i dokumentacijsko društvo http://www.hidd.hr/).

The study programs where examined and analyzed from January till June 2009. Hereafter we will present preliminary results, which include the analysis of 472 study programs (out of 963 published programs, 49%) that are offered on 71 faculties and other organizational units (academies, departments) at 6 Croatian universities (Dubrovnik, Osijek, Pula, Rijeka, Split, Zadar). Although the results are preliminary, they show a clear absence of IL in Croatian higher education curricula. The survey has hereto surfaced following results:

1. existing curricula do not explicitly offer information literacy or an integral information literacy subject
2. a number of subjects (70) contain isolated elements of information literacy, predominantly within diverse subjects relating to scientific literacy, labelled as: Methodology of scientific work, Introduction to scientific work, Introduction to research etc.
3. those different labelled subjects are offered at various levels of study (undergraduate, graduate, postgraduate), have different status (elective, obligatory), and bear various credit points (in the range from 0 to 20).
4. these subjects are conducted as lectures, lectures and seminars, and lectures with exercises, but the predominately form are either lectures or lectures with seminars
5. descriptions of the offered subjects do not indicate that the faculty librarian is somehow involved in planning or delivering the course (except in a few examples, where librarians have academic status).

**Discussion**

These preliminary results indicate the main issues, inadequate perceptions and fragmentary approaches not only to the IL concept but general to the development of generic information competencies as a prerequisite for lifelong learning and information handling at the workplace.

The overall absence of an integrated approach to IL is certainly an issue, particularly having in mind the anomalies occurring in contemporary educational processes. Existing limited approaches which focus on finding scientific information are neither comprehensive (41 out of 71 units are offering such contents) nor consistent (various levels, different status). The analyzed descriptions of the offered subject show that they are mainly delivered through lectures and comprise a large portion of contents relating to:

- Choice and Statement of Research Problem,
- Techniques of the research execution,
- Design of experiments and apparatus,
- Execution of Experiments, Analysis of Experimental Data,
- Basic principles of scientific categories,
- Classification of papers,
- Searching for the data in literature and scientific documents.

Science literacy significantly overlaps with the conception of information literacy. Teaching students in the scientific method and culture has long been recognized as an important part of education for those entering scientific professions. Nevertheless, IL is a much wider concept and goes beyond information retrieval and accessing scientific information. It encompasses question like:
Why is information important in our contemporary society? How do I actually inform myself? Where does information come from? Who determines that it is published? What is the difference between a scholarly journal article and a webpage and are these differences still important in a world of converging information? What are the problems and benefits that are caused by anonymity and collective knowledge creation? Why do I have to pay for some information? What is metadata, and how can it help me? What are the implications of electronic searching and electronic documents for the way we do research? How do we evaluate what we have found? What are the legal and ethical considerations?
This sample of questions indicates that IL has much wider implications and is efficiently transferable to all learning situations during ones academic career, but to workplace situations as well. It is crucial for graduated students to come to the workplace and perform adequately in the realm of information handling, information management etc. A university education is not purely about gaining specific subject knowledge; it must challenge students to view their learning as something which isn't bounded by their time at university but is part of their everyday world (Bent, 2008). Taking this assumption as a premise, one can come to the conclusion that the Croatian higher education sector has not recognized the need for information literacy and that the existing elements are not adequately represented or integrated. There is a particular need for integrating IL at the undergraduate level, since students entering higher education are usually not familiar with the research process or with plagiarism and are at the same time overwhelmed by the amount of information available at their fingertips. The preliminary results show that existing approaches are neither sufficient nor systematic and new, information literacy focused strategies are needed to face the described challenges.

## Conclusion

Information literacy should be perceived as a strong plank of educational processes, be it in the print era (expressed in the notion of "the library as the heart of university"), it in the digital era or in the context of the Web 2.0. Today's students are supposed to use all these different information realms simultaneously, therefore IL which offers conceptual insights into all these different environments is crucial for learning and for avoiding pitfalls generated by these new environments. This assumption leads to the expectation that information literacy as a key competency would gain the status of a global educational outcome and

that educational institutions would express their commitment for curriculum integration of IL.

However, preliminary results from a national survey about the inclusion and integration of information literacy elements into Croatian new higher education curricula (i.e. Bologna programs) show a poor understanding of IL within this sector. Individual IL elements are scattered throughout the subject "Methodology of scientific work", which itself is offered unsystematically and at a minor number of studies. Having in mind the anomalies occurring in contemporary educational processes (cut and paste, plagiarism, accessing and evaluating information, reading with understanding, academic writing etc.) it is necessary to conceptualize new and rethink existing approaches to integrate IL as a crucial cluster of competencies into Croatian higher education curricula.

## References

ALA. Presidential Committee on Information Literacy: Final Report. 1989. http://www.ala.org/ala/acrl/acrlpubs/whitepapers/ presidential.htm . (2008-10-20)

Badke, W. A rationale for information literacy as a credit-bearing discipline. // *Journal of information literacy.* 2 (2008)1. http://jil.lboro.ac.uk/ojs/index.php/JIL/article/view/RA-V2-I1-2008-1 (2009-06-30)

Bent, M.; Gannon-Leary, P; Webb, J. Information Literacy in a researcher's learning life: the seven ages of research // *New Review of Information Networking* 13(2007)2; 81-99.

Bent, M.; Stockdale, E. Integrating information literacy as a habit of learning - assessing the impact of a golden thread of IL in the curriculum.// Journal of Information Literacy. 3(2009)1; 43-50. http://ojs.lboro.ac.uk/ojs/index.php/JIL/article/view/PRA-V3-I1-2009-4 (2009-06-30)

Brabazoon, T. The University of Google: Education in a (post) information age. Aldershot: Ashgate, 2007.

Cheuk, B. W.-Y. Information Literacy in the Workplace Context: Related Concepts, Challenges and Issues. 2002. http://www.nclis.gov/ libinter/infolitconf&meet/papers/cheuk-fullpaper.pdf (2009-07-10)

Jenkins, H et al. Confronting the Challenges of Participatory Culture: Media Education of the 21st Century. 2006. http://digitallearning.macfound.org/atf/cf/%7B7E45C7E0-A3E0-4B89-AC9C-E807E1B0AE4E%7D/JENKINS_WHITE_PAPER.PDF (2009-08-04)

Lloyd, A. Information literacy: Different contexts, different concepts, different truths? // Journal of Librarianship and Information Science 37(2005); 82-88.

Maughan, P. D. Assessing information literacy among undergraduates: A discussion of the literature and the University of California-Berkeley assessment experience. // *College & Research Libraries,* 62(2001)1; 71-85.

Maurer, H.; Kulathuramaiyer; N. Knowledge gathering as it changes with new technologies // Proceedings of e-learning 2009. Vol. 1. / Nunes, M.B; McPherson, M. (ed). IADIS, 2009, 23-30.

Nadelson, S. Academic Misconduct by University Students: Faculty Perceptions and Responses // Plagiary 2(2007); 67-76.

OCLC. College Students' Perceptions of the Libraries and Information Resources: A Report to the OCLC Membership, OCLC, Dublin, OH. 2006. http://www.oclc.org/reports/perceptions college.htm (2009-08-02)

University College London (UCL) CIBER group. *Information behaviour of the researcher of the future.* London: University College London. CIBER Briefing paper; 9. 2008. http://www.jisc.ac.uk/media/documents/programmes/reppres/gg_final_keynote_11012008.pdf (2009-07-15)

Original scientific paper

# Evaluation of Digital Collections' User Interfaces

Radovan Vrana
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
rvrana@ffzg.hr

**Summary**

*Digital information resources play a very important role in today's world. They eliminate or reduce prior constraints of distance, fragility of resources, or limited physical access to resources premises a freedom and flexibility in information access unprecedented in human history (Bates, 2002). Their existence would be impossible without information systems that enable their functioning. The life cycle of an information system consists of the standard sequence of development phases. This sequence of development phases includes the process of evaluation or assessment of the information system. Only thorough and frequent evaluation of the information system and its components will ensure its flawless functioning. During the process of evaluation, special attention is given to user interfaces, access points to the content of online information resources. The process of evaluation can be carried out by application of several different methods. This paper puts focus on selected methods for evaluation of information systems and their user interfaces that can be also applied to digital collections (and their user interfaces) available on the Internet. It also presents results from the research of the Croatian national heritage digital collections available on the Internet and common characteristics of their user interfaces.*

**Key words:** digital libraries, digital collections, evaluation

## Introduction

Digital information resources play a very important role in today's world. They eliminate or reduce prior constraints of distance, fragility of resources, or limited physical access to resources premises a freedom and flexibility in information access unprecedented in human history (Bates, 2002). Despite the system developers' efforts, design of information systems that make possible functioning of online digital information resources is rarely without flaws. To discover weaknesses of such information systems, and to improve their functioning, designers and developers use different methods of evaluation of the whole information system or its respective components. This paper puts focus on one such component – front end of online information resources i.e. user interface of

digital collections on the Internet. User interfaces are important because they serve as access points to the content of the online information resource. In case of digital collections and digital libraries, user interfaces can be quite complex and require substantial knowledge from user for their use. Their design should be carried out in compliance with the current user interface development standards and with end users and their needs in mind. End users are not always included in the process of the information system development, but are rather imaginary category excluded from the direct involvement in that process. As a result, components of the information system are built separately with no knowledge of the user interfaces of other components they may be composed with and this can result in component-based applications with inappropriate, inconsistent interfaces. Problem with use of such user interface may appear due to the user’s level of expertise, the task and role being performed, and user’s personal preferences. Every new additional component that may be integrated, can introduce further problems or inconsistencies to the overall application interface (Grundy and Hosking, 2002). The idea behind every user interface is to make access to digital content as easy as possible, and that is the goal of an information system developer. Despite the existing usability standards, a certain number of user interfaces of online information resources is still difficult to use or doesn’t have all the necessary functions the user would expect from a modern user interface. In such cases, evaluation process can indicate weak points of tested user interfaces and help information system designers to improve those inadequate parts. Result should be better understanding of the process of design of user interfaces and their easier use.

## Design and structure of user interfaces

The user interface is a part of the computer and its software that people can see, hear, touch, talk to, or otherwise understand or direct (Galitz, 2002). Proper user interface design will provide a mix of well designed input and output mechanisms that satisfy the user’s needs, capabilities and limitations in the most effective way possible. The best user interface is one that it is not noticed, one that permits the user to focus on the information and task at hand, not the mechanisms used to present the information and perform the task (Galitz, 2002). A well-designed user interface can help users to use the system more easily by reducing the effort to identify a particular object on the screen, or providing smooth navigation among screens (Thong, Hong and Tam, 2002).

Behind every user interface is a conceptual model. One such model was proposed by William Arms in his book on digital libraries. His model describes the manner in which the system is used. His conceptual model contains 4 layers (Arms, 2001): interface design, functional design, data and metadata and computer systems and networks.

Interface design encompasses what appears on the screen and how the user manipulates it (fonts, colors, logos, keyboard controls, menus and buttons). Func-

tional design specifies the functions that are offered to the user (selecting parts of a digital object, searching a list or sorting results, obtaining help, and manipulating objects that have been rendered on the screen). Enumerated functions are made possible by the data and metadata that are provided by the digital library and by the underlying computer systems and networks.
The user interface development would be very difficult if not impossible without fundamental design principles that apply to the structure of the user interface and all its parts (Dennis, Wixom and Roth, 2006):

- Layout: the interface should be a series of areas on the screen that are used consistently for different purposes: navigation (top area), input and output (middle area) and system status (bottom area)
- Content awareness: user should be aware of where they are in the system and what information is being displayed
- Aesthetics: interface should be functional and inviting to users through the careful use of white space, colors and fonts
- User experience: some users will prefer ease of learning and some will prefer ease of use
- Consistency: it enables users to predict what will happen before they perform a function
- Minimize user effort: the user interface should be simple to use.

Despite the existing principles of user interface design, we are still aware of very different user interfaces we use every day. Generally speaking, the structure of a user interface includes three fundamental parts (Dennis, Wixom and Roth, 2006):

- Navigation mechanism: the way in which the user gives instructions to the system and tells it what to do (buttons, menus)
- Input mechanism; the way in which the system captures information (Web forms etc.)
- Output mechanism: the way in which the system provides information to the user or to other systems (reports, Web pages)

These fundamental parts are starting points for evaluation of user interfaces of various types of online information resources.

## Evaluation of digital collections

Digital collections are key parts of digital libraries and are central point of this paper. Digital libraries are revolutionizing the ways library services provide access to digital information (i.e. data or articles) through their collection, repackaging, and online distribution via local or international networks, such as the internet (Sutradhar, 2006). Digital libraries are therefore networked information space in which users can discover, locate, acquire access to and use information (Greenstein, 2000). Their functions are similar to those in conventional libraries, but they differ in storage and retrieval, where digital libraries are dependent

almost exclusively on computer and electronic network systems (Waters, 1998). Digital libraries can be also perceived as sets of electronic resources (i.e. digital collections) and associated technical capabilities for creating, searching, and using information. They are extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks (Borgman, 2003). To make the assessment of the current state of development of digital collections and digital libraries, they should be constantly evaluated.

The term evaluation has many connotations ranging from highly focused and well-defined product testing to the highest form of cognitive reflection (Marchionini, 2000). In case of online information systems, one can perceive as if the content (information) itself is badly prepared and presented rather than to put the blame on the poorly designed front end of the information system. According to Norman, who compared evaluation of information and information systems, information cannot by itself be good or bad; it can only do so within the context of a person being informed. However, an information system can be assessed —issues of usability, speed and reliability are open to objective measurement and are largely independent of the context of an information transaction (Norman, 1997). When speaking about evaluation of digital libraries Chowdhury and Chowdury point out that they may be evaluated from a number of perspectives, such as: system, access and usability, user interfaces, information retrieval, content and domain, services, cost and the overall benefits and impact (Chowdhury and Chowdhury, 2003).

## Evaluation of digital collection user interfaces

Evaluation of user interfaces can be a very difficult task. As the information and communication technology develops, so change the user interfaces that help us access the digital content on the Internet stored in digital collections. When referring to user interfaces of digital collections available on the Internet, we usually refer to Web interfaces, and less frequently to other types of user interface that exist too (e.g. JAVA applications etc.).

According to Dennis, Wixom and Roth the objective of user interface evaluation is to understand how to improve the user interface design. User interface evaluation should begin in the design phase when design problems can be identified and corrected (Dennis, Wixom and Roth, 2006). User interface evaluation is necessary because it is difficult if not impossible to design an user interface that for all users and all tasks on all occasions will function perfectly (Tedd and Large, 2005).

Savage (Savage, 1996) compared three methods for user interface evaluation:

- Expert reviews: conducted in the presence of human factors specialists and consist of a combination of standard inspection methods (in this case, heuristic evaluation, cognitive and pluralistic walkthroughs, and consistency and standards inspections) all bundled into one inspection session

- Reviews: conducted by end users
- Usability Testing.

Dennis, Wixom and Roth added another four approaches to for (user) interface evaluation (Dennis, Wixom and Roth, 2006):

- Heuristic evaluation: examines the interface by comparing it to a set of heuristics or principles for interface design; as this approach does not involve the users, it is considered the weakest type of evaluation
- Walk-through evaluation: it is a meeting conducted with the users who will ultimately have to operate the (information) system who go through various parts of the interface; the users identify improvements to the interface
- Interactive evaluation: the users work with the prototype of the user interface with the project team behind the information system
- Formal usability testing: it is done with the help of commercial software products; the user participates in one-on-one session in which he or she works directly with the software to accomplish given tasks; the evaluation is conducted in a special lab equipped with video cameras and special software that records each keystroke and mouse operation so they can be replayed to understand exactly what the user did;

Another common and widespread evaluation approach is evaluation against the set of pre-selected criteria. Set of criteria should encompass most vital parts of the evaluated user interface. Tedd and Large (Tedd and Large, 2005) suggest five evaluation criteria that can be applied to any interface: the time it takes to learn how to use the interface properly; the speed at which the interface performs actions requested by the user; the rate of errors committed by users at the interface; the ease with which users can remember the interface and its features from one session to the next session and the level of individual satisfaction that users derive from their experience with the interface.

The final part of the paper will present the results of the comparison of user interfaces of digital collections that are part of the Croatian national cultural heritage against the set of pre-selected criteria.

## Research: Comparison of digital collection user interfaces

This part of the paper introduces comparison of digital collections user interfaces that are part of the portal about the Croatian Cultural Heritage project, a national project for the digitization of archival, library and museum material (Croatian cultural heritage, 2007) available on http://www.kultura.hr. This Web site offers information about current digitization projects in Croatia and it was used as a staring point for the creation of the list of digital collections accessible on the Internet that will be compared against the list of criteria. The final list for the comparison was created from the list of all registered projects of digitization in all regions of Croatia (the complete list is available at: http://www.kultura.hr/

hr/zbirke/po_regijama). On April 26th 2009, 214 digital collections were registered on this Web site. It must be noted that not all of 214 digital collections were accessible online, since some of them are available for use only in the premises of libraries, archives and museums where collections are stored. Web pages of 66 out of 214 digital collections were reachable online at the moment of the creation of the list of digital collections that were used in the comparison.
The hypothesis for this research was that most digital collections available on the portal of the Croatian cultural heritage share common screen / user interface elements which make their use easier. The second hypothesis was that user interfaces of digital collections available in the Croatian Web space are still underdeveloped. The aim of this research was to collect the data which would confirm or reject these hypotheses. Based on the results of this comparison, software developers can make the necessary improvements to those digital collections user interfaces that are found to be underdeveloped or inadequate in some areas.
The list of comparison criteria will be based on work of Xie and Cool (Xie and Cool, 2000) who selected six tasks which users have to achieve in order to accomplish their search tasks in online information retrieval systems and which are realized in an user interface as functions: database selection, query formulation, query reformulation, access to help function, organization and display of results and delivery of results. These tasks are common in digital libraries today.
The list of criteria suggested and used by Xie and Cool (Xie and Cool, 2000) will be expanded for the evaluation of interfaces of digital collections:

- Category 1. Access: browsing capabilities; searching capabilities (simple and advanced)
- Category 2. Query formulation: simple; complex (AND, OR, NOT operators), query reformulation
- Category 3. Help: general; contextual
- Category 4. Organization and display of results: sorting capabilities; limiting number of results
- Category 5. Delivery of list of results: file, print, clipboard, e-mail
- Category 6. User interface language choice.

These tasks are essential for the successful completion of users’ tasks. For instance, the search and browse tools that a site provides to its users are increasingly important as users become more and more sophisticated in their search strategies and, at the same time, become less inclined to spend a lot of time learning the ins and outs of a Web site (Shiri and Molberg, 2005).

## Results and discussion

The results will be presented jointly in each category.

Category 1. Access (n=66)

| | Browsing | Searching: simple | Searching: Advanced |
|---|---|---|---|
| N | 57 | 8 | 7 |
| % | 86,36% | 12,12% | 10,60% |

The comparison of digital collections user interfaces in this category shows that the large percentage of them are built with the browsing function in mind. The amount of content in a particular collection is very small, and therefore it is easier for users to browse the collections instead of searching; this is especially true for new users.

Category 2. Query formulation (n=66)

| | Simple | Complex | Reformulation |
|---|---|---|---|
| N | 6 | 3 | 5 |
| % | 9,09% | 4,54% | 7,57% |

The results in previous and this category indicate that searching is not included frequently as a function of digital collection user interface. Very few collections (9,09%) offer even simple query formulation while even smaller percentage collections (4,54%) offer complex query formulation (use of AND, OR, NOT operators). As searching and browsing are two main access points to digital content available not only in digital collection in archives, libraries and museums but also on the Internet, this results should be taken into consideration when implementing the search function into user interfaces of future digital collections.

Category 3. Help (n=66)

| | General | Contextual |
|---|---|---|
| N | 4 | 0 |
| % | 6,06% | 0% |

Novice users of digital collections would find themselves in very difficult position if they are not familiar with the content of digital collections registered at the Ministry of culture of the Republic of Croatia. Help option was found in 6,06% of digital collections included in this comparison. Digital collections in the comparison do not offer any instance of contextual help that is necessary when the collection user moves from one part of the collection to another. In case of change of type of content or type of content handling, the collection user cannot count of any type of help. As digital collections available on the Internet grow in number, and are still very different, general and contextual help will

become necessary parts of every digital collection. The problem is even more delicate when one thinks about including the digital collection in educational process in schools or at universities.

Category 4. Organization and display of results (n=66)

| | Sorting | Limiting no. of results |
|---|---|---|
| N | 1 | 1 |
| % | 1,51% | 1,51% |

This category is dedicated to the management of search results. Organization and display of the search results help users to find results that suite best to his or her needs. In case of the 66 compared digital collection, only 1 collection offers mechanism for sorting and limiting number of the search results.

Category 5. Delivery of list of results (n=66)

| | File | Print | Clipboard | E-mail |
|---|---|---|---|---|
| N | 1 | 0 | 0 | 1 |
| % | 1,51% | 0% | 0% | 1,51% |

In addition to results from previous category, only 1 collection offers delivery of search results in file or by e-mail. These options have been standard for information retrieval systems for decades. They help user to transfer the search results to their personal computers.

Category 6. User interface language choice (n=66)

| | User interface language choice |
|---|---|
| N | 17 |
| % | 25,75% |

In today's multilingual world, the possibility of change of language of a user interface is vey important, as users come from different parts of the world. Only one quarter of collections offers such a possibility which is not enough if web want to internationalize digital collections of the Croatian heritage.

## Conclusion

Evaluation of user interfaces of digital collections available on the Internet is an important, complex and necessary activity during the process of development and use of digital information resources available on the Internet. As the number of available digital collections grows, so grow the expectancies of their users. User interfaces are undergoing changes and new types of user interfaces are being introduced frequently and thus are becoming a new challenge to digital collection software developers as well as to the users of these digital collections. Their evaluation should be carried out frequently in order to improve the access

to the content stored in online information resources. The results of the research in this paper show similarities between user interfaces of digital collections which is not unexpected since use of the previous knowledge and experience from interaction with digital collections worldwide can help users to access the content of reasonably high number of new digital collection they will encounter on the Internet. The results may also motivate digital collections developers to compare the user interfaces of their digital collections with other heritage digital collections available on the Internet, and to make the necessary improvements.

## References

Arms, W. Digital libraries. Cambridge, Mass., London : The MIT Press, 2000.

Bates, M. J. The cascade of interactions in the digital library interface. // Information Processing and Management. 38(2002), 3; 381-400.

Borgman, C. L. The Invisible Library : Paradox of the Global Information Infrastructure. // Library Trends. 51(2003), 4; 652-674.

Chowdhury, G.G.; Chowdhury, S. Introduction to digital libraries. London : Facet publishing, 2003.

Croatian cultural heritage. 2007. http://www.kultura.hr (26.4.2009.).

Dennis, A.; Wixom, B. H.; Roth, R. M. Systems analysis design. Hoboken : John Wiley and Sons, 2006.

Galitz, W. O. The Essential Guide to User Interface Design. New York : John Wiley and Sons, 2002.

Greenstein, D. Digital Libraries and Their Challenges. // Library Trends. 49(2000), 2; 290-303.

Grundy, J.; Hosking, J. Developing adaptable user interfaces for component-based systems. // Interacting with computers. 14(2002), 3; 175-194.

Marchionini, G. Evaluating digital libraries: a longitudinal and multifaceted view. // Library Trends. 49(2000), 2; 304-333.

Norman, F. Digital libraries—a quality concept. // International Journal of Medical Informatics. . 47(1997), 1; 61-64.

Savage, P. User Interface Evaluation in an Iterative Design Process: A Comparison of Three Techniques. 1996. http://www.sigchi.org/chi96/proceedings/shortpap/Savage/sp_txt.html. (15. 4.2009.)

Shiri, A.; Molberg, K.. Interfaces to knowledge organization systems in Canadian digital library collections. // Online Information Review. 29(2005), 6; 604-620.

Sutradhar, B. Design and development of an institutional repository at the Indian Institute of Technology Kharagpur. // Program: electronic library and information systems. 40, 3(2006), 244-255.

Tedd, L. A.; Large, A. Digital libraries : principles and practice in a global environment, München : K.G. Saur, 2005.

Thong, J. Y. L.; Hong, W.; Tam, K.-Y. Understanding user acceptance of digital libraries: what are the roles of interface characteristics, organizational context, and individual differences?. // International Journal of Human-Computer Studies. 57(2002), 3; 215-242.

Waters, D. J. What are digital libraries?. // CLIR Issues. 4(1998). http://www.clir.org/pubs/issues/issues04.html (15.3.2009.).

Xie, H.; Cool, C. Ease of use versus users control: an evaluation of Web and non-Web interface of online databases. // Online information review. 24(2000), 2; 102-115.

# Wikipedia's Influence on the Evolution of Encyclopedia

Sara Librenjak, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sara.librenjak@gmail.com

Zdenko Jecić
Lexicographic Institute Miroslav Krleža
Frankopanska 26, Zagreb, Croatia
zdenko.jecic@lzmk.hr

Damir Boras
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
dboras@ffzg.hr

**Summary**

*This article deals with the users of Wikipedia and their usage and opinions regarding Wikipedia in comparison to printed encyclopedias. A representative sample of high school and university students from Zagreb was surveyed.*
*It is investigated why students tend to use Wikipedia significantly more than printed encyclopedias, even though they are considered more accurate and reliable, which is important in educational usage.*
*The results of the survey may serve as guidelines for both users and professionals, i.e. encyclopedists. Various issues with Wikipedia which also need attention in this context, e.g. the development of Croatian Wikipedia, are briefly addressed.*
*The goal of this work is to broaden the awareness of recent phenomena in encyclopedic, or generally, information science, in order to aid in improvement of encyclopedic products.*

**Key words:** traditional encyclopedia, online encyclopedia, encyclopedics, survey, Wikipedia

## Introduction

In recent years we are witnessing substantial growth of digital media usage. What follows from this fact is that in our school or work environment printed media is used less frequently. This migration to the "digital world"[1] has brought various changes to our perception and usage of information and knowledge. Thus, it is a relevant topic in both information science and encylopedics[2].

Shores defines encyclopedics as "the art and science of selecting and disseminating the information most significant to mankind" [4]. Although almost half a century old, this definition stresses the *significance* of information as the criteria for its selection. Since the first (known) encyclopedias in the Roman time, the selection of information has been made by the authors. Traditional encyclopedia is compiled and edited by a group of experts, which has, especially in past times, caused some subjectivity in the decision making process: which are those information that are significant to mankind? This problem is less present since the development of the objective science work in encyclopedics, but it is always difficult to speak in the name of the *whole mankind* when a minority is making this decision.

The issue with traditional encyclopedia which is more difficult to mend is due to objective reason: it is considerably difficult to update information in the paper format. Traditional encyclopedia takes years to develop and print, and reprints and corrections are expensive and time consuming. This can cause recurring errors which are hard to spot and improve, and outdated information which appears in subsequent editions.

Digital format, especially online published, grants significant improvement in this field. Data is easily updated, checked and compared. It allows more editors to work at the data, at the same time, and even the usage of language technologies in spell and grammar checking (we may expect more of this field in the future).

The most prominent example of online encyclopedia is omnipresent Wikipedia[3]. It embodies the good sides of the digital format, as well as some drawbacks.

---

[1] See McLuhan [2] for theory of printed media influence ("Gutenberg galaxy") and early predictions of "electronic age" emergence.

[2] A branch of information science; scientific discipline which deals with principles and practices of assembling an encyclopedia.

[3] Mituzas, one of the Wikipedia's system administrators shared the following metrics about Wikipedia usage in 2008 [3]:

50,000 http requests per second
80,000 SQL queries per second
7 million registered users
18 million page objects in the English version
250 million page links
220 million revisions
1.5 terabytes of compressed data

We can view the matter of open editing, one of the frequent complaints about Wikipedia, from two perspectives. It can remove the subjectivity and control the mistakes, because the mistakes and inappropriate content will be removed faster. On the other hand, it is prone to vandalism and, more subtly, it is also biased in a different way. At the moment, most of the Wikipedians[4] belong to the (sub)culture of Internet users, and as participants of specific culture, the information selection and presentation might also be culturally biased. Unlike the paper formatted encyclopedia, Wikipedia is up-to-date, but not all topics are updated as promptly as recent news and popular culture, in which Wikipedia resembles a news portal.

This research will try to answer why Wikipedia is so popular among users, in what manner and why are its shortcomings ignored, is it replacing the traditional encyclopedia or is the format of encyclopedia just evolving with the social changes.

## The survey: demographics and background

In an online survey[5], a sample of 123 Wikipedia users answered questions about their usage of Wikipedia, stated opinions about its reliability and compared it to a traditional encyclopedia. The sample consisted of high-school (13%) or university students (59%) and employed young people (28%). It represented both genders equally, with 48% of men and 52% of women subjects. Most of the subjects resided in Zagreb (84%), and the others were from large (10%) or smaller cities (6%). Subjects are all Internet users, who use it every day for various purposes. They all had some or considerable experience with Wikipedia. Graphs 1 and 2 summarize general statistics about Internet and Wikipedia.

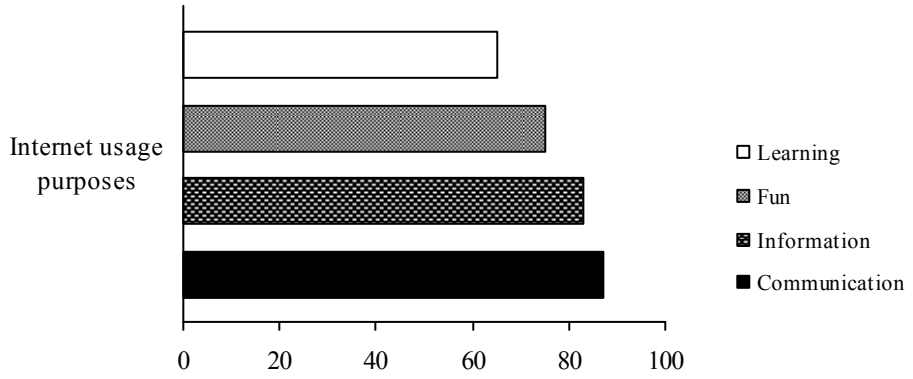
The subjects are from the demographic group which uses Internet the most, and to whom the education plays one of the central roles in life. Since encyclopedia's primary purpose is to inform and educate, the sample consisted mostly of students. Also, since Wikipedia is an Internet phenomena, chosen subjects range from average to experienced Internet users. It is observed that English Wikipedia is used more than Croatian, which will be commented later in the article, but at this point informs us of young people's English proficiency.

The results of the survey are divided in three topics. Firstly, general opinions of Wikipedia are addressed: reasons and manner of usage, reliability and its role in personal education. Secondly, the subjects compared Wikipedia with the traditional encyclopedia. The last topic is the opinion on the development of the Croatian version of Wikipedia, which is connected to the traditional vs. online encyclopedia debate.

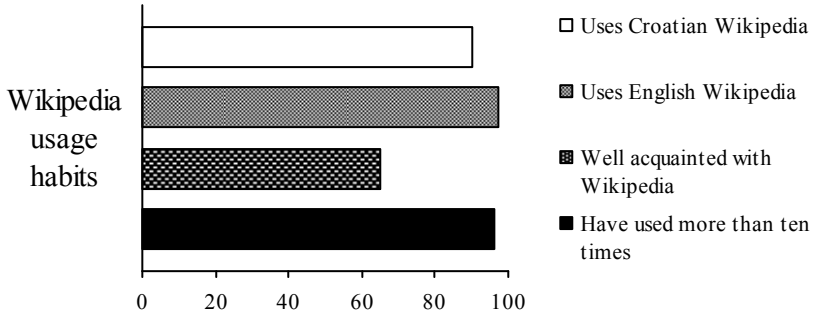
[4] Term used to denote people who write and edit articles on the Wikipedia.

[5] The survey was assembled, conducted and processed by the authors on the address http://osnovephp.kset.org/varsara/anketa, which is no longer available online.

Graph 1: Internet usage purposes




Graph 2: Wikipedia usage habits




## Opinions on Wikipedia, its usage and role

The survey showed that most subjects use Wikipedia because of the vast information available, and facilitated availability. Users stated that they mostly use the search box and click the links in the article when trying to find some information on the Wikipedia. Average surveyed subject does not use Wikipedia's category system, and most users were satisfied with searching and data retrieval options. Table 1 sums up these answers with statistical data[6].

[6] For both questions, more then one answer was accepted. Answers are sorted in ascending order according to the percentage of each answer.

Table 1: Reasons for Wikipedia usage and browsing manners

| **Why do you use Wikipedia?** | |
|---|---|
| Wide choice of subjects | 74% |
| Ability to find desired information quickly | 70% |
| Free of cost | 68% |
| Available in multiple languages | 35% |
| Open editing | 28% |
| Quality of content | 26% |
| **How do you search for the information in Wikipedia?** | |
| Using the search box | 81% |
| Following the links in the articles | 51% |
| Browsing categories | 19% |
| I rarely use Wikipedia for searching or browsing | 9% |

Wikipedia also seems to play a role in subjects' education. But, according to the answers, it is used more as a simple introduction to a topic, then to broaden knowledge about some previously known facts. 81 percent of the subjects agree that Wikipedia is useful as an introduction to some topic. 74% uses it as an aid in writing a school or university paper (more or less reluctantly), and 52% uses Wikipedia for learning. Although they are using it, that does not mean that the users are considering Wikipedia reliable. 26 percent said they consider Wikipedia absolutely reliable, 3% stated they do not consider it reliable at all, and the majority (71%) answered that Wikipedia is more or less reliable. More specific answers to the reliability issue can be found if we question what (un)reliability entails in the case of Wikipedia, and what the reasons for the lesser quality of article content are.

Table 2 explores the issue of reliability, showing the percentage of subjects who agreed with the statements. The first issue is connected to Croatian version of Wikipedia, and shows a separate problem: less content means less quality. This does not mean that the content – quality ratio is an exponential function, because vast content does not guarantee quality, but considerable lack of content simply causes a specific version of Wikipedia to be unusable. Second issue that is mostly agreed upon is that reliability depends on the article topic. This issue is explored later. More than half of the subjects agree that there should be some content control by experts, which points out the need for a reliable method of quality control. The fact that users re-check the facts from Wikipedia in other sources seconds that conclusion.

Another question dealt with the reasons for the lack of reliability. According to the 35% of the subjects, topics concerned with subjective interest (politics, religion, history or business) are least reliable. Next reason for lower quality of the content are objective difficulties concerning the topic, which would need professional writing (18%). Scientific and technical topics fall in this category. Poor choice of style and article organization are also mentioned as one of the problems (9%).

Table 2: Opinions on the Wikipedia's reliability issues

| | |
|---|---|
| English version of Wikipedia is of higher quality and I use it more frequently then Croatian version | 75% |
| The quality of Wikipedia article depends on its topic | 57% |
| Wikipedia should be edited and checked by professionals | 52% |
| Whenever I use some data from Wikipedia, I check it in other sources | 46% |
| Open editing option leads to vandalism and misinformation | 39% |
| When writing, it is better to leave Wikipedia out of references list, even we use material from it | 37% |
| Open editing option makes Wikipedia more reliable, because misinformation and errors can be corrected quickly | 26% |
| Content control would considerably slow down the content growth, so I don't find that good idea | 15% |

The article style is one of the points where Wikipedia can be compared with the traditional encyclopedia. It also reflects how the users perceive Wikipedia – as a real encyclopedia, or as a tool for quick and simplified information about various popular subjects. Table 4 shows the answers to question about the current style of Wikipedia's articles, and the style the users think Wikipedia should employ. Most prominent points are the need for the comprehensive style of articles (which is at the moment fulfilled), and the spelling or grammar errors free article. One third of subjects also feel the style should be more like that of a professionally edited encyclopedia, and appropriate for citing. The fact that these qualities are not strongly stressed by users tells us that Wikipedia is not perceived as a counterpart to a traditional encyclopedia, but as an entirely different kind of information source.

Table 3: Style of Wikipedia's articles

| **How Wikipedia's articles…** | **are written** | **should be written** |
|---|---|---|
| So that anyone could understand the content | 93% | 83% |
| Appropriate for citation in papers | 15% | 38% |
| Mostly of questionable quality | 8% | - |
| Professional style, as in traditional encyclopedia | 11% | 32% |
| Without spelling or grammar errors, a role-model | 20% | 62% |
| The style is irrelevant | - | 11% |

## The comparison of Wikipedia and traditional encyclopedia

The opinion that Wikipedia should not be classified as an encyclopedia, or even that Wikipedia in fact deceives its users by calling itself so, is present with some professionals. The former editor-in-chief of *Encyclopedia Britannica* and one of the Wikipedia's critics, Robert McHenry, writes [1] that an average user of encyclopedia has surprisingly low expectations. What he seeks is just a quick answer to his questions. According to McHenry, a more serious user would be satisfied only with the accurate information, but this statement is put to a test in a fast world, overloaded with information. In the survey the subjects were asked in which situation they prefer the traditional encyclopedia, which they consid-

ered more accurate and reliable, to Wikipedia. As table 4 shows, most subjects agreed that they would use a traditional paper encyclopedia when they use the encyclopedic information as a reliable knowledge source, not as a casual information source. It is shown that users tend to turn to traditional encyclopedia when they expect that the information cited will be read and examined by some other party.

Table 4: Situations in which the traditional encyclopedia is preferred

| | |
|---|---|
| When writing a school/college paper or article | 46% |
| When I need to check Wikipedia information | 31% |
| In any situation | 8% |
| Never | 8% |
| Only if the traditional encyclopedia is new and updated | 4% |

When asked why they would use one or the other in specific situations, three recurring points emerged amongst various answers. Subjects explained that they do not want or are not allowed to cite a Wikipedia in a more demanding or serious writing work. Once again they stress that Wikipedia is not very useful when they need accuracy and reliability. A subject gives an example of a betting situation: he or she would turn only to a traditional encyclopedia when they need to check the answer to a general knowledge question on which they made a friendly bet. Wikipedia has almost no authority in such situations.

The second point made about preference of a traditional encyclopedia falls in the domain of national encyclopedias. Quite a few subjects stated that they do not use Wikipedia when they need information about anything specifically Croatian: history, geography, famous people and so on. Croatian Wikipedia is poorly developed in comparison to English version, which is understandable and applicable to almost any smaller national Wikipedia. Since, at the moment, only several authors and editors actively work at the Croatian version, the development is slow, and Croatia related content could be richer. Users that were surveyed recognized that, and they tend to avoid Wikipedia when in need for Croatia related information.

Finally, some users stated an interesting usage for the paper encyclopedia which looks at their format as a certain advantage, not just as a setback. Although it is possible to view editing history on a certain Wikipedia article, the interface is less then comprehensive and shows the changes users have made, and not the actual history of knowledge and thought on the subject. Comparative reading of older paper encyclopedia grants us a look into history of knowledge and spirit of those times. This historical dimension is lost in the paperless world.

## What about the Croatian version?

As we have seen, the users we surveyed mostly agree that the Croatian version of Wikipedia is underdeveloped, and even though 90% have used it, 75% prefer

the English version. They were asked if they think that Croatian version should be more developed, and why. 68 percent of the subjects think it is important that the Croatian version of Wikipedia develops more in the future. In addition, 41% noted that the size and quality of a national version of Wikipedia speaks about the country (or language) itself, and represents it in the digital world.
In order to be more developed, the group of motivated users/volunteers has to work on the articles. Most of the subjects did not take part in the development of Croatian Wikipedia. 38% said that they never thought about editing or writing Wikipedia articles, and 21% said they are not interested in participating in development of Croatian Wikipedia at all. 27% did write or edit an article (two thirds of that number just edited), and the rest stated that they find it too complicated, technically or content-wise. Although most surveyed users agree that we should develop Croatian Wikipedia, they feel it is a job someone else should do. This fact is in opposition to the free and open approach that Wikipedia promotes, so it is unclear why so little motivation exists. Table 5 shows most important reasons for or against the development of Croatian Wikipedia which were stated in the survey[7]. The main reason stated is to make Wikipedia more accessible to those who do not speak foreign languages, and the reasons connected to Croatian identity or Croatia-related facts follow.

Table 5: Should we develop the Croatian version of Wikipedia?

| | |
|---|---|
| I find it important | 68% |
| I don't find it important | 32% |
| Yes, for those who do not speak English or other foreign language | 52% |
| Yes, because it represents Croatian identity | 26% |
| Yes, to add more information about Croatia | 14% |
| Yes, because we need Croatian perspective for some events | 8% |
| No, there is no need | 11% |
| No, the English version will always provide more quality and content | 5% |

Regardless of the Croatian version issue, the points which were made about the difference between traditional encyclopedia and Wikipedia are still valid. Even those subjects who have no problem with reading the English version, and did not need Croatia-related information do not use Wikipedia as an exclusive, reliable source of general encyclopedic information, let alone the specialized information for a certain profession. This does not mean Wikipedia is useless, should be ignored or that its development should stop.

[7] More than one reason was acceptable.

## Conclusion

The survey showed the distinct boundary in perception in usage of Wikipedia and traditional encyclopedia. On the debate which one should be used, the answer mostly depends on the motivation for the usage. But more importantly, the results of the survey show that Wikipedia is used more than traditional encyclopedia, despite the awareness of its shortcomings. When deciding which encyclopedic work to use, Wikipedia's vast choice of topics, quick searching and open access are shown to be more important than information accuracy. These results incline us to think about migrating our traditional, reliable, encyclopedia to a digital space, and learn from Wikipedia's positive sides. If there was a "traditional encyclopedia" (traditional in sense of professional editing and proofread, accurate content) which was as accessible and updated as Wikipedia, it would satisfy both casual and demanding users. The process of evolution of encyclopedia perception and usage is visible throughout the results. Thus, it is argued that the traditional values of the encyclopedic science should be reconsidered. For example, a Wikipedia critic may argue that Wikipedia errs in calling itself an encyclopedia; but we can reply that the definition of encyclopedia might have changed, and that encyclopedist must consider this when designing his or her next encyclopedic work.

## References

McHenry, Robert. The Faith-Based Encyclopedia. 11/15/04. http://www.tcsdaily.com/article.aspx?id=111504A (08/05/09)

McLuhan, Marshall. The Gutenberg Galaxy: The Making of Typographic Man. Toronto: University of Toronto Press, 1962.

Miller, Rich. A Look Inside Wikipedia's Infrastructure. 06/24/08. http://www.datacenterknowledge.com/archives/2008/06/24/a-look-inside-wikipedias-infrastructure (08/05/09)

Shores, Louis. Encyclopedics. // *Reference as the Promotion of Free Inquiry,* Littleton, CO: Libraries Unlimited, 1976, p. 162.

Professional paper

# Objective Journalism or Copy-Pasted Press Releases: A Preliminary Media Content Analysis

Siniša Bosanac, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sbosanac@ffzg.hr

Bojana Mandić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
bmandic@ffzg.hr

Andrija Sprčić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
andrija.sprcic@gmail.com

**Summary**

*The volume of information that is transmitted to the public is constantly growing. Unfortunately, the quantity and quality of information in mass media are often in inverse proportion. Objectivity in reporting and investigative journalism are becoming more of an exception than a rule as journalists invest less effort in their work. Reasons for this are various, ranging from economic to ethical, and we tried to present some of them in this article. The aim of the analysis was to find out what is happening in the process of news creation at a stage between the source of information; in this case the public relations manager, and the end product; a news article written using a press release as a source. For this purpose we have analysed a selected volume of press releases and news articles, and compared their content. The results show a considerable overlap, not just regarding the identical information, but also verbatim copies of whole sections of text. Our preliminary findings call for establishing of a quality-control mechanism that would remedy the situation.*

**Key words:** plagiarism, press release, media article, comparison, paraphrase

## Introduction

The right to receive and give information and the freedom of expression are fundamental rights of every human being. The public has the right to know objective facts and other people's opinions and from this right follow the obligations and rights of journalists[1]. They are obliged to comply with the highest professional and ethical standards, but despite of that we are witnesses that this is not always done or even possible to do.

The public likes to be informed of everyday events on a regular basis. The people who they mostly trust and from whom they get the news are journalists from their favorite newspapers or magazines. To gain and obtain such trust, a journalist must at every time give exact, and most importantly, verified information. After recapitulating these principles, we asked ourselves a few questions: "Do they really have enough time for investigative journalism and what is the quality of articles we are reading? Is the source most of the time the same, or do they contain objective observations?". This urged us to do a research of the Croatian media by comparing press releases issued by a public relations agency with the articles published in the media after the PRS is distributed. Clarifying the communication process between the public, journalists and public relations managers in this article we will give an overview of the situation we encountered when doing an analysis of newspaper articles.

Based on the available information, we found that several similar researches were done. Barbara Bearns, a well respected professor at the University of Berlin, back in 1985 wrote her hypothesis on determination in which she stated that public relations determine journalism. Her study showed that most of the published news come from different sections of public relations offices. According to her, almost two thirds of the articles that reported on politics of the German province of North Rhine – Westphalia came from information given to the journalists by the press office, adding that journalist did not consult other sources on received announcements that were based on public relations materials. It was also said that if there were any other researches, that they are mostly focused on requesting statements and postures on events.[2]

Another research which is important to mention was done in 2007, by Julije Katančević, a consultant at "Dobre komunikacije", in which he investigated the use of public sources of informing and concluded that in most cases journalist believe and publish information without investigating it a little bit further.[3] Both

[1] Verčič, D.; Zavrl, F.; Rijavec, P.; Tkalac Vercic, A.; Laco, K. Media relations. Zagreb: Masmedia, 2004.

[2] Kunczik, M. Public relations: Concepts and theories, 4. ed., Böhlau Verlag, Köln, Weimar and Vienna, 2002.

[3] Malović, S., Sušanj, D., & Jusić, D. (2007, March 9). ICT journalism and public relations. Zagreb, Croatia.

Katančević and Bearns affirmed that journalists investigate less and they increasingly more often use completed materials.

To our knowledge, Croatia does not have any system or law in force that verifies and controls the media's accuracy and to what amount it is allowed to copy an article, even if it's from a PR manager. We can only say that the principle posture agreed upon by the journalists is not to copy, but the reality differs.

According to the Merriam-Webster Online Dictionary, the definition of the verb *to plagiarize* is: "to steal and pass off (the ideas or words of another) as one's own: use (another's production) without crediting the source, to commit literary theft: present as new and original an idea or product derived from an existing source"[4].

The word derived from Latin verb *plagere* which means *to kidnap*, *steal*. It implies the act of appropriation or inclusion of someone else's writing or other creative work in their own, in whole or in part, without adequate recognition of another's work. Plagiarism, unlike counterfeiting in which is questionable the authenticity of the work itself, is concerned with false attribution of work. Plagiarism can also happen unintentionally[5].

The Croatian Copyright Law clearly states that "Author's work is an original intellectual creation in literary, scientific and artistic fields, which has individual character, regardless of the manner and form of expression, type, value or purpose"[6]. The Croatian Copyright Law provides information that copyright does not include discoveries, official government texts such as laws, regulations, logs, court decisions, standards, daily news and other news that can be recognized as regular media information. This information was confirmed after an official at the State Intellectual Property Office of the Republic of Croatia (SIPORC) assured us that no manner of public information could be regarded as intellectual property, unless it has individual character. Individual character of one's writing is something that highly praised journalists develop during years of writing and reporting. Their own opinion on the subject is cleverly woven in the textual structure of an article by an unique, individual approach.

There are several different types of plagiarism. According to Plagiarism.org, we can regard plagiarism in form of *sources not cited* and *sources cited* (but still plagiarized). Although the source provides more than enough information on the types of plagiarism, the most commonly found in the results of our research were results of: *another's work*, *copied word-for-word*, *as their own* (most

[4] Merriam-Webster Online Dictionary: Plagiarize. URL: http://www.merriam-webster.com/dictionary/plagiarizing (2009, August 25)

[5] Wikipedia, the free encyclopaedia: Plagijat. (2009, June 3) URL: http://hr.wikipedia.org/wiki/Plagijat (2009, June 21)

[6] Narodne novine (2003) *Zakon o autorskom pravu i srodnim pravima*. Zagreb: Narodne novine d.d., (167)

commonly referred to as the *verbatim copy*); *copying significant portions of text from a single source, without alteration*; *tweaking the sentences to make them fit together while retaining most of original phrasing and paraphrasing most of the text from other sources and make it all fit together*.

**Hypothesis**

Journalism and public relations can undoubtedly be put in the same professional branch, but even if the same technique is used, these are two different professions with great differences. The role of a PR manager in today's time is indispensable from the company's point of view, because it relies on their activities to communicate with the public. From the journalist's point of view, they provide them with timely information about certain events.

Petra Dorsch (1982), professor Emerita of Social Communication and Media Research at the University of Munich (LMU), found in her research that the lack of public relations managers would cause information gaps, because the actual function of PR managers is to disburden journalist to a certain point.[7] If we lay our complete trust in PR managers, not saying this should be in any way bad, that does not mean that investigative journalism should be ignored. A good article is defined as having accurate data provided by subject of whom the journalist is writing about, and having an additional dose of objectivity and knowledge. On top of that, they should be inputting their own comments if they are going to sign the article. Many of them withhold the information of the source of the information and without any shame sign an article that they have completely copied from the source, taking credit for the work. People believe that because they do not have any reason to distrust the people whose job is to provide information.

According to the HND[8] secretary general, Vladimir Lulić, and HUOJ[9] secretary, Boris Hajoš, and their experiences in practice, it is acceptable for a journalist to copy certain formulations from a press release there aren't hidden ads behind it, if it is about a brief news announcement or if there is limited space available for publishing the information in the publication. Other exceptions also include relaying the announcements from the police, court or state's attorney, or those received minutes before broadcasting an informative program. The general opinion is that PR managers and their clients would be much happier to get dozens of different articles regarding their company than ten articles that are verbatim copies of the same text[10].

[7] Kunczik, M. Public relations: Concepts and theories, 4. ed., Böhlau Verlag, Köln, Weimar and Vienna, 2002.

[8] Croatian Journalists' Association

[9] Croatian Public Relations Association

[10] Lulić, V. Private message. (2009, August 21); Hajoš, B. Private message. (2009, August 22)

We anticipated that the majority of the articles found in printed media, and especially on internet portals are verbatim copies of the original press release sent to the journalist by a PR manager.

## Methods

Our analysis is aimed at recognizing two types of plagiarism. The first type is the *verbatim copy*, which is in recent times colloquially referred to as "*copy-paste*". The second type, which is somewhat harder to detect, is *paraphrasing*, which is defined as "restatement of a text or passage giving the [same] meaning in another form".[11]

In theory, the procedure for spotting verbatim copies is relatively simple. It is only necessary to compare the content of the two documents, and determine whether their content overlaps. Spotting paraphrases is done by close reading of suspected documents, and comparing the meaning of individual sentences or larger parts of texts. Manual comparison is suitable and relatively accurate when fewer documents are examined.

### Data Collection

Before carrying out an analysis, it is necessary to choose a set of documents which are suspected to contain plagiarized material. In this article they will be referred to as *derived documents*. After that, a set of documents for which is suspected that the plagiarized material originated from is selected. These documents will be referred to as *source documents*. In our analysis, this was relatively easy due to the fact that we were dealing with a specialized type of documents which are in fact created for the purpose of being a source for media articles.

In a larger collection of documents, this becomes extremely resource-demanding, and would need to be done automatically using search engines and large databases of documents that are potential sources for the content that was plagiarized.

Suspected *source documents* were press releases granted to us by a market communications agency specialized in public relations that produced them for the ICT-related company.

The pool of suspected *derived documents* was created by collecting media articles using a media monitoring service "Press Clipping" and selecting articles from July 2008 – July 2009 period that mentioned a certain ICT- related company. All data provided by the media monitoring service was in .PDF format, so it was necessary to convert it to an editable format, and to extract individual articles. This was done using optical character recognition software ABBY Fine Reader.

[11] Merriam-Webster Online Dictionary: Paraphrase. URL: http://www.merriam-webster.com/dictionary/paraphrase (2009, August 25)

As a result we got 217 media articles as suspected *derived documents*, and 26 press releases as suspected *source documents*. All of the documents were published in the same one-year period. Articles from these four types of publications were included in the analysis: daily newspaper, weekly newspaper, monthly magazine, and Internet news portal article. The average length of articles was around 400 words.

**Pairing of documents**

The first step of the analysis was to pair *source* and *derived* documents. This was done in three steps. The first was comparing the publication dates and discarding *derived documents* with publication dates earlier than *source documents*. Documents with small interval in publication were paired together during the first pass. The next step was to pair documents according to titles. Where there was ambiguity, we compared the two documents by preliminary comparing the content of two documents with regards to topic.

Out of 217 articles, we were able to link 131 of them with a corresponding press release, providing an average 5.03 articles per press release. The other 86 articles were mostly false positive search results, or covered the ICT-related company in a way that was completely unrelated to topics covered by available press releases.

In order to show more reliable results, only four pairs of documents that had *source – derived* document ratio higher than 10 were selected for an in-depth analysis.

**Analysis**

The analysis was done on a sample of four *source* documents, and 60 *derived* documents.

The documents were prepared for semi-automatic comparison by trimming additional content except the title and content. The comparison itself was done using the *Compare documents* feature in Microsoft Word 2007. After the tool marked all the words that occurred in the same places in both documents, and manually correcting the results, the word-count tool was used to measure the size of the copied content.

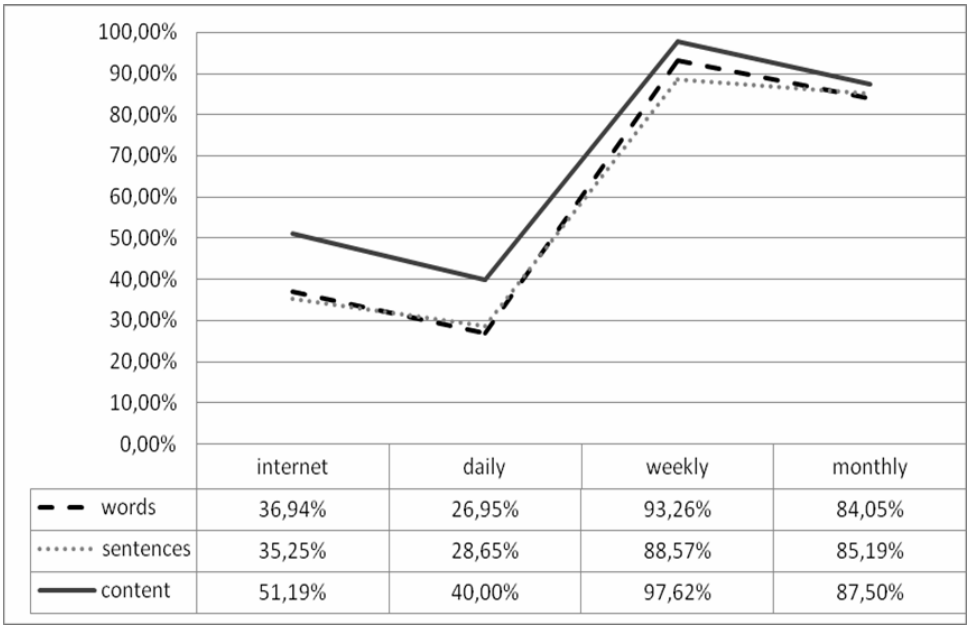
During manual correction, only word strings longer than three words were counted as copied. Names of organizations and products, formal titles, and common phrases were not regarded as copied content.

The number of completely and partially copied sentences was determined by manual inspection. The number of paraphrased sentences was determined after close reading of the remainder of the document.

## Results

The results show that the highest percentage of copied words, sentences and paraphrased parts of the PRS is in weekly (88.5% / 97.6%) and monthly (85.2%

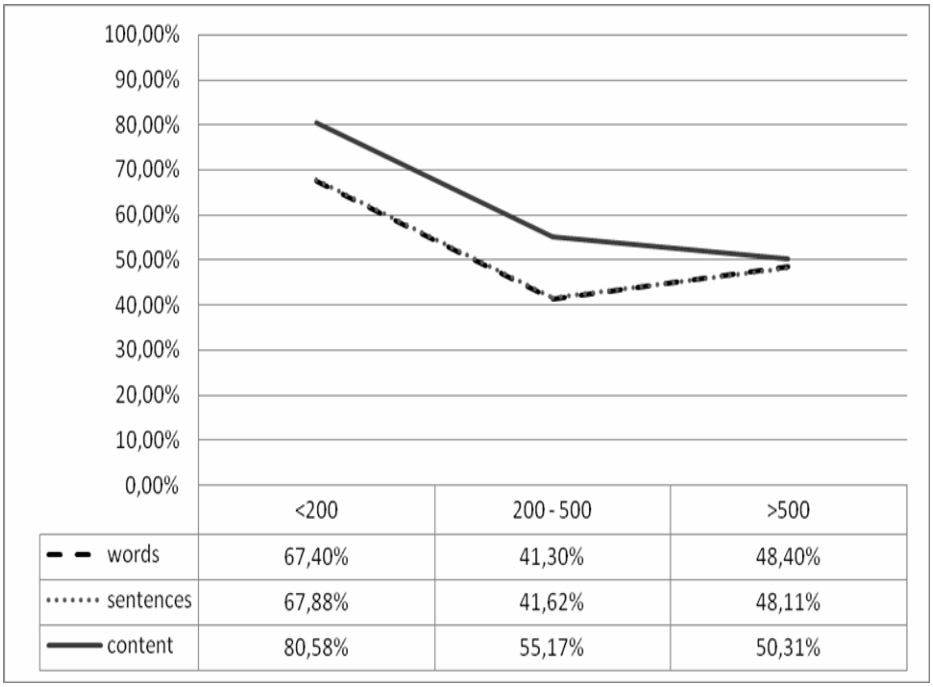
/ 87.5%) editions analysed. We should accentuate that the number of weekly editions is not representative in this analysis so this should be verified by a more focused inquiry. The Internet editions copied as high as 35.25% / 51.2% and the smallest percentage has been reported in daily newspapers (25.75% / 41.41%), business newspapers – usually daily editions (34.44% / 37.2%).
The number of words in an article should also be taken into consideration because although the monthly and weekly editions have the greatest percentage of copied content, they have by far the smallest amount of words per article. Weekly editions have an average of 274.66 words while the monthly editions have an average of 318.41 words. Internet editions have an average of 344.41 words; daily editions have an average of 460.5 words while business editions have the highest average of 503.66 words per article.




| | internet | daily | weekly | monthly |
|---|---|---|---|---|
| words | 36,94% | 26,95% | 93,26% | 84,05% |
| sentences | 35,25% | 28,65% | 88,57% | 85,19% |
| content | 51,19% | 40,00% | 97,62% | 87,50% |

Figure 1: Percentage of copied content according to type of publication

On average, the percentage of copied content is highest in articles with fewer than 200 words in size, with 6 of 11 articles entirely consisting of whole sentences extracted from press releases. In total 16 of 45 articles, more than one third are written by copying or rephrasing the press releases, i.e. without any original contribution by the author.




| | <200 | 200 - 500 | >500 |
|---|---|---|---|
| words | 67,40% | 41,30% | 48,40% |
| sentences | 67,88% | 41,62% | 48,11% |
| content | 80,58% | 55,17% | 50,31% |

Figure 2: Percentage of copied content according to size of articles

## Interpretation of results

The aim of a public relations manager is that his press release is used by the journalist to have truthful, complete and utterly reliable information. His/her services do not stop when he distributes the information; he/she is available for any other inquires. A good relationship between a journalist and public relations manager is based on long term experience and a good business relationship after years of working together. To public relations managers the most important thing is to preserve their credibility and reliability in their work. A public relations manager’s reputation is everything in his job. Losing it will be job threatening, depending on the scale of his mistake. For a journalist, a worse situation can happen, if he is known to be low-quality, non-creative and non-professional, or in other words, loses his professional integrity. Although, we can fall in a trap of blaming solely them for copying an entire article, we must consider that the same article has to go through the hands of editor in chief, or column editors before it gets published. There is a whole hierarchy through which the article has to be run through, and it is evident that it fails to detect plagiarisms.

Although it is logical that every PR manager is pleased when his press release gets published in whole, and when a journalist mentions his client in a positive way, there is a breach of professional ethics in taking credit for someone else’s written work. The results show that the same press release gets published word by word in more than a few articles. To what extent does this practice contribute

to the credibility of the information, the objectivity of the journalist, and their professionalism? Is the real intent really accomplished?! Such text should be used as a starting point for further investigations and journalists should not completely rely on it to use it and present it as a finished article. Sometimes the reason lies in journalists' nonchalance, or it is done when they are struggling with a deadline and are under pressure, not having enough time to do research for the article. At times they have to write multiple articles for the same newspaper issue, or they are unfamiliar with the topic. Often the problem lies in the lack of specialised journalists in the newspaper and/or low fees they get paid for their job. Often we come to realize that a specialised approach is needed when dealing with topics that require a certain level of understanding of the topic. In many cases, unfortunately, the lack of professionalism has roots in more than just one problem.

## Conclusion

The aim of this research was to show the current situation in printed media in terms of originality. While investigating the reasons that could explain the results of our study we noticed that one of the problems lies behind the financial issues that editors have at hand. Publishing many of the monthly editions is costly and editors mainly rely on big advertisers as most of their profit comes from them. Journalists reportedly practice self-censorship to protect the economic interest of owners and major advertisers. Lacking the sufficient money to employ more people they are liable to search cheap work force in inexperienced students. And when that doesn't help they are forced to completely rely on press release to fulfil their columns and pages. Even though the results show a high percentage of "copy-pasting" they cannot be viewed without taking into consideration the outer and inner factors that might have influenced the publishing.

In order to conduct this type of analysis more easily, it is necessary to develop adequate tools that would enable the comparison of a much higher volume of articles in order to get more reliable results and ensure a more frequent analysis as one of the methods of quality control in journalism which would help in creating an information society that is more objectively informed.

## References

Kunczik, M. Public relations: Concepts and theories, 4. ed., Böhlau Verlag, Köln, Weimar and Vienna, 2002.

Lulić, V. Private message. (2009, August 21); Hajoš, B. Private message. (2009, August 22)

Malović, S., Sušanj, D., & Jusić, D. (2007, March 9). ICT journalism and public relations. Zagreb, Croatia.

Narodne novine (2003) *Zakon o autorskom pravu i srodnim pravima*. Zagreb: Narodne novine d.d., (167)

Verčič, D.; Zavrl, F.; Rijavec, P.; Tkalac Verčić, A.; Laco, K. Media relations. Zagreb: Masmedia, 2004.

Wikipedia, the free encyclopaedia: Plagijat. 2009, June 3 URL: http://hr.wikipedia.org/wiki/Plagijat (2009, June 19)

# See Also: Auto Generated Recommendations

Mislav Cimperšak, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mcimper1@ffzg.hr

Marija Tkalec, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mtkalec@ffzg.hr

Siniša Jovčić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sinisasanseverija@yahoo.com

**Summary**

*Wikipedia is one of the most used encyclopedias today. It earned its status for more than one reason: easy accessibility, regular and quick updates and useful hyperlinks within the articles. Apart from linking to the other articles within the main body of an article, Wikipedia also uses a section called "See also" in which similar or related articles are listed.*
*The main idea behind this research is the creation of automatic recommendation for the "See also" section based on soft clustering of document similarities with the addition of the hyperlinks within the main body of an article as the observed objects. The research is conducted on the Croatian version of Wikipedia, with short articles omitted. In the article it was concluded* that the procedure cannot be regarded as being successful enough for an unsupervised implementation on articles in Croatian Wikipedia.

**Key words:** Wikipedia, See also, document clustering, soft clustering

## Introduction

Today Wikipedia[1] is a very reliable source of information which is accessible to everyone around the world. One of the biggest Wikipedia's advantages is that it

[1] Wikipedia, 2009.

is the most up-to-date online encyclopedia which means that its articles are in most cases most prompt of all articles from other online encyclopedias. People who want to extend their knowledge daily visit Wikipedia because of its quick and easy accessibility.
It is also necessary to mention Wikipedia disadvantages. The most talked about disadvantage is uncertainty of accuracy of information provided, but it should also be taken in consideration that articles with a lot of sources are in most cases more accurate. Because anyone can start a new article, topics that usually would be considered too obscure for a general encyclopedia are included leading to may very specific articles which are often very short.
Among many different sections, Wikipedia includes a section called "See also" which includes the list of similar or related articles to current article which urges users to continue browsing and reading articles on the page itself.
In classic encyclopedias entries often have further reading sections that list a handful of related articles or even entire books, making encyclopedias extensive references for scientific research. The same idea is built inside the Wikipedia using the connections between articles within the article text and the "See also" section on its own.
The main topic of our research described in this paper is the creation of an automatic recommendation system for the "See also" section based on soft clustering of documents.
The research was performed on 5,012 articles taken from the Croatian version of Wikipedia.

## Thesis

Within the body of an article text Wikipedia's users can add connections to other articles on Wikipedia providing a useful way of finding out more about an unknown term within the article. The idea behind the research was that users on similar topics create connections to same articles. That would mean that by comparing two articles connections we could conclude how similar those two articles are. Since there are usually not that many connections per article based on which we could do the comparison, we compared the whole text with added extra weight on the article connections making them more important than the rest of the article text.

## Database

10,100 articles were originally collected and run through a specific cleaning process. The cleaning process included the removal of all HTML tags (except connections to other Wikipedia articles within the text of an article), general parts of Wikipedia such as navigational components, articles with the same title, but in different languages. "See also"/"Related articles" parts were also removed since they would impose their Wikipedia connections to the overall re-

sults. After the cleaning process there remained 5,012 articles that we used as our main database.
Articles were collected between 12th and 27th August 2009 using the application "Wget" freely available online and preinstalled in most Unix-like operating systems. During the process of collecting some of the Wikipedia's webpages were omitted such as category pages since they contain only connections to articles without any real text on their own, further more editing pages, user pages, user's talk pages and Wikipedia help pages.
Articles shorter than 4,000 characters, after the above mentioned cleaning process, were discarded because they do not contain enough information needed for reasonable clustering.

## Research

Based on a previous research[2] in which the optimal procedure for one-pass document soft clustering[3] was determined, we used tokens as vector features and document similarity threshold of 0.5.
We compared our results to human created connections to similar articles (where they were available) on Wikipedia itself taking in consideration the objectivity of selected connections (there are well known cases of private companies editing Wikipedia under the mask of ordinary users and adding connections to the article about their company and/or product creating for themselves a sort of free advertisement).
Connections within Wikipedia were treated as separate tokens which were given extra weight when comparing the articles.
Out of 5,012 articles, 509 clusters were created. Average size of a cluster was 14.12 articles per cluster and a single article was places in averagely 2.4 clusters.
Analyzing the final clustering results we note that clusters can be classified in three categories. First category includes clusters which have no real value, i.e. clusters which contain totally incoherent articles of quite different areas. Second category includes partly relevant clusters containing some articles of the same area and others of some different topic area. And finally the third category includes well-formed clusters whose articles are completely based around the same area and therefore it is obvious that they would be very good candidates for one's "See also" section.

### Clusters with no real value

In a certain amount of cases, generated clusters were not usable, because the subject of those articles is in a completely different theme area. Explanation:

[2] Cimperšak, Tkalec, 2009.

[3] Jain, 1999.

because of the added weight on the connections to the other Wikipedia articles the algorithm grouped together articles that were at first glance not related, but were for an instance connected through a random number (whether a number on its own or a year within a date etc.) or through a country name as the case was in the example of cluster number three which contains four members.
There is an apparent case of total disconnection of articles placed in that cluster. The topic of article number 4929 is St. Peter, article number 4450 Saint-John Perse, article number 1697 General Staff of the Armed Forces of the Republic of Croatia, article number 1709 French Guiana and article number 3027 Marine mammals. At first sight, it's obvious that all of the above mentioned articles are not related and they are not covering similar themes. Another example of such a cluster was cluster number 11 which consists of three members where the topic of the first article are Eurasian Avars, the second is Psychology and the topic of the third article are birds.
With this term, clusters with no real value, we could also describe clusters which contain too many articles, such as clusters containing more than 30 articles, and there were a fair number of clusters of that type. Since our research was conducted on a "smaller" Wikipedia[4], we consider that it is unwise to take into consideration clusters of such enormous size, because it is impossible that articles inside those clusters are connected through the same topic, and hence most of the articles within the cluster are absolutely unconnected. As an example, we enclose cluster number 171 which contains 39 articles and cluster number 530 which consists of 201 articles from which is obvious that it is not possible that all of them are closely related through the same topic considering Croatian Wikipedia.
The number of useless clusters is unfortunately larger than anticipated which takes us to the conclusion that our algorithm for finding related articles isn't creating satisfactory results in this case.

**Partially relevant clusters**

Some articles within this kind of clusters are thematically related, while the remaining articles are not bound with the same subject or they don't involve the same or similar area. A good example of such a cluster is cluster number 529 where the subject of an article number 1908 is the Croatian Football Federation, the subject of articles 1909 and 1911 are Parliamentary elections, 1919 Orthography, 1914 Presidential elections, 1768 Croatian Academy of Sciences and Arts and finally 2816 Loyalist (American Revolution). Analyzing the example above, everyone can notice that one part of articles share a similar theme, because they talk about elections (1909, 1911, 1914), some about Croatia (1908

[4] By smaller we mean smaller in the amount of articles available on Croatian Wikipedia when compared to some others.

and 1768 directly and 1909, 1911, 1914 indirectly) while the remainder includes articles no more related to the same election theme.
Based on the given results, we can conclude that among all of the created clusters, partially relevant clusters are the most common.

**Well-formed clusters**
Clusters that explicitly contain articles connected to the same subject are too rare for our likings. Cluster number 432 which speaks about the Olympic Games is a very good example of a well-formed cluster. Articles placed in this cluster involve Olympic Games in Tokyo, London; Barcelona, Atlanta, Athena and Beijing. The subject of one article is Berlin and it contains connection to the 1936 Summer Olympic Games. To determine if the cluster is well formed, we evaluated our generated results to user created "See also" section of those articles in the Croatian Wikipedia. We observed that within the Olympic Games article there is a list of Winter Olympic Games created by users where all list items are shown as connections leading to articles where the subject of each article is the Winter Olympic Games hosted by another city.
Cluster number 357 includes articles whose common topic is football teams and every article belonging to that cluster describes a different team.
Cluster number 511 involves articles speaking of different varieties of Airbus airplanes. While the number of such articles is six, it also includes an article about a Boeing 747 for which is clear that the main topic is still an airplane. One of articles is about Ahmed Sékou Touré and it contains a connection to the article about France which was probably crucial when placing that article under the above mentioned cluster since the centre of the European air company Airbus is in Toulouse in France and articles about Airbuses also include connections to the article about France. When we compare our results to user created "See also" section in the article about Airbus 380 there are connections to articles about Airbus A350 and Boeing 747 (and to three more airplanes, but their articles were not in our database since they were shorter than 4000 characters) which demonstrates our initial concept.

## Observations
During the course of our research we noticed that Wikipedia users (by users we mean contributing users) more often create connections on more general and, at first glance, more obvious terms such as dates (whether complete dates or just names of the months or years), geographical terms (cities, countries), objects of general use etc. For example, article number 4329 about Roman-Persian wars contains a great number of connections leading to articles about Iranian empires, Mesopotamia, Arab Muslim armies, Euphrates, North Africa etc. which are all general terms. More seldom they create connections to highly specific or specialized terms or terms less known in general public (such as a name of an

unfamiliar town), because they are not aware that the article about that term was even published.

## Conclusion

Based on the results presented here, the procedure cannot be regarded as being successful enough for an unsupervised implementation on articles in Croatian Wikipedia. Unfortunately it would be necessary to manually inspect each generated recommendation.

Most likely the algorithm would be more successful in a strictly supervised encyclopedia where connections to all possible terms would exist and not just on those which are more familiar to an average Wikipedia user.

## References

Cimperšak, Mislav, Tkalec, Marija. Utvrđivanje optimalnog postupka za raspršeno grupiranje jednim prolaskom. Zagreb; Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu. 2009

Jain, Anil K., Murty M. Narsimha, Flynn, Patrick J. Data Clustering: A Review. ACM Computing Surveys. Vol. 31, 1999. No. 3; 264-323

Ljubešić, Nikola; Agić, Željko; Bakarić, Nikola. Document Representation Methods for News Event Detection in Croatian. Zagreb : Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu, 2008. 79-84

Wikipedia. Wikipedia. http://en.wikipedia.org/wiki/Wikipedia (19th August 2009)

Original scientific paper

# Social Software: Teaching Tool or Not?

Marija Matešić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
marija.matesic@gmail.com

Kristina Vučković
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
kvuckovi@ffzg.hr

Zdravko Dovedan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
zdovedan@hotmail.com

**Summary**

*Social software, in means of software that enables people to connect with one another, exchange opinions and information online, has now become a recognizable phenomenon at the early stages of Internet evolution. Significant increase in use of various applications which these social services provide has been noted in the last few years. The goal of the survey was to find out how many teachers and students at universities in Croatia use social software and for what purpose. The research was also aimed at how often they use it and to check what are the advantages and flaws of social software in means of an auxiliary tool in education. Research was conducted during the spring of 2009 in form of an online survey for teachers and students at universities and academies in Croatia. The survey investigates respondent's familiarity with five types of applications: 1) social networks (Facebook, MySpace and LinkedIn); 2) media sharing site (YouTube, Flickr and SlideShare); 3) social bookmarking or tagging sites (Delicious and StumbleUpon); 4) wikis (Wikipedia); 5) blogs and microblogs (Twitter). 368 people participated, 100 teachers and 268 students.*

**Key words:** education, social software, social networks, media sharing site, social bookmarking, tagging sites, wikis, blogs, microblogs, teaching tools.

## Introduction

Our everyday lives are shaped by technology development. The flow of information and knowledge enabled by different types of media has changed throughout the history. This change was influenced by the technological progress that simultaneously reorganized the way media allowed information flow. For some time now, the networked media have been recognizable phenomenon of our society. The appearance of new forms of networked media, known as social software, can be seen as the following step in the information and knowledge exchange process.

Social software is a media with primary purpose to enable its users to connect and communicate in a networked environment. Such an environment changes the patterns of private and business communication, but also learning models and information and knowledge flow.

Universities are traditionally considered to be sources of new ideas and knowledge and places that gather unlimited potentials of young people. At the same time they are places organized on some traditional principles of sharing knowledge, and although incredibly innovative are slow to changes of any kind.

It is obvious how the development of technology brings changes in media and how it influences the work of publishers, librarians, press release experts and marketing experts. The main question for us here is how this influences the work of a university. This paper will research the ways in which university teachers and students in Croatia respond to changes that social software brings regarding possible reorganization of the traditional models in the teaching process and the traditional forms of communication in that model.

## Social software

In pre-technological era the vast transfer of inormation was possible only through oral communication. Printing press has changed all that alowing multiple copies to be easily prepared for their distribution. The next big step was made with the emergence of networked media that allowed completely new ways of information flow. This new media lets one be an information user and its creator at the same time thus changing the nature of an interaction between an information and its user.

Social media, known also as a social software, appears as a special form of newtorked media. It is a set of tools, aplications and/or services that enables its users online interaction, information (or knowledge) sharing and exchange of opinions.

The development of social media is intertwined with the history of Internet. The first forms of social software were mailing lists, chatrooms and instant messaging that appeared in the 1970 (Boyd, 2008). However, the average user did not have an access to it until 1993 when the first browser, Mosaic, was built and the popularity of World Wide Web started to rise.

The peak of social software development and its broader usage happened in 2004 when the second phase of Internet development, known as web2.0, has started (O'Reilly, 2005). At that time, some new forms of social softare started to appear. Among them are social networks, media sharing services, social bookmarking and tagging services, content discovery services, wikis, blogs and microblogs.

**Social networks** are a new genre of social software. In the past few years it has recorded a lagre increase in use. Its predecessors are considered to be online dating sites (Boyd, 2008). Uniqueness of this genre lies in the following properties: 1.) creation of public or semipublic user profiles within the system; 2.) creation of personal groups of contacts (one to one, one to many or many to many) with whom the user to some extent shares the same views; 3.) browsing the profiles of others within the group or within the system. This category merges features of all the other genres of social softare. Its most representative[1] examples are Facebook (2004), MySpace (2003) and LinkedIn (2003) all of which are included in this research.

**Media sharing services** have a primary role to enable its users exchange of different types of data. In this paper, this category is represented by the following services: Flickr[2] (2004), YouTube[3] (2005) and SlideShare[4] (2006).

**Wikies** are systems or programs with selforganized structures that allow its users to brows, create and edit different contents. The most representative example of this category is Wikipedia (2001).

**Blog** is a personalized network site written in the form of a magazine or journal, i.e. a system of published posts displayed in the opposite chronological order created by an author or a group of authors. This type of social software enables individuals to publicly express their opinions about certain views. It is, to some extent, identified with an amateur journalism. Collection of all the blogs makes the blogosphere which is again a type of social network.

**Microblogs** are a new form of blogs that enable its users to enter 140 characters long entries which are than displayed in the real time and are visible to their group of contacts. We have used Twitter (2006) as the most representative form of a microblog.

Certain shifts happened also in the area of **bookmarking, tagging and categorizing items** on the web sites. The set of key words or tags that describe each item are selected by the users in real time. In this way they themselves catego-

[1] In respect of its number of users.

[2] Flickr is a service that enables its users to store, organize, search and share photos, add comments and leave notes beside the photo.

[3] YouTube is a service that enables its users to share videos ranging from educational to entertaining content.

[4] SlideShare is a service that enables its users to share slides.

rize the available contents. This procedure is known as folksonomy and it is an opposite form of a taxonomic approach to content categorization. The main representative of this category that we have included in our research is Delicious (2005).

**Content discovery** services allow the user to find the web content based on the given parameters. The pages that have been marked as positive by the users are the ones that are displayed. The main representative of this category is StumbleUpon (2006).

## Research

### Sample selection

The research was conducted during the spring of 2009. The total of 368 teachers and students of the Universites of the Republic of Croatia were included in the survey, 100 of which were teachers and 268 were students.

Age of student group ranges from 18 to 27 years. Most of them are students of technical sciences (38.43%) and social sciences (30.97%), followed by natural sciences (16.42%), humanities (12.69%), biomedical sciences (1.12%) and biotechnical sciences (0.37%). Regarding the gender, this group is made of 55% male and 45% female respondents.

The teacher group includes mostly young resarch and teaching assistants which leads us to conclusion that higher-ranking teachers eather find the online surveys too demanding to use or not scientific enough. The respondents from this group come from the social sciences (42%), humanities (31%) and technical sciences (22%) and only small number from natural sciences (5%). In contrast to the students' group, this group is made of only 37% male and 63% female respondents.

### Questionnaire

Separate online surveys were designed for teachers and students. The first set of questions were designed to investigate respondents' familiarity with different types of social software named and described in the previous chapter. The second set of questions was offered to find out how often our respondents use the social software and for what purposes. The last set of questions was to see what in their opinion would be the advantages and flaws of using social software as an auxiliary tool in education.

## Results

We have divided our results in four separate sections that we will present here in the following order: familiarity, usage, social software as a teaching tool, and advantages and flaws of social software in education process. Graphical representations will be given where appropriate.

**Familiarity**

When asked if they are familiar with different types of social software and whether or not they use it, the answers within each group were the following: all but one student are familiar with social software and of that 86.09% actualy use it while 13.53% of the students is familiar with social software but do not use it; 94% of teachers are familiar with social software, but within this group only 59% use it while even 35% although familiar with it do not use any of the above mentioned applications. Among students only one respondent (0.38%) is not familiar with social software while 6% of teachers are not familiar with it.

**Usage**

The most used services among the student group are YouTube (76.12%), Wikipedia (70.52%) and Facebook (64.43%). The most used applications among teachers are the same three services but in different order of popularity with Wikipedia on the first place (53%) than YouTube (42%) and Facebook (27%) followed by blogs (15%). In contrast to Facebook, usage of other two social networks is very rare among our respondents. Only 11.94% of students and 1% of teachers use MySpace and 6.34% of students and 7% of teachers use LinkedIn.

Resons for using social software are given in Table 1 for both groups. It is very clear from the Table 1 that teachers and students both use the social software the most often for the data retrieval and information extraction and for the educational purposes. This is more popular way of researching and data extracting among students than it is among teachers.

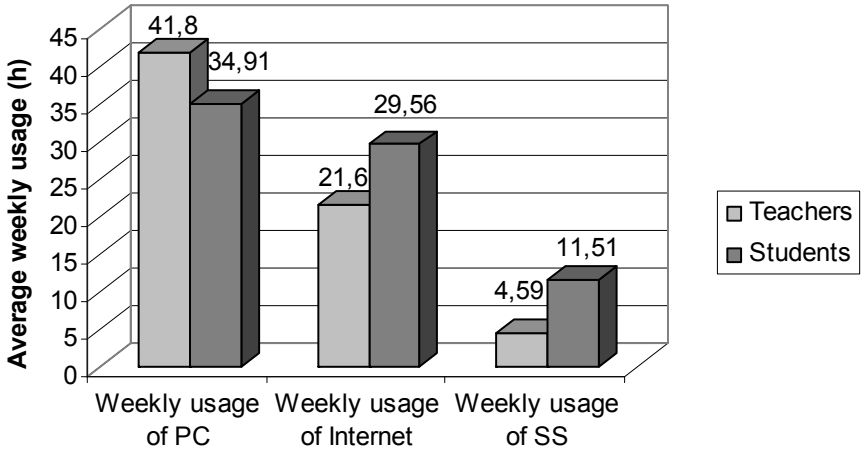
High percentage of teachers and even higher percentage of students use the social software as a new comunication tool for reconecting with old contacts, making new contacts and maintaining the existing contacts. Students use all three types of communication more than their teachers. For reconecting with old friends this tool is used 29.25% more by students, for maintaining the existing contacts 31.57% more and for acquiring new contacts 21.21% more.

Table 1

| Reason for using social softwaer | Teachers | Students |
|---|---|---|
| Data retrieval and information extraction | 52% | 70.15% |
| Education | 36% | 50.37% |
| Selfpromotion | 4% | 19.40% |
| Reconecting with old contacts | 20% | 49.25% |
| Maintaining contacts | 30% | 61.57% |
| Conectiong with new contacts | 12% | 33.21% |
| Organization and planning of social events | 12% | 34.33% |
| Other | 4% | 5.22% |

The next reason for using social software is selfpromotion or organisation and promotion of social events. This is also more popular way of promotion among students than it is among teachers. Around 5% of participants in both group use the social software for some other purposes as well.
Figure 1 shows the distribution of responses on the weekly usage of social software compared to the weekly usage of computers and the Internet in general. Students spend an average of 34.91 hours per week behind the computer, 29.56 hours use the Internet and 11.51 hours some social software application. Teachers on the other hand spend an average of 41.80 hours per week behind the computer, 21,60 hours use the Internet and 4,59 hours use social software applications. Although students spend more time using the Internet and social software than their teachers, the average number of hours both groups spend behind a computer is close to the number of working hours in a week.

**Figure 1**




**Social software as a teaching tool**
96% of teachers answered the questions on possible usage of social software in teaching process. This group includes those teachers who are familiar and already use social software, and those who are familiar with it but do not use it.
18% of teachers do not consider social software to be of any use for educational purposes. Only 14% of teachers already use some social software in teaching process, mainly Wikipedia and YouTube (10%), followed by Delicious and blogs (2%), Facebook and Flickr (1%). Although 57% of teachers consider it useful, they still do not use it for educational purposes. However, they believe that online encyclopaedia Wikipedia (50%), video sharing site YouTube (31%) and blogs (16%) could be useful tools in teaching process. These types of social software are followed by Facebook (8%), Delicious (7%), MySpace (6%), Twitter and SlideShare (4%), Flickr and LinkedIn (3%) and StumbleUpon

(2%). Remaining teachers did not feel comfortable at the time to answer on the questions concerning the educational usage of social software due to the lack of knowledge about these services.
Although students do not take part in creation of teaching process, we were interested to see their opinions on this subject from the user's point of view. It was interested to notice that their answers were similar to those of their teachers. Thus, 64.02% consider possible educational usage of social software as a potentially good idea. Software that students would like to use in education are Wikipedia (54.48%), YouTube (44.40%) and blogs (22.76%) followed by Facebook (18.66%), SlideShare (7.84%), Delicious (6.34%), Flickr, MySpace and LinkedIn (4.85%), Twitter (4.48%) and StumbleUpon (3.36%).
Surprisingly, 24.24% of students think that using social software for educational purposes is not a good idea, while 11.74% of students were unable to give an answer to this question due to the lack of knowledge about these services.

**Disadvantages of social software – students**
Students give six main disadvantages to using social software as a teaching tool (Table 2). The top disadvantage is further computerization of teaching process which would significantly reduce the quality of the same. In students' opinions teaching process should be as simple as possible and based on traditional teaching models without implementation of technological innovations. This way they believe that the gap between these who are computer literate and these who are not would be decreased.[5]
Some students believe that the main and only purpose of social software is entertainment and not education and as such can not be used for educational purposes. Others believe that social software would bring on further social isolation. They feel that implementing social software in teaching process would radically shift both students and teachers "from the real world to the virtual world" that would further reduce the already visibly reduced interpersonal communications between people in general.
There are also students that in social software usage recognize a problem of authorship rights and unreliable content. According to this group, problems of social software are unclear boundaries between public and private, relevant and irrelevant, visible and invisible audience, lack of temporal, spatial and social boundaries, etc. Their view can in a way be linked to the privacy issues that are seen by the next group of students as the main disadvantage to using social software. Since social networks do not allow individuals the full control over their personal information, it is only natural to ask oneself if the right of indi-

[5] Although, survey results for the computer literacy of working population in Croatia show that only 3% (n=631) of respondents age 17 to 24 do not use computer. (Algebra, 2009)

vidual to protect his/hers own privacy is now dead. Or are we just embracing a more transparent society in which all aspects of privacy will no longer exist?[6]
And the last disadvantage of social software educational usage is the insufficient education of teachers about these new technologies.

**Advantages of social software – students**

We can divide the advantages of using social software as a teaching tool, as seen by students, into three main groups of answers.
The first group is comprised of students who point out that the further computerization of teaching process will significantly improve its quality. Of course, this kind of teaching process demands good planning and understanding of possible usages of social software. It is also necessary to determine which type of social software is most appropriate for certain type of educational use. YouTube, Wikipedia and Delicious are examples of social software that students would like to use as a part of teaching process.
The second group of students indicates that information retrieval and sharing and a possibility for further discussion on given topic can only be seen as an advantage of social software in education. This would enable students to, within limited space of certain type of social software, publish student's papers, share additional educational resources that teacher has confirmed as relevant and continue discussion on given topic if such a need would present itself.
The last group of answers is given by students who believe that social software can provide them with more systematic monitoring of previously used materials but also as a motivation for learning via new networked media that puts them beyond the traditional academic methods of acquiring knowledge.

**Table 2 Advantages and disadvantages of social software (students)**

| Disadvantages of social software (students) | Number of respondents |
|---|---|
| Further computerization significantly reduces the quality of the teaching process | 26 |
| Primary purpose of social software is entertainment and not education | 19 |
| Social isolation | 16 |
| Problem of authorship rights and unreliable content | 8 |
| Privacy problems | 5 |
| The insufficient education of teachers about new technologies | 5 |
| **Advantages of social software (students)** | |
| Further computerization significantly improves the quality of the teaching process | 41 |
| Information sharing and retrieval, discussion on a given topic as a help in learning process | 19 |
| A more systematic monitoring of previously used materials | 7 |

[6] Visibility of personal information in certain types of social software is only partly a decision of an individual that decides which personal information will be visible to other members.

**Disadvantages of social software – teachers**

Disadvantages of using social software in education as reported by the teacher group (Table 3) are almost identical to those reported by the student group. These are the lack of theoretical background about certain types of social software, an unreliable content, data security and authorship rights problems. Although the possibility of useful and relevant materials is not excluded, teachers strongly believe that information created in certain types of social software (especially Wikipedia) is inaccurate, simplified, non-systematicaly organized and retrieved from unreliable sources.

As a disadvantage this group also notes an existence of e-learning platforms (such as Moodle) which are not too open and too wide as it is the case with the social software and thus present much safer environment to work within. Thus, they have no need for other, in their opinion, less safe and data questionable environments.

Some teachers believe that further computerization would reduce the quality of teaching process and stress out the importance of avoiding the misuse of new technology products.

**Advantages of social software – teachers**

Contrary to the previous group of teachers, this group indicates that further computerization would only increase the quality of teaching process (we had the same contradictory opinions in the student group as well). In their opinion some forms of social software can contribute to the dynamics of a teaching process if closely related to the course contents. They see the necessity in exploiting the new information-communication technologies in order to improve the traditional teaching and learning processes.

Some respondents have recognized social software as possible tools for information retrieval and sharing, possible discussions on given topics among teachers, among students and also among teacher-student groups.

**Table 3 Advantages and disadvantages of social software (teachers)**

| Disadvantages of social software (teachers) | Number of respondents |
|---|---|
| Unreliable content, data security and authorship rights problems | 10 |
| Teachers lack of theoretical background about certain types of social software | 10 |
| Existence of closed e-learning platforms | 8 |
| Further computerization would only reduce the quality of teaching process | 6 |
| **Advantages of social software (teachers)** | |
| Further computerization would only increase the quality of teaching process | 16 |
| Information retrieval and sharing, possibility for further discussion on given topics | 5 |

## Conclusion

The goal of our research was to learn about Croatian University teachers’ and students’ attitudes towards using social software as a teaching tool. Although our respondents in both groups (teachers and students) were largely younger people, we were surprised to find out that they still do not use social software in the teaching/learning process as much as it would be expected in the present time. Our data suggests that this is mainly due to teachers’ lack of knowledge about social software possibilities as a teaching tool.

Social software allows the student to be in the center of the dynamic learning process which is something that every educator should aspire to. At the moment, it is maybe the best to see it as an important enhancement to standard teaching process and to learning menagament systems such as Moodle, and not as their substitute. The potentials of social software in education and its benefits to both students and teachers are identical to those of the semantic web as explained by (Koper, 2004). Just by changing our view towards social software, or maybe better said, educational social software, we open up new approaches towards both teaching and learning as it can be seen in Dalsgaard (Dalsgaard, 2006).

To see only danger in reading web content and thus forbid using web as a source of information would be same as to forbid reading books since some of them may carry content that is inapropriate or even false. What we actualy need and want to teach our students is how to find information and needed knowledge on the web just as we used to teach them how to find them in a library or a bookstore.

The times we live in are full of fast changes, especially changes in knowledge. What we can only hope for is that these changes are reflection of our own growth and improvement and that the tools we use will only be used to prosper.

## References

Boyd, Danah Michele. Taken out of context. // *PhD Thesis*. http://www.danah.org/papers/TakenOutOfContext.pdf (24.06.2009), 2008.

Dalsgaard, Christian. Social software: E-learning beyond learning management systems. http://www.eurodl.org/index.php?p=archives&year=2006&halfyear=2&article=228 (20.6.09), 2006.

Koper, Rob. Use of the Semantic Web to Solve Some Basic Problems in Education: Increase Flexible, Distributed Lifelong Learning, Decrease Teachers’ Workload. // *Journal of Interactive Media in Education, Special Issue on the Educational Semantic Web*, 2004 (6). http://www-jime.open.ac.uk/2004/6/koper-2004-6.pdf (2.6.09), 2004.

Bejune, Matthew; Ronan, Jana. Social Software in Libraries. // *SPEC Kits 304* / George, Lee Anne (ed). Washington: Association of Research Libraries, http://www.arl.org/bm~doc/spec304web.pdf (24.6.09), 2008.

Učilište Algebra. Survey about usage of computers and education of working population. http://forum.algebra.hr/Istrazivanje_inf.pismenosti.pdf (24.6.09), 2009.

Wikipedia Contributors. Wiki. Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Wiki (24.6.09), Aug 16, 2009.

O’Reilly, Tim. What is Web2.0 : Design Patterns and Business Models for the Next Generation of Software. O’Reilly Media, Inc. http://oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (24.6.09), 2005.

Professional paper

# The Integration of Library Users into the European Cultural and Scientific Space through Searching Electronic Information Resources

Vesna D. Župan
The "Svetozar Marković" University Library
Bulevar kralja Aleksandra 71, 11000 Belgrade, Republic of Serbia
E-mail: buzupan@rcub.bg.ac.rs

**Summary**

*A period of transition doesn't imply only economic changes in a country or an organization but also a scientific, technological and cultural reversal. Its' success will depend on the way disposable resources are being used. The Learned Society is unimaginable without indispensable information resources for users of different educational vocations. Technical skills are being developed in accordance with new software on the world market. Information literacy became a precondition for attaining intellectual freedom. Contemporary tendencies enforce the adjustment of libraries to technological change as well as to the needs of users who are conquering an electronic city. Internal and external, real and potential library users tend to penetrate the European cultural and scientific space as fast as they can. They also tend to make positive influence on its development. Librarians have a key role in preparing users for further work. Their creativity in almost all activities will depend very often on appropriate search and the use of the electronic information resources.*

**Key words:** e-information resources, library, development, Europe, integration.

## Introduction

Integrative process at the Old Continent goes on intensively thanks to the tendencies of his inhabitants to share economic and political reality and also participate in technical progress which contributes positively to their cohesion. Europe includes almost 50 states. There are 27 members of the European Union among them (1).

For famous poet Homer, Europe was not a geographical designation but a mythological queen of Crete (2). It is her name that connects symbolically the Old Continent and Ancient Greece which was being very long the core of the West European cultural development. As there are numerous nations in Europe, their cultural influence on one another is obvious in almost all fields of human life and activity.

According to a study of the Faculty of Economics, Finance and Administration at Singidunum University in Belgrade (3), it is from 1992-2000 that Serbia had lost potential social product of over 100 billion euros. Every employed citizen was losing 300 euros monthly e.g. almost 29.000 euros as a total amount. According to the same resource pensioners were loosing an average amount of 150 euros monthly in comparison with 1992, so it is more than 14.000 euros. It was almost impossible to get medicaments and patients used to go very often to hospitals with their own towels, anesthetic ampoules and other indispensable material. Several thousand young people tried to find happier destiny abroad.

## The role of library users in the European integration process

If Serbia would keep postponing the entrance to the European Union it would make this country less interesting and attractive for foreign investors. New investments offer more chances for the increase of employment. Serbia paid very high price for the fact that it didn’t become member of the EU yet although the library users approach it quickly, everybody in their own professional field and everyone in his own way. They access cultural and scientific sites but also promote a dialogue with their colleagues and friends using the advantages of the Information Society.

The end users belong to diverse generations and professions. Furthermore, they perform different activities. Except this, their interests are not identical. Real and potential users make great influence on the development of communication among the countries of the European Union by contacting individuals who are also willing for professional cooperation no matter whether they are located in a library or out of it. The interactivity of sites helps the users a great deal to remove obstacles for penetrating cultural and scientific space of the contemporary world.

New technologies make the development of scientific thought more intensified and the preservation of civilizational achievements more successful. Therefore the users get included into these streams with the assistance of the librarians. They participate together in the Information Society relying on their professional knowledge, cultural interests and experience.

An academic library user is not just a citizen interested in the content of his inbox. He may be a scientist, an expert, a publicist etc. Everything that he performs will be done much better if he keeps following contemporary activity of his colleagues worldwide and particularly from Europe. Librarians are present in order to instruct, give a piece of advice, focus users’ attention on collections, programs and services.

## Electronic information resources

Big contemporary libraries do not ask whether e-information resources should be used or not. The question is just which e-information resources should be used, under which conditions and who should be the one to address and instruct

the users having in mind their needs and interests as well as the results of technical progress. As far as top management is concerned it is very important to access the librarians in the adequate way in order to prepare them for further work. Top management which has an insight into the activity of the librarians, their obligations and responsibilities, can make influence on their perfectioning more successfully. New technologies imply investments into staff, equipment and space.

Many countries of the European Union are the members of OECD e.g. Organization for Economic Cooperation and Development. This paper concentrates on the countries of Europe. They cooperate very intensively with developed and other countries in the field of culture and science either directly or in an electronic way.

The development of culture and science is one of priorities for all those governments which make efforts in order to establish peace and progress in the world. The automation of libraries makes globalizing more efficient. That is particularly important for countries in transition. They unavoidably follow the tendencies in OECD countries which are more developed and also very progressive in the implementation of the information technologies.

According to actual resources "OECD brings together the governments of all countries committed to democracy and the market economy from around the world to:

- Support sustainable economic growth
- Boost employment
- Raise living standards
- Maintain financial stability
- Assist other countries' economic development and
- Contribute to growth in world trade" (4).

A precondition for the creative economy is a librarianship which accepts the development of information literacy as an imperative. Creative economy is based on useful information. The users search e-resources because they need more actual data and professional papers also. Search sessions may be carried out if the user has certain level of technical skills. If this is not the case, the librarian is present as a person the user may rely on. So, technical progress refers to both: librarians and users as well.

"Information transfer has become a significant economic and social activity, which is critical to the society's well being. As information processing has become the focal point of economic activity such issues as intellectual property rights, privacy, accuracy, access continue to challenge our global society." (5)

Human resource is the most important for performing library and information activity in an adequate way, then for its' improvement, the development of marketing as well as library image in relationship towards users in general. (6)

According to statistical research carried out in Australia by Michael Middleton, following skills are important in information sector:

- how to use traditional and automated reference resources in locating required information,
- adequate evaluation of users' information needs,
- efficient search of diverse databases in order to locate an indispensable information,
- defining a strategy of searching in accordance with users' requirements,
- successful oral and written communication with users,
- helping users to evaluate information resources and information,
- searching Internet in a useful way,
- overtaking initiative in order to develop permanently technical skills and knowledge,
- efficient communication through presentations,
- the implementation of knowledge in the field of bibliography and informatics in order to discover the origins of information,
- everyday evaluation of personal professional work,
- successful addressing users into library materials in different languages,
- presenting materials orally to small groups of users,
- overtaking the initiative to solve the problems which appear in group work,
- solving problem in connection with the access to information,
- making education possible in small groups of users etc. (7)

Every library can point out an access to its' own work. One of those which is well known and doesn't neglect any segment in working process is TQM (Total Quality Management) access. It contributes a great deal to the efficiency of work in organizations which implement contemporary equipment and employ staff ready to be trained.

In a library which implements Total Quality Management (TQM) there are no departments which can be privileged at all. They should all be included into the process of development. In a library which has accepted "TQM", personnel is very concentrated on services for the users, there is an indispensability to act in accordance with their reactions as well as to learn permanently so that higher quality of services may be achieved. (8)

The offer of the e-services for the users' needs will depend on the acquisition policy in the library. That policy will be successful if it relies on the expectations of the users which are to be followed permanently through surveys and focus interviews. Marketing is to be intensified in technically progressive libraries having in mind their space conditions, financial and intellectual resources. Of course, normative acts are unavoidable, so their authors should be able to follow technical progress in a Learned Society.

Consortium for Coordinated Acquisitions in Serbia gives the opportunity to its' libraries-members to use indispensable electronic services for the work with users. Thanks to the aggregates of databases the readers may search more than 35.000 e-journals through articles, to receive their full-texts and use them for further work. Except this, the users may search Internet in those libraries which have such a centre. They may also use the advantages of an electronic cooperative catalogue either from home or from their working places.

## Work with the users of library services

Marketing orientation in the libraries of all types imposes focusing on the users, their expectations, needs and requirements. Only loyalty to librarianship can lead to an efficient work with the users who have different observations and interests. Librarians may find cultural inspiration for their work in actual events, particularly among jubilees. Such examples may stimulate users to carry out non linear search of databases, in fact, search of e-information resources in general. The users sharpen their technical skills in this way. This is the first step to prepare them for future electronic education no matter in which field. Although the word of mouth is usually the best, the electronic education is sometimes unavoidable.

The users make positive influence on the library work searching disposable electronic information resources and implementing new technologies. Some questions and wishes that previously were being rarely pronounced became usual and welcome such as: "Please, I would like to download this sound file and listen to it if possible", or, "I would like to send this animated file by e-mail to my friend", or, "can you help me, please, to create this computer presentation? I need it for an exposé at the second year of my studies", and so on. Library staff is being adjusted to such questions and desires. It is library staff which leads internal and external users through the electronic environment with a clear intention – to help them enter and stay in the Open Society. However, feedback with the users shouldn't be underestimated. In a way, users develop clear orientation among librarians towards contemporary tendencies in informatics.

Instructing the users is usually being carried out directly and in some cases online. In a university library, work is easier if there are specialists in each scientific field. Specialists with long professional experience are usually able to help the users from all scientific fields. A precondition for an electronic course is that the users possess PC with indispensable software but also certain level of technical knowledge. An open dialogue between a librarian and a user makes a topic more clear with the aim to find the most appropriate materials for a potential paper.

In Belgrade, for example, there are several universities. State university has the longest tradition but there are also private universities which keep following international standards and tendencies in their professional and scientific prac-

tice. However, they do not adjust themselves equally to the need of establishing a contemporary library. As far as this is concerned some of them are just in a preliminary phase. Their students address themselves to the "Svetozar Marković" University Library which is a budget institution with a long tradition. It has contemporary equipment. Except this, it is on its' way to become a library 2.0. However, this will require more work with its' staff and the users. Those library users whose technical knowledge is on high level are able to promote web 2.0 space by creating their own interactive sites, up-dating certain web pages by useful contents, as well as by realizing some other activities in the electronic environment.

Classic or traditional catalogues are being replaced by the electronic ones. Library 2.0 would be very difficult to imagine without blogs or wikies which make the work on web more successful and efficient. Text, sound and picture synchronized cause remarkable progress in comparison with previous modest results of searching classical library collections in Serbia. Users focus their attention more and more to multimedia resources of information and materials wherever they are.

Libraries in Serbia do not make much effort to become libraries 2.0. They do not implement often new technologies in promotional activities as there are limiting factors in their everyday work. Financial problems, organization, the training of staff are some of these limiting factors. In such a situation, libraries of Serbia cannot reach the level 2.0 easily.

The Municipal Library of Belgrade makes efforts in order to create useful contents and it implements Web 2.0 technology. These activities are remarkable in Serbia but in comparison with the libraries of developed countries they are still very modest. (9)

## E-library and e-city

There are many examples of good practice among electronic libraries. Cooperative electronic catalogue itself is insufficient to illustrate the entire cultural and scientific treasure of a nation. Museums and archives "hide" also very important materials. As urbanization goes on, cities tend to present themselves efficiently in an electronic way also.

Nowadays, cities use web space in order to attract potential investors and tourists by their informative materials. The cities present their advantages to the world as well as the truth about their own history, everyday life and development. In addition, the cities address their inhabitants and Internet users into diverse valuable contents giving them the opportunity to contact municipal bodies because of all problems concerning the quality of life and work in the city. Librarians know very well and accept that "libraries are @ the heart of the Information Society". They should be at the heart of e-cities also. However, the cohesion between libraries and e-cities is not so strong as it should be. There are many examples of successful city sites:

http://www.zagreb-convention.hr
http://www.ljubljana.si/en/
http://www.beograd.rs
http://www.visitlondon.com/accomodation/
http://www.london.gov.uk
http://www.florence.ala.it
http://www.barcelona.com

However, searching them citizens may notice that the creators of sites rely mostly on museums trying to make users familiar with cultural life in the city. Therefore, libraries are obviously neglected. So, top managers in the libraries as well as leaders in library associations should make efforts in order to focus the attention of municipal authorities to these organizations, their collections, services and programs. Libraries should become an unavoidable component in the promotional activities of an e-city.

## Conclusion

The citizen becomes an e-citizen and the library user becomes an e-user. Librarians do not try to give tasks to the users. They should be willing to help library members as well as potential users. It is technical progress itself which imposes the rhythm of social events in the electronic environment. The users can penetrate the European cultural and scientific space:

- by searching the e-catalogues of leading libraries which contain very rich paper and on-line collections,
- by searching and using interactive sites for their personal and professional aims,
- by using the advantages of famous projects which offer free access to certain library materials,
- by searching the aggregates of databases with full texts of articles in electronic form but having in mind the type of the library they are members of as well as their personal and professional needs.

Complete texts in free access are a way to keep cultural and scientific treasure in nations' and worlds' memory. Library programs, collections and services are being described across sites in Serbia and elsewhere. Serbia tends to become the member of the European Union. Therefore it keeps adjusting itself to the European and world standards in economy as well as in non-profit organizations. Such efforts are being made in order to improve the quality of citizens' life. On the other side, libraries tend to improve the quality of their collections, instruct the users in the best possible way and to offer them good programs in order to meet their expectations.

## Resources

1) Europa.eu (July, 2009.)
2) En.wikipedia.org/wiki/Europe (July, 2009.)
3) http://www.fefa.edu.rs (July, 2009.)
4) http://www.oecd.org (July, 2009.)
5) Achleitner K., Herbert. Information transfer and ethics. In: Intelektualna sloboda i savremene biblioteke : zbornik radova sa međunarodnog naučnog skupa održanog u Beogradu od 25. do 27. septembra 2003. godine, str. 107.
6) Župan, Vesna. Marketing u bibliotekama. Beograd, Svet knjige, 2001., str. 135.
7) Middleton, Michael. Skills expectations of library graduates, New Library World, 104, 2003, 1184/1185, p. 47.
8) Graham, John. Le management de la qualité totale à la Bibliothèque d’Etat de Nouvelle-Galle du Sud, Bulletin des Bibliothèques de France, Paris, 1998, XLIII, 1, p. 57.
9) Sofronijević, Adam. “Kurs C/08-Web 2.0 i Biblioteka 2.0”. Beograd, Zajednica biblioteka univerziteta u Srbiji, str. 7. (3 August, 2009.) http://www.unilib.bg.ac.yu/zajednica01/obavestenja/kursevi-2008/kurs%20C_Web20.doc

Original scientific paper

# New Access Structures to Scientific Information: The Case of Science 2.0

Sonja Špiranec
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sspiran@ffzg.hr

Ana Babić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
anbabic@ffzg.hr

Ana Lešković, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
aleskovi@ffzg.hr

**Summary**

*Since the early 1980s, the scholarly community has been witnessing a considerable increase in the use of information and communication technologies (ICT). Specifically the use the Web has led to qualitative changes in the research community. With the advent of the Web 2.0 a new level of possible functionalities for science has been reached, leading to the concept of Science 2.0. Will the new research technology 2.0 change the way research is done and what aspects are already visible in current structures of scientific communication are questions this paper tries to answer.*

*Several clusters of expectation emerge from the prospect of applying the principles of Web 2.0 to scientific communication, like the opening of science communities towards public and the acceleration of dissemination of scientific research through new communication and collaboration tools. In the first part the authors will comment how the Web 2.0 challenges some traditional and known structures of scientific communication and explore possibilities of applying Web 2.0 principles (collaboration, collective validation, access and generation of information) to scientific work. In the second part the authors will present results gathered through analysis of Web 2.0 services that have been integrated*

*into academic databases and vice versa, the analysis of scientific information spaces that have been created within the Web 2.0.*

**Key words**: Scientific communication, Web 2.0, Science 2.0, academic databases

## Introduction

ICT and the Web have forced old-line institutions to adopt whole new ways of thinking, working and doing business. In the last decade a new version of the Web has emerged that has changed the premises of new developed models that comply with the Web 1.0 world, by transforming the very nature of knowledge, information and learning and by blurring the boundaries between using and producing, by connecting people and by expanding through user contributions like comments, texts, AV content, pictures, tags etc. This new cluster of practices, habits and business imperatives has influenced a wide range of domains and sectors and initiated a flood of buzzwords like *Business 2.0, Marketing 2.0, Libraries 2.0, Education 2.0* etc. Since scientific work has very much to do with information and knowledge, and since these are in the very heart of shifts initiated by the Web 2.0, it is time to analyze whether the restructuring of information spaces and creation of new communication patterns and information cultures do have an impact on scientific processes. What innovative configurations of scientific discourses are possible within this new and restructured Web 2.0 world and is it possible to talk about the concept of Science 2.0?

## Web 2.0: new forms of online practices

The phenomenon known as Web 2.0 can be characterized by its technological aspects, but they are just a support for its conceptual nature which is what distinguishes it the most from the "old Web" (Banek Zorica et al, 2007, 93). Web 2.0 uses many new approaches for dealing with information including wikis, weblogs, aggregators, RSS and mashups. These often require active participation of users. Words that determine the Web 2.0 environment are participation, collaboration, sharing and communication. Applications like Instant Messaging and Peer-to Peer networks can be perceived as precursors of the Web 2.0 hype. But with the Web 2.0 the focus has shifted from communication between two users towards communication within groups of users (many-to-many relationships). New are also multiple communication paths, where the user contributions are not any more a collection of isolated monologues, but represent a matrix of dialogues (Maness 2006). Web 2.0 and social software tools are directed towards online cooperation and user communication; they facilitate communication in virtual networks and enable collaborative work on specific undertakings. With the application of widespread Web 2.0 services like blogs, wikis, social bookmarking services, media sharing or academic paper services – to name just a few – traditional educational and scientific institutions and information agencies transform themselves from a place of passive information consumption

to a dynamic, participative, collaborative and creative knowledge production space. Many of the very words that describe Web 2.0 phenomenon's – communication, critiquing, suggesting, sharing ideas – are in the heart of science and bear the capacity to transform it.
Another facet of the Web 2.0 paradigm with the potential to transform classical perspectives in science is the notion of collective intelligence. The application of collective intelligence in science results in the breakdown of traditional assumptions about scientific expertise and the transformation of rigid scientific processes by more open-ended processes of communication in cyberspace. P. Walsh suggests that the expert paradigm (that dominates traditional "Science 1.0"), uses rules about how you access and process information, rules that are established through traditional disciplines. By contrast, the strength and weakness of a collective intelligence (that determines Web 2.0) is that it is disorderly, undisciplined and unruly. This certainly affects the way science is done (Jenkins, 2006, 53). The above discussion shows that central conceptions of the Web 2.0 coincidence with ideas that determine science, yet the question remains how can science harness this new possibility. Several clusters of expectation emerge from the prospect of applying the principles of Web 2.0 to scientific communication; according to Weller et al., the relations between Web 2.0 and scientific work require a differentiation into several dimensions:

- new ways of public relations for scientific and research activities (blogs, podcasts etc)
- collective knowledge generation and management
- new structures of scientific communication (dissemination und discussion of scientific contents, finding and access of scientific information (2007).

## The new anatomy of science: Science 2.0

Science has always been looking for solutions which will minimize organizational and technical routine tasks occurring during research processes and simplify scientific workflows. Recent endeavors undertaken in this direction are known by the generic term e-science.[1] The term e-science, as it was designed by John Taylor, refers to "global collaboration in key areas of science, and the next generation of infrastructure that will enable it". (Taylor, 2001) This original definition implies that e-science consists not only of tools and technologies, but depends on pooling resources and connecting ideas, people and data. It has to do with information management as much as with computing.
Therefore, the concept *Science 2.0* is complementary with the idea of e-science and may be defined as a mean for realizing the principles of e-science. Although efforts in this direction are still too scattered to be called a movement, an

[1] The terminology is not globally uniform, while in Europe the use of the term e-science is quite common, the equivalent term in the USA is "Cyberscience".

ever growing number of researchers are beginning to harness wikis, blogs and other Web 2.0 technologies as a potentially transformative way of doing science.

**Potentials**

*Unifying the roles of scientific producers and consumers*

The common perception of the user – in our case a scientist – is mostly defined by passive and one-directional information consumption. In this perception, the scientist can participate in his or her science communication system by following two options: (a) using informal channels (personal communication) or (b) preparing a critical article and publishing it via a formal channel, i.e. a scientific journal. (a) is effective in deed, but limited, and (b) is a time-consuming process. Through diverse Web 2.0 services it is possible to formalize informal channels and transform the scientist as a reader into the scientist as a prosumer, a person who simultaneously *pro*duces and con*sumes*. (Stock, 2007, 97)

*Acceleration and multiplying of scientific communication processes*

An often critiqued aspect of scientific work relates to its traditional communication tool, the journal. Journals are either about one-way communication, indirect communication (references, letters to the author or editor) or one to one communication (author and reader). This static structure of journals forces the inherently dynamic scientific discourse to decelerate. Assimilating Web 2.0 tools with scientific work helps to overcome the shortcomings of linear and static processes provided by journals and leads to the creation of dynamic, circular, modular, multidirectional and decentralized contexts.

*New communication channels: blogs, wikis etc.*

Communication is the driver of scientific processes. Only published (i.e. communicated) research results "exist" for the scientific community. Therefore, blogs as an online communication infrastructure that provides bi-directional communication and real dialogs are the perfect mean to speed up and revitalize existing static structures provided by journals. Wikis on the other hand are often used to create collaborative websites or to power community websites. Within a research group wikis can function as information- document- or project management systems. They are also suitable for discussion lists. Constructing a wiki within a community allows real and high-quality teamwork and constant updating. It creates conversations between researchers, lets them discuss the data and connect it with other data that might be relevant. Blogs and wikis permit users to make information available in ways that create a conversation.

*Access points to knowledge, collaborative organization of resources*

An achievement of the Web 2.0 is that users do not only actively participate in the provision of content but also are employed to organize and index (or tag) it. The set of terms a group of users tagged content with are not a predetermined

set of classification terms but rely on user experience. No longer do the experts and professional catalogers have the monopoly in this domain. As users continue to add tags, grassroots organizational scheme begins to emerge which is usually referred to as folksonomy. Such systems currently predominate in the private area for users to manage photos, videos, link collections etc. For scientific use the collaborative managing of web links and bibliographic data could and already does represent a special niche of interest. In most cases, users define the tags primarily for their own use and for organizing and accessing their own collection, but at same time there is the possibility to exploit approaches and alternative perspectives of other users.

With the aid of folksonomies the end user in formal science communication, the reader, is able to contribute to the indexing of scientific documents. This is a significant and up to this moment insufficiently explored area, at least in the science domain[2]. Prior to the Web 2.0 and folksonomies, scientific databases, the common access points to scientific information, lacked effective mechanisms to gain insight into user perspectives. This inhibited the process of retrieving information since the ones who usually created the databases were not the ones who actually used them. The information users were not able to easily question the information- gatherers or organizers. On the other hand, user-created tags are searchable for everyone beside the interpreter-created controlled terms and the author-created text words and references.

The described enrichment of scientific databases through folksonomies shows a lot of benefits: tagging represents authentically the use of language inside scientific communities, it allows for multiple interpretations from different disciplines or different schools, it can help to recognize neologisms and new scientific results fast. Folksonomy is by no means opposing controlled vocabularies; it should be clear that the development will profit from tagging, because tagging provides a rich source of authentic term material (Stock, 2007, 99).

*Alternative forms of evaluating and pre- reviewing scientific works*

Besides folksonomies the Web 2.0 offers further opportunities for accessing content, like social navigation or collaborative filtering. These are based on directly or indirectly derived user judgments, reviews or comments. The most famous expression of social navigation mechanisms are probably recommendation systems. With their help users can determine most popular or best rated articles, or within a bookmarking system identify those websites that users with similar profile have bookmarked and tagged.

Those systems could also be very valuable in a reconceptualized scientific peer review or a quality-control system. Classical peer review has many shortcomings (quality, delay, bias). The described recommendation systems provide an

[2] In other areas a critical mass of solutions and applications already do exist, e.g. Library 2.0, Catalog 2.0

useful alternative tool for evaluating scientific work, as well as some other popularity markers (the equivalent of citations), like download counts, number of tags etc. The more scientists tag a document, the more relevance does this article seem to have for this people. This would lead to a new scientific "currency" besides citations, voting systems etc.

**Risks**

Beside the analyzed potentials of Science 2.0 there are as many risks and controversial points. Diverse issues that arise within Web 2.0 discussion refer to the domain of science as well. The collaborative model of knowledge production, mash-up practice and anonymity creates information spaces where authenticity, trustworthiness, authority and reliability have to be continually questioned. Misinformation emerges, is worked over, refined or dismissed before a new consensus emerges. A particular risk for scientists refers to the global *copy-past culture*: scientists who put preliminary findings online risk having others copy or exploit the work to gain credit or even patents; particularly in hypercompetitive fields where patents, promotion and tenure can hinge on being the first to publish a new discovery (Waldrop, 2008).

Another problem of current social networks and other community portals are that many of the systems depend on active participation (starting with the registration) of community members. Trustworthiness and diligence in dealing with personal data are the minimum requirements to ensure this participation. To ensure the exchange of knowledge within scientific communities a basis of mutual trust must be created in order to prevent misuse or abuse of private data.

## Science 2.0: how far are we?

A small but growing number of researchers (and not just the younger ones) have begun to carry out their work via the wide-open tools of Web 2.0. And although their efforts are still too scattered to be called a movement–yet–their experiences to date suggest that this kind of web-based "Science 2.0" is not only more collegial than traditional science but considerably more productive (Waldrop, 2008).

Several clusters of expectation emerge from the prospect of applying the principles of Web 2.0 to scientific communication, like the opening of science communities towards public and the acceleration of dissemination of scientific research through new communication and collaboration tools. In order to determine the actual degree of integration of Web 2.0 services into formal and informal scientific communication channels, an ad-hoc analysis of potential forms of interlinking 2.0 services with science has been conducted in the first half of year 2009. The review was made in two directions:

1. Have Web 2.0 services have already been integrated into academic databases?

2. To what degree have scientific information spaces have been created within the world of Web 2.0?

**Integration of Web 2.0 tools into scientific databases**

Currently professional science databases of the "old" information industry (e.g., CAS, INSPEC, MEDLINE, BIOSIS, Web of Science) are at the beginning of exploring potential interactions between Web 2.0 tools and the services they provide. Existing features of this interaction will be illustrated through several databases and hosts: EBSCOhost, Project MUSE, Citeseer and EiVillage.

*EBSCOhost*

EBSCOhost provides the functionality of bookmarking. The system enables users to collect and save references on some social bookmarking site such as del.icio.us, Technorati, dig etc. A further and important step EBSCO host has undertaken in the direction of Web 2.0 is the new interface named EBSCOhost2.0. This interface incorporates Web 2.0 user controls such as a date slider, mouse-over previews, and modals making the interface more dynamic and complying with the 2.0 philosophy of rich-user experience.

*Project MUSE*

MUSE supports bookmarking/tagging for sites including CiteULike, Connotea, del.icio.us, Facebook. In that way it provides the community an insight into its resources. Syndication via RSS Feeds provides a mechanism for users to subscribe to important information regarding Project MUSE journals, journal issues, and announcements.

*Citeseer*

CiteSeer$^{X}$ offers personalization and Web 2.0 features such as personal collections, tagging for articles, error correction and document submission (user-created content). Users can monitor specific papers for metadata updates via email and create bibliographies by marking and downloading specific records. The system itself offers bookmarking through several services such as Connotea and Bibsonomy, so in that way it includes its resources into social web.

*EiVillage*

EiVillage (includes the databases Compendex and Inspec) is probably the first host that has started to work with folksonomies. EiVillage enables the user to assign tags to documents. He can choose to make those tags accessible to colleagues, peer groups and even to all the users in the Engineering Village community. Users may also decide to keep tags private for personal use. Engineering Village databases have traditionally relied upon records being classified by experts using structured indexes. Now, by adding record tagging the power to classify records and create content has been extended to users. Users can tag re-

cords based on how they define a record's relevance and importance. By choosing to expose those tags, Engineering Village users' community is provided with a powerful way to identify engineering content other users find meaningful.

**Science in Web 2.0 information spaces**

From this part of analysis individual blogs written by scientists where excluded since they are already common and exist in a great number. Attention was paid to more complex and sophisticated services as well as bookmarking services.

*2coolab*

2coolab[3] represents a platform for cooperation. It allows using bookmarks or tags and sharing Internet resources, from articles to video clips. It allows building networks and find, evaluate and initiate contact with new people The major characteristic is that 2coolab is a completely free, open and independent service. Anybody can open a user account and add or save bookmarks, or even create groups. Members of groups can evaluate these resources (by adding ratings and comments. Also, it is possible to use RSS feed to receive notification’s about members, groups and users.

*SciTopics*

SciTopics[4] is a free expert-generated knowledge-sharing service for the scientific community. It serves as an information and collaboration tool for researchers. It is designed as a wiki that invites the scientific, technical and medical community to participate by posting comments, feedback, questions and discussion items. In order to contribute, interested users need to register as a SciTopics member and identify themselves with name and affiliation. In addition, users may find contact information for authors who they have identified as being potential collaborators on future research projects. Quality of the contributions is safeguarded by non-anonymous postings.

*Bookmarking services*

Bookmarking services are one of the mostly used 2.0 applications in scientific work. As previously mentioned the most known bookmarking services in the scientific community are Connotea, Bibsonomy and CiteUlike[5]. It is interesting to note that all these services have been created by the scientific community for

[3] http://www.2collab.com/nonLoggedInHomePage

[4] http://www.scitopics.com/

[5] http://www.connotea.org/, http://www.citeulike.org/, http://www.bibsonomy.org/

the scientific community.[6] They are usually based on already existing popular services like del.icio.us. Since they where primarily created for scientific and research needs these services have special features for this particular audience.
Every of the analyzed service has its own distinctive features, like CiteUlike which combines Web 2.0 tools with traditional software for the bibliographic handling of information, or Bibsonomy which aims to integrate the features of bookmarking systems with team-oriented publication management. All of the services continuously expands existing functionalities and includes new ones, like new resources for the systems automatic recognition and processing (Connotea), the improvement of tagging tools, group and private bookmarks, or work on the convergence with the Semantic web by concentrating on machine understandable tagging and new methods for extracting semantics from folksonomies (Bibsonomy).

### Trends

Big databases are starting to incorporate in Web 2.0 trends and work primarily with services that focus on information discovering and handling: RSS, bookmarking, tagging, annotation and reviews. On the other hand, the web itself consist of the various number of different Web 2.0 services which are created specifically for (and by) the scientific community. Unlike the nonscientific version of these services which are based on anonymity, more and more attention is dedicated to the prevention of the aforementioned risks that are usually caused by anonymity. Therefore, such services are oriented towards developing mechanisms for the identification of scientist, user groups/scientific communities etc.

## Conclusion

The Web 2.0 with its tools and technologies has begun to alter science. The comparison of older or classical scientific practices with those carried out in the Web 2.0 environment shows a transformation and shift that improves many facets of scientific work and even allows science the return to values that where the supposed hallmarks of science, but bring about some ambiguous and controversial aspects.
The ad hoc analysis of scientific information spaces conducted here shows that today the technical and infrastructural preconditions that have the potential to raise the functionalities and improve the features of scientific processes are available. Anyway, as with many other domains, technology does not drive change as much as our cultural response to technology does. Therefore the ac-

[6] Connotea was created by the Nature Publishing Group, Bibsonomy by a team of experts and students at the Institute of Knowledge and Data Engineering in Kassel, Germany and CiteUlike at the University of Manchester.

ceptance and emergence of a Science 2.0 would require a big change in academic culture.

Despite the discussed pros and cons, Science 2.0 is beginning to proliferate; current applications and articulations of Science 2.0 allow scientists to built networks and communities, market themselves, improve multidirectional communication, publish or disseminate materials at the point of need, criticize and discuss, find and access better, organize better. But for Science 2.0 advocates, the real significance is the technologies' potential to move researchers away from a focus on priority and publication and overcome the shortcomings of the traditional review process. The conducted analysis of Web 2.0 services within scientific communication processes has shown that 2.0 tools have been integrated into academic databases and vice versa, that scientific information spaces have been created within Web 2.0 environments. Having in mind that prominent and respectable research organizations, commercial and non-commercial, have begun to effectively leverage Web 2.0 practices, this domain is certainly going to advance and prosper.

## References

Banek Zorica, M; Špiranec, S.; Zauder, K. Collaborative tagging: providing user created organizational structure for Web 2.0. // *INFuture 2007: Digital information and heritage* / Seljan, S; Stančić, H. (ed). Zagreb: Odsjek za informacijske znanosti, 2007, 193-203.

Jenkins, H. Convergence culture: where old and new media collide. New York, London: New York University Press: 2006.

Maness, J.M. Library 2.0 Theory: Web 2.0 and Its Implications for Libraries //,*Webology* 3(2006)2. http://www.webology.ir/2006/v3n2/a25.html (2009-05-15)

Stock, W.G. Folksonomies and science communication: a mash-up of professional science databases and Web 2.0 services. // *Information Services and use*. 27 (2007), 97-103.

Taylor, J. Defining e-science. http://www.nesc.ac.uk/nesc/define.html (2009-08-04).

Waldrop, M. Science 2.0: great new tool or great risk? // Scientific American. Jan, 2008. http://www.scientificamerican.com/article.cfm?id=science-2-point-0-great-new-tool-or-great-risk (2009-08-06)

Weller, K et al. Wissenschaft 2.0? Social Software im Einsatz fuer die Wissenschaft // *Information in Wissenschaft, Bildung und Wirtschaft. Proceedings der 29. Online-Tagung der DGI* / M. Ockenfeld (ed). Frankfurt am Main: DGI, 2007, 121-136.

Professional paper

# Usage of Print and Electronic Resources at the Faculty of Humanities and Social Sciences' Library, University of Zagreb – Analysis and Comparison Based on the Usage Statistics

Dorja Mučnjak
Library, Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10 000 Zagreb, Croatia
dmucnjak@ffzg.hr

**Summary**

*Many pieces of research have shown that different subject disciplines make different usage of professional literature. The results usually point to differences in the usage of sources of information between physical sciences, biomedical sciences and engineering on the one side, and social sciences and humanities on the other, stressing the fact that the former make more use of electronic resources than the latter. The results of this research should show us whether this is the case at the Faculty of Humanities and Social Sciences' Library.*

**Key words:** humanities, social sciences, literature use, print resources, electronic resources

## Introduction

Significant differences in the presentation of scientific discoveries are most clearly seen in the scholarly communication. The information in physical sciences, engineering, and biomedical sciences becomes outdated much faster than the information in social sciences and humanities. That is why discoveries and achievements in physical sciences, engineering and biomedical sciences are normally published in journals, while books are still very much present in social sciences and humanities (Nederhof,[1] Hiller[2]). A survey of users also revealed different behaviours in the very process of searching through referential elec-

[1] Nederhof, Anton J. Bibliometric monitoring of research performance in the Social Scinces and the Humanities : a review. Scientometrics, 66(2006), 1; p84

[2] Hiller Steve. How different are they? A comparison by academic area of library use, priorities, and information needs at the University of Washington. // Issues in Science & Technology Librarianship, 33(Winter 2002); 13p. http://www.istl.org/02-winter/article1.html (5 August 2009)

tronic resources.[3] Since these differences are quite prominent, it was highly probable that differences in the usage of professional literature also exist between social sciences and humanities. Research conducted in recent years indicates that there are also differences in the usage of literature (print versus electronic resources, books versus journals) between different fields of social sciences (Liu,[4] Kriebel et Lapham[5]…). However, a piece of research conducted this year has shown that, contrary to the existing information, there is no connection whatsoever between the scientific field and the treatment of electronic resources.[6] This discovery remains to be verified.

Several pieces of research have been conducted in order to study the differences in the usage of sources of information between social sciences and humanities (Nederhof,[7] Dilevko et Gottlieb[8]…) The results have shown that there is a minor difference in the usage of a certain kind of source of information – humanities rely more on books, while social sciences makes more use of journals.

This research is intended to analyse, according to the statistical data gathered at the Faculty of Humanities and Social Sciences' Library, the information about the frequency of the usage of traditional (e.g. print books) and electronic resources (e.g. electronic journals) and to investigate whether there is a considerable difference in the frequency of this usage between the humanities and the social sciences. Unfortunately, due to the fact that the ILS (i.e. KOHA) was acquired only recently, it was only possible to use the statistical data gathered between March and September 2009. The databases taken into consideration were Project MUSE, JSTOR, and relevant bases from the EBSCOhost aggregator.

---

[3] E-journals: their use, value and impact : a Research Information Network report : April 2009. 3 July 2009. http://www.rin.ac.uk/files/E-journals_use_value_impact_Report_April2009.pdf (24 August 2009)

[4] Liu, Ziming. Print vs. electronic resources : a study of user perception, preferences, and use. // Information processing and management, 42(2006); 583-592

[5] Kriebel, Leslie; Lapham, Leslie. Transition to electronic resources in undergraduate social sciences research : a stuy of honors theses bibliographies, 1999-2005. // College and Research Libraries, 69(2008), 3; 268-283

[6] Gerke, Jennifer; Mannes, Jack M. The physical and the virtual : the relationship between library as place and electronic collections. // College and Research Libraries, preprint – accepted April 8, 2009; anticipated publication date November 2009. http://www.ala.org/ala/mgrps/divs/acrl/publications/crljournal/preprints/Gerke-Maness.pdf (20 August 2009)

[7] Nederhof, Anton J. Bibliometric monitoring of research performance in the Social Scinces and the Humanities. 2006.

[8] Dilevko, Juris; Gotlieb, Lisa. Print sources in an electronic age : a vital part of the research process for undergraduate students. // The Journal of Academic Librarianship, 28(2002), 6; 381-392

The purpose of this research is to see whether there is a considerable difference in the frequencies of usage of printed and electronic resources as seen from the perspectives of humanities and social sciences.

## Print vs. Electronic

### The Problem

Based on the aforementioned research, the issue of the usage of printed and electronic resources at the Faculty of Humanities and Social Sciences' Library has arisen. From the statistical information we have at our disposal, it is possible to gather the data about the number of borrowed books and about the frequency of access to the referential electronic resources (e.g. electronic journals) that the Faculty of Humanities and Social Sciences' Library has access to. Unfortunately, we do not have any information about the usage of print journals yet because no records thereof have been kept so far.

The aim of this research is to determine the relation between the number of borrowed books and the access to referential electronic databases in order too see whether there is a significant difference in the usage of print books and electronic journals with regard to scientific fields being taught and researched at the Faculty of Humanities and Social Sciences, University of Zagreb. As we have previously mentioned, we will not be able to include the information about the usage of print journals, due to the impossibility of gathering that information. This is left to be done in future research!

### Scientific Disciplines and Fields and Faculty of Humanities and Social Sciences' Library Subjects

The Faculty of Humanities and Social Sciences, University of Zagreb, encompasses only some fields of humanities and social sciences. The classification of sciences in this work has been made according to the valid Pravilnik o znanstvenim i umjetničkim područjima, poljima i granama (Regulation of Scientific and Artistic Disciplines, Fields and Branches).[9] The scientific fields from the discipline of social sciences that are taught and researched at the Faculty of Humanities and Social Sciences in Zagreb are:

- Information and communication sciences
- Sociology
- Psychology
- Pedagogy

The fields from the discipline of humanities that are taught and researched at the Faculty of Humanities and Social Sciences in Zagreb are:

- Philosophy

[9] Pravilnik o znanstvenim i umjetničkim područjima, poljima i granama. // Narodne novine, 78(2008). http://narodne-novine.nn.hr/clanci/sluzbeni/340161.html (15 July 2009)

- Philology
- History
- Art history
- Archaeology
- Ethnology and anthropology

The FHSS Library subjects are formed according to fields of study and study programs, so that call numbers indicate which scientific domains a certain subject encompasses. (Table 1.)

Table 1. Faculty of Humanities and Social Sciences' Library subjects and scientific disciplines and fields

| Call numbers | Subjects | Scientific disciplines and fields[10] |
|---|---|---|
| | SOCIAL SCIENCES | SOCIAL SCIENCES |
| BB | Methodology | All social sciences |
| BC | Psychology | Psychology |
| BD | Sociology | Sociology |
| BF | Pedagogy | Pedagogy |
| BG | Information sciences | Information and communication sciences |
| | HUMANITIES | HUMANITIES |
| BA | Philosophy | Philosophy |
| CA | History | History |
| CB | History of art | History of art |
| CC | Archeology | Archeology |
| CD | Ethnology | Ethnology and Anthropology |
| BE | Anthropology | Ethnology and Anthropology |
| D | Slavic languages and literatures | Philology |
| EA | Phonetics | Philology |
| EB | Linguistics | Philology |
| EC | Comparative literature | Philology |
| ED | Classical philology | Philology |
| EE | Indology | Philology |
| EF | Chinese language and literature | Philology |
| EG | Japanese language and literature | Philology |
| EH | Hungarian language and literature | Philology |
| EI | Turkish language and literature | Philology |
| EJ | Scandinavic languages and literatures | Philology |
| FA | English language and literature | Philology |
| FB | German language and literature | Philology |
| FC | Dutch language and literature | Philology |
| FD | Italian language and literature | Philology |
| FE | French language and literature | Philology |
| FF | Spanish language and literature | Philology |
| FG | Portuguese language and literature | Philology |
| FH | Romanian language and literature | Philology |

[10] Ibid.

## Statistical Data

*Books*

The FHSS Library was opened on March 11th 2009, and that is when the KOHA integrated library software (ILS) was implemented. Previous to this, books were being borrowed through library cards, which made it quite difficult to collect any statistical data. According to the circulation logs drawn from the call numbers, 7,141 books in social sciences and 33,334 books in humanities were borrowed until September 1st 2009.

Since subjects were formed according to the fields of study and study programs, it is possible that in every subject there are also books belonging to the other scientific discipline. Seeing as this situation exists in all subjects, we did not take it into consideration.

*e-journals*

The databases included in this research are the ProjectMUSE, JSTOR, and the EBSCOhost aggregator, due to their relevance for the domains of social sciences and humanities, and also because of their availability.

The classification of sciences was effected according to the aforementioned Regulation.[11] The problem that occurred in the process of gathering the data was that of how to classify the journals in the databases according to the scientific domain they cover (that is, how to distinguish journals for humanities from those for social sciences). We opted for manual sorting of journals into two groups as the solution. The criteria we sorted the journals by were the descriptions of each journal in the database. The journals that could not be unequivocally sorted into one of the two groups (multidisciplinary journals that cover both domains) were excluded from this research.

The data about the access to the relevant databases (ProjectMUSE, JSTOR, EBSCO aggregator) were taken from the statistical reports on the access to full-text articles, either in PDF or HTML format.

In the same period, according to the access logs, full-text articles in the referential databases were accessed 39,257 times in the discipline of humanities, while full-text articles covering the discipline of social sciences have been accessed 16,155 times.

## Discussion

The results of the research have shown that the literature from both humanities and social sciences is more often accessed electronically than by borrowing print books. (Table 2.)

[11] Ibid.

Table 2. Percentage of full-text access and book loans

| | Humanities | Social sciences |
|---|---|---|
| **E-resources (full-text access)** | 54.0% | 69.2% |
| **Print resources (book loans)** | 46.0% | 30.8% |
| **Base** | 72,491 | 23,329 |

Regarding the access to full-text articles in the referential databases, we have learned that social sciences make more use of e-resources than humanities do: the ratio is 69.2% of total literature usage in social sciences to 54.0% in humanities.

As expected, e-resources were used more in social sciences than in humanities. This can be explained by the fact that the literature becomes outdated in social sciences faster than in humanities, so the social sciences scholars are traditionally more incline to journals, therefore to e-journals as well. Also, the simplicity and ease of using the e-sources can just help the social sciences scholars and students to use e-journals frequently.

The results of this research show that the usage of e-resources is surprisingly high in humanities comparing to aforementioned researches. That can be explained by several reasons:

- **Selection of the referential databases** - EBSCOhost is financed by the Ministry of Science, Education and Sport, while Faculty of Humanities and Social Sciences has financed ProjectMUSE with its own funds since 2006. Apart from those two, there is also a database Jstor which is specialized mainly in humanities. Faculty of Humanities and Social Sciences has acquired Jstor in 2009 after multiple requests from the faculty, and this acquisition has definitely increased the number of accesses to e-journals, after a long period of waiting and anticipation.
- **Higher level of computer and information literacy** – in the last few years new information and computer technologies have penetrated deeply in everyday life, and the consequence is that the users are more computer and information literate, due to frequent use of computer or information retrieval. Faculty of Humanities and Social Sciences' Library has played an important role in this process with continuous efforts and educational programs that are conducted for all faculties, including humanities scholars and teachers as well. It should be mentioned that in few last years great number of young scholars, who are more computer literate, have become faculty members so that can also be the reason of higher percentage of the usage of e-resources.
- **Accessibility** – Faculty of Humanities and Social Sciences' Library has secured 200 computers with Internet access for users in new facilities

since March 2009. Apart from the improvements in the library building, it is possible to access and search the databases from home via proxy-server that helps the faculty to work at home because the Faculty facilities cannot provide them enough space to work at the Faculty. Such results justify high investments in the acquisition of the e-resources at the Faculty of Humanities and Social Sciences for last few years, and give the further stimulation for librarians to continue the good work. Nevertheless, some questions remain to be answered. It is necessary to conduct further researches to see whether the acquisition of Jstore has significantly influenced on the data. Another research should definitely include the data about the use of print journals.

## Conclusion

As previous researches dealing with the usage of print and e-resources have shown, this one also points that there is a statistically significant difference in the usage of print and electronic resources, respective of scientific domains. Related to the topic of this research, social sciences make more use of e-resources than humanities (expected hence their nature).

Overall, e-journals are used more than print books in both - social sciences and humanities. It is interesting to see that they are used more than print in the humanities as well (different from previous researches). The reason behind this could be the right selection of the specialized and high-quality e-journals Faculty of Humanities and Social Sciences has access to. The policy of investment in referential databases (ProjectMUSE, Jstor), additional to the one financed by Ministry of Science, Education and Sport (EBSCOhost) turned out to be valid, allowing the faculty and the students to consult the most relevant information. Furthermore, the library users have become more computer and information literate due to everyday use of information and computer technology, as well as the Faculty of Humanities and Social Sciences' Library efforts to inform and educate the users to use those databases. The third reason is the accessibility of referential databases from Faculty facilities as well as from home via proxy server.

## References

Dilevko, Juris; Gotlieb, Lisa. Print sources in an electronic age : a vital part of the research process for undergraduate students. // *The Journal of Academic Librarianship*, 28(2002), 6; 381-392

E-journals: their use, value and impact : a Research Information Network report : April 2009. 3 July 2009. http://www.rin.ac.uk/files/E-journals_use_value_impact_Report_April2009.pdf (24 August 2009)

Gerke, Jennifer; Mannes, Jack M. The physical and the virtual : the relationship between library as place and electronic collections. // *College and Research Libraries*, preprint – accepted April 8, 2009; anticipated publication date November 2009. http://www.ala.org/ala/mgrps/divs/acrl/publications/crljournal/preprints/Gerke-Maness.pdf (20 August 2009)

Hiller, Steve. How different are they? A comparison by academic area of library use, priorities, and information needs at the University of Washington. // *Issues in Science & Technology Librarianship*, 33(Winter 2002); 13p. http://www.istl.org/02-winter/article1.html (5 August 2009)

Kriebel, Leslie; Lapham, Leslie. Transition to electronic resources in undergraduate social sciences research : a stuy of honors theses bibliographies, 1999-2005. // *College and Research Libraries*, 69(2008), 3; 268-283

Liu, Ziming. Print vs. electronic resources : a study of user perception, preferences, and use. // Information processing and management, 42(2006); 583-592

Nederhof, Anton J. Bibliometric monitoring of research performance in the Social Scinces and the Humanities : a review. // *Scientometrics*, 66(2006), 1; 81-100

Pravilnik o znanstvenim i umjetničkim područjima, poljima i granama. // Narodne novine, 78(2008). http://narodne-novine.nn.hr/clanci/sluzbeni/340161.html (15 July 2009)

Original scientific paper

# OA Repositories @ Special and Academic Libraries in Zagreb

Ivana Hebrang Grgić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
ivana.grgic@ffzg.hr

Ana Barbarić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
abarbari@ffzg.hr

Iva Džambaski, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
idzambas@ffzg.hr

**Summary**

*The introductory part of the paper deals with the "serial crisis", the open access movement and the characteristics of the Croatian scientific community. The most important causes of the "serial crisis" are explained. The open access movement is defined and the most important initiatives, documents and projects are described. Two ways of achieving open access (open access journals and open access repositories) are defined as well.*
*Research on special and academic libraries in Zagreb constitutes the main part of the paper. An electronic online questionnaire has been sent to the libraries of all research institutions in Zagreb that have not yet established their institutional repositories. Librarians were asked whether their institutions should have open access repositories and, if the answer was affirmative, how they think their future repositories should be organized (which software should be used, which formats and types of documents should be deposited, whether they have to be OAI-PMH compliant, how to deal with copyright issues, etc.)*
*The aim of research is to determine how many librarians of special and academic libraries in Zagreb think that institutional repositories are necessary and find out librarians' opinions on establishing open access repositories. The results show librarians' awareness of the importance of the open access move-*

*ment. The results also indicate if there is a need for some national-based plans for action and the funding of institutional open access repositories.*

**Key words:** academic libraries, institutional repositories, open access, special libraries

## Introduction

Since the launch of the first scientific journals in 1665, the number of journals has been growing rapidly. Scientific journals have been recognized as the best way of formal scientific communication. Owing to quality control by means of peer-review, the pyramid of journals has been established. At the bottom of the pyramid there are a great number of low quality journals, and at the top of the pyramid there are only few high quality journals. After World War II some commercial publishers saw a money-making opportunity in publishing high quality journals. Increasingly many publishers started publishing scientific journals and they kept increasing subscription prices without fear of losing subscribers. Libraries had to reallocate their budgets in order to meet their users' needs for high quality scientific journals.

The “serial crisis”culminated in the 1990s as a result of the increasing prices of journal subscriptions, which had an adverse impact on formal scientific communication. Rising journal prices resulted in the decreasing accessibility and visibility of scientific output. The crisis was a very serious threat to scientific communication, as well as to science in general.

At the same time, as a result of technological development electronic publishing became a new way of making research results available to the scientific community and wider public.

The possibility of publishing on the Web seemed to be a way out of the “serial crisis”.

## The Open Access Movement

Three initiatives important for defining and popularizing Open Access (OA), the so-called 3B initiatives, are the *Budapest Open Access Initiative* (2002), the *Bethesda Statement on Open Access Publishing* (2003), and the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* (2003).

*The Budapest Open Access Initiative* (BOAI) was the first document to define Open Access and two ways of achieving it. According to the BOAI, OA is free and unrestricted online availability of scholarly literature (primarily peer-reviewed journal articles, but it can also include unreviewed articles). Open Access would increase visibility, readership and impact of scholarly literature. Two ways of achieving OA, which are recommended by the BOAI, are self-archiving (depositing scholars’ refereed journal articles in open electronic archives or repositories) and open access journals (a new generation of journals that are openly available online, without any restrictions). The BOAI also em-

phasized that the overall cost of providing open access should be lower than the costs of traditional forms of publishing.[1]

The *Bethesda Statement on Open Access Publishing* defines an Open Access publication as one that meets two conditions. The first one is that the author (and/or copyright holder) grants to all users free, irrevocable, worldwide and perpetual access to the work and a license to copy, use and distribute it, as well as the right to make small numbers of printed copies for personal use. The second condition is that a complete version of the work is deposited immediately upon its initial publication in at least one online open access repository.[2]

The goal of the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* is the realization of the vision of a global and accessible representation of knowledge. The web has to be sustainable, interactive and transparent; content and software tools must be openly accessible and compatible.[3]

Open Access was very soon recognized in Croatia as a valuable initiative for achieving better visibility of both global and Croatian research results. Today (May 2009) more than 200 Croatian scientific journals are OA journals (available either on Hrčak, the portal of Croatian scientific journals, or on their own websites). The second way of achieving Open Access, OA repositories, is less common in Croatia. Only three institutional repositories have a tendency to be OA repositories, and they are all established at the faculties of the University of Zagreb. Do academic and special librarians in Croatia think that OA repositories are necessary? Have they considered the possibility of establishing them? Are they aware of all the problems connected with choosing appropriate software, metadata model, and types of documents? Are they aware of the problems connected with copyright issues?

## Research methods and sample

The aim of our research was to examine librarians' attitudes about launching institutional open access repositories. We created an online questionnaire for the libraries of research institutions without repositories and sent it by e-mail to 46 addresses – to each of the 25 libraries of the University of Zagreb faculties, as well as to all the libraries of research institutes in Zagreb (21 libraries). The list of research institutions has been obtained from the Ministry of Science, Education and Sports website.[4] The questionnaire was created using Formdesk web-

[1] Budapest Open Access Initiative. 2002. http://www.soros.org/openaccess/read.shtml (3-4-2008)

[2] Bethesda Statement on Open Access Publishing. 2003. http://www.earlham.edu/~peters/fos/bethesda.htm (3-4-2008)

[3] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html (3-4-2008)

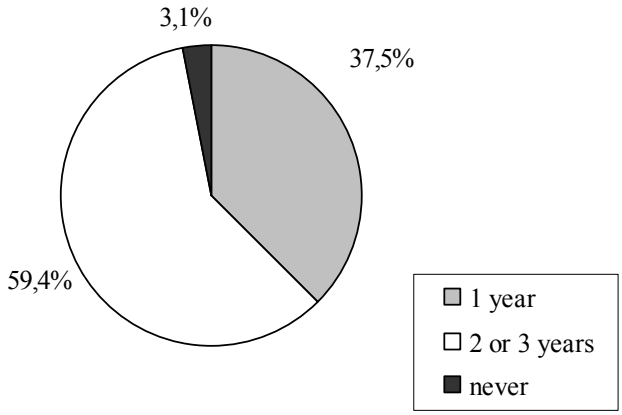
[4] Ustanove iz sustava znanosti. Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske. http://pregledi.mzos.hr/Ustanove_Z.aspx (14-1-2009)

site forms which offer various useful features such as e-mail auto responses, statistics, result downloading, password protection, and secure data transfer.[5]
We received 32 responses (69.6 % response rate).
The questionnaire consisted of 14 questions. We asked the librarians whether their institutions should have OA repositories and, if the answer was affirmative, how they think their future repository should be organized (which software should be used, which formats and types of documents should be deposited, whether they have to be OAI-PMH compliant, how to deal with copyright issues, etc.) We also asked them who should initiate the launch of their institutional repository, who should maintain it, who would be its users and, finally, if they had any plans for establishing a repository.

## Results

A great majority of respondents (31 or 96.9%) find it necessary to establish a digital repository at their institution.
12 librarians (37.5%) think that a repository should be established within a year. In our opinion, this expectation is impossible to fulfill because of all the work that must be done before a repository becomes operational. Nonetheless, some of the respondents later answered that they had already begun setting up a repository and might be able to finish the work in one year's time. 19 librarians (59.4%) answered that their repository should be set up in 2-3 years' time. One respondent (3.1%) answered that no repository should be established (Chart 1).

Chart 1: How long will it take to set up an institutional repository?

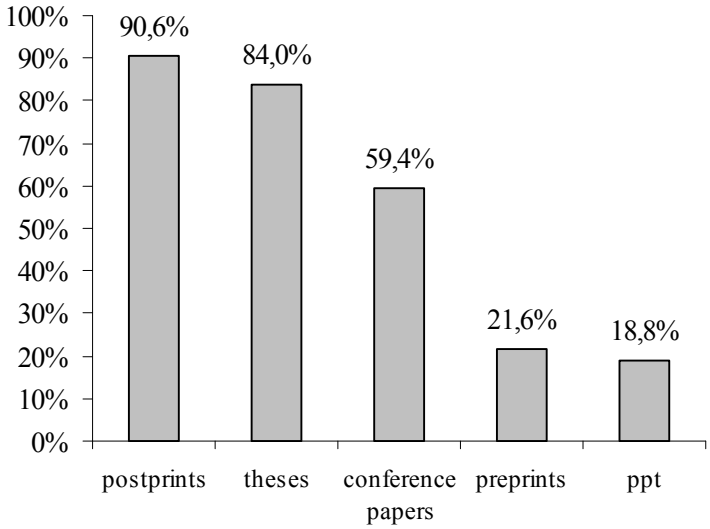



[5] Otvoreno dostupni digitalni repozitoriji. April 2009. http://www.formdesk.com/grgic/oa_repository (17-4-2009)

In the third question librarians were asked who they think should initiate the launch of their repository. 34.4% of them think that it should be initiated by the Ministry of Science, Education and Sports, 31.3% answered that it should be initiated by the library, and 28.1% of the answers were – the institution. The answers indicate that many librarians are aware of their own role in launching an institutional OA repository, but also a great number of them know that repositories could be established much more easily with the help of the Ministry.
The answers to the fourth question support that conclusion – 80.7% of librarians think that their repositories should be funded by the Ministry.
In the fifth question librarians were asked who should have access to the repository. 93.8% of them think that repositories should be OA repositories and 6.2% think that only the employees of their institution should have access to the repository.
How many persons should be responsible for maintaining a repository? 65.6% of librarians answered that one person would be enough and 31.3% answered that two to three persons should be in charge of repository maintenance. Maintaining a repository is hardly a job that can be done effectively by one person. The majority of librarians are not aware of the complexity of work that must be done if they want the repository to be updated on a regular basis.
When asked about software that should be used, 75% of librarians answered that open-source software should be used and 18.8% think that commercial software would be a better choice. The majority of librarians know that OA repositories around the world use open-source software. The two most important preconditions for an OA repository are that open source software is used and that it is OAI-PMH (Open Access Initiative – Protocol for Metadata Harvesting) compliant.[6]
Librarians think that scientific papers and doctoral theses are the most important material for their users. 90.6% of librarians would like to deposit peer-reviewed scientific papers (postprints) and 84% of them would deposit doctoral theses. They also find conference papers very important (59.4%) and 21.6% of them would deposit preprints (papers not yet peer-reviewed but submitted for publishing in a scientific journal). Chart 2 shows answers to the question about material types that should be deposited in an institutional OA repository.
As to the most appropriate format for depositing documents, the PDF format received the highest percentage of answers (93.7%), while other formats such as RTF, Tiff, HTML, DOC and PPT obtained lower percentages. This leads to the conclusion that librarians find the format similar to printed publications the most appropriate and that they have not given a lot of thought to the potential advantages of other formats.

[6] Corrado, E. M. The Importance of Open Access, Open Source and Open Standards for Libraries. // Issues in Science and Technology Librarianship. 2005. http://www.istl.org/05-spring/article2.html (11-12-2008)

Chart 2: Which material types should be deposited in a repository?




In the tenth question we asked librarians who should archive papers in a repository. The highest percentage of answers, i.e. 56.3%, refers to self-archiving by authors themselves (34.4%) or with the assistance of a librarian (21.9%). Since the question was a multiple answer type, the solution where librarians alone should archive also prevailed with a significant percentage (43.7%). 12.5% of answers refer to self-archiving by authors assisted by someone other than a librarian. It seems that librarians are aware that authors, especially those self-archiving for the first time, might need some assistance. They also believe that librarians are professionals who have to and are trained to help them. Based on the percentages of answers shown above, it can be claimed that librarians understand that self-archiving requires continuous cooperation between authors and librarians and that the archiving process, even within the same repository, can be organized in several ways. The mentioned percentages are shown in Chart 3.

The majority of surveyed librarians, i.e. 77.4%, consider that consent for archiving should be sought from copyright owners. Librarians are aware of the importance of copyright issues knowing that any copyright infringement could cause problems for their institution (Chart 4).

Answers to the twelfth and thirteenth question also indicate that librarians are conscious of the environment of the institution the library is part of. Namely, a high percentage of librarians (96.8%) believe that self-archiving in an institutional repository should be obligatory for all the employees of the institution. An institutional self-archiving policy is necessary in order to define self-archiving and the repository itself. Librarians obviously understand that their cooperation with the institution is crucial for establishing and maintaining a high-

quality digital repository. Librarians know that the efficiency of such cooperation would be guaranteed by the existence of a formal document (i. e. the self-archiving policy). Such a document would also contribute to the better recognition of the institution in a wider scientific community. Nevertheless, the above-mentioned awareness of librarians of the importance of copyright issues points to the fact that librarians understand that obligatory self-archiving cannot be done "automatically", but requires continuous cooperation with copyright owners.

Chart 3: Who should be in charge of archiving material in a repository?

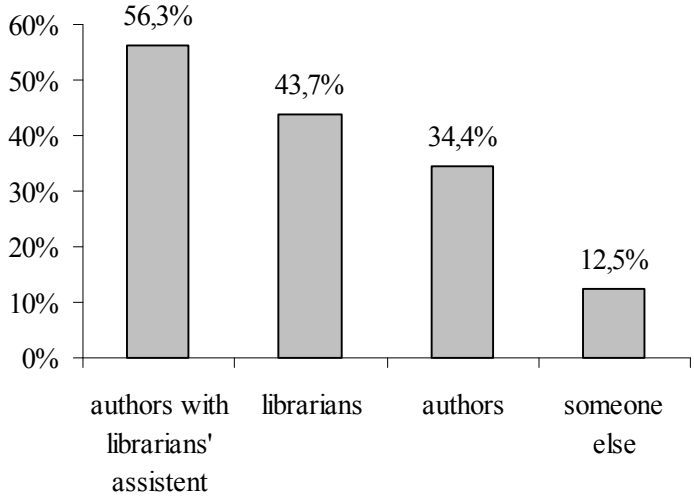
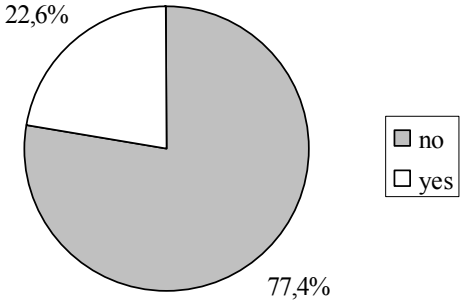



Chart 4: Should consent for archiving be sought from a copyright owner?



Answers to the fourteenth question indicate that the reasons why an institutional repository has not yet been established are the lack of obligatory self-archiving (58%), followed by the lack of staff in charge of its maintenance (48.4%), the lack of financial means (38.7%), the indifference of management (25.8%) and indifference of employees (19.3%), as shown in Table 1. By pointing out the fact that self-archiving is not obligatory as the main reason why institutional repositories do not exist, librarians implicitly express their opinion that their users (employees of an institution) have not adopted and/or recognized the main principles of the OA movement.

Table 1: Main reasons why an institutional repository has not been established

| Reason | Percentage |
|---|---|
| Non-obligatory self-archiving | 58.0 |
| Lack of maintenance staff | 48.4 |
| Financial problems | 38.7 |
| Indifference of management staff | 25.8 |
| Indifference of employees | 19.3 |
| Nothing of the above | 9.7 |

Exactly half of surveyed librarians answered affirmatively to the question if they have any plans for establishing a repository, while the rest of them answered in the negative. These results are far from satisfactory. It can be assumed that librarians who currently do not plan to establish a repository are waiting for a formal initiative, either from the Ministry of Science, Education and Sports or from their institution. They believe that such an initiative could solve financial problems, the problem of obligatory self-archiving as well as the problem of the lack of staff who would be reponsible for its maintenance. However, this thesis leads us to the fundamental question – who should initiate the establishing of repositories? Although some respondents (31.3%) believe that the initiative should come from the library, the majority believes that the initiative should come from the Ministry of Science, Education and Sports (34.4%) or from the institution the library is part of (28.1%), probably having in mind the depositing and funding problems.

## Conclusion

We can draw a conclusion that the librarians of special and academic libraries in Zagreb are relatively well acquainted with the issues relating to establishing and maintaining OA repositories. The fact that the majority of respondents believe such a repository is necessary for their institution is commendable, but the fact that only half of surveyed libraries (or their institutions) have plans for setting it up gives cause for concern. Certain misleading assumptions of part of the surveyed librarians, e.g. those that a repository can be established within a year and that one person is sufficient for its maintenance, can be interpreted as the lack of

practical experience or knowledge of how to set up and maintain a repository. Librarians are also aware of copyright issues and related problems.

The most important problem and the main reason why OA repositories have not been established yet is the lack of self-archiving policies. Funding is also an important issue. The majority of librarians think that the Ministry of Science, Education and Sports should take part in both initiating and funding institutional OA repositories. Also, libraries should have detailed plans for setting up and maintaining an institutional repository. Most librarians are aware of the importance of their help to authors in the self-archiving process.

In conclusion, we can say that special and academic librarians in Zagreb are very well informed about OA repositories and are ready to take an active part in their establishing and maintenance.

## References

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html (3-4-2008)

Bethesda Statement on Open Access Publishing. 2003. http://www.earlham.edu/~peters/fos/bethesda.htm (3-4-2008)

Budapest Open Access Initiative. 2002. http://www.soros.org/openaccess/read.shtml (3-4-2008)

Corrado, E. M. The Importance of Open Access, Open Source and Open Standards for Libraries. // Issues in Science and Technology Librarianship. 2005. http://www.istl.org/05-spring/article2.html (11-12-2008)

Otvoreno dostupni digitalni repozitoriji. April 2009. http://www.formdesk.com/grgic/oa_repository (17-4-2009)

Ustanove iz sustava znanosti. Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske. http://pregledi.mzos.hr/Ustanove_Z.aspx (14-1-2009)

# Libraries in Web 2.0 Environment

Mihaela Banek Zorica
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mbanek@ffzg.hr

Ana Eremić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
aeremic@ffzg.hr

**Summary**
*Changes in the Web environment have influenced all aspects of human professional and leisure behaviour. As libraries main purpose has always been to respond to its users' information needs the transition currently affecting information environment has posed new challenges on them. Paper presents overview of the definitions and insights into the Library 2.0 concept looking at its both positive and negative aspects.*

**Key words:** library, web 2.0, library 2.0

## Introduction – current information space and emergence of web 2.0

Thought the history libraries have always been social and cultural centres aiming to respond to the users' information needs. The advancement of technology and the emergence of the internet transformed the information space and enabled access to information also outside the library walls. Emergence of the next level of web - Web 2.0 - influenced all aspects of human life including the library environment. Philosophy of Web 2.0 environment, mainly dependent on its users and their social interaction, is based on "gravitational core" that ties together set of principles and practices lacking any hard boundaries. (O'Reilly 2005) In practice Web, 2.0 offers access and reuse of data and services that were once, in Web 1.0 environment, "locked" on various web sites. Web 2.0 is about interactive systems i.e. applications that enable users to gather information resources, add comments, adapt retrieved items to their own needs, as well as publish them and create their own information space (Špiranec, Banek Zorica, 2008). This new information environment dependant on the social aspect posed itself as a "new problem" as well as a challenge for the Being that Web 2.0 provides so many new possibilities for the information sector, a lot of libraries have accepted and integrated it into their system thus becoming Libraries 2.0. (Kelly, Bevan, Akerman, 2008).

Current research literature offers various definitions of the Library 2.0 – L2.0 (Casey and Sastinuk 2006; Mannes, 2006; Miller, 2006; Holmberg, 2009 …) and being a relatively new subject, there is a lack of a worldwide standard leaving the libraries to cope and struggle by themselves in implementing Web 2.0 tools. On the other hand setting a standard in the context of Web 2.0, would be problematic as its definition is "constant beta". i.e. not a fully developed product. Still consistency in Library 2.0 definitions exists as it is an environment oriented on and developed by the users who by participation and feedback become co-creators. Implementation of Web 2.0 tools in the library setting varies depending on how Library 2.0 is defined. For some the term Library 2.0 means the incorporation of blogs, wikis, instant messaging, RSS, and social networking into library services while, for others it suggests involving users through interactive and collaborative activities such as adding tags, contributing comments and rating different library items (Aharony, 2008)

Traditionally, library's main function was oriented towards their users and their information needs. One should remember Ranganathans Five Laws of Library Science which confirm this statement. Therefore, we should emphasize that current transition is in the information environment and user types i.e. a user-centered change (Casey, Savastinuk, 2005). Traditional services oriented towards users growing up in the pre-digital environment can, unfortunately, not survive and respond to the "new needs" of current and potential library users living a working in the changed information environment. Prensky (2005) defines two groups of today's users: generation X any Y. *Generation X* which are the library's "old users" born before the emergence of the digital world and the *Generation Y* or *Digital natives* i.e. generation born in the digital world fluent in technology use. Responding to and adapting services for these different types of users together with the management of the structured information space now becomes the main task of Library 2.0.

Library 2.0 is therefore a logical step in responding the users' needs and it certainly does not mean breaking up with the traditional models but is more a response to the transformed information environment. As it presents more a philosophy of new information behaviour and represents an innovative view on the solution of current situation there is a lot of critique of Library 2.0 found in the library community. Habib (2006) gathers them around two crucial ones:

1. The term "Library 2.0" is confrontational in that it declares, or implies, all prior library services obsolete and in need of replacement;
2. The term "Library 2.0" is meaningless in that it provides nothing new to the professional discourse. It essentially means nothing more than progressive librarianship.

First critique can easily be disputed as definitions of Library 2.0 have not claimed that libraries should end their traditional services but open up to the user-centred environment and create responding combination of their services. This could be confirmed in Miller (2007) definition of Library 2.0 as Library +

Web 2.0. Furthermore, statement like the one claiming that relation of library and Library 2.0 is the same as the Web 1.0 vs. Web 2.0 are unsustainable as in case of web environment new version makes the first one obsolete while in library settings all aspects are encompassed. What Library 2.0 represents is a subset of new library services that are occurring because of the changes brought on by Web 2.0 services.

Casey and Savastinuk (2006) emphasize that even traditional libraries can be Library 2.0 if their services successfully reach users, are evaluated frequently and make use of users input. Moreover, they define Library 2.0 as a model for library service that encourages constant and purposeful change, invites user participation in the creation of both the physical and the virtual services they want, supported by consistently evaluating services and attempts to reach new users as well as better serve current ones through improved customer-driven offerings. Combination of these factors is what constitutes the Library 2.0. There are four essential elements constituting Library 2.0 (Maness, 2006):

- *The library has to focus on its users* - The users actively participate in the creation of content and services available on the libraries web sites, OPAC, etc. The consumption and creation of content is extremely dynamic and this is why the lines between the roles of librarian and user are sometimes blurred. The Librarian 2.0 can offer help and support, but in the Library 2.0 he is not solely responsible for the creation of content.
- *The library has to offer a multimedia experience* - Both the collections and the services offered by the Library 2.0 can contain both video and audio components.
- *The library is socially diverse* - The presence of the library on the Web also entails the presence of the users as well. There are synchronous (like IM) and non- synchronous (like wiki) ways for the costumers to communicate between themselves or with librarians.
- *Libraries as the innovators of a community* - Libraries offer their services to a community of people, but as communities change so they affect the libraries to change as well, so the libraries have to let the communities to change them. The library has to continually change its services, find new ways in which whole communities, not only individuals, can search, find and use information.

Accordingly, the best way to visually represent this concept would be by utilizing the Web 2.0 meme map and adapting it to the library setting thus creating a Library 2.0 meme map (Figure 1). It presents a transformed library environment which tries to bind different services, traditional and modern, physical and virtual which co-exist in today's library environment. Similarly to the Web 2.0 philosophy, Library 2.0 has its gravitational core and set of principles and practices floating around this core.

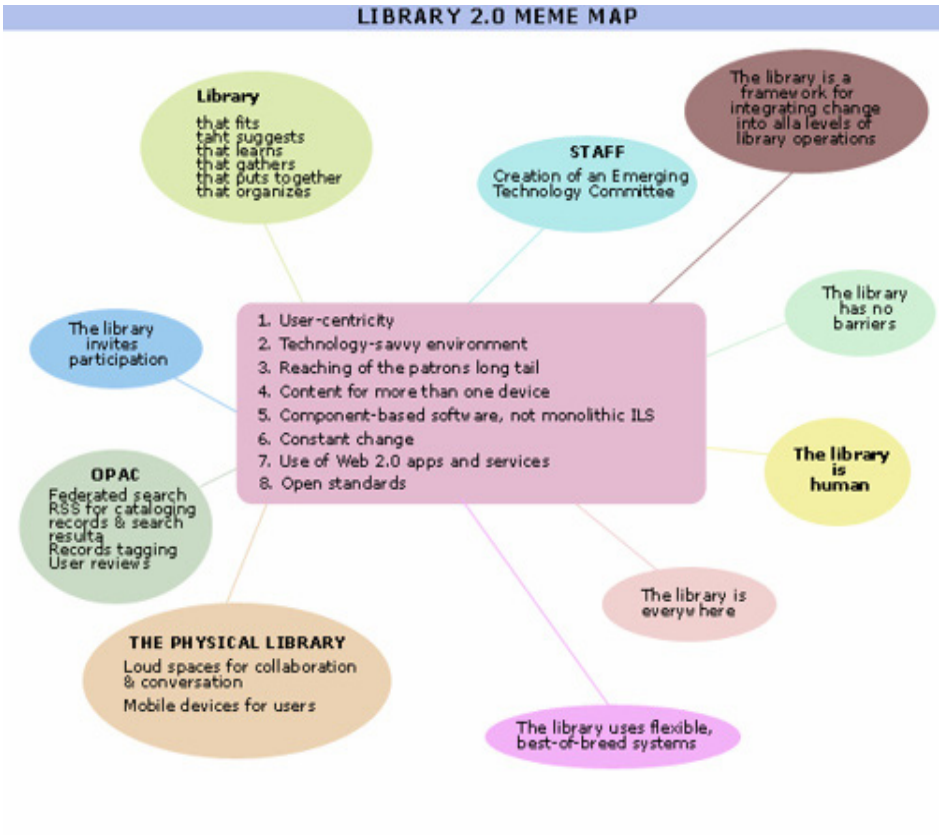



Figure 1. Bonariabiancu 2006 http://www.flickr.com/photos/bonaria/111856839/

In order to consolidate these statements, critiques and definitions in one comprehensive insight into the problem and in order to standardize set of guidelines a model of what constitutes a Library 2.0 is necessary. Holmberg et al. (2009.) proposed a model of Library 2.0 (Figure 2) based on the library community point of view. New model takes into account all the aspects of both traditional and new library environment defining seven building blocks of Library 2.0: interactivity, users, participation, libraries and library services, web and web 2.0, social aspects, and technology and tools.

From these building blocks an empirical definition can be drawn. Library 2.0 presents a *"...change in interaction between users and libraries in a new culture of participation catalysed by social web technologies..."* (Holmberg et al., 2009.)

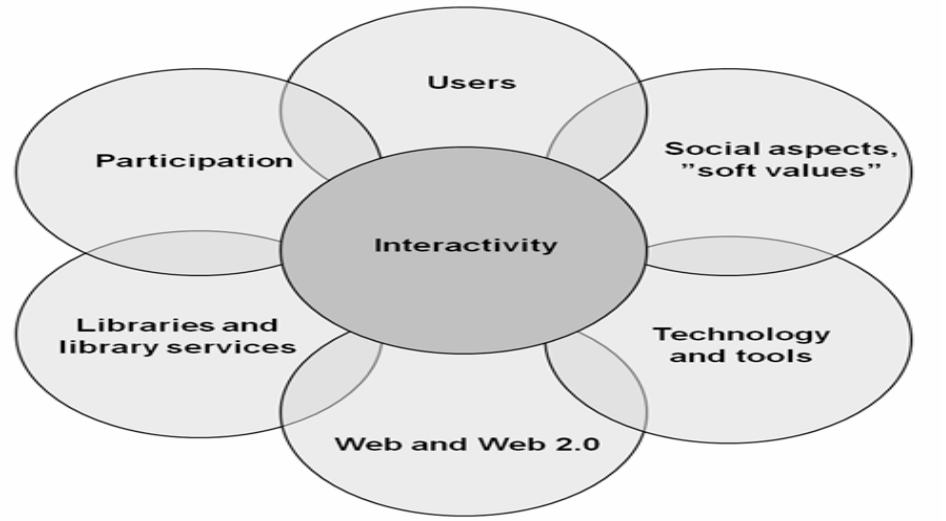



Figure 2. Model of Library 2.0 (Holmberg et al., 2009.)

## Responding to the challenge

Historically the discovery and location processes were tied to each other in the catalogue. Where somebody discovered something elsewhere (citation, bibliography …) they would then inspect the catalogue. Of course, we want to be able to find out what is in the local catalogued collection, but to what extent should that be the front door to what the library makes available? (Dempsey, 2006) Rather than being locked inside the library system, data can add value to the experience of users wherever they are, whether it is Google, Amazon, the institutional portal, or one of the social networking sites such as MySpace or Facebook. By unlocking data and the services that make use of it, the possibilities are literally endless, and it is here that efforts such as those around the construction of a library 'Platform' become important. (Miller, 2007)

Users of Amazon and other consumer sites are becoming used to a 'rich texture of suggestion', which leads into the mobilisation of user participation - tagging, reviews - to enhance the discovery experience. There is a general recognition that discovery environments need to do more to help the user by: *ranking* (using well-known retrieval techniques with the bibliographic data, or probably more importantly, using holdings, usage or other data which gives an indication of popularity), *relating* (bringing together materials which are in the same work, about the same thing, or related in other ways), and *recommending* (making suggestions based on various inputs - reviews or circulation data for example). (Dempsey, 2006) These new technologies brought changes in the library environment enabling library catalogues to become more opened to the users by

enabling them to communicate inside the catalogue and offering them different access options. The new and improved "editions" of catalogues are now called Catalogue 2.0 or OPAC 2.0.

Guidelines on how to create catalogues more appealing to the users and similar to other services found on the web and especially in the Web 2.0 environment were presented on the Librarything blog in 2006. Suggestions for a Catalogue 2.0:

- Provide blog widgets and RSS feeds so patrons can show off what they're reading and what they thought of it.
- Let people find what they want, but let them also get entertainingly lost. Encourage exploration, serendipity and lost-ness.
- Give authors, subjects, languages, tags and other facets their own pages. That stuff's interesting, and can lead one delightfully astray.
- Allow patrons to interact with the catalogue via tags, ratings and reviews. (And would it kill you to give them patron pages?)
- Link outward. The web is fun. Point to it.
- Allow (static) inbound links. What are you, a bouncer?
- Let patrons access your data via API. Some clever patron will do something fun you hadn't thought of.
- Give patrons a reason to check in every day—something about the books, and ideally about *them* and the books, not some "trick" like free movie passes.
- Talk to patrons in their own language (eg. with tags), not in some crazy argot, where "cooking" is "cookery" and "the internet" is "the information superhighway."
- Give patrons fun, high-quality recommendations.
- Give patrons enjoyable metadata. I don't intend to read any of the books in today's *NYT Book Review*, but I loved reading about them.
- Let users interact socially around the books they read. (Obviously, anything social needs to be voluntary.)
- Make it usable and finable too.

Examples of this practice can bee seen in various library catalogues all over the world. Unfortunately, this is still not a standard with its general application but rather a movement where transition is applied on single libraries like *Ann Arbor District Library* (http://www.aadl.org/catalog) who's classic OPAC evolved into Social OPAC or SOPAC. The adjective "social" is due to the new possibilities of interaction and collaboration offered to the users. The applications which are normally found on social networks outside of libraries have been integrated in the library's catalogue. These applications give the users the possibility the rank, comment, tag and review specific objects in the catalogue. Second example is the *Scriblio* at the Lamson Library, Plymouth University (http://library.plymouth.edu/read/223702) which gives users more options of

browsing and searching and what is more important, an opportunity to mashup the information as it suits them. The mashup of information is, according to some experts, a new, "online" way of thinking and classifying which gives a better overview and lining of information. The creator of this system states the flexibility of the content as the most important feature of Scriblio. On the other hand OCLC initiative to reach their users in their social nework was realized in the Wordcat[1] project and creation of application implemented in the Facebook social environment.

## Negative aspects or what to keep in mind when creating Library 2.0

One of the interesting aspects of the last couple of years is the emergence of several large consolidated information resources (Amazon, iTunes, Google …) which have strongly influenced behaviour and expectation. Unlike these resources, the library resource is very fragmented: it is presented as a range of databases, places, and services. In other words, libraries do not aggregate supply very well. (Dempsey, 2006) Transformation of the information space has put high expectations on the library service. In theory, the implementation of the Web 2.0 system into the library systems is a relatively easy and good idea, however, the best results and consequences are visible only in practical use.

Still, some crucial questions need to be asked: How exactly is this collaborative knowledge to be used in libraries? Should social software be included in the library catalogues? (Pedersen, 2007). One of the problems that the new kind of content brings into the hybrid libraries is certainly the evaluation of content. Several years ago, it was much simpler to compare the digital object with its published version or study the credibility of the author of the web site which contains the content in question. However, nowadays a lot of digital content is digital in origin and the basic assumption of the Web 2.0 technology is that every visitor of a web site is permitted to modify its content. How to evaluate a digital document? There are several potential problems which can arise in the incompatibility of software and formats and in the inadequate design of a web site. The question of education also arises, the education of librarians in using the new technologies and the education of users, which is partially possible through tutorials on the Internet and various projects of educations and seminars. Although the degree of computer literacy has risen dramatically in the last decade, there still are limitations. Not every household is supplied with electricity or owns a computer or has access to the Internet, meaning that there are people who heavily rely on traditional libraries and traditional library management. A vast majority of users has grown accustomed to "older" technologies (Web 1.0) so education is a more that vital issue. For libraries education is crucial,

[1] WorldCat.org is the world's largest network of library content and services. It enables search of the books, music, videos, research articles and digital items (like audiobooks) in numerous collections of libraries around the world.

being that librarians have to be well acquainted with the system in order to help the users and deliver information more efficiently.
Nevertheless, negative aspects of this constantly changing user oriented environment and finding best solutions and create guidelines for its implementation. Kelly, Bevan and Akerman (2008) emphasize main risks in the implementation of web 2.0 technologies are Sustainability - a great risk relying on external commercial associates, especially in the era of an economic crisis, when a lot of companies fail, and the services and data which were entrusted to them become endangered. Preservation – occurs when dealing with online digital objects, where their preservation depends on a rapid and ever changing technology, and there are several organizational, legal, technical and financial problems which occur in such cases. The human factor - when the people participating on some on line tool, like a blog, lose interest, making the site outdated or it's updating is cancelled. *Accessibility issues* - content has to be accessible for users with special needs.

## Conclusion

Libraries are increasingly focusing on their users, as is the Web 2.0 technology; the library managements are aware that the users are the most important link in the chain. The main task of libraries is to deliver good quality information and the new and improved technologies like the Web 2.0 simplify this task by close interaction with the users through which the library can mould its modus operandi. In theory, the implementation of the Web 2.0 system into the library systems is a relatively easy and good idea, however, the best results and consequences are visible only in practical use.

## Reference

Aharony, N. Web 2.0 in U.S. LIS Schools: Are They Missing the Boat? 54, 2008 http://www.ariadne.ac.uk/issue54/aharony/ (11.06.2009.).

Bonariabiancu Library 2.0 meme map.2006 http://www.flickr.com/photos/bonaria/111856839/

Casey, M. E. & Savastinuk, L. C. Library 2.0: Service for the next-generation library. *Library Journal*. September, 2006. http://www.libraryjournal.com/article/CA6365200.html (11.06. 2009).

Habib, Michael C. Toward academic library 2.0: development and application of a library 2.0 methodology. A Master's Paper for the M.S. in L.S degree. November, 2006. 49 pages. http://etd.ils.unc.edu/dspace/bitstream/1901/356/1/michaelhabib.pdf (11.06.2009).

Holmberg, K. Huvila, I. Kronqvist-Berg, M. Widén-Wulff, G. What is Library 2.0 Journal of Documentation. 65, 4, 2009. pp.668 – 681 (20.06.2009).

Kelly, B., Bevan, P., Akerman, R. (et. al.). Library 2.0 : balancing the risks and benefits to maximise the dividends. *Bridging Worlds 2008 conference*, 16-17. (2008) URL: http://opus.bath.ac.uk/12109/2/bridging-worlds-2008-kelly_opus.pdf (11.06.2009).

Macan B. Tehnologije Web 2.0 i njihova primjena u knjižnicama : iskustva knjižnice instituta "Ruđer Bošković"s posebnim osvrtom na njezin blog. *Iz naših knjižnica*. (2009) URL: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=57497 (20.06.2009).

Maness, J. M. Library 2.0 Theory : Web 2.0 and its implications for libraries. // Webology. 2. 2006 URL: http://www.openj-gate.com/Search/SearchResults.aspx?SearchTerm=Library%

202.0%20Theory&Field=All&res=10&type=0&sub=All&update=None&from=-1&to=2010&pr=2. (13.06.2009).

Miller, P. What Happens When We Mash The Library? Ariadne. 50, 2007. http://www.ariadne.ac.uk/issue50/miller/ (20.06.2009).

Miller, P. Web 2.0: Building the New Library. Ariadne. 45, 2005. http://www.ariadne.ac.uk/issue45/miller/(20.06.2009).

Monday, December 04, 2006. Is your OPAC fun? (a manifesto of sorts) Thingology Blog. http://www.librarything.com/thingology/2006/12/is-your-opac-fun-manifesto-of-sorts.php

Oberhelman, D. D. Coming to terms with Web 2.0. // Reference Reviews 7. 2007.

Pattern, Dave. Are you happy with your Opac? // Update magazine. CILIP.6 (10) October 2007 http://www.cilip.org.uk/NR/rdonlyres/5FFA1788-AEC8-4282-BBBF-C23C5E318EFA/0/DavePattern.pdf (01.07.2009)

Pedersen, C. V. Who are the oracles – Is Web 2.0 the fulfilment of our dreams? 2007. URL: http://portal.acm.org/citation.cfm?id=1370663.1370670&coll=GUIDE&dl=GUIDE&CFID=41351456&CFTOKEN=86414025 (11.06.2009).

Technology trends for a 2.0 world. *Library Technology Reports*. 2007. URL: (15.06.2009).

# Catalogue 2.0 and Bibliography 2.0: Collaboratively Created Structured Resource Lists and their Aggregation

Lobel Machala
National and University Library
Hrvatske bratske zajednice 4 p.p. 550, 10 000 Zagreb, Croatia
lmachala@nsk.hr

Krešimir Zauder
National and University Library
Hrvatske bratske zajednice 4 p.p. 550, 10 000 Zagreb, Croatia
kzauder@nsk.hr

**Summary**

*Paper presents how web 2.0 and its tools affect traditional library services and products such as catalogue and bibliography. Using set of social tools and their integration into the library services the user may be given a new means for personalization of traditional library tools. A whole new range of possibilities and usage scenarios open up not only for citations but also for multimedia and hyper textual interactivity with bibliography lists. These tools for bibliographic record management became easy to use but, in the same time, powerful enough to be accepted by library professionals. The paper presents key issues and possibilities surrounding the catalogue 2.0 and bibliography 2.0, points out the differences and similarities between them and suggests the possibility of integration or a creation of a network of these tools.*

**Key words:** Catalogue 2.0, Bibliography 2.0, Library 2.0

## Introduction

"*... looking is easy; finding what you want presents the difficulty!*"
(Hildreth, 1995)

We are aware of definitive changes in our information environment in general and Web in particular as the latter grows from the network of essentially printed media in digital form to a digital medium that greatly differs from traditional ones and which is starting to realise its full potential. Some of the more obvious of these changes became transparent in roughly 2004/05. and were collectively dubbed and popularised as "Web 2.0". At the core of the "2.0" concept, is a

move towards the social web, a move from a read-only medium to a read/write medium which brought a new paradigm into the architecture of a web page: a move from a page as a fixed document (a primarily printed construct) to a page as an application (a primarily digital construct). As we pair these changes with the growth of digitally available information, online identities, greater user participation, user generated content, ever increasing sizes of users' personal collections, new paradigms in web business models and design patterns and so on, it becomes obvious the new medium is having a great impact on information use and management as we understand it. In addition, one should not forget that we are looking at increased permanent connectivity via a plethora of devices that have increasing processing power and storage capabilities (Dempsey, 2009), some of which (i.e. e-paper, some smart phone uses) are specifically tailored to replace paper for reading texts but also in acquiring them.

In library world, the move of the library as an information space to embrace the new paradigm is frequently referred to as "library 2.0", a collage term denoting the libraries using the tools of the social web (blogs, wikis, social networking, tagging and personal collections, commenting and so on) and, more importantly, ideas, mechanisms and trends behind them. Some of the most important components of library 2.0 include increased user participation in the digital information space provided by the library, unique identification and tracking of users' actions and in communication with the library staff and other users online as well as increased librarian participation in online communities.

One of the most important mechanisms used to enable this is personalization, a creation of one's identity online in order to create one's own corner of the digital world. User might then use this online identity for activities such as communication in various contexts but also as a unique reference to a virtual information space unique for this user which might or might not be available to the general public. This space is then frequently used for the storage of digital resources. This way users are able to create own collections of resources in a certain context (e.g. books the user read or a personal bibliography of scientific articles). What should be noted is that users are frequently collecting not the full resources, but references to them. In other words they are structuring their own resource lists and dealing mostly with metadata. This makes sense in the web environment as resource lists also serve as direct access points to resources.

We will return to this in more detail in discussion of possibilities offered by catalogue/bibliography 2.0 as this idea is central to both. It is worth, however, to first explore what is "2.0" in this constructs.

## Catalogue 2.0

"For at least the last 200 years, no service or image has been more closely associated with the library than its catalogue. Whether as a hand-written book, in card format, or as a digital entity, the catalogue both defined and represented the library. ... The whole point of a library was not just assembling the world's

knowledge, but assembling it in a manner which made it relatively easy to find, retrieve, and use." (Kohl, 2007). This made the library catalogue an icon, the foundation of library identity. And, in the printed era, it has been so as the library functions and its position in the society have been devised and maintained by librarians. In today's hybrid world, which shows rapid development towards a "mostly digital" one, library functions and its tools are changing rapidly and thoroughly.

We can differentiate three generations of online library catalogues. The first one resembled the card catalogue. The second, more sophisticated, one appeared in mid 80s and brought many new possibilities but with a high price in usability. In mid 90s an idea dawned for a third generation with increased functionality. Researchers (Hildreth, Borgman, etc.) pointed out their visions how new catalogues should look and feel like. Hildreth (1995) even called these catalogues $E^3$OPAC, a name that represents: enhancing (functionality and usability), expanding (indexing, data records, collection coverage, "full-collection" access tool) and extending (through links, networks, and gateways to additional library collections, information systems and resources).

On the other hand, Borgman (1996) reports on poor functionality and difficult use of catalogues for over 20 years. Every change that was made on library catalogue interfaces was just a scratch on the surface and did not improve core functionality that would truly affect the users' experience. There was increased need for the catalogues to become more user friendly by implementing natural-language searching, browsing, key ranked results, expanding coverage and scope, feedback methods ("more like this" or "not interested"), user-popularity tracking and all kind of different aids (spell corrections, synonyms, automatic term conversion etc.) and so on. However, except of some prototypes, the third generation catalogues did not appear till now.

In the meanwhile, web search engines developed what users missed in library catalogues. Web offered users easier and quicker (if not "better") ways of finding information. Users got accustomed to natural-language searching and type multiple search terms without Boolean operators. They came to expect the same functionalities in library catalogues. Although users still see the catalogue as a trustworthy, well-organized and impressive tool (OCLC, 2005), the advanced features of catalogues are suitable more for the well-qualified librarian than for the average user. Today, a large and growing number of students and scholars routinely bypass library catalogues in favour of other discovery tools, and the catalogue represents a shrinking proportion of the universe of scholarly information (Calhoun, 2006).

While the features of the third generation OPACs are still being planned and sporadically implemented, ideas for some new features appeared out of web 2.0 pioneers such as del.icio.us and Flickr. The idea is that each user can create own collection of resources for future reference and that this collection is available online for the user to use from any device connected to the internet. Each user's

collection is tagged via natural language keywords called tags to facilitate organizational scheme. The collections are interconnected so one user may access others' resources, add other users to friend lists, have his tags suggested from ones others used and so on. These tools came to be known as "social bookmarking" and are beyond this article. For discussion see Hammond et al. (2005); Macgregor, McCulloch (2006); Banek Zorica, Špiranec, Zauder (2007).
This idea as an addition to library catalogue is the most central "2.0" component in "catalogue 2.0". Library may allow users to create profiles in order to personalise the catalogue and services. Using the profile a user may keep track of saved searches, save references to items to own collections, be referred to "recommended" resources, connect with other users and librarians, browse other users' (who may also be librarians) collections and so on. The combined actions of many users will then create a new layer on the catalogue: an access layer of user tags and collections may be used to both search and browse the library resources but also for analysis of users use and organization of resources (i.e. what do they read, what do they save for later reference and which terms do they use for organization). On this layer additional services may be built, such as automatic recommendation of tags, resources, and users with similar interests. Also, unlike other parties offering this kind of features, librarians are in a unique position of already possessing a large amount of quality metadata.
The most problematic part here is the software which needs to either be developed on top of already existent OPAC or implemented in existent commercial products. Breeding (2005) suggests that libraries will continue to use commercial library systems. The open source library systems usually cannot compete with the stability and support of the commercial ones which set high criteria. A possible development is in the direction of open source solutions which will provide additional possibilities to already present systems (Breeding, 2007). A live example is VuFind. Under the slogan "the library OPAC meets web 2.0", VuFind is a tool for creating library resource portals which support much of the features mentioned in this article. It can connect several commercial systems and using them create added value and pave the way for active users' participation in content and metacontent generation as well as for social networking and social bookmarking.

## Bibliography 2.0

"Bibliography is about books", claims R. B. Stokes before the age of the dot.com. As Encyclopaedia Britannica puts it, the tasks for compiling bibliography consist of finding materials on particular subject, describing them item by item and assembling results entries into useful arrangements for reference and study. Today, when "Shift happens" (Fisch, 2008), the context of bibliographing changed rapidly. Changes in ICT made the realm of information access and control impossible to separate from the new technologies. Major reference works appear only online, most electronic publications are searchable in full-

text, publishing world is shrinking through consolidation and mergers, self-publishers are growing in numbers, and the list goes on (Parent, 2007). However, the core of the bibliographic work, gathering, assigning metadata and presenting, remains unchanged.

On the web, such activities may be noticed since the early days as various structured and semi-structured resource lists have been active and range from efforts such as Yahoo! directory or simple lists of "related links" to subject gateways such as Intute. These lists serve bibliographic functions and are often published by people without specialized skills in information organization who follow non-rigorous selection and organization procedures. Still, this publishing activity is fundamentally important since it structures information locally, creating a patchy network of secondary access points for human users but which are also frequently used by large-scale web search engines such as Google as input for their indexing and ranking algorithms.

Producing this kind of resource list is, in its essence, bibliographic work, albeit very often amateur in nature. The new possibilities of collaboration in the digital environment, whether via automatic connection of aforementioned constructs or via user contribution to the "bibliography" has led some researchers to call this type of work "collaborative bibliography" (Hendry, Jenkins, McCarthy, 2006). A large number of these constructs (the "related links" pages of many sites) have been made by individuals for use of others, and differ greatly in style and fullness of description while the structure is most frequently a flat list of resources. It is in this respect that the "2.0" paradigms make a difference from these types of "collaborative bibliographies" to "bibliography 2.0". The "2.0" features are mostly the same as those displayed by catalogue 2.0, and they include a single system in which individual's reference lists are tied to his or hers profile making the list an individual's own personal collection (structured by tags) of references to resources. In this type of approach, the single system may provide the basic description elements, ensuring that it is uniform across users' collections, and automatic metadata harvesting possibilities, ensuring metadata of some quality. The collaborative part in this respect is mostly gained through network of these collections manifesting itself through interconnectedness of user profiles, tags and resources. The main difference from the catalogue 2.0 is that user's collections are not an additional layer on the catalogue of a library but that the service allows references to resources of certain types (e.g. scientific texts) without regard to ownership of those resources, which we think to be, as explained later in the text, to be more characteristic of the digital medium and expect it to be more prominent in future systems.

There are a few products developed and freely usable which would fit the description of a bibliography 2.0 tool. Examples include: LibraryThing (a service to help people have structured lists of books online) and CiteULike or Bibsonomy (services applying the idea of social bookmarking to scholarly texts). Some other products are primarily designed for personal desktop use, but fea-

ture an online component which facilitates remote synchronisation and collaboration between users. They share some similarities with bibliography 2.0 tools in that they support networked personal structured resource lists. An example of this kind of software is Zotero, a popular personal reference management tool.

## Catalogue 2.0 and Bibliography 2.0: Unification possibilities

As more resources are accessible online in their entirety, the users' distinction between catalogue and bibliography begins to blur. The physical location begins to matter less in the digital world as users wish to satisfy their information needs as efficiently (or at least as quickly and effortlessly) as possible. In other words, in a world of ubiquitous openly accessible information sources, "information gateways" play an increasingly more important role to the user than "information warehouses", as the actual ownership of a resource begins to matter less and less to the user providing it's in open access. Although we currently live in a hybrid world, and users need to ask themselves "Can I find this in digital form?" and "Where can I access this?/Who will buy it for me?", given the amount of freely available information on the web and the "googlization" of the information world, it is no wonder a web search engine is frequently a typical user's first stop for solving his or hers information need, whether that engine is a viable choice as a tool for satisfying it or not. Also, it is to be expected that the amount of digitally available resources will only grow as various digitalisation projects bear fruit and as digital-only resources become the norm for some areas of human knowledge and activity and as the reading technology changes (e.g. e-paper) the current problem some have with reading long texts or screen might very well disappear bringing even more convenience to usage of digital resources.

Concerning the user, he/she is more and more frequently walking in the shoes of the information expert as user collections of various information constructs (textual documents, photos, multimedia files, various reference lists such as favourites, reading and listening lists, references to scholarly work and so on) grow to the size of that previously owned only by institutions and select individuals and as new types appear. In the digital world, which is bombarding the user with the amount of accessible information, such collections and listings might be necessary as users need to model their own information environment to suit their needs. User collections of references to items are prominent especially due to the nature of the medium where having a unique reference to a resource is frequently like having a resource itself (at least where time needed to access the resource is concerned; legal, security and preservation issues aside) and where resources' content is changing fluently, not in discrete editions (e.g. a link to a wiki page vs. a locally stored wiki page).

What we are dealing with in both catalogue 2.0 and bibliography 2.0 is a system of networked tools that provides additional value as it creates a layer of access and automatic reasoning (e.g. "recommended resources" feature, gained through

a network of users, tags and resources) which, unlike traditional web search tools, has user decision making at its core and supports additional discovery mechanisms such as serendipity in browsing (not unlike traditionally browsing the shelves but with different placement), identification of users with similar interests (which then function as a recommendation mechanism), detection of popular items and so on. One should bear in mind that while most users are amateurs when it comes to information organization, and may need assistance in the field, they may very well be excellent subject experts in the field from which they are collecting resources.

**Personalised structured resource lists as online library services?**

Both catalogue and bibliography may be viewed as structured resource lists mainly differing in body of written knowledge from which the references are derived, in fullness of description and in organizational approach. A significant difference in the printed world is also that the catalogue, unlike bibliography, may be used to facilitate access to some possessed instance of the resource. However, as these tools move online some of the distinctions begin to blur as both may be used for a direct access to resources. The differences organizational and access possibilities are also diminished as new ways of structuring the lists may be done automatically and the structure of a list may change on demand for a single user. One of the most significant abilities of the digital medium in this respect is to tie these tools together and provide additional layers of possibilities for the users of the system (both library users and librarians) without necessarily modifying the original data.

From the users' perspective, personal information management tools are needed more as the need for modelling one's own information space raises. No single system may currently serve all the user's needs, mainly due to the fact that new types of information constructs users need to manage are still rapidly appearing and it is questionable which should be included in the library systems. However, literature lists in the broad sense are surely one of the needs of the new generation and libraries are in a unique position of already possessing quality metadata and employed information experts which could provide an enviable basis for quality implementation of the "user collections" or "personalised structured resource lists" idea. By mashing up these tools and approaches (catalogue, bibliography, personalised resources lists, automatic metadata harvesting, user networks and so on) a library as an online information space may be a lively place which can attract users, a tool to be used for various user analyses and, in general, a construct which suits the new media.

**Benefits and problems of personalised structured resource lists**

Using these tools, the library can offer the user a service for personalising his library information space or information space in general, depending on the service, the community with an additional access layer to library or other informa-

tion resources and the librarian with the data about users’ use and organization of resources. In addition these tools may make a digital library a lively place with a (hopefully) active community, make the library valued beyond “literature warehouse” and serve to keep the librarian in the position of an information professional both in printed and digital worlds.

There are a few problems in implementation of described services in libraries. These systems require an active development which requires funds and teams able to pull it off both conceptually and technically. Also, collective intelligence can be achieved only when a critical mass of participation is reached. There has to be sufficient number of frequent users using the service to enable the service reach its potential and become valuable (Anderson, 2007). Some researchers even claim that library communities are too small to achieve that critical mass (Wenzler, 2007). In addition, while libraries are always late in use of new technologies, other hand users move very quickly toward another source if they are not instantly gratified. Another problem is motivation of users to participate actively in library catalogue. Do we have users that are willing to help altruistically or will they participate only when they can also fulfill their private incentive. This problem is somewhat alleviated by the fact that this systems may be built in a way that implements collaboration on the level of aggregation, so that the service gains value if even it has a community where each user works for herself only.

To sum up, these services are currently out of the scope for many libraries due to lack of funds, active users and/or teams who can pull it off both conceptually and technically. One of the solutions to the problem is in products which may be implemented as components on top of current library solutions such as VuFind, which are developed by a certain community and then released for use and customisation by others. This may not solve the problem of attaining a critical mass of users, but it will at least lessen the amount of resources needed to implement the service and serve to attract the users with the new possibilities offered and with being in trend.

## Conclusion

A modern day library functioning as storage is but an information island outside the network. This doesn’t mean everything should be online, but it does mean that, in the internet era, library websites (or hubs to libraries) should be lively places as they are the face of the library that more and more users will first see and use. Besides, a social library website presents the librarian with many new possibilities.

A modern day user quite frequently does not need just the hard copy of written knowledge but a reference to a recommended piece to retrieve for which he or she may or may not need the library. Given the amount of written information around, a social approach may be an interesting bibliography tool for dealing with online resources: access, evaluation, personal information tool.

Given the number of resources and current search problems, two points are to be made. First, a help in the selection of quality resources plays an important role as current web search tools are great for known-item retrieval but subject based searches are much more problematic. It is here that the user will need most help, whether via direct tutoring or advice or via resources (subject gateways, bibliographies) either specially prepared by professionals, gained through "collective intelligence" or compiled with the combination of these approaches. Second, as users' collections continue to grow both in number and in size, they offer an important device for the single user, who has a device for keeping his collected information resources in one place, for the community, which can benefit from another type off access layer, and for the librarian, who can benefit from having data about users' use and organization of resources.

In both points, information literacy is of paramount importance and presents one of the possible challenges for libraries in the future: if these types of activities are done in the (digital) library this presents an opportunity for the library to educate the community. Also, given the quality of library metadata and staff already employed on "quality control" in various guises, librarians are uniquely poised to provide metadata of greater quality than most other institutions and to add value to users' collections working as behind-the-scenes information professionals.

The success of catalogue/bibliography 2.0 depends on both parties involved: libraries need to design social tools that are attractive, intuitive and useful, and users need to contribute and use the services provided by the catalogue. In order to realize their true potential in the digital world and, libraries need to bring convenience, trends and quality close together. To put it figuratively: Libraries should be more "tree-focused" rather than "forest-oriented" in developing software tools for users. In the end, well tended trees will produce fine forests.

## References

Anderson, Paul. What is Web 2.0? : ideas, technologies and implications for education. // *JISC Technology and Standards Watch.* February, 2007. URL: http://www.scribd.com/doc/300024/What-is-web-20-Ideas-technologies-and-implications-Paul-Anderson?autodown=pdf (20.06.2009.)

Banek Zorica, M., Spiranec, S., Zauder, K. Collaborative Tagging: Providing User Created Organizational Structure for Web 2.0 // *The Future of Information Sciences : INFuture2007 - Digital Information and Heritage.* Sanja Seljan and Hrvoje Stancic (ed). Zagreb : Department of information sciences, Faculty of humanities and social sciences, 2007, pp 193-202

Bates, Marcia J. An exploratory paradigm for online information retrieval. // *Intelligent information systems for the information society : proceedings of the Sixth International Research Forum in Information Science (IRFIS 6), Frascati, Italy, September 16-18, 1985.* / Brooks, B. C. (ed.). New York: North-Holland (Elsevier), 1985, pp 91-99.

Borgman, Christine L. Why are online catalogs hard to use? : lessons learned from information retrieval studies. // *Journal of the American Society for the Information Science.* vol. 37 (1996) no. 6 ; pp 387-400.

Breeding, Marshall. The new landscape of the automation business. // *Computers in libraries.* vol. 25 (2005) ; pp 40. URL: http://www.librarytechnology.org/ltg-displaytext.pl?RC=11590 (20. 06.2009.)

Breeding, Marshall. The sun sets on Horizon. // *Computers in libraries*. vol. 27 (2007) ; pp 38. URL: http://www.librarytechnology.org/ltg-displaytext.pl?RC=12736 (20.06.2009.)

Calhoun, Karen. The Changing Nature of the Catalog and its Integration with Other Discovery Tools : final report for Library of Congress. March 17, 2006. URL: http://www.loc.gov/catdir/ calhoun-report-final.pdf (20.07.2009.)

Fisch, Carl. Shift Happens : did you knew 3.0. 2008. URL: http://www.youtube.com/watch?v= jpEnFwiqdx8 (29.08.2009.)

Gatenby, Janifer. Today’s information consumer tapping into international library services: making it a reality. November 2006. URL: http://www.oclcpica.org/content/1400/pdf/article-informationconsumer-internationallibraryservices.pdf (20.08.2009.)

Hammond, T. et al. Social Bookmarking Tools (I) : a general review. // *D-Lib Magazine.* vol 11 (2005), no. 4. http://www.dlib.org/dlib/april05/hammond/04hammond.html (07.08.2009.)

Hendry, David G.; Jenkins, J.R.; McCarthy, Joseph F. Collaborative bibliography. // *Information Processing and Management: an International Journal*, vol 42 (2006), no 3.

Hildreth, Charles R. Online catalog design models: are we moving in the right direction? : a report submitted to the Council on library resources in August 1995. New York : Palmer School of Library and Information Science, Long Island University, 1995, URL: http://myweb.cwpost.liu.edu/childret/clr-opac.html (10.08.2009.)

Keen, E. Michael. Designing and testing an interactive ranked retrieval system for professional searchers. // *Journal of Information Science.* vol. 20 (1994) no. 6 ; pp 389-398.

Kohl, David F. Alas, poor catalog, I knew thee well---. // *Journal of academic librarianship*. vol. 33 (2007), no. 3; pp 315-316.

Macgregor, G., McCulloch, E. Collaborative tagging as a knowledge organisation and resource discovery tool. // *Library Review.* vol. 55 (2006), no. 5; pp 291-300.

Merčun, Tanja ; Žumer, Maja. New generation of catalogues for the new generation of users : a comparison of six library catalogues. // *Program: electronic library and information systems*. vol. 42 (2008), no. 3; pp 243-261.

OCLC (2005), Perceptions of libraries and information resources : a report to the OCLC membership. Dublin, OH : OCLC, 2005, URL: http://www.oclc.org/reports/2005perceptions. htm (01.07.2009.)

Parent, Ingrid. The Importance of National Bibliographies in the Digital Age. // 73rd IFLA General Conference and Council, 19-23 August 2007, Durban, South Africa, URL: http://www.ifla.org/iv/ifla73/index.htm (21.06.2008.)

Stokes, Roy B. Bibliography. // *Encyclopedia of Library and Information Science*. 2nd ed. New York : Marcel Dekker, 2003.

Wenzler, John. LibraryThing and the Library Catalog: Adding Collective Intelligence to the OPAC. // A Workshop on Next Generation libraries held in San Francisco on September 7, 2007, URL: http://online.sfsu.edu/~jwenzler/research/LTFL.pdf (05.08.2009.)

# Curricular Approach to School Libraries Education Program

Jasmina Lovrinčević
Department of life-long learning, Faculty of teacher education
Lorenza Jagera 9, Osijek, Croatia
jlovrincevic@net.hr

Dinka Kovačević
Elementary school "Antun Mihanović"
Mihanovićeva 35, Slavonski Brod, Croatia
dinka.kovacevic@sb.htnet.hr

Marija Erl Šafar
City and University Library Osijek
Europska avenija 24, Osijek, Croatia
merlsafar@net.hr

**Summary**

*This paper reports on the model of the curricular approach to teaching within the school library program. In 1980's, school librarians have started to look for the ways and possibilities to educate pupils. Since then, school librarians have used their own programs in direct educational process. Those decisions and also their implementation, were on the level of revolutional changes, but today the library education programs participate in knowledge management within the school curriculum.*
*In that way, school libraries have shown the way and the application models of information resources in education which were forerunners of today's new ways of learning.*
*Methodology: The authors will use comparative analysis and statistic indicators based on perennial practical research.*
*Results: Based on practice, literature reviews and comparative analysis, this paper shows the contribution of the school librarianship to learning which is based on information resources within the National and school curriculum frameworks. The key result is the Librarianship and information management education aiming at expectable progresses, that is, at learning results which are visible and measurable.*

**Key words**: school curriculum, knowledge management, information resources, education, school library

## Introduction

The basics of modern knowledge concepts in general, and new ways of knowledge achievement are marked by the attitude that education has not prepared people for workplace because they haven't achieved enough qualitative knowledge. Those wide spread and often mentioned attitudes seem like consequences of dissatisfaction and unadapted people who were educated and now they work. That (very natural and logical) imbalance is mostly led in tight connection to very fast progress of technology which has brought radical changes in the society. According to that or just to it, there is a need for equally swift and drastic changes in knowledge achievement, even in knowledge contents. According to such conclusions, educational process wants to make novelties by using new strategies, reforms and big steps which were taken before the former preceding steps were fully usable in the educational process.

Following such logic and analysing contents, methods, manners and strategies of knowledge achievement, all innovations (which are always welcome) within the education system could be carried out far more rationally and with less shock for the process participants esp. students and teachers. The Croatian education system has always followed educational needs of society and individuals. The newest in the series of suggestions which should offer more qualitative education is National curriculum frameworks. Keeping in mind that every novelty brings something new, better or/and more qualitative, this paper would also be a valuable contribution. Nevertheless, it can be claimed, with audacity based on experience, that long time ago school librarians found and since then practised curricular approach in education.

## School library and curriculum

In 1980s in some schools there were only outlines of the curricular[1] approach to education, but in that approach was clearly visible. The proof is in existence of school libraries which in early 1980s started the transformation of classic school library into Library-information centres whose main idea was the school library as a part of educational process. By focusing on information in educational work in school library, the project was started with programs of "learning how to learn". Integration of that knowledge opened the door to students and teachers to learning based on information and knowledge resources. The defined part of school activity program, which defines the frameworks, makes the firm basis which every school shapes in its most appropriate way.

[1] The term was at first used in Anglo-Saxon countries and then in Germany, later in other European countries and also in Croatia. Originally, the term "curricular" means *sequence, way* (to reach something), for example, *the way of planning during education process.* It is interpreted differently, for example, as learning content, as substitute for term "syllabus", in the sense of system, educational program, planning, preparing and evaluation (*development curriculum*), in the sense of *curriculum* programming access. That is the reason that there are *curriculum theory* and *curriculum practice.* Source: Antić, Stanko (2000.)

That is the reason why it is so important that the annual plan and activity program are adopted, planned and made by those who are going to implement it. Except considering the main frameworks of the subjects and time-table for regular classes, it is also possible to add and change all other activities according to capabilities of school or needs of the students. A team of experts (professionals like educator, psychologist, defectologist and school librarian, also parents, if necessary) chooses activities and create time-table according to activities which are going to be presented during the school year. Flexibility of the annual plan and school activities program reflects its diversity every year. Changes, which are clearly visible in the program, help us to follow real intentions of a school to improve educational process. Breadth and diversity during planning and programming give a picture of employees and school profile. Annual plan should reflect school's identity, which means that it should only be partially similar to other plans. Most of the annual plan has to be different from every school to school. That would indicate that schools do care for modern approach to education and their needs and preferences – a student in the middle of the education, the quality of learning as a need, key competences for lifelong learning as a process and learning outcome as a final goal.[2]

The real school curriculum is a reflection of such sequence. In Croatian school libraries such learning approach is in practice since 1980s. The foundations of the National curriculum frameworks are compatible with those described within the organisation in the school library.

## Towards curriculum guidelines

The Croatian school librarians tabled *Library education of students* program which includes two fields: *information literacy and promotion of reading.* The program became a part of the National plan and program for primary school (more important than other subjects, not any more within the Croatian language classes, section media culture) and is implemented in all Croatian primary schools. Considering other subjects, the program is divided in three parts: *topic, key terms and educational achievements.* The program baselines are not some-

"In the recent times, the term "curriculum" has been used very often. This term is not uniquely defined except the consensus that it refers to process and it is wider than syllabus. If we talk about *curricular or cross-curricular approach* in a specific subject, then it refers to content interconnection of specfic (and all) subjects – "processing" of specific contents (or aspects) of all subjects, that is, full approach to realization of National plan and program, with other school subjects and activities." (Kurikularni pristup promjenama u gimnaziji. Zagreb: Ministarstvo prosvjete i športa Republike Hrvatske, 2003.)

"The term *curricular* is nothing else but didactics which was developed in Anglo-Saxon speaking area. Curricular cycle is targetting at objectives, procedures, methods, learning strategies, evaluations, and all that are didactic issues" (Jurić, "Didaktika u kurikularnom krugu". Školske novine br.13. (2006.)

[2] According to D. Kovačević; J. Lasić-Lazić; J. Lovrinčević: "Školska knjižnica – korak dalje".(2004.)

thing that students learn during one school year, but their knowledge, skills and competences were determined across different educational levels and in different fields and learning contents. The main task is to create an active reader and an user of different information resources who is capable of independent life-long learning. By implementing the new National plan and program, the Library education of students program was adjusted to other subjects. The following example shows how it works in practice:

Table 1: Library education of students and Croatian language classes

| | |
|---|---|
| School library: Information literacy<br>Topic: Reference collection<br>*Key terms:* encyclopaedia, thesaurus, dictionary, spelling book, atlas<br>*Educational achievements:* to comprehend the reference collection and how to use it aiming at expanding knowledge; to recognize reference collection on different media; to know how to find, choose and apply information | Croatian language – Media culture<br>Topic: school library – using of dictionary and spelling book<br>*Key terms:* dictionary, spelling<br>*Suggestions for methodic interpretation:* to learn about different editions of dictionaries and spelling books that are available in the library; to encourage students to express linguistic doubts and to solve them by using dictionaries and spelling books |

Reference collection is used in the third (level of recognition) and fourth grade (beginning of learning in how to research information in specific field). The example shows that it is possible to plan and implement classes in school library through cooperation between the Croatian language teacher and school librarian who have a common goal: to present to students important reference books (dictionary and spelling) and at the same time to check acquired knowledge (educational achievements) by using a specific text, which means to encourage and develop reading skills which are the criteria to enter the world of information research. Without qualitative professional cooperation between teacher, school librarian and co-workers the Program can not be implemented, because it is not just a teacher's/school librarian's tool, but also a student's tool in the process of independent learning and information research.

## Library-information education program

School library is accessible to students in the period of their most intensive knowledge acquisition and learning, and in the period of their attitude and behaviour development which is important for their future lives. The school library is not only support to education but also supports personal creative progress of every student who, in that way, develops permanent need for life-long learning. Partnership and team work enable the school library and librarian to participate actively in accomplishing interdisciplinary and multimedia approach to classes which emphasize methods and readiness to use knowledge and its testing and refining. In such conditions the school library enables realizing individual and collective educational, information, cultural and social needs of its

patrons, especially students. Information theory and practice of teaching program in the school library reflect legislative guidelines (IFLA/ UNESCO School Library Manifesto: The school library in teaching and learning (2000)) which emphasize the basic tasks of the school library, which are important for basic literacy, computer- and information literacy development, and professional ethics (the school librarian as information professional and educator). For the first time, in the history of Croatian school librarianship, the implementation of a unique teaching program in school libraries enables vertical educational connection and it can be expected that this vertical connection is going to be realized on all educational levels. School library will be indirectly included in school curriculum through the Library-information education modulus and directly, within interconnected subjects. The Library-information education will be realized within three fields: reading, information literacy, and cultural and public activity from the 1st primary school grade to the 4th high school grade, in five educational cycles:

Primary school: I. ($1^{st}$-$4^{th}$ grade); II. ($5^{th}$-$6^{th}$ grade); III. ($7^{th}$-$8^{th}$ grade)
High school: IV. ($1^{st}$-$2^{nd}$ grade); V. ($3^{rd}$-$4^{th}$ grade)

**Reading**

In search for new approaches to learning in library, librarians apply reading & understanding methods which should be connected to the classes' contents. Instructions to subject comprehension are divided into five learning methods: connectivity, experience, application, cooperation and transfer to new contents. Connectivity is related to learning in the context of life experience which is the pith of learning (research, retrieval and finding). At the beginning of education, students are encouraged to reading, improving their reading skills and reading habits (retelling, writing, dramatization, singing, drawing). It is important to strengthen student's self-esteem during solving given tasks and finding library resources. By accomplishing those activities, students start to understand importance of reading and learning in everyday life because they enrich their vocabulary and develop written and verbal communication.

**Information literacy**

The best way to develop information literacy is the team work of teachers and other professionals in school library and methodic planning based on existing students' skills and needs, with already known modules of good practice which trace developmental school plan. In this context, it is important to promote multidisciplinary field known as human information behaviour. On the information field, it includes process of information recognition, searching, evaluation and use of information. Those are the reasons to research information behaviour. Continuous professional education and knowledge acquisition, and life-long learning impact information behaviour the most.

### Cultural and public activity

One of the important components in school librarians' work are cultural and public activities which actualize important events in school or in its surroundings (important anniversaries of events and persons, promotion of cultural events, development of ecological consciousness etc.). Contents of cultural and public activity are components of annual plan and program of school library and librarians, and also components of educational process of the school in general. They are also a stimulus to conduct school projects on specific topics which are initiated and coordinated by a school librarian in cooperation with teachers. Cooperation between the school librarian and cultural institutions – libraries, museums, theatres – aims at education of an individual with developed cultural needs and habits.

## Expected achievements in library-information education

According to the form of educational cycles, which was given in the draft National curriculum framework for pre-school and compulsory education in primary and high school, the workgroup for Library-information education program proposed expected achievements of students, already mentioned in three fields: reading, information literacy, and cultural and public activity. Those expected achievements, as a final learning result, are the starting point of topics and contents of the future Library-information education program, according to working arrays and educational cycles. The following examples show how it looks in practice:

EDUCATIONAL CYCLE (1st-4th grade, primary school)

*Reading*

Students:

- are familiar with the school library and book lending rules
- independently choose books
- participate in different activities which stimulate reading and developing reading culture (retelling, writing, dramatization, singing, drawing)
- understand importance of reading in their lives, compare situations and characters from literary works to everyday life; they communicate with literary text on the level of recognition
- know what is a children's magazine; recognize columns which are educational and those which are for fun
- independently choose and read books and children's magazines in order to develop reading skills and to achieve reading habits
- reading enriches their vocabulary and develops written and verbal communication, they understand and react to simple and complex questions

- understand the value of creative achievements
- accept library as a place for learning, talking and having fun

*Information literacy*

Students:

- know the difference between book and non-book materials
- recognize children's magazine as a part of the library collection and as information resource; they use indexes to find wanted contents
- know the difference between literary-artistic texts and popular-scientific ones
- are familiar with book parts and know how to find specific information in a book (title page, foreword, afterword, note about author)
- are familiar with different information resources in library and use them according to their age
- use thesaurus and encyclopaedia in order to expand their knowledge and to develop information skills; they know how to search the library collection by using alphabet, indexes and marginal words
- know the use of dictionary and spelling book; they are familiar with and respect the spelling standards of the Croatian language

*Cultural and public activity*

Students:

- are familiar with children's department of the city/public library and as their patrons they use the library's services in order to learn and to spend there their quality time
- are familiar with and they visit cultural institutions
- are familiar with children's rights, they respect them and have positive attitude towards themselves and others
- accept differences

I. EDUCATIONAL CYCLE

(1st -2nd grade – high school and triennial vocational and art school)

*Reading*

Students:

- have developed reading interests and skills
- have developed the need to be familiar with popular-scientific and professional journals
- are able to read and understand popular-scientific and professional texts
- know how to connect personal experience, pre-knowledge and new information from different information resources.

*Information literacy*

Students:

- are able to find, evaluate and use different information resources
- independently form queries in order to find information
- know how to search library catalogues and Internet resources
- are able to find information in reference and professional literature
- are used to writing notes
- know how to list used bibliographic references
- understand and critically review information, they use them in the right way and creatively
- understand professional terminology
- are familiar with cooperative and research learning

*Cultural and public activity*

Students:

- are familiar with and respect local and the Croatian cultural and natural heritage
- have a positive attitude towards cultural and natural heritage of other people
- have a need to participate in different cultural events
- have a positive attitude towards nature and environmental protection
- accept differences and other people’s opinions and attitudes.

## Conclusion

The main objective of every school and its library is learning. The quality of the school library program is tightly connected to the quality of education. The old model “one teacher in one class” is now history. Information society calls for new models of teams of teachers who have different qualifications and competences to create a new cooperative environment. Workgroup plans everything, aiming at developing research abilities and students’ comprehension of school subjects, but also at encouraging cooperative learning, reading with understanding and developing of social skills. Presented programs and ways of working and learning in and with the school library speak for themselves about theoretical ingenuity and practical competences of the school library in achieving high standards in learning and teaching.

## References

Kovačević, D., Lasić-Lazić, J., Lovrinčević, J. 2004. *Školska knjižnica –korak dalje*. Zagreb: Filozofski fakultet, Zavod za informacijske studije Odsjeka za informacijske znanosti: Altagama.

Plan razvoja sustava odgoja i obrazovanja 2005. – 2010. Zagreb: Ministarstvo znanosti, obrazovanja i športa Republike Hrvatske (2005). Dostupno na URL: http://public.mzos.hr/Default .aspx?sec=2420. (28.06.2006)

Prijedlog Nacionalnog okvirnog kurikuluma za predškolski odgoj i opće obvezno obrazovanje u osnovnoj i srednjoj školi http://public.mzos.hr/Default.aspx?sec=2685

Shaw, Marie Keen. Block Scheduling and its Impact on the School Library Media Center. Greenwood Professional Guides in School Librarionship, Greenwood Press. Westport, Connecticut. London, 1999.

Znanjem do znanja: prilog metodici rada školskog knjižničara. 2005. Jasmina Lovrinčević, Dinka Kovačević, Jadranka Lasić Lazić, Mihaela Banek Zorica. – Zagreb: Zavod za informacijske studije Odsjeka za informacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu.

# VIRTUAL ENVIRONMENT IN EDUCATION

Original scientific paper

# Virtual Learning Spaces: Example of International Collaboration

Jadranka Lasić-Lazić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
jlazic@ffzg.hr

Mihaela Banek Zorica
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mbanek@ffzg.hr

Senada Dizdar
Department of Comparative literature and Librarianship,
Faculty of Philosophy, University of Sarajevo
Franje Rackog 1, 71000 Sarajevo
senadadizdar@gmail.com

Jasmin Klindžić
Department of informatics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
jklindzic@ffzg.hr

**Summary**

*Paper presents results of international university collaboration in developing virtual learning environment at the Faculty of Philosophy University of Sarajevo. Based on the success of the Omega system implemented at Faculty of Humanities and Social Sciences University of Zagreb collaboration in creating a similar virtual learning environment on Faculty of Philosophy University of Sarajevo has started. Results of the implementation of BISER e-learning system are shown. Special emphasis is put on the initial usage statistics as well as future development plans.*

**Key words:** Virtual learning environment, international collaboration, OMEGA system, BISER system

## Introduction

Development and advancement of information-communication technology has influenced all aspects of our everyday lives. One of the major changes happened in the communication process and information retrieval. This led to the emergence of two different user groups. On one side are users born in this digital environment based on internet and technology *colloquially millenials* or *digital natives*. The other group of users are *digital immigrants* or *baby-boomers*, born in the *old* times i.e. before the emergence of internet, who have gradually adopted developed technology. In today's educational environment, these two groups coexist and collaborate as teachers and students. Today's students (Prensky, 2001) represent the first generations to grow up with this new technology. They have spent their entire lives surrounded by and using computers, videogames, digital music players, video cams, cell phones, and all the other toys and tools of the digital age. This has forced educational institutions to change their policy and direct their research and action into creating learning spaces that will respond to the needs of today's users. Britain and Oleg (2004) in the excessive research two main motivators for ICT implementation in educational institutions:

- to enhance the quality of teaching and learning by allowing teachers to use pedagogies that are not possible with large numbers in a face to face environment;
- to manage the delivery and administration of programmes of learning as well as groups of students through an electronic (on-line) medium.

Research projects together with the advancement of ICT created a very strong e-learning community oriented towards improvement of educational process so that it corresponds the needs of gradually changing users. Although in the early stages the primacy of commercial solutions was notably present, developments in the open source sector and its affordability provided an opportunity for much larger number of institutions to get involved in research and contribute. Strengthening of the open source movement opened possibilities for wider collaboration and created an opportunity for a wider number of users and collaborators to get involved in the process creating quality learning spaces. Spread of affordable technological solutions enabled institutions with less financial backup and opportunities to get involved in the process of creating and further developing virtual learning environment. Furthermore, possibility of self-management system, as opposed to commercial service providers on which universities depended, attracted more universities to switch to open source solutions.

Nevertheless, developments in the information landscape and variety of new media and communication channels have created an abundance of digital materials impacting all aspects of educational process (teachers, students, library, administration etc.). Change of the library services, need for creation of repositories as collections of digital materials, organization of digital educational materials, user support, creation of various e-services like e-portfolios, etc. forced

universities to integrate these services and their virtual representations in form of virtual learning environments.
Creation of virtual learning environments as designed information spaces supporting the educational process was therefore the next logical step. Virtual learning environments present settings that encompass different static and interactive tools used in teaching and learning like: communication tools such as email, forums and chat rooms; collaboration tools such as intranets, electronic diaries and calendars; tools to create online content and courses; online assessment and marking; integration with management information systems; controlled access to curriculum resources; student access to content and communications. Research in usage of VLEs and its efficiency shows that while using VLEs student experience is being enhanced through improved delivery of teaching materials and course announcements, improved access to learning resources and better communication. (Britain and Oleg, 2004) Integration of various scattered services, tools and technical solutions under one comprehensive setting, such as virtual learning environment is fundamental for a successful and up-to date universities.

## OMEGA – an example of best practice

Successful implementation of the Omega e-learning system has started back in September 2002 under the project OIZEOO[1] funded by the Croatian Ministry of Science, Education and Sport. One of the project objectives was to investigate, evaluate and implement the best suited technical solution for creating virtual learning environment at the Faculty of Humanities and Social Sciences (FHSS) University of Zagreb. Our goal was not simply to create a self-sustained e-learning solution, but to implement or develop virtual learning environment that supports blended learning. Intention was to create a virtual representation which will provide support to our teaching staff and students in enabling blended learning. Goal was not to create purely virtual campus for distant learning but a solution that will support educational process in facilitating a shift from teacher-centered to student-centered learning. On the other hand it was to help us in dealing with the difficulties of overloaded schedule (problem of time and space).
After testing and evaluating both open source and commercial solutions decision fell on open source course management system Moodle which was then implemented in 2004. The fact that MOODLE is a free, easy-to-use system (i.e. everyone with basic computer literacy can easily use it) with simple and understandable interface was the main reason for implementing it. Furthermore, its large variety of modules and the ability of implementing new modules; SCORM

[1] OIZEOO - Organization of Information and Knowledge in the Electronic Learning Environment (Organizacija informacija i znanja u elektroničkom obrazovnom okruženju - http://infoz.ffzg.hr/oizeoo)

compliance; authentication via LDAP or IMAP (e- mail) user accounts issued by the Faculty; were a major advantage in choosing which product to implement in a large heterogeneous institution like ours is, baring in mind the large number of staff not very familiarized with the idea of using ICT in their teaching. After translation and implementation OMEGA e-learning environment was created.

First stage of our implementation was to test the system on several courses held at the Department of Information Sciences. After positive user evaluation OMEGA (Figure 1.) was ready to extend on other Department, integrate more courses and users (teacher and students) and start functioning as FHSS virtual learning environment. In October 2006, there were 20 departments offering their courses through OMEGA with total number of 180 courses, more than 170 professors and TAs working on their courses and more than 6 GB of produced learning and teaching materials. Now in August 2009 numbers have grown and we have 24 departments with total number of 5000 users out of which 245 are professors and TAs working on 597 courses and more than 63GB of produced learning and teaching materials. As one can see, the number of courses, data, professors and students has largely increased and we expect them to grow even more until we have all the courses and students signed in this VLE.

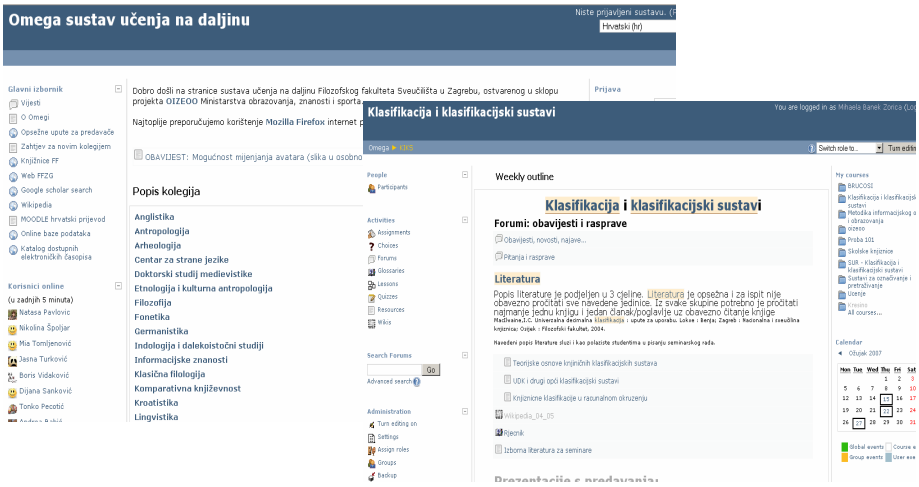



Figure 1. OMEGA VLE

Currently, we are using Moodle v.1.9 with the Croatian language interface and some other minor modifications and customizations, based mainly on end users’ requests or needs. Comparing Moodle module usage statistics (Table 1.) over the three year period constant progression is clearly seen.

Table 1. Omega module statistic

| Module | 2006 | 2009 |
|---|---|---|
| *Resource module*<br>(text, html, PDF ... etc.) | 3,071 | 12,760 |
| *Forum module*[2] | 436 | 905 |
| *Assignment module*<br>(students can send them online, offline or via document upload, making the paper submission and grading both easy and transparent) | 298 | 1,131 |
| *Quiz module*<br>(online quizzes with numerous capabilities - random questions, timed sessions, question databases, variety of question types, access control etc.) | 212 | 358 |

Apart from this four, now widely used modules progression in usage of other modules has also happened. Noticeable growth has happened with the following modules:

- question module –consists of only one question, usually used for immediate response in class - 201 instances
- dictionary module – can be done by teacher but also collaboratively by students and teacher - 93 instances
- lesson module – set of connected html pages with a question at the end of the path and possibility of choosing the path of learning -52 instances
- wiki module – 47 instances

Thanks to the modular and open-source nature of Moodle, the teachers can use only those modules that they really need at the present time, and if they need some additional type of resource or activity, it can be easily implemented (sometimes in the matter of hours).

We can say that we have the largest single Faculty virtual learning environment at University of Zagreb and in Croatia. Strong Moodle movement that was initiated at our Faculty influenced other higher education institutions to implement Moodle. Currently, the largest Faculties in Croatia are favouring Moodle and the total number has grown up to 25 higher education institutions (faculties and colleges). Still new implementations are emerging. One of the factors of such an outspread is the fact that our Croatian language package was made available on the Moodle site as our act of giving back something to the e-learning community.

[2] Since there can be more than one forum on the course, and the access to the forums can be customized easily, the teachers are using them both for communication with student body and among themselves.

## BISER implementation

Success of Omega e-learning system was recognized by the Faculty of Philosophy in Sarajevo and after several meetings collaboration on creation of similar VLE was initiated. The main objective was to improve the quality of teaching and create a virtual learning environment. As well as many European universities, so has University of Sarajevo passed through curricular reform and its adjustment with European Credit Transfer System. It was perceived that the successful implementation and regular usage of VLE could improve the performance of high-quality curricular programs.

Implementation started in February 2009 and Department of Comparative literature and librarianship courses were chosen as a test bed. Due to the lack of high quality ICT infrastructure at the Faculty we have agreed to support the whole system through FHSS server located in Zagreb. New learning environment was named BISER (Bibliotečki Sarajevski Elektronski Repozitorij) meaning *Sarajevo librarian electronic repository* (Figure 2).


Figure 2. BISER virtual learning environment

The first phase of the project was to install the software on server and create test courses that would serve as an exemplaratory for both teachers and students. After the introductory presentation for both teachers and student, due to the more complicated interface and numerous possibilities that the system offers, we have organized separated workshops for teachers. Next step was organization of several courses to be held in spring semester. Part of the courses was or-

ganized in a similar mode as the ones organized at our institutions while the other part was left for teachers to explore and improve on their own pace.

## Biser usage statistics

Although this is relatively new system and there is a need to spawn its usage on the other Departments, with the ultimate goal of being accepted by the whole Faculty, initial positive evaluation of the system can be seen from the statistics. Currently, in August 2009, there are 142 student users, 19 teachers and TAs organizing 28 courses. Majority of used modules are: resource module - 166 instances: forum module - 43 instances; assignment module - 22 instances; and chat/dictionary/choice module - 4 instances. Usage of the modules is typical for such an early phase when users are still experimenting with the system but we expect a growth in all modules and user during the next academic year.
Interviews with the student showed satisfaction with usage of the system and improvement in their educational space. Usage statistic (Figure 3.) shows that students have regularly accessed the system during the semester and until the end of exam period (July 2009). Numbers represent a good indicator for the Biser success but there is still much work to do.

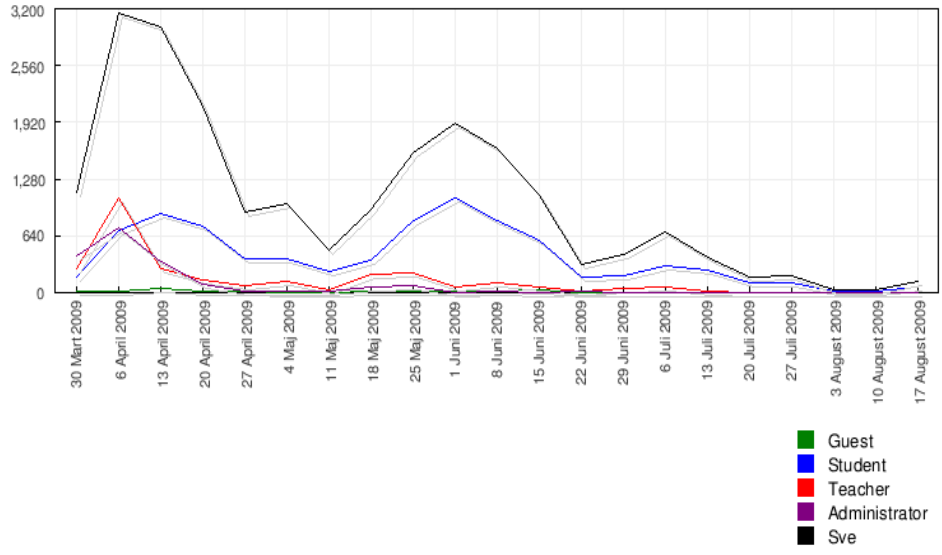



Figure 3. Usage statistics

Current total number of data is 700mb. At the first glance it might seem small, but one should bare in mind that policy regarding upload of data is taken form FHSS's Omega. Recommendation for teacher is to keep in mind the duplication of data i.e. not to put the actual files on system if they exist on web or in library

and could be accessed by students, but rather to link them. It is necessary to emphasize that the system should not be overwhelmed by the excessive data that can be found on the web or in the library. Successful VLE should integrate all supporting elements of the educational process i.e. library and educational materials. Therefore, second part of the project will be to create a digital repository of learning materials based and managed by the faculty library. Collaboration with faculty library will be necessary for creating and managing sustainable digital repository of learning materials.

## Conclusion

The technology has brought changes in society which started the transformation of the educational environment and sets of educational reforms that European universities are undergoing. Implementation of new technology in the traditional environment resulted in creation of virtual learning environments and change of modes of teaching from traditional teacher-centred to a student-oriented research based learning enhanced through the usage of technology.

Implementation of ICT in various discipline related environments does not result in same success. Baring in mind that within the humanities and social sciences the benefits of e-learning applications show lower rates of acceptance, which is mainly caused by prevailing traditions of teaching the respective subject, and the considerable pressure from within the university, particularly from part of the staff, to preserve the status quo, the popularity of OMEGA as a faculty-wide VLE seems surprising. The reason for the positive attitudes could be explained through the phenomenon of blended learning, which does not require a complete break-up with traditional learning but the complementary application of the traditional and technological paradigm.

Driven by the positive results of the implementation of technology at Faculty of Humanities and Social Sciences, University of Zagreb a project of creating VLE Faculty of Philosophy in Sarajevo has started. Initial results, implementation on only one Department, show user satisfaction presenting a good foundation for expanding BISER on the other Departments and creating a unique Faculty-wide system.

## References

BISER: Bibliotečki Sarajevski Elektronski Repozitorij http://biser.ffzg.hr/ (20.8.2009)

Britain, S. Liber, O. A Framework for the Pedagogical Evaluation of Virtual Learning Environments. http://www.cetis.ac.uk/members/pedagogy/files/4thMeet_framework/VLEfullReport (20.8.2009)

Lasić-Lazić, J. Banek Zorica, M. Špiranec, S. Klindžić, J. Using Open source Learning Management System for educating information professionals // *Current Developments in Technology-Assisted Education* / Mendez-Vilas,Antonio.(ur.). Badajoz : Formatex, 2006. pp. 88-92

Moodle - A Free, Open Source Course Management System for Online Learning. http://moodle.org/ (10.08.2009.)

OIZEOO - Organizacija informacija i znanja u elektroničkom obrazovnom okruženju http://infoz.ffzg.hr/oizeoo (20.8.2009)

OMEGA sustav učenja na daljinu http:/omega.ffzg.hr (10.08.2009.)

Prensky, M. Digital Natives, Digital Immigrants. On the Horizon. MCB University Press, Vol. 9 No. 5, October 2001.

Wyles, R. Evaluation of Learning Management System software: Part II of LMS Evaluation. March 2004. https://eduforge.org/docman/view.php/7/17/Evaluation%20of%20LMS%20-%20Part%20II.pdf (20.8.2009)

Original scientific paper

# Organizational Design Strategies in Higher Educational Institutions in Accordance with Electronic Learning and Teaching Environment

Mislav Balković, Danijel Kučak
University College for Applied Computer Engineering
Maksimirska 58a, Zagreb, Croatia
mislav.balkovic@racunarstvo.hr

**Summary**

*New approaches to learning and teaching, introduced by electronic environment are present in higher educational institutions through the world. Still many institutions are either purposely unaware of them or they do not have organizational infrastructure and willingness to fully adopt them. A rigid internal governance structures that is strongly influenced by academic council in most institutions tends to retain status quo. At the same time Laissez-faire approach to development of training materials and adoption of electronic teaching methods tends to increase expenses with no serious results and no dissemination of knowledge that is gathered from those, in most cases, isolated and enthusiastic projects. Advance in technological development is for many institutions just insignificant external factor that only in a way influences their internal organizational change. Still, in the course of this article it will be shown that only strong willingness to adopt full organizational change in the fields of: governance, organizational model, funding and internal culture change, through new vision statement and detailed strategic plan can fully prepare institution for electronic teaching and learning environment. This assertion will be elaborated both from organizational as well as from financial aspect and some best practices solutions and recommendation will be drafted in order to propose concrete change in governance structure, organizational model, funding and internal culture.*

**Key words:** organizational design, e-learning, distance learning, teaching supported by technology

## Introduction

Advance in technological development is for many institutions just insignificant external factor that in a way influences their internal organizational change. There are also institutions that are directed towards market and business excellence and at the point when that kind of institution fully recognizes positive ef-

fects of technology they can become strong internal factors that models many key organizational segments such as:

- Employees and a way they are working
- Location and geographical accessibility of resources
- Product (curriculum)
- Tasks and tools used to solve them

In the course of this work analysis of existent referent experiences in organizational change due to introduction of electronic teaching environment in higher educational institutions will be discussed. Also, strategic guidelines will be drafted in order to achieve full implementation of education supported by technology, with decreased spending and organizational problems in the way. As higher educational institutions should be focused on customer (student) rather than on the product (curriculum, materials, technology) the emphasis will be also put on the influence of technology on the learning itself.

## Influence of technology on learning

Aforesaid mentioned changes influences significantly the way teaching is conducted in electronic environment. Opposed to teaching that still dominates in most of higher educational institutions, that was in methodological structure imposed in the 19$^{th}$ century by Thomas Huxly and Von Humbolt, new approaches introduced significant comparative advantages as follows;

### Increase in learning quality

Quality of learning can be observed from the learning outcomes as well as from the teaching process standpoint. Both standpoints show quality increase. Learning outcomes are positively influenced by the opportunity to use e-learning content and to reuse stored teacher's presentations from the repository of audio/video materials that is available to student at any time and for as long as needed. Quality of teaching process at the other hand can be significantly improved by development and use of interactive video simulations in order to familiarize students with processes and equipment otherwise unavailable to them or by involvement of top of the class international trainers and experts via web conference infrastructure.

### Increase in training availability

Today training and learning are indeed lifelong processes that enable individuals to stay current on the labor market. To be able to support that demand, higher educational institutions are in many ways involved in LLL career programs as well as in educational programs for students that are employed. New electronic environment such as e-learning courseware, on demand audio/video materials that were filmed during the lectures as well as on-line interactive

simulations and possibility to access real hardware and software solutions in training purposes via Internet, increases training availability.

**Increase in cost effectiveness of education**

Investments in electronic training infrastructure and content in order to build quality higher education supported by technology are significant. It is also important to stress out that costs of that kind of training is increased compared to classical training. Still, investments in e-learning and web conferencing infrastructure can increase income due to increases in number of international and distant students to whom the education is now available. It can also cause savings if huge number of students can be trained fully or partly using the technology instead of arranging more classrooms and trainers. Also, possibility to reuse already prepared e-materials by inferior students instead of redoing classroom training for them can cause savings.

## Present organizational structures, situation and problems

In higher educational institutions two approaches among teachers can be witnessed; there is a group of enthusiasts who supports introduction of new technologies as a support tools for training and learning. At the other hand resistance to change is also present and is often augmented in the following way: "We have to use technology only for the blind belief that technology is good for us. If we do not accept to use technology in teaching, students will regard us as old-fashioned and will lose their credence in us." (Bates, 2004) Aforesaid opposite attitudes point to the complexity of a problem as well as to the scale of influence technology has on teaching and learning.

**Organizational structures in higher education**

Organizational structure and governance model in most of the higher educational institutions hasn't changed for centuries and is therefore today almost completely inadequate not only to impose new technologies in teaching but also to underrun any significant enough change. According to (Žugaj, Schatten, 2005) it can best be described as hybrid hierarchical organizational structure that is almost common to old industrial organizations like one imposed in 19$^{th}$ century by Henry Ford. That categorization can be advocated for number of shared elements;

- Work division (different tasks are divided to different groups of workers; teachers do teaching, accountants do accounting, …)
- Hierarchical governance model (depending on the institution's size governance and hence responsibility is cascaded from Rector or Dean down to Faculties or departments)
- Organizational units are formed according to business functions (accounting unit, maintenance unit, …)

- Standardization and high level of bureaucracy is imposed (starting from admission procedure to collegiums definitions all procedures are standardized)
- Economy of scale (huge investments are justified to prepare collegiums in a way that will later reduce delivery costs due to number of enrolled students)

Opposed to pure industrial or post industrial organizational structures, higher educational institutions have specificities which makes them even more rigid and inappropriate for change. They are as fallows;

- Method that is used to develop teaching process and train new teachers is similar as in preindustrial, agricultural society, where farmer was in charge of whole process from sowing to sales of crops. Thus in higher educational institutions teacher is in charge of whole collegium from the design of curriculum up to teaching materials and delivery. The same situation can be witnessed in selection and training of new teachers. As in aforesaid agricultural society teacher in most cases alone picks and trains his successor.
- Governance in higher educational institutions is specific although it slightly differs due to institution's size. In large institutions such as universities Rector is formally in charge but in most cases he or she only controls overall university's budget and some development projects. University components (faculties) are highly independent and are controlled by Deans and most of their decisions have to be supported by academic council of each faculty. That way it is almost impossible to gather all faculties to the same vision and priorities. In such heterogenic and uncoordinated system any decision and specially one to make significant change is either blocked or fades out. In smaller institutions such as colleges and independent Faculties, Dean controls the institution and the budget but academic council is still tough to persuade in support of change.
- Organizational culture in academic institutions as well as value system is specific. Academic independence is almost a dogma which models teacher's mindset in a way that most of the them finds themselves being independent of any but scientific obligations. Even teaching for some of them poses burden since their career path and promotions is far less influenced by teaching quality than by number and quality of scientific work they publish.

In such culture, resistance to change and new technology introduction is to be expected for many reasons. Firstly teachers feel independent to develop their collegiums as they did in the past and as they learned from their mentors. Furthermore being independent and being empowered to influence all strategic institutional decisions through academic council gives solid grounds to support status quo.

**Present situation**

It can be witnessed that in some institutions there are projects, mostly led by enthusiastic teachers in order to implement some electronic teaching materials and infrastructure. That kind of projects are in most cases either done solely by teachers and his or hers students or are financed by department and in some rare cases faculty or university. Aforesaid Laissez-faire approach dominates as most commonly found model of introducing electronic teaching environment when e-content is concerned. At the other hand computer and networking infrastructure that is mostly used for institution's business functions (collaboration, document sharing …) are much more developed. It is not uncommon for present institutions to even invest in hardware and software that can be used as electronic teaching infrastructure (i.e. "smart" whiteboards, web conferencing infrastructure) but that infrastructure is not in use or is in use only by a small portion of teachers.

**Problems**

Aforesaid status is formally supported by not having vision and strategic development plan on the institutions level that recognizes need to change and modernize teaching process. Furthermore, in always insufficient funding environment where more trained students for less money is more emphasized each year, courage and envision to start changes can rarely be founded. Laissez-faire approach used to introduce electronic teaching environment is unfortunately in support of that because it increases expenses with little or no results at all. Most of the projects, even if successful, are not promoted and there is no dissemination of knowledge gathered during the project on the institution level. Even bigger problem is poor quality of produced content since one is build by amateurs (teachers and/or students not trained as content developers) and therefore in most cases poorly accepted by students.

## Proposed solutions

In order to fully introduce teaching and learning supported by technology institution have to undertake significant change in organizational model, governance structure, organization's culture and funding. Changes of that proportion can be executed only in time and supported by majority of staff. They are much easier to conduct in smaller institutions than in huge universities. To start changes governance has to:

- Share a vision of new era in teaching and learning and also of institution's positioning, primarily to the teachers.
- Prepare vision statement on the level of each department that will represent direction in which teachers and staff from that department see development of the department. After that, first phase, component's (faculty) governance has to gather team that will be populated by representatives of each department in order to prepare faculties' vision statement. Fi-

nally, vision statement on the university level is formulated by a team populated by Deans of each institution that forms university.

- Prepare strategic development plan for the institution that is detailed, founded on the vision statement and that envisage new trends in technology and introduction of electronic teaching and learning environment. To produce strategic development plan that is indeed applicable, many parameters has to be encountered, time and money has to be spent and a team of experts representing each institution or department have to be involved.

At the point when strategic development plan is in place changes are much easier to execute. Firstly because sole process of vision statement and strategic plan production made teachers and other staff more sensible to change and most importantly because decisions that are in line with the plan are no longer to be approved by academic council.

**Proposed changes in organizational model**

From hybrid industrial model (mixed with agricultural society model) higher educational institutions have to outgrow to modern post industrial model of matrix organization (Žugaj, 2007). That organizational model is represented by:

- High level of specialization and professionalism for key business process. First of all teaching has to be recognized as that kind of process in higher educational institutions. To achieve that, changes in organization's culture that will be elaborated later, have to be executed. In line with that, support teams of experts in teaching as well as in production of e-content and use of e-environment (i.e. web conferencing, "smart" whiteboards, …) will have to be formed (i.e. centre or department for teaching and learning). Those teams will help avert amateurism in content production making investment in content more worthwhile. Furthermore they will gather and disseminate knowledge and expertise in content development at the institutional level. Support team for teaching and design of teaching materials will help teachers learn methodics, didactics and possibly andragogy in order to improve their teaching skills, ones they missed to learn during their formal education. That way even selection and training of new teachers will be more professional and controlled process.
- Introduction of project management approach through the organization that will encourage development projects and that will change organizational model to matrix one. That way each significant enough expense will be observed as a project. Each project proposal will have to first be checked against strategic development plan and approved, upon being submitted to institution's development office. Each project proposal, in order to be eligible, will have to be already approved by department's head or Dean in the case of university prior to submission. Approved projects will be financed from institutions development budget, according

to the planed annual priorities. For each approved project contract will be signed with project leader regulating copyright of materials produced during the project, dedication of project's expenses, division of income if project's deliverables are later commercialized, schedule of milestones and payments, project's scope and due date as well as reporting obligations to institution's project management office (PMO).

**Proposed change in governance structure**

It is important to appoint person that will be in charge of implementation of technological support to teaching and learning to as high position in governance structure as possible. That person will act as project sponsor for most of the development projects. In the case of smaller institutions (i.e. college) appointment of vice-dean for technological support to teaching or at least vice-dean for development of teaching and learning is strongly proposed. In larger institutions such as university appointment of the vice-chancellor for academic and technological development is a good first step. Still one man can hardly change university so he or she should have up to three deputies; for technology in business, technology as a support tool in teaching and teaching itself. Also, it is recommended to form advisory council for technological support of teaching that consists of; few Deans (department heads in the case of college), representative of centre for teaching and learning, representatives of teachers and person responsible for technology.

**Proposed change in organization's culture**

Academic community respect's principles and values which are partly opposed to today's dynamic and market oriented trends. That principles and values are outcome of scientific work being almost only relevant parameter in career development and advancement of teachers. To change that culture which partly neglects or at least does not promote quality of teaching, institutions have to change priorities and promote as desirable attitude and manners development and excellence in teaching and learning. That cannot be done only on declarative level or by honourable mentions. Instead it has to be done by adequate scoring of such activities in proceedings of advancement and reappointment of teachers.

**Proposed change in funding**

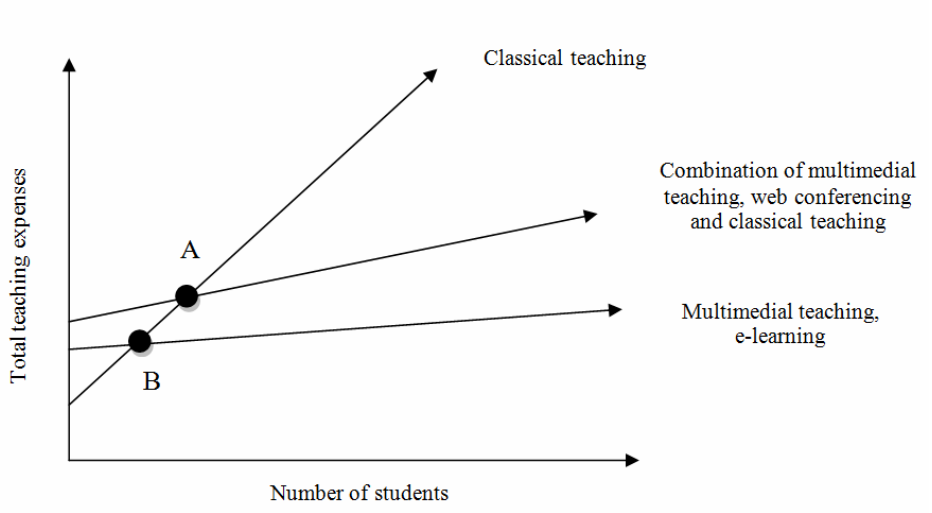
Today most of higher educational institutions invest in information technology if not to develop and support teaching then as a means of support to administrational and business activities. Still, despite sometimes millions spent they often do not have clear conception of results that aforesaid investments made (financial nor educational). In order to propose funding strategy that will be in favour of introduction of new teaching and learning supported by electronic technology it is first important to disseminate expenses.

*Production expenses*
Are considered to be all expenses that arise during the production of training content or are done to purchase hardware and software equipment required to support that kind of teaching and learning. Those expenses are in most cases relatively high and they pose good grounds to blench from teaching supported by technology.

*Variable expenses*
Are considered as all expenses that arise in order to deliver teaching to students. They are called variable since their amount will vary due to number of students enrolled. In this kind of expenses one must count teacher's expenses, training premises expenses, expenses of book procurement, etc.

*Total expenses*
Are sum of production and variable expenses. They are shown in figure 1.

Figure 1: Bearing of total expenses and number of enrolled students




Source: Bates, 2004.

It can be observed in the picture that initial total expense for teaching that is supported by technology is higher compared to classical training, as expected. Also breakeven points A and B show number of enrolled students which, when overcome will yield savings compared to classical training. Point B for pure multimedia teaching and e-learning occurs at lesser number of students while point A occurs for more students. Still, modern trends in higher education advocate combination of multimedia and classical teaching (blended model) that is more expensive to introduce instead of pure multimedia content (Garrison,

Vaughan, 2008). Altogether it can be stressed out that, in order to make savings and to justify expenses in development of teaching supported by technology some minimal number of student have to be enrolled to the module. That number differs on the technique of technology supported training that is used. For the aforesaid reasons institutions should analyse number of enrolled students and other input parameters to decide which for modules (courses) it is feasible to prepare teaching supported by technology.

*Origin of funding*

In today's world where more trained student are expected for lesser resources each year institutions have to come up with a strategy to introduce technology to their teaching which is in most cases expensive. Proposed strategies are:

- Seek for donations and development project funds financed by state or international donator (i.e. EU Funds such are Tempus, Erasmus…). Be sure to distribute and allocate fund donation in a way that after money is spent project can live alone or financing is continued by institution's resources at least until full results and deliverables are completed.
- Charge newly developed training materials and resources to external students, distant students or ones involved in LLL programs that are using the same materials.
- Redistribution of internal resources in a way that part of the resources are budgeted for development of teaching supported by technology. It is reasonable to budget approx. 5% of total training expenses annually for development projects. (Barr, 2005)
- Centralization of resources can help make savings. It is especially feasible in bigger institutions since most of the faculties maybe already have their support departments and sometimes even departments have some people employed for support. Introduction of service desk concept and also internal invoicing for the services that are delivered to internal customers helps reduce costs and also improve quality. Quality increase is inherent since internal customers do tend to demand quality once they know that even internal services are invoiced to their project or department's budget.
- Strategic alliances can help reduce costs and even improve sales. Institutions that are not market rivals can decide to develop some common courseware that are used in common modules (i.e. basics of mathematics) making significant savings that way. Also, academic institution can make alliance with professional company that offers their services of i.e. courseware production on the market. Being significant strategic client and a reference polygon, higher educational institution can gain better commercial terms and such production expenses will be in most cases lower compared to internal expenses that would institution make working alone.

## Conclusion

Introduction of electronic teaching and learning environment in higher educational institutions is inevitable purely because new generation of students already habited to technology in their basic or secondary education and everyday life will hardly be willing to condone higher educational institutions for their life in some other time. Pressure of those students fighting for their careers will require higher educational institutions to be more and more relevant in every aspect. It is almost certain that some of the institutions will not be able and willing to accommodate to aforesaid trends in time, mostly for their rigid organizational structure and ineffective decision making mechanisms. That kind of legging in market driven higher educational systems would possibly strongly influence institution’s mere existence.

## References

Barr, Nicholas. Financing higher education. // *Finance and development magazine – International Monetary Fund* / Volume 42 (June 2005), number 2

Bates, Tony. Upravljanje tehnološkim promjenama. Zagreb : CARNet, 2004

Garrison, D. Randy; Vaughan D. Norman. Blended learning in higher education, Framework, Principles and Guidelines. Organizacijsko oblikovanje, San Francisco : John Willey & Sons, 2008

Žugaj, Miroslav. Organizacijsko oblikovanje. Osijek : Ekonomski fakultet, 2007

Žugaj, Miroslav; Schatten, Markus. Arhitektura suvremenih organizacija. Varaždinske toplice: Tonimir i FOI, 2005.

Original scientific paper

# Influence of ICT on Working Style Used Within Frames of Lifelong Education

Ivan Pogarčić
The Polytechnic of Rijeka, Business Department
58 Vukovarska, 51000 Rijeka, Croatia
pogarcic@veleri.hr

Tatjana Šepić
The Polytechnic of Rijeka, Business Department
58 Vukovarska, 51000 Rijeka, Croatia
tatjanas@veleri.hr

Sanja Raspor
The Polytechnic of Rijeka, Business Department
58 Vukovar ska, 51000 Rijeka, Croatia
sraspor@veleri.hr

**Summary**

*The development of technique and technology significantly influences the shaping of business activities in all fields of human work. Therefore new professions are being developed that consequently require new combinations of knowledge, skills and competencies. Final and by all means the most important consequence is a strong need for constant learning and supplementary knowledge. Perhaps it is too bold to say that comprehensive professional education no longer exists. Although the need for continuous education cannot be observed as a novelty, its importance however is more pronounced. The influence of ICT on profession and working activities can be expressed in different modes. It is indisputable that at present ICT plays an important role in education as well as in performing daily business activities. This paper indicates a tendency of blurring the distinction between ICT application in education and work, during working life and a possible need for education supported by ICT in post-work period. The hypothesis of this research lies in analysing the style of performing working assignments and the need for constant supplementary knowledge – a Lifelong Education.*

**Key words:** education, ICT, student, individualisation, style, paradigm, teaching

## Introduction

The development and implementation of ICT in all areas of human activities significantly influences them in different ways. Within frames of a particular business activity ICT can be involved in many ways while its influence can be analysed from different aspects. Indirectly and directly, the consequence of technological development in general is its influence over educational process in qualitative and quantitative sense. Education is observed as a process of gaining knowledge, skills and competencies that necessarily requires time and space for its realisation. Time, as an educational determinant, is defined by a series of different, primarily pragmatic factors. Formal education is usually defined by the need to acquire basic knowledge and skills which will further on be used to develop specific knowledge and skills defined by the profession that shapes an individual for a longer period of time as well as his position in the society. The ICT significantly influences specificity and definition of particular professions. This influence is not exclusive and is often combined with technological progress and technical solutions in certain economical branches. The quality and quantity of these solutions indisputably influence distinctiveness of particular professions. This way, some professions die or are reborn as new or significantly modified. This process forces an individual to constantly implement and broaden his knowledge within frames of his own profession as well to acquire new skills. It is not a rare occasion that one has to professionally redirect or develop in completely new circumstances.

The recognition of the needs and changes mentioned is presently defined as a need for a Lifelong Education. Nowadays we believe these needs are more expressed and although they do not represent a novelty they are presented as such. The ICT and especially the web provide a possibility to access a huge amount of information and knowledge required in certain situations. Although access and quantity do not represent quality as well, an individual has an opportunity to access information that is appropriate for him and his needs in a specific moment and for a specific period of time.

This paper tries to research in what ways and in which situations the ICT can influence the working style. The answers presented have been obtained through polling a specific number of individuals of different profiles and professions but within the same field of work. The age of examinees has been observed only from the aspect of aspirations toward the ICT usage and information science education.

## Style and a way of acquiring one

The term style is frequently used in different contexts. When we deal with performing professional assignments and realisation of particular activities, the competition itself may require certain knowledge and skills expressed in specific interests and styles. In these situations the working style, management and communication styles are frequently named.

What does a style truly represent? Each term is defined and specified by the closest family term together with specific differences. Naturally, the exceptions refer to these terms that have original or axiomatic meaning. A style can be connected to a mode by meaning and importance. A mode on the other hand represents a way of realising one or more activities. The way of realising a specific activity or a series of activities is generally predefined and in certain measure determined. The level of determination defines the possibility for any kind of variations of a predetermined sequence of activities and the mode of their realisation. Variations as such have stochastic characteristics that can define differences in the mode of performance. Expressed differences in qualitative and quantitative sense represent a style. In this sense, our own life style is what makes us different from the usual life style in the community (in a more restricted) and the society (in a broader sense) that we belong to. Similar definitions of the term style can be found on the web in the dictionaries of scientific institutions (wordnetweb.princeton.edu/perl/webwn, 30.07.2009)

As an example we can mention that a style is a manner: how something is done or how it happens or if we use it in other meaning as an expressive style: a way of expressing something in language, art, music and architecture that is characteristic of a particular person or group of people or period of time. Style (lat. stilus - pencil) is way of expressing character by all those features that differs it from others (hr.wikipedia.org/wiki/Stil, 30.07.2009).

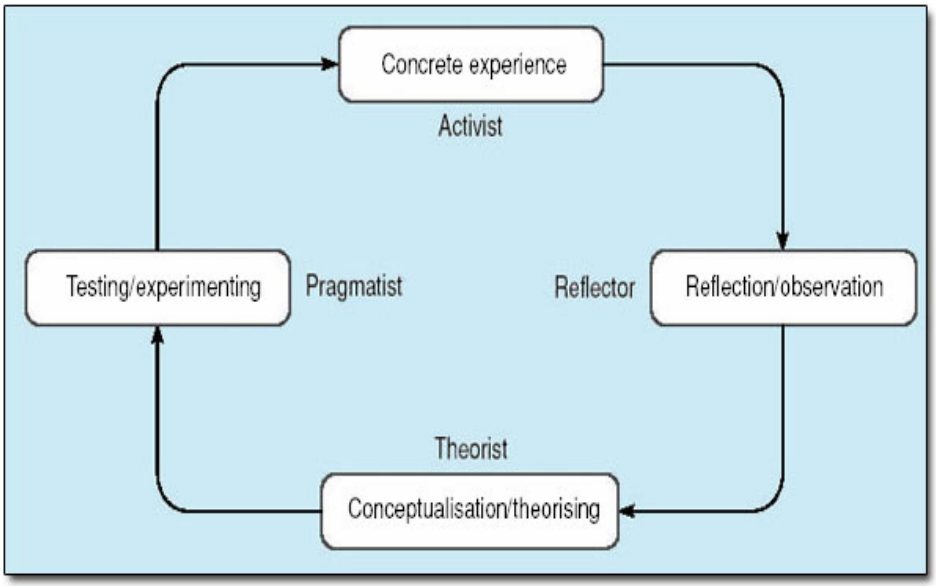



Figure 1. Entrepreneurial work style – Entrepreneurial behaviour (Source: http://openlearn.open.ac.uk/mod/resource/view.php%3Fid%3D212980)

All meaningful activities have an outcome in a desired or expected result. In all these activities a human being is involved in different ways whether as a creator, performer or consumer of the results of these activities. For the same reason, a human being is the one who determines the way of realising activities by directing mode of their realisation and modification that can be desirable or even necessary. A style can be determined by an individual's character (See: Entrepreneurial work style: http://openlearn.open.ac.uk/mod/resource/ view.php %3Fid%3D212980) but also by the character of a group or an organisation (See: Leadership Styles http://openlearn.open.ac.uk/file.php/3038/B722_1_004i .jpg) which through its internal communications create their own character, a combination of individuals' characters. Therefore one usually insists on characteristics such as readiness for a team work, readiness to accept new solutions, agility, self-esteem, consistency etc. It is important to mention that a style does not require the presence of a human being in order to be determined but it does require him in order to perceive and qualify it.. For example, certain natural phenomena, such as natural disasters, can have a style but its nature is determined by humans.

From this perspective it is necessary to determine the way we define style, i.e. whether it is determined independently, or is influenced by someone or something. Since the basic frame of the research is a Lifelong Education and a need for a Lifelong Education, it is important to define whether we can and under which circumstances change the working style. Within the same context it is important to determine the readiness to change style in compliance with one's own perceptions or external influences and warnings.

## Style and its indicators

What kinds of indicators are used in defining a working style? The abovementioned examples indicate that a working style is defined by observing individual's behaviour in a specific environment.

By describing and evaluating behaviour in accordance with a specific group of parameters we can define the style of behaviour and consequently a working style. At the same time different authors and different techniques of evaluation apply detailed or less detailed set of parameters.

Still between them we can single out a certain number of mutual factors. In an example (http://assessment.insala.com/centermark/Includes/SampleWorking Styles.pdf, (30.07.2009.)) parameters are grouped by defining four basic working styles. The basic behavioural characteristics that define individual style are described in Figure 2. The authors (Jackson&McCarthy, 2003) use a Sigma Assessment method to evaluate features such as: impatience, anger, work involvement, time urgency, job dissatisfaction and competitiveness.

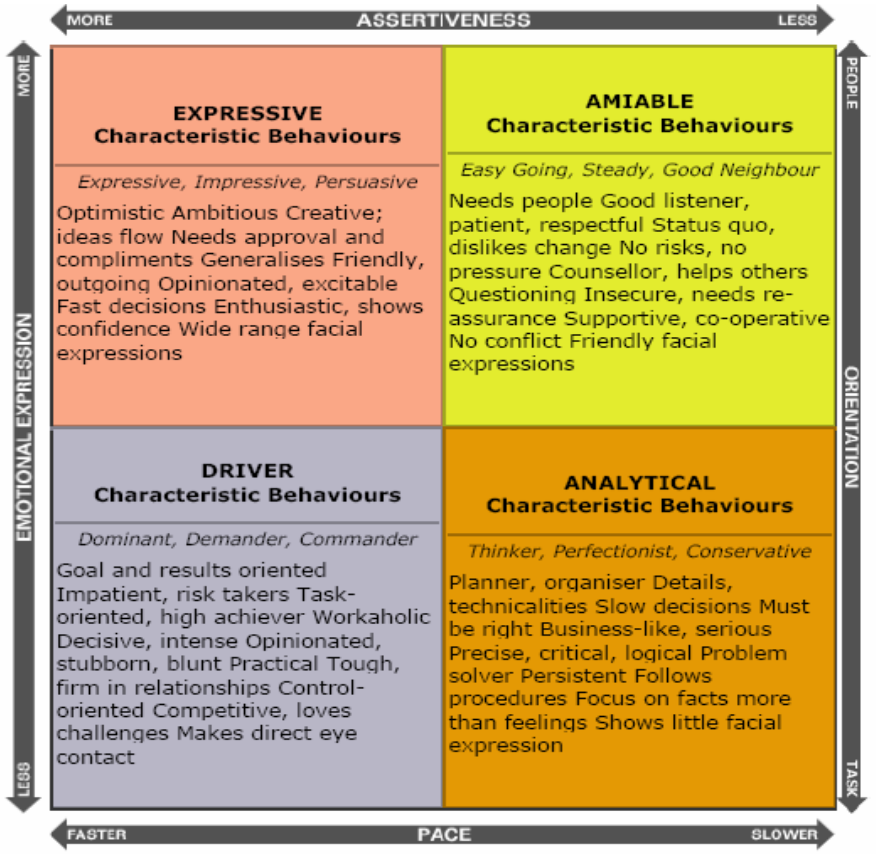



Figure 2. A Map of the Four Working Styles (Source: http://assessment.insala.com/centermark/Includes/SampleWorkingStyles.pdf)

On the basis of which they define type/style and predict potential problems that such a style can cause e.g. heart problems caused by the change of the working style. (Bolton&Grover Bolton, 1996) emphasize the need for business efficiency and harmonization of the working style with these needs. They suggest the identification of a working style through an identification of preferred working styles of employees. The research conducted by the group of authors (Hiroki at all, 2005) is based on the fact that during the last couple of year the diversification of the types and patterns is being intensified which changes professional aptitudes of individuals and the structure of the employment exchange. The consequence is the change in the human resources management which is reflected on the social infrastructure, stability of professions, development of an individual's career and the safety of employment network that corresponds to that diversification. The most detailed way of defining work style can be achieved by using MBTI method, i.e. by determining Myers-Briggs Type Indicators (Hammer, 2008). The MBTI instrument is based on the theory of person-

ality types described by Carl Jung and Isabel Briggs Myers and Katharine Briggs.

## Education and style

The abovementioned overview clearly indicates that all evaluations and definitions of working styles are connected to individual's behaviour or a group in general. A behaviour implies a pattern determined upon experiences or current circumstances. This fact indicates connections between working styles and acquired knowledge and skills directly and indirectly in combination with teaching and educational styles.

If education is perceived as a synergy of teaching and learning then it is obvious that working style is a consequence of adopted and applied paradigm in the same fields. Therefore it may seem that a working style is specifically connected to behaviourism as a teaching paradigm. The more intensive application of ICT in education, on the other hand, more significantly promotes constructivism as a teaching paradigm. Within conditions of a Lifelong Education, especially under conditions of informal education, constructivism is enforced as a prevailing paradigm though it cannot or does not exclude other paradigms.

Considering the connection between learning, teaching and working activities (http://www.creativelearningcentre.com/Products/Working-Style-Analysis/Pyramid-Model.html, (30.07.2009)) the authors have developed pretty unified way of determining teaching, learning and working styles. It is executed through valorisation of indicators organised in the form of a seven-class pyramid. Six crucial areas are being observed: Brain Processing, Sensory Modalities, Physical Needs, Environmental Preferences, Social Aspects and Professional Attitudes.

Each higher class represents in a certain way sublimation and supplementation of a lower class. Each class accepts individual characteristics of a human being and his inclination towards specific paradigms. At the bottom of the pyramid, the lowest class, attitudes are being determined through valorisation of indicators: Motivation, Persistence, Conformity, Responsibility, Structure and Variety. The next class evaluates social characteristics by determining inclination towards individual work, pair, group or team work, or inclination towards authority. The third class defines inclination towards specific type of working environment: noise, lightening, temperature, type of furniture etc. The fourth class specifies preferred physical activities such as mobility during the work, intake and a specific part of the day. The fifth class evaluates sensitive capabilities of an individual such as: listening, visualisation, tactile and kinetic capabilities. The sixth class evaluates inclination of an individual towards a certain way of perceiving and analysing business problems and activities. The highest level tries to determine prevalence between reflexive and impulsive mode of working and decision making.

The described methodology is acceptable both in determining the learning style and working style. The set of indicators ensures individual approach considering all specific qualities of an individual: biological, acquired or conditioned.

## ICT and its influence on the working style

Starting presumptions of this research are:

- ICT influences educational style through different modes of e-Learning
- Information science knowledge increases inclination towards the usage of ICT in education and work
- More intensive usage of ICT in all modes of business is an imperative of adjustment to new conditions of work – PC on the working place
- ICT application increases and improves possibility of profession identification or adjustment to the present needs
- ICT ensures potential independence of place and time when executing working assignments
- ICT allows virtualisation of work and working place

The target population are employees working in a large Croatian service company. The research has been conducted through a questionnaire on the sample of 100 examinees. It comprises employees from all hierarchical levels in the company, working in different company sectors conducting a variety of working assignments. Furthermore, the examinees belong to different age groups and have secondary school, college or university education. Their level of computer skills varies from low to advanced.

The examinees have been questioned via e-mail, providing anonymous answers to 20 questions. They were familiar to pollsters but the questionnaire itself was anonymous. The questions have been grouped according to six main fields applied in method of defining a working style with detection of ICT usage and inclination of examinees towards ICT application in a specific main field. The goal of the research is to confirm starting presumptions and define factors within individual main area that will imply more significant influence of ICT.

The questionnaire was made with a tool http://inovacije.eu/ankete/admin and is available at the following address http://www.veleri.hr/~pogarcic/WSQinF09.doc. The results of the survey are available in table 1.

Objective disadvantages of the questionnaire are:

- Relatively small group of examinees – the size was determined by the choice of the field of work (telecommunications) that employs examinees and the size of their organizations (big organizations with physically dislocated parts and adequate ICT backup in every sense)
- Professional heterogeneity of the group – employees belongs to different sectors of the company such as procurement, warehouse, investments and development, human resources etc.

- Assumption of the existing developed ICT infrastructure – the access to the Internet in all places
- Questions and answers can be inadequately elaborated considering heterogeneity of group

Subjective disadvantages of the survey:

- Personally acquainted with working environment and a larger part of the examinees. This can be regarded as a disadvantage only partly since the questionnaire was conducted anonymously and collected data were processed automatically by the PC.
- Data on the working environment and examinees are historical so the final judgment may seem subjective.
- Although it was not suggested directly but through questions, we expected results that would imply continuous readiness of the participants to anticipate the assignments and to define a style to comply with ICT possibilities.

Advantages of the survey:

- Group of examinees is heterogeneous in terms of age that was one of the basic reasons for conducting this research,
- The author is personally acquainted with the majority of examinees as well as with their information science education in fulfilling work assignments, which can confirmation the credibility of the answers
- Good infrastructural ICT backup and safe network
- Certain experience in the usage of ICT and certain information science education

## Results of the questionnaire

The analysis of the questionnaire has resulted in following findings presented in Table 1.

The questions that were asked tried to encompass all crucial areas (http://www.creativelearningcentre.com) and examine a possible influence of the PC in these areas. The examinees belong to different groups according to their commitments and the way of performing them. The level of responsibility puts an individual into certain position in the chain of performing business activities which redefines his possibilities of using the PC. Still, one can notice a prevalent readiness for the ICT application and adjustment to new modes of performing a business and new solutions in doing the same. We believe that the reason for conducting the research about the influence of the ICT on the working style (http://www.creativelearningcentre.com) or, better yet, profiling the modes of work are acceptable. This is even more acceptable when considering the fact that the survey can include the whole population regardless of its age and dependence upon conditions of a Lifelong Education.

Table 1. The questionnaire and the results

| Questions | Results |
|---|---|
| 1. How often do you use PC in executing work assignments? | weak majority (45%) uses PC occasionally in fulfilling work assignments |
| 2. How strongly does PC motivate you in fulfilling work assignments? | more than half of examinees (52%) believes that PC partially strengthens their motivation in performing business assignments |
| 3. Do you believe PC supports you in your dedication in performing work assignments? | the majority of examinees (63%) believes that PC doesn't influence their commitment in performing business assignments |
| 4. How does PC influence your creativity in performing work assignments? | relative majority (41%) considers PC helps them in improving the way of performing business activities |
| 5. Does PC influence rules that determine "discipline" of work assignments? | PC has no influence over "disciplinary rules" – (arrivals, interruptions, departures) when performing business assignments (84%) |
| 6. Do you believe PC strengthens impact of your responsibility? | PC doesn't strengthen impression of responsibility (78%) (deadlines for finishing working assignments) |
| 7. Does PC help you in creating concepts for fulfilling work assignments? | relative majority (49%) believes PC has no influence over a concept of working assignments |
| 8. What mode of making work do you prefer? | question about preferred mode of making a business doesn't have an adequate answer and will be commented later * |
| 9. Does PC influence your independence? | relative majority (44%) believes PC makes them more independent in performing business assignments |
| 10. Can you imagine yourself in one of the Internet user groups? | the majority of examinees (45%) uses Internet for finding new information about their own profession |
| 11. Do you believe PC usage requires special conditions in environment (lightening, temperature, furniture)? | the majority (89%) believes usage of PC requires special spatial and time conditions |
| 12. Does relatively quiet music disturb you while working on PC? | the majority of examinees (57%) doesn't bother relatively quiet music when performing their work |
| 13. Do you believe PC usage ensures mobility – independence of place where you execute assignments? | the majority of examinees (43% and 46%) claims that PC influences their mobility and dependence upon location and time of performing working assignments |
| 14. Do you believe PC usage ensures timely independence of executing assignments? | |
| 15. Do you believe PC usage requires some other conditions in physical sense such as additional nutrition? | the majority (41%) cannot define whether there are some other conditions of physical type neither if there is a necessity for specific type of sense (53%) or kinetic requirements |
| 16. Do you believe that PC usage prefers certain type of senses, such as touch? | |
| 17. Do you believe PC usage requires certain kinetic conditions such as additional body activity? | |
| 18. Which of the following work style do you prefer? | the majority has a holistic approach (79%), but depending upon their type of work they are willing to apply analytical approach (63%) while in making decisions most of them (82%) uses combined approach. |
| 19. Do you believe that PC usage can influence your work style described in the last question? | |
| 20. Can you imagine yourself in following groups? | |

## Conclusion

The nature of the work itself mostly specifies a way of performing business or executing the business activities obligations. An individual with his personality and characteristic features adds certain qualities to the mode which in turn gets a certain specific shape – style – that makes it recognizable. It is logical that a style depends on the complexity of work since opportunities for expressing a personality grow together with their complexity. More simple, repetitive tasks do not provide a possibility of style expressing. In these cases automation is possible but it excludes stylization. The development of the technique and technology makes us free from such work and their complete automation.

At the same time, technique and technology create new possibilities and demand definition of new professions or more detailed modification of the existing ones. The time period of such changes and the creation of new demands are getting shorter, together with the time required for acquiring new knowledge and skills. The abovementioned calls for adjustment of educational system and emphasizes the need for continuous supplementation of earlier knowledge through adjusted forms of education. Today this process is known as a Lifelong Education. The application of ICT through the basic usage of the PC on working places, through the Internet approach and e-mails ensure possibility for education through work and while working. Laptops and mobiles remove spatial and time restraints and ensure complete continuity where necessary.

This research has been oriented towards an individual and the definition of his style but it did not take into consideration the influence of the group and other activities when there is a human connection or mutual dependence between assignments and their sequence. Further researches in defining the influence of ICT on a working style could be conducted in that area.

## References

Bolton, R., Grover Bolton, D.: People Styles at Work, AMACOM, ISBN 0814477232, 1996

Hammer, A.L.: Work Styles Report: Enhancing two-way communication in organizations, CPP, Inc., 800-624-1765, http://www.cpp.com, 2008

Hiroki, S. at al.: Diversification of Working Styles and Safety Nets: Focusing on Capability Development and Work-Life Balance, Japan Institute for Labour Policy and Training, R75, 2005

Jackson, D.N., McCarthy, J.M: Survey of Work Styles (SWS): Development Report, Sigma Assessment Systems, Inc., 2003

Steiner, J.: The art of space management: Planning flexible workspaces for people, Journal of Facilities Management, Vol. 4, page 6-22, 2005, Emerald Group Publishing Limited

## Links

http://assessment.insala.com/centermark/Includes/SampleWorkingStyles.pdf, (30.07.2009.)

http://hr.wikipedia.org/wiki/Stil, (30.07.2009.)

http://wordnetweb.princeton.edu/perl/webwn, (30.07.2009.)

http://www.creativelearningcentre.com/Products/Working-Style-Analysis/Pyramid-Model.html, (30.07.2009)

# Teaching Digital Collections Management: Issues and Priorities for the Future

Terry Weech
Associate Professor, Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, IL, USA
501 East Daniel Street, Champaign, IL 61820, USA
weech@illinois.edu

Eve Gaus
Adjunct Reference and Instruction Librarian, Elgin Community College Library
1700 Spartan Dr. Elgin, IL 60123, USA
evegaus@gmail.com

**Summary**

*The paper contrasts traditional collection management instruction with that which is necessary in the context of digital collections. A brief review of the transition from print oriented collection management issues and priorities to those of digital collections is provided. The role of digital collections management in library and information science and in specialized digital library curricula is analyzed. Specific issues related to digital collections, such as models of ownership and access, negotiating contracts with vendors and digital content providers, the role of consortia and cooperative agreements in obtaining cost-effective collection content, are explored. These and other issues will be prioritized in terms of their significance for the future education of library and information science professionals. Model syllabi of digital library collection management are reviewed and critiqued with suggestions for core elements that should be part of the competencies of anyone working with digital collections in the 21st Century.*

**Key words:** library and information science education, collection development, digital libraries

## Introduction

There was a time when collection management was considered one of the core courses in Library and Information Science (LIS) education. For years it was one of the required courses in the LIS program at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. (GSLIS-UIUC). Both the authors of this paper obtained their master's degrees from this program, although some 40 years exist between the time each

of us were granted our degrees. Thus we can provide a personal account of some of the similarities and differences in teaching Collection Management at one of the top LIS programs in North America. Until the last quarter of the 20th century, a course in collection management was one of four required courses at GSLIS-UIUC. The other three required courses were cataloging, reference, and library administration. In the 1970s, the curriculum at Illinois, as at many other LIS programs, was updated and revised to include the establishment of a smaller number of required courses, which were to provide an introductory overview of what had been the four "core courses." These changes came about because of the recognition of the growing role of technology within the profession as well as other social and cultural issues, which affected access to information. These changes required a tailored approach to course work in the one year master's degree program to meet a greater variety of job opportunities in the expanded information profession. The content of four core courses continued to be taught as advanced courses, but was no longer required of all students. At Illinois most students, often 80% or more in the master's degree program, took cataloging, reference, and library administration as part of their course work for the master's degree. However, the number of students who took the collection management course steadily declined. This may be because faculty had argued that the concepts of collection management should be integrated into the other courses, and specifically courses in library administration and reference expanded their syllabi to cover collection management issues. In fact, some specialized courses were developed under the titles of legal issues and related concepts to cover the copyright and censorship content of what had been included in collection management courses.

All of this was happening at a time when greater emphasis was placed on expanding the scope of the LIS curriculum beyond that of traditional jobs in the institution of the library. In fact, there were movements to "deinstitutionalize" Library and Information Science education as it was argued that in the future fewer graduates of LIS degree programs would work in traditional library settings and more in would work in alternative careers. While efforts to promote alternative careers for LIS graduates has made considerable progress in Europe and other parts of the world, in North America there is little evidence of much movement to these alternative careers. But ironically the literature and the course syllabi in the area of collection management suggests that with the evolution toward electronic and digital library collections, the role of educating the new professional for collection management may again be an important factor in the future of the profession.

The terms "collection management" and "collection development" are often used interchangeably, although in the U.S.A, the preferred term seems to be "collection development." Kennedy (1998) suggests that one of the reasons for this preference may be the need for the American Library Association to identify a group of specialists working in the library that have responsibilities for

selection policies, collection evaluation, user needs assessment, selection of materials, collection maintenance and weeding, and planning for resource sharing. Kennedy notes that there may be many reasons for the transition from the term collection management to collection development, but he suggests that the most likely reason for the adoption of collection management as an umbrella term for the activities involved in library collection maintenance grew from a change in academic libraries. Previously faculties were primarily responsible for the selection of materials in their subject disciplines and librarians were responsible for managing the collections. More recently, however, professional librarians have assumed the selection as well as the management role. The reason for this change, Kennedy argues, was the growth of higher education institutions in the latter half of the twentieth century. This growth caused an increase of students and faculty as well as an increase in library budgets. As more materials were acquired, the task of selection was "professionalized" by the collection development librarian, or in larger institutions, the collection development unit. In fact, the subject bibliographer position, which had existed in many libraries, was transformed into the more multifaceted position of collection development librarian. Kennedy argues that electronic resources led to the return of the collection management position because in the electronic environment, collection managers not only spend much of their time negotiating licenses for electronic access, but also create and disseminate electronic documents in institutional repositories. (Kennedy, 1998, p. 3)

## Prior studies of Teaching of Collection Development

For most of the 20th century the focus in collection development courses in LIS education was on the selection of "good books" that would make a difference in the lives of readers. Part of this tradition had its origin in the development of "readers advisory" services in public libraries and the extended education mission of school and academic libraries. Within this context was the Bibliotherapy movement what suggested that books might be "prescribed" by professional librarians to meet the medical and psychological needs of readers. (Bibliotherapy Education Project, 2009). There were debates from time to time as to what kinds of "popular" reading materials might be appropriate to hold in a collection. These books were sometimes justified as "bait" to hook the potential library user on reading anything so they eventually could be upgraded to "good" literature consistent with the quality assurance standards of the library and of the professional librarians that were educated in such selection guidelines. Jean Weihs, in a 2008 article, reminisces about her experiences as an LIS student taking a "Book Selection" course in a Canadian library school in the mid-20th century. (Weihs, 2008, p. 9) The text for the course was Haines, *Living with Books,* which was a standard text in LIS education for many years in the mid-twentieth century. Robert B. Downs, Dean of the Library and Library School at the University of Illinois at Urbana-Champaign, was active in writing a series of

volumes on the impact of good books, including *Books that Changed the World.* This approach to collection development applied not just to print materials, but also to other media, such as sound recordings, films, and other non-traditional resources. While such assumptions were challenged from time to time, with "rebel" librarians encouraging young people and others with graphic novels, Harlequin romances, and other "popular" or sometimes characterized as "trash lit" one could argue that the true challenge did not materialize until the advent of digital materials when the universe of resources expanded well beyond the self-contained "good book" collections.

Kennedy, in his 1998 paper, includes a study of the "conditions" of education for collection management/development in the United Kingdom, Australia and North America. (Kennedy, 1998). His review of the literature in the last decade of the 20th century established the changes that occurred in LIS education for collection development. Specific courses in collection development are no longer required in (most?) LIS schools and within collection development classes the trend is to recognize the more managerial aspects of collection development tasks, such as financial management and financial negotiation skills. Fund raising and knowledge of preservation are also considered more important for collection development than in earlier decades. Ultimately, Kennedy argues against those who predict that local collections, and thus collection development activities in libraries will be made obsolete by the expansion of "universal" collections made possible by the Internet. Writing before the Google book project was put into full development, he dismisses such speculation as unlikely because of the significant financial, logistical, legal, and constitutional obstacles. (Kennedy, 1998, p. 7) Of course, in 2009 we are seeing Google and others confronting these obstacles and overcoming some of them.

In 2007, the Collection Development Education Committee of ALA's Reference and Adult Services Association CODES (Collection Development and Evaluation Section) described a 2006 study of collection related course offerings in ALA accredited LIS programs as indicated from the websites of the schools. They found that all but six programs had one or more courses related to collections. That suggests a continuing strong commitment to collection related courses in U.S. and Canadian LIS programs as of 2006. (American Library Association, RUSA, 2006)

## Elements of collection development courses shared by pre-digital and post-digital course offerings and elements that are unique to each

For purposes of analysis we are defining "pre-digital" collection development courses as those developed prior to 1990. Post-digital will include courses developed after 2005. This somewhat arbitrary distinction is chosen to reflect the fact that the transition to digital libraries took place in LIS schools between 1990 and 2005. Although there are many instructors who incorporated digital library issues in collection development courses prior to this, by 2006 the im-

portance of digital library collections was clearly recognized in the course materials and in the literature of library and information science. The "pre-digital" collection development courses had a variety of titles, ranging from "Introduction to Collection Management" to "Materials Selection." In the earlier days of 20th Century LIS education, the term "Book Selection" was often used with the caveat that the term "book" was meant to be used "generally" to indicate any type of information medium.

**Traditional "pre-digital" collection development courses included the following content**

1. Identify and evaluate the various reviewing sources.
2. Obtain data relating to the information needs of users.
3. Collection development policies and procedures.
4. Cooperation and networking among information agencies.
5. Evaluate and select resources in all formats and for a variety of user needs.
6. Issues related to intellectual freedom.
7. Relationship of copyright laws to collection development.
8. Resource sharing, collection evaluation, and networking.

The syllabi of two post-digital collection development courses at GSLIS at the University of Illinois at Urbana-Champaign were examined. The syllabi represented two different courses on collection development taught by two different instructors. One course (LIS 590CD-"Collection Development" clearly attempted to cover collection development in all types of libraries and the other course (LIS 590CD2-"Current Topics in Collection Development"), which focused on collection development issues, clearly had an academic and research library orientation.

In the case of the first course, all the elements found in a pre-digital collection development course were maintained, but a number of topics were added. These included units on Acquisition Procedures, Budgets, Licensing, Vendor Negotiation, and Access vs. Ownership.

The second course was more issue oriented and was more focused on academic and research libraries. It was not a more advanced course than the first in so far as the first course was not a prerequisite for the second. The second "issues" course covered most of the content in the pre-digital course, but explicit references to intellectual freedom and legal issues were not evident in the course syllabus. In addition to the units found in the first analyzed post-digital course (units on Acquisition Procedures, Budgets, Licensing, Vendor Negotiation, and Access vs. Ownership) this issues oriented class included units on scholarly communication, institutional repositories, and the open access movement.

Clearly there are changes in the content of LIS courses on collection management in the course offered at this one LIS program. This pattern of content seems to follow the prediction of Kennedy and others as to the future trends of

teaching collection development in LIS programs. Aspects of electronic and digital collection management are clearly incorporated into the course syllabi in the course work at the University of Illinois at Urbana-Champaign. But the next question is, what is the role of collection development in the digital libraries program at the same University?

**Inclusion of post-digital collection development elements in digital library courses**

The GSLIS-University of Illinois at Urbana-Champaign digital library program is a sixth year post-master's degree program. The program description and list of required and elective courses can be found at: http://www.lis.illinois.edu/programs/cas-dl.html.

The four required courses are: LIS453 "Systems Analysis and Management", LIS590DIL "Introduction to Digital Libraries", LIS590IML "Information Modeling", and LIS590MD "Metadata in Theory & Practice". Of the four required courses, only LIS 590DIL has elements of content found in collection development courses. LIS 590DIL (Digital Libraries Research and Practice) includes units on Intellectual Property, Security, and Privacy. It should be noted that the "Current Topics in Collection Development" course that is discussed above, is one of the suggested electives, but not a requirement, for the advanced degree in Digital Librarianship. Thus is it possible for someone at Illinois to complete a degree in digital libraries without being exposed to the fundamentals of collection development beyond he few elements in LIS 590DIL. The focus of most of the courses in the program is on the technical side of developing and maintaining digital libraries rather than on the theoretical and managerial side of building collections.

How does this compare to digital library courses in all ALA accredited programs? Pomerantz and others in 2006 published an analysis of the digital library course syllabi in ALA accredited programs and found that collection development was third in the frequency of reading topics found in digital library syllabi. (Pomerantz, 2006, Figure 2: Distribution of readings across topics). Only "Project Management" and "Architecture" exceeded the Collection Development readings.

This finding seems to be puzzling given the lack of readings and focus on collection development found at Illinois in the digital library courses. Two possible explanations come to mind. 1) Illinois may not emphasize collection development in digital libraries as much as other programs; or 2) there may be a difference in the frequency of readings in those courses offering collection development topics in digital libraries compared to other topics. Since the literature of collection development has a longer history than that of digital libraries, perhaps the measure of readings is reflecting that richness of resources rather than the actual content of the digital library programs. More investigations would be

required to determine whether this second explanation has merit. (Pomerantz, 2006)

It should also be noted that in the Framework for a Digital Library Curriculum (2008) "Collection Development" is broken down into the following categories:

3-a: Collection development/selection policies
3-b: Digitization
3-c: Harvesting
3-d: Document and e-publishing/presentation markup
3-e (7-e): Web (push) Publishing
3-f (7-f): Crawling

Of these six categories, most librarians would consider only 3-a to be directly related to collection development. This might also explain the larger number of collection development readings found by Pomerantz in 2006.

## Digital Library model curriculum

The Digital Library Curriculum Development Project, funded by the National Science Foundation and established as a joint research project at Virginia Tech and the University of North Carolina, Chapel Hill has as its goal the development of a model curriculum for digital library education. The project began in 2006 and was funded through 2008. While many of the curricular modules have been completed, one that was not completed by the end of the project was "Collection Development – 3a Collection Development/Selection Policies" As noted above, while this is just one component of a collection development course taught in general LIS programs, it suggests that among those instructors and researchers involved in digital library curriculum development, collection development is of a lower priority that other elements of the curriculum. This may be the case because it is assumed that the theory and principles of collection development will be obtained from other coursework in the LIS curriculum. But regardless of the reason, the gap left in the model curriculum development for digital libraries is a concern for those who believe that collection development is a very important issue in the digital library world.

## Conclusions and future research directions

Clearly, courses in collection development are adjusting to the digital age and are upgrading content to cover the necessary competencies needed for maintaining digital library collections. More attention is being paid to the financial side of collection development, including vendor negotiations. While instruction in collection development has made considerable accommodation to digital material, the question is, have programs of study in digital librarianship recognized the importance of collection development to their instructional mission?

It is the intent of this paper to stimulate the discussion raised of what should be in a model curriculum for collection development in a world of digital libraries. A related question is whether assimilation of the content of collection develop-

ment courses into digital library courses is the best solution. And behind both these questions is perhaps an even more important one: Will the profession tolerate a continued separation of digital and non-digital education content as represented by those programs that are established as independent or advanced programs in digital librarianship?

It is our belief that in the future, digital library education programs will be integrated into the basic professional education of librarians and information science. This integration will answer the questions posed above and hopefully will result in the recognition that all librarians and information scientists must be equally exposed to the principles of collection development and effective collection management.

## References

American Library Association, RUSA (2006), "Collection-Related Courses in AA-Accredited Master's Programs: A Description." Reference & User Services Quarterly, vol.47, No. 1, pp. 42-43.

Bibliotherapy Education Project – University Libraries, University of Nevada, Los Vegas. http://www.library.unlv.edu/faculty/research/bibliotherapy/ (Accessed 87/09)

Digital Library Curriculum Development Project (2008a) "Tracking of module development & wiki evaluation status," http://curric.dlib.vt.edu/~dlcurric/ModuleTracking.20080516.pdf (Accessed 8/9/09)

Digital Library Curriculum Development Project (2008b) "Tracking of module development & wiki evaluation status," http://curric.dlib.vt.edu/~dlcurric/ModuleTracking.20080516.pdf (Accessed 8/9/09)

Digital Library Curriculum Development Project (2008c) "Framework for a Digital Library Curriculum (2008/08/23)" http://curric.dlib.vt.edu/DLcurric_images/ModuleFramework2008-08-23.pdf (Accessed 8/9/09)

Kennedy, John (1998) "Education for collection management: Ending before it ever really started, or only just beginning?" Education for Information; 1998, Vol. 16 Issue 1, p45, 12p

Pomerantz, J., et al. (2006). The Core: Digital Library Education in Library and Information Science Programs. *DLib Magazine*, 12 (11).

University of California – Santa Barbara. "Collection Manager's Manual" 2002. http://www.library.ucsb.edu/collman/ (Accessed August 7, 2009)

Weihs, Jean (2008). *Technicalities,* vol. 28, No. 4, pp. 8-11)

# Teaching Quality Management at the Course Level

Krešimir Pavlina
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
kpavlina@ffzg.hr

Mihaela Banek Zorica
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mbanek@ffzg.hr

Ana Pongrac
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
apongrac@ffzg.hr

**Summary**

*This paper presents quality assurance systems used in the countries of the world and presents processes for internal and external quality assurance in institutions of higher education. By analysis of existing systems for control of the quality of teaching, it has been found that all observed the quality of teaching only to the course level. The central part of this paper presents developed model for quality management at the course level that is based on continuous assessment during semester. Evaluation is conducted at the level of individual teaching topic using on-line questionnaire. The model presents a way of statistical analysis of the results obtained by evaluation of teaching quality. This has enabled improvement in teaching practices within each individual teaching topic, as well as the improvement of the entire course. Based on the model presented in this paper, researchers created on-line quality management system.*

**Key words:** quality management, quality control, higher education, course, teaching topic

## Introduction

In last two decades higher education has been rapidly expanding in the number of institutions. By creating Bologna process European countries are trying to harmonize their higher education system to be comparable between different universities and different countries.

One of major requirements of Bologna process is promotion of student mobility, which implies that student can listen part of his study programme as guest student at the foreign university. To create network of universities of comparable quality it is necessary to develop efficient quality management systems that will ensure models for measuring and managing quality of teaching. (Husbands, Fosh, 1993)

## Quality Management Models

Quality is by definition the extent to which a product or service meets a complex requirements. It ensures that product or service is designed and manufactured to meet customer requirements.

Quality in higher education ensures that educational service has been designed and conducted by the requirements set by society. Different societies have different needs, so we cannot speak of quality management system which could be globally acceptable, rather every country should create its quality management system which best suite its needs.

Generally quality management can be divided into basic two quality management levels: internal quality management and external quality management.

Internal quality management enables institution to independently control and improve quality without external pressure.

External quality management is usually conducted by official state agencies which try to manage quality of teaching at state level. Their main objective is to assure that all universities fulfill minimal quality conditions set by state. (Brennan, Shah, 2000)

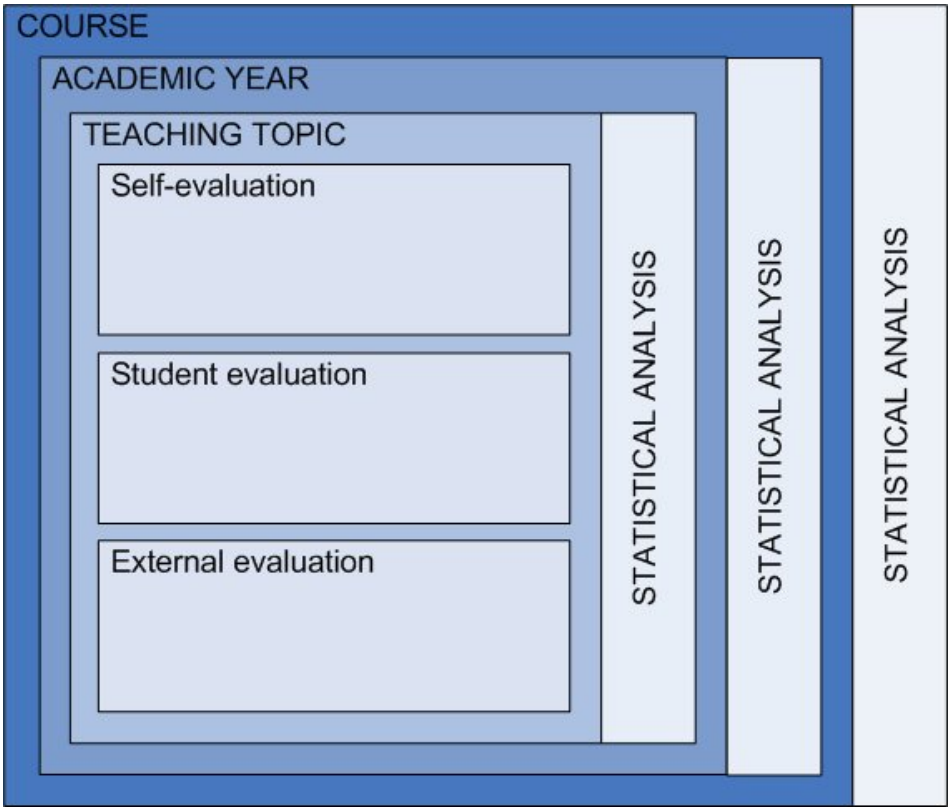
Because of wider extent, external quality control is usually conducted every several years so it is relatively slower than internal quality management which usually has shorter quality management cycle. Internal quality control surveys are usually conducted at the end of semester, so that only future generations of students can experience improvement of teaching which resulted from conducted quality control survey. (Marsh, 1987) It would be beneficial to use quality management with shorter cycle because it would enable quality improvement during semester.

## Course Quality Management Model

Theoretical model of quality management at the course level advocates quality control at the lowest level, which is the level of teaching topic. It is a dynamic quality control, which allows improvement of teaching practices during course execution within the same academic year. This represents great progress over the existing models which do not carry out quality control during the semester, but only at the end of the semester, allowing students to miss the improvement in the quality of teaching, because the improvements can only be made in the following academic year.

This model is based on three groups of evaluators: teachers, students and external evaluators. All three groups fill the same quality control survey which enables the comparability of results obtained from different evaluator groups. Valuable information can be gained from comparison of self-evaluation surveys, student surveys and external surveys, because sometimes teacher's self-perception and student perception of teaching can be very different.
Statistical analysis of collected data provides valuable information for quality management. It is possible to compare quality of topics lectured during current semester which can show us current trend in quality of teaching during current semester. Since all topics during semester aren't equally attractive to students this should only be used for estimating trend in quality of teaching. Model fills this gap by enabling comparison of topic lectured this year with survey results from the same topic from previous years. This is good indicator for improvement in quality of teaching, because these topics are equally attractive, and can easily be compared.

Figure 1. Logical structure of the model for managing the quality of teaching at the course level



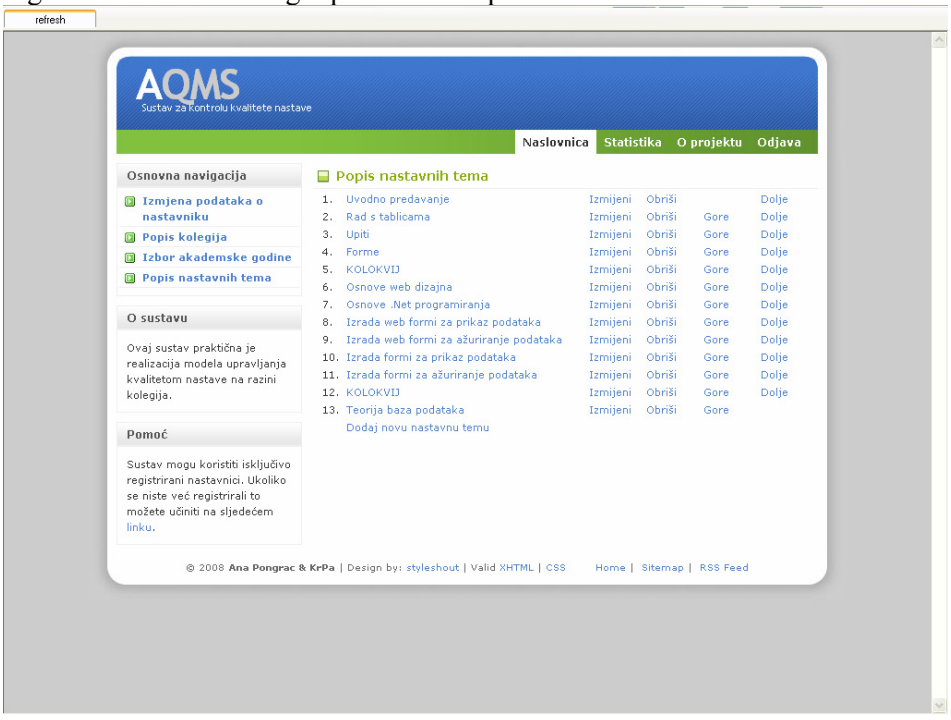
Good teaching quality management should always provide students with feedback. That motivates students to continue to help improvement of quality by providing objective and constructive answers in quality control surveys. Existing quality management models face great difficulties because of lack of good communication strategy and students are often very demotivated because they don’t see the point in participation in quality control surveys if their opinion is not respected and implemented in teaching practice. This model recommends that teachers publish results of their quality control surveys to students, so they can be active participants in quality improvement process. This model also enables students to enjoy improvement in quality of teaching during semester which should motivate them for further participation in surveys.

## Academic Quality Management System

Theoretical model of quality management at the course level is practically realized in the form of web application that allows teachers to manage quality of their courses. The system can be found at following web address:

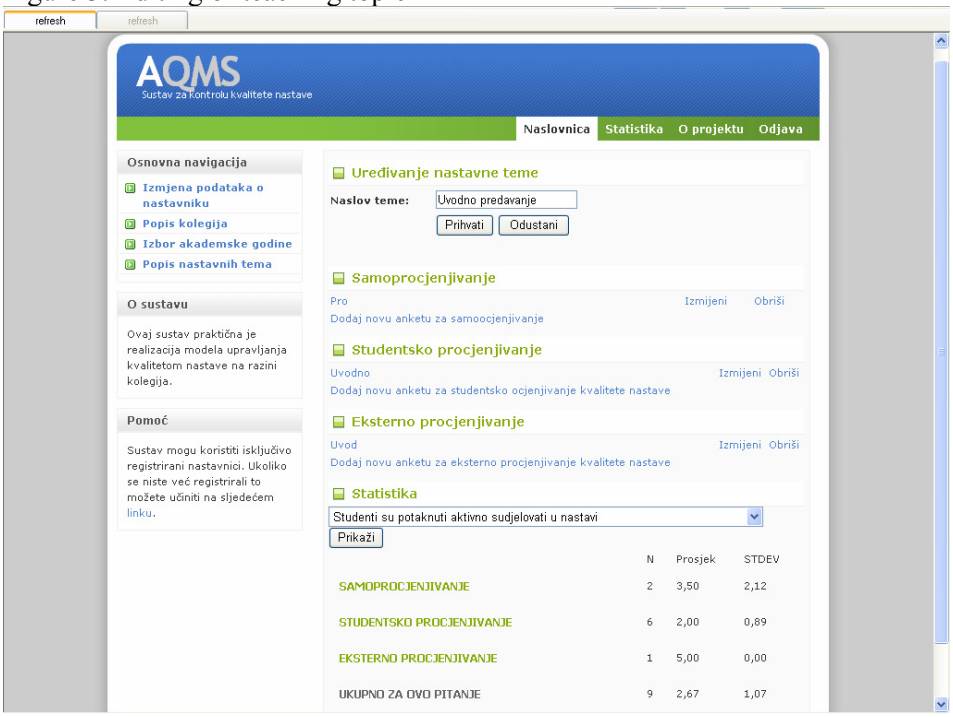
**http://infoz.ffzg.hr/quality/**

Figure 2. List of teaching topics within a particular course



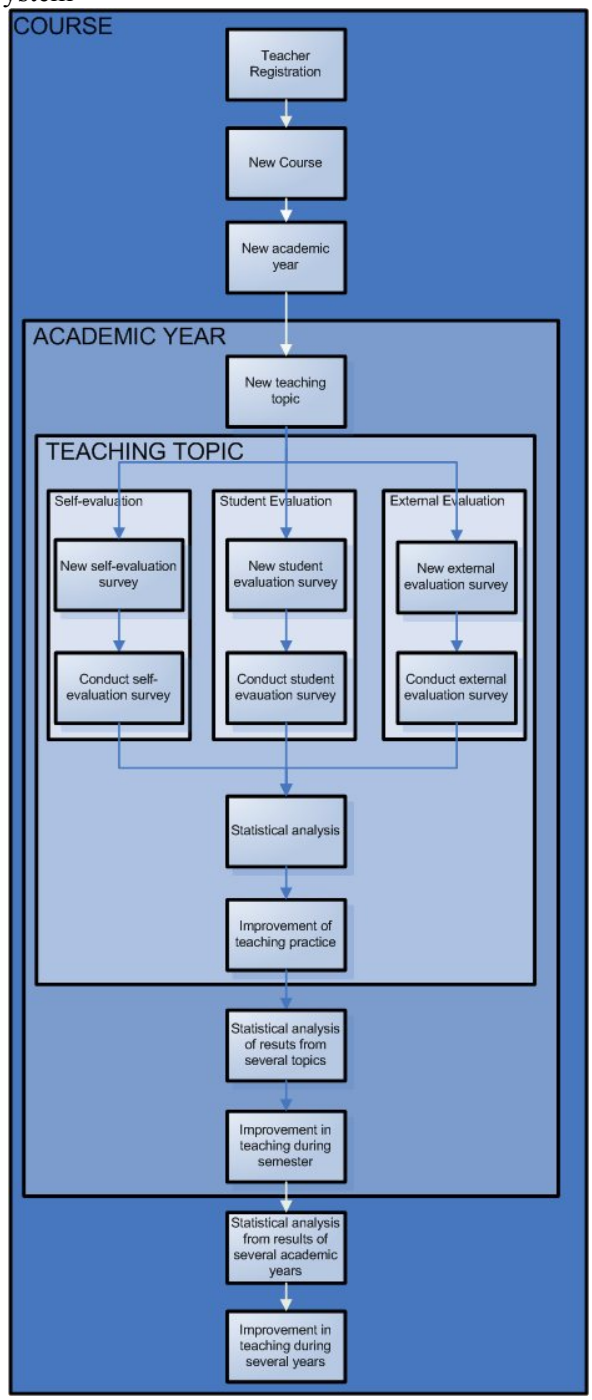
Anyone can use the system free of charge. The system allows high level of privacy protection for students and teachers. Students are absolutely anonymous in completing the survey and teachers are given the choice whether they wish to publish the results of assessment of the quality of teaching.
Student anonymity improves objectivity of their answer in conducted survey. (Leckey, Neill, 2001) Academic Quality Management System have built in safety mechanisms that can prevent same student to fill same survey several times and that way impair collected results. One of methods allows only one access to survey in certain period of time i.e. 30 minutes from same IP address. Other method allows teacher to randomly distribute to students anonymous access codes which can be used to fill survey only once. There is also option to allow access only after authentication where system provides total anonymity for students, but students are little skeptical about anonymity so it can affect objectivity of their answers.

Figure 3. Editing of teaching topic




To use Academic Quality Management System teachers need to fill simple online registration form. Every teacher can participate in one or more courses. Teacher can add a course in which he lectures. After adding a course, the teacher adds a new academic year in which the course is teaching. Within academic year teacher needs to define teaching syllabus by simply listing teaching

Figure 4. Flow diagram of available activities when using the Academic Quality Management System



topics. Teaching syllabus can be changed from year to year but existing quality control data won't be lost, so that it is easy to compare results to previous years.
Teacher can create several surveys within same teaching topic. This allows teacher to monitor quality in several student groups, or to conduct self-assessment before and after teaching certain topic.
Academic Quality Management system uses official quality control survey questions used at the University of Zagreb. The survey consists of 20 questions that examine various aspects of quality of teaching and an open field in which it is possible to write a comment.
It is possible to compare results gained from different sources. By comparing results gained from self-assessment, student assessment and external evaluation, teacher can view wider picture of his quality of teaching. Figure 3 shows teaching topic edit form where teacher is presented with basic statistical variables for every question in a conducted survey: number of respondents, arithmetic mean value and standard deviation.

## Conclusion

Major problem with existing quality management models is long quality control period, where surveys are usually conducted at the end of semester. Because of this, it is only possible to improve teaching for next generation, but unfortunally not for current generation of students.
Another commonly found problem in existing quality control models, where quality is improved from generation to generation, is that students rarely receive any feedback about actions taken to improve the quality of teaching that would show them that their suggestions are implemented in teaching practice.
Presented quality control model at the course level advocates quality control at lowest level – teaching topic. By conducting surveys after every topic it is possible to improve quality during semester so that present generation of students can enjoy benefits of quality improvement. This also improves student motivation for objective assessments because they receive continuous feedback and can observe improvement in quality of teaching.

## References

Brennan, J., Shah, T., "Quality assessment and institutional change" // Higher Education 40, 3 (2000), 331-349

Husbands, C. T., Fosh, P., "Students Evaluation of teaching in higher education: experiences from four European countries and some implications of the practice", Assessment and Evaluation in Higher Education, Vol. 18 (1993), Issue 2

Leckey, J., Neill, N., "Quantifying Quality: the importance of student feedback", Quality in Higher Education, Vol. 7, No. 1 (2001)

Marsh, H.W., "Students' evaluations of university teaching: research findings, methodological issues and directions for future research, International Journal of Educational Research, (1987)

# Education in Virtual Environment

Mihaela Banek Zorica
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mbanek@ffzg.hr

Antonija Lujanac, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
ahorvat2@ffzg.hr

**Summary**
*Students are our most precious resource, and as such should be cultivated with the utmost care in well-designed educational environments. If our goal is to assist the learner in developing the knowledge, skills, and attitudes to join society as a functioning member, we should be oriented towards the future creating successful learning systems. Virtual environment potentially provides such a interesting, instructive, interactive environment also for students in Croatia, but there are some possible detrimental educational consequences. Anyone who is engaged in education must be aware that many of the necessary skills people learn through direct interaction with others even without technological intermediaries. This paper will give a general overlook on the characteristics of education in virtual worlds, give insight in pilot study on implementing virtual world in university classes and discuss both positive and negative sides of education in virtual environment.*

**Key words:** virtual environment, Second life, education

## Introduction

For more than three and a half centuries European and worldwide educational systems were based on the book as the sole medium for storage and transmission of information. Today, young people and adults live in a completely different media environment. Most of them possess a computer connected to the Internet, which is used for their work, education and leisure. Learning in a virtual environment poses a very practical and effective way of learning, but we should bear in mind that when implemented in the primary and secondary education it also has some negative consequences like insufficient physical activity and reduced immediately socializing with peers. This was confirmed by the survey of 3,833 primary and secondary pupils conducted in April, 2002 in Za-

greb area and central Croatia[1]. According to the results 43.70% of children attending primary school have personal computers in their homes and use them together with their friends still maintaining social interaction. However, students of high schools of which 60% regularly use the Internet reported that they spend more time on the Internet than with their friends which shows a decrease in their social interaction. Some respondents stated that they had to end many friendships due to spending too much time at the computer. On the other hand usage of technology had many benefits on educational process, like course and student administration which was confirmed in another research of the same the author. In teachers opinion use of computers loosens school curriculum, reduce the number of textbooks to be carried daily by the students, facilitates teachers administrative tasks, to name just a few.

## Forms of education in a virtual environment

Today's ICT supported education is implemented in different types of classrooms: multimedia classrooms, interactive multimedia classrooms and virtual environments. Multimedia classrooms are usually equipped with TV, speakers and an LCD projector thus creating a first level multimedia environment that tries to respond to the student needs. Teaching mode is still oriented towards teacher as a "information keeper" i.e. teacher centred and does not respond to the learners individual needs. The only advantage of the multimedia classrooms versus classical classrooms is in students' experience which is enhanced due to the simultaneous activation of several perceptual organs ultimately leading to a positive knowledge transfer. Combination of these electronic devices with computer could lead to usage of enriched educational materials instead classic textbook. Offering internet connection and creating a network of computers enables creation of interactive multimedia classrooms which can offers teachers a good technological base for transformation of educational process into a student centred one this enabling and encouraging meaningful learning. It enables teacher to define the learning objective and create learning tasks that allows student to learn in, for them, the most suitable mode. By doing so learning becomes more interesting and responds to their personal needs and capabilities. One of the main challenges is the loss of control over the individual information needs of each student and possible information overload. Such a personalized learning environment requires work in smaller groups which in current school environment harder to ensure. Furthermore, teachers need to stimulate both individual and collaborative work. They need to develop social skills and learn how to evaluate and compare each ones work.

[1] Matijević, M. Internet, multimedij i cjeloživotno učenje. // Zagreb : Hrvatsko andragoško društvo, 2002, pp. 267-276.

Interactive multimedia classrooms are the easiest form of a virtual environment that endorses full interactivity. The most complex type, virtual classrooms (Figure 1), present a combination of interactive multimedia classrooms connected to the Internet and enriched with advanced audio-visual devices and virtual representation of the world. In these environments high level of interaction is obtained and physical location of students and teachers is in not of importance.


Figure1. Virtual classrooms should be connected [2]

Virtual classrooms should be created in the following manner:

- team of experts with the help of advanced technology generate virtual reality for every teaching situation and store it in the educational information system in order to make it accessible to all users,
- the teacher chooses the appropriate situation according to the teaching plan, and organizes educational environment (computer, audio-visual helmet, sensor gloves, etc.) which enables him to develop a learning program
- student interact with the system

The teacher chooses a learning objective and students themselves select the learning paths that the system generates. Students can create new situations, study them, change the method of trial and error and finally finish the learning task.

High quality virtual learning environments and intense activity of the student's perceptive organs create a "reality" in which students create cognitive and experiential effects in the process of learning. One of the many advantages of learning in such virtual environment is the interdisciplinary teaching scenarios in

[2] Virtual classroom // Virginia department of fire programs http://www.vafire.com/higher_education/virtual_classroom.htm (15.08.2009)

which knowledge is set in a broader context and becomes more comprehensive this of better quality.

## Implementation of virtual environments in Croatia

Croatian initiatives in implementation of virtual environments of in school children education could be found in programmes for the gifted children. One successful example is “Worlds apparent reality” a IT and robotics programme that has resulted in creation of virtual world “Croatia”. It is based on 3D Construction Kit, which is completely free and available online. Using this program, many children develop numerous architectural buildings (buildings, wind power, holiday resorts), set up exhibitions of pictures in museums, etc. by using 9000 previously defined objects as well as different types of 3D shapes and surfaces. Objects can be created in any 3D program (True Space, Imabot, Xelagot, even Visual Basic). Stimulating children participate in and develop virtual environment develops not only spatial perception, but also creativity, and due to communicate with people online has shown an increase emotional intelligence, today much appreciated of the once popular IQ (Figure 2).

Figure 2. Anita, hostess in a virtual world, “Croatia”. Gestures and mimicry bots in virtual reality worlds are at a very high level.

Furthermore, some children also participate in the robot making courses using ROBOLAB system that expands the range of LEGO sensors, computer-controlled motors and RCX module (Figure 3) combining LEGO blocks with the right industrial microcontrollers (Figure 4).
Many children who have already participated in these courses were thrilled by the possibility to express their creativity in such a way.

Figure 3. Modul RCX

Figure 4. Creating a robot

## Implementing virtual worlds in higher education

Under the project *Knowledge organization, management and sharing in electronic learning environment* financed by the Croatian Ministry of Science, Education and Sports an analysis of the existing game based environments and virtual worlds was undertaken. The decision of testing fell on the currently most popular one – Second life. Project goal was to test this environment and create an extension to the current Faculty of Humanities and Social Sciences (FHSS) e-learning environment specifically for the part-time students and their distant learning courses. Extending current electronic educational environment to the virtual 3D space was motivated by the fact that avatars and virtual worlds could partially substitute the real life classes and interaction both between students and between students and teacher. Moreover, interoperability between FHSS virtual learning environment OMEGA (based on Moodle) and the Second Life platform was also to be tested as a foundation for administration of classes held in real and virtual world.

Figure 5. Virtual learning space of Faculty of Humanities and Social Sciences, University of Zagreb in Second Life

Pilot study was based on one elective course *Information in electronic environment* offered to part-time LIS students. Introductory meeting was organized and students were informed on the procedures, tasks and learning objectives. They were divided in 3 groups of five, in order to give each student full support and individual teachers backup, while the course content was divided in three parts: introduction to Second life, teaching how to create object and decorate meaningful information space, and retrieval and evaluation of information found *in*-world. At the end of the course students were asked to evaluate course content, teaching methods and learning environment. Course was evaluated as a successful and proposition to hold similar courses was made. Students had learned how to act in virtual world and were able to utilize previously learned skills like information literacy, programming, reference service etc. As most of the student, attending this course, were employed in the school or university libraries, additional outcome is their introduction to new technology which they can now implement in their work environment.[3]

## Educational side of virtual education

Often is emphasized potential negative impact of virtual environments on youth thinking together with violent computer games or the exposure of young to inappropriate web content. Many articles were written on these subjects so we would like to focus on some of the possible consequences of virtual education in youth education. For example, in Germany out of 11 million students 700 thousand experience that their behaviour interferes with their learning. Out of the 100 students they observed nearly 23% reported various forms of aggressive behaviour[4]. Teachers and professional staff often complain about the lack of time and would very gladly spend time doing more to raise young. Many students complain about the overcrowding learning material that remain on their cognitive level, but do not penetrate more deeply into their emotional lives. Moreover, materials contribute to the informational overload so relevant issues can not be discussed due to the fact that student concentration is reduced due to the lack of energy. Right there could we utilize the advantage of virtual environments in creating deeper connections to educational content and transfer knowledge not only memorizing facts. Multimedia and virtual classrooms can be perfectly used for faster adoption of educational and less important content, or for the deepening of knowledge and emotional experiences of educational and important facilities. Interdisciplinary classrooms that support reduced material cause positive emotions such as curiosity which will open the way to the

[3] Banek Zorica, M. Spiranec, S. Pavlina, K. Immersive worlds as educational environments. In Research, Reflections and Innovations in Integrating ICT in Education, MICTE 2009. Lisabon: Formatex, 2009

[4] Winkel, R. Djeca koju je teško odgojiti. Zagreb : Educa, 1996, pp. 26-27.

emotional world of students. Virtual representation of the students can easily environment where they connect with the concrete practice and in which students develop motivation and positive emotions such as feelings of success and usefulness. Positive emotions that arise when learning certainly do not encourage aggressive behaviour which often occurs due to dissatisfaction of students who do not see the usefulness of material exhibited by frontal teaching. Research shows that the concentration of students in the last twenty years has reduced by more than 70% together with the knowledge they need to master multiple magnification[5]. Many teachers fear allowing students to use technology in the classroom can cause complete chaos and lack of control, but they forget that the new generation of students must learn to use technology and school. Educational institutions should offer students knowledge, skills and emotions that will be required after completion of classes, in their professional lives.

Implementation of computer technology is not about teaching but about learning[6]. In relation to the book it allows endless variations in the realization of the educational tasks while respecting the principles of multiple intelligence and humanly organized schools. Learning in a virtual environment enables a new kind of communication between participants. It enables a teaching process that encourages students for greater openness in communicating in a way that seems useful, interesting, and for crucial for the contemporary society. Virtual education acquires work habits and self-confidence and develops a sense of responsibility and creativity.

## Conclusion

Comparing the advantages and disadvantages of learning in a virtual environment it has become clear that today's generation wants and needs information-based education using new media. All the negative consequences of virtual education can be avoided when including virtual learning environment in the regular school system, because then students will learn self-knowledge and skills they need in order to be equal members of information society. In Croatia, initiatives for providing creative learning virtual reality and virtual worlds have started. Through virtual education, it is possible to achieve a better relationship with young people who will surely be grateful for a sign of desire to understand their modes of communication as transformation of traditional learning environments.

[5] Borba, M. Building Moral Inteligence. San Francisco : A Wiley Company, 2001.

[6] Drucker, P. Nova zbilja. Zagreb : Liber, 1992, p. 221.

## References

Banek Zorica, M. Spiranec, S. Pavlina, K. Immersive worlds as educational environments. In Research, Reflections and Innovations in Integrating ICT in Education, MICTE 2009. Lisabon: Formatex, 2009

Borba, M. Building Moral Inteligence. San Francisco : A Wiley Company, 2001

Drucker, P. Nova zbilja. Zagreb : Liber, 1992

Encarnacao, J.L; Leidhold, W; Reuter, A. Expertenkreis Hochschulentwicklung durch neue Medien. http://www.bertelsmann-stiftung.de/bst/de/media/xcms_bst_dms_13133_13134_2.pdf (15.08.2009)

Gardner, H. Disciplinarni um. Zagreb : Educa, 2005

Makanec, B. Kako u darovitih učenika poticati kreativnost kroz informatiku i robotiku. Zagreb : Centar za poticanje darovitosti djeteta, 2005.

Matijević, M. Internet, multimedij i cjeloživotno učenje. Zagreb : Hrvatsko andragoško društvo, 2002

Winkel, R. Djeca koju je teško odgojiti. Zagreb : Educa, 1996

Original scientific paper

# E-portfolio for Recognition of Prior Learning Assessment in Continuing Education for Librarians in Croatia

Dijana Machala
National and University Library
Hrvatske bratske zajednice 4, Zagreb, Croatia
dmachala@nsk.hr

**Summary**

*Paper presents theoretical issues related to e-portfolio as lifelong learning tool. E-portfolio supports ongoing learning/professional development, formative and summative assessment, reflective writing and collaborative, active and deep learning. It is powerful tool for professional development planning or career planning by fostering intrinsic motivation of a learner to maintain portfolio on an ongoing basis throughout formal classes, programs of non-formal learning or just in demonstrating professional or personal growth over time. Focus is on e-portfolio as tool for evidencing prior learning for further assessment or recognition. Different types of digital technology are in use, and diverse portfolio purpose and use are reflecting on assessment. Paper focuses on modeling a conceptual model of e-portfolio system as a tool for assessment and recognition of prior learning in continuing education of librarians in Croatia.*

**Key words:** e-portfolio, recognition of prior learning, continuing education

## Introduction

Portfolio is a written record of the skills, achievements and learner's development over time. There are several types of portfolios; most commonly we think about artist's portfolio which is consists of artwork that the artist can take to job interviews, conferences, galleries, to give others an idea of what type of genre the artist works in. Art portfolio, sometimes called "artfolios", can be a variety of sizes, and usually consists of approximately ten to twenty photographs of the artist's best works. Artists could maintain multiple portfolios for different types of work, one for technical illustrations and another for paintings.

Portfolio doesn't stand instead of formal qualification. Its purpose is to give evidences of someone's professional or personal competencies or learning experience gathered over some period of time. Portfolios can be of different types: artistic, learning, research, institutional, career, financial and so forth; and could be maintained on different formats, such as written, electronic or web-based

portfolios. LinkedIn and such social portfolio web sites have become very popular for presenting personal credentials and connecting with peers.
The aim of this paper is to examine a purpose of an electronic portfolio as a lifelong learning tool. Paper focuses on modeling a conceptual model of e-portfolio system as a tool for assessment and recognition of prior learning in continuing education for librarians in Croatia. Paper is reflecting on findings of research project "Lifelong learning for librarians in Croatia" financially supported by The National Foundation for Science, Higher Education and Technological Development of Republic of Croatia.

## Learning e-portfolios

Empirical researches on use of e-portfolio are very limited and focus more on its technical development. Helen C. Barrett (2005) gives a review on theoretical issues related to e-portfolios, and their use in education or academic research. Definition of learning e-portfolio differs in scope of its purpose, use, intended users or technical constraints.
Herman and Winters (1994) define well-designed educational portfolios as "representing important, contextualized learning that requires complex thinking and expressive skills. Traditional tests have been criticized as being insensitive to local curriculum and instruction, and assessing not only student achievement but aptitude. Portfolios are being heralded as vehicles that provide a more equitable and sensitive portrait of what students know and are able to do. Portfolios encourage teachers and schools to focus on important student outcomes, provide parents and the community with credible evidence of student achievement, and inform policy and practice at every level of the educational system."
Barrett and Wilkerson (2004) proposed a new taxonomy of electronic portfolio systems already in use in HEI:

- portfolio as a digital archive of learner's work
- portfolio as a learner-centered electronic portfolio and
- portfolio as an institution-centered database, or assessment management system, to collect administrative assessment data based on tasks and rubrics.

Portfolio system's market is every growing and Batson (2002) describes "e-portfolio boom" as follows: "We seem to be beginning a new wave of technology development in Higher Education. Freeing student work from paper and making it organized, searchable, and transportable opens enormous possibilities for re-thinking whole curricula: the evaluation of faculty, assessment of programs, certification of student work, how accreditation works. In sort, ePortfolios might be the biggest thing in technology innovation on campus. Electronic portfolios have a greater potential to alter higher education at its very core than any other technology application we've known thus far."

Kimball (2005) surveys trends in the e-portfolio boom, comparing technical constraints of portfolio systems to portfolio pedagogy. He states that majority of standardize database-driven portfolio systems is lacking in technical functionalities for meaningful reflection over learning artifacts. Self-reflection protects e-portfolio to turn to be a humped with of various elements, or just an enumeration of facts. Database portfolio systems employing either too much standardization or too much flexibility, and in both cases portfolio risks to missing the pedagogical target. Making a portfolio should be an imaginative, creative and rhetorical act, not merely a form to fill out. On the other hand, too much flexibility would fail to give adequate pedagogical guidance. Kimball concerns about privacy and ownership over electronic or web portfolios. The movement toward portability and persistence of created portfolios throughout creator's lifecycle, rises problems concerning access rights and ownership over portfolios.

## Portfolio for assessment

Portfolio is more that just a tool for recording professional or personal progress over time. 'Suitcasing' of qualifications is just not enough in highly demanding information era where professional and personal competencies develop daily. E-portfolios have advantages over hard copy portfolios by being easily accessible, having the capability to store multiple media, being easy to upgrade, and allowing cross-referencing of learner's work (Johnson et al., 2006). Lin (2008) reviews theoretical findings to foster development of e-portfolios as a learning strategy, as a reflective and also technical tool. In Higher Education e-portfolio has been used as an alternative assessment tool. Principles of learner-active, experiential learning were central to the portfolio approach and foster the use of portfolios for the purpose of assessment and personal development plan.

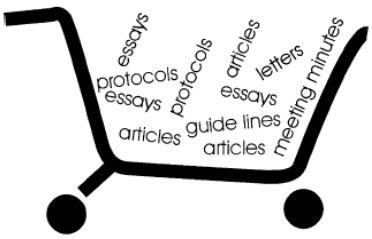
Envisaging assessment methods, Endacott et. al. (2004) concluded that portfolios are a system under development. This development affects portfolio assessment. Analyzed data revealed four approaches of the structure and use of portfolios: *shopping trolley*; *toast rack*; *spinal column* and *cake mix*. All four structures (Figure 1) imply the effectiveness of portfolios in assessing learning and competence.

The *shopping trolley* is similar to suitcase, repository for artefacts collected during the course. There is little cohesion evident in the portfolio, and little attempt to link evidence to learning outcomes or competencies. *Toast rack* is made up of discrete elements (the toast); assessing different aspects of practice and or theory, for example, sills log or reflective account. This elements remained separate even if the binder simply acting as a convenient device for keeping the elements in one place. There is no overarching narrative to connect the various sections, and different people may participate in the assessment of the various sections. The portfolio itself may or may not be assessed or reviewed. The *spinal column* is structured around practice competencies or learning outcomes (the 'vertebrae' making up the central column), and evidence

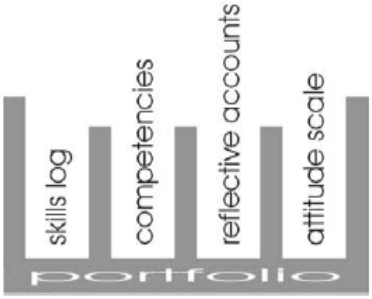
is slotted in, to demonstrate how each competence is been met. Reflective accounts could consider over more than one competency, and act as a linking flesh. The emphasis is on the original work of learner, whiles the evidence was used to support or illustrate the case being made. The *cake mix* approach is chosen when evidence from theory and practice is integrated into the portfolio and the whole (‘cake’) is assessed. Narrative form combines elements. Reflective commentary is aimed to demonstrate the learner’s critical and analytical skills by considering how they achieved what they have, how the evidence supported this, and what they had learnt. Form of a *cake* is a sum of its individual parts, and it is the whole that is assessed rather then the ingredients. Reflectivity, practice and professional development is likely to be features of this model.

Figure 1. Four models of portfolios by Endacott et al. (2004)

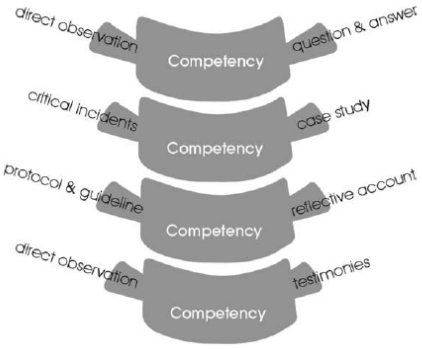
The shopping trolley model




The toast rack




The spinal column




Cake mix


Source: Based on R. Endacott et al. (2004, 252-253)

Use of an e-portfolio in Higher Education has been a great technological innovation. For learners, e-portfolio fosters deep-learning, emphasizes reflective learning practice and involves learners to take more active role in learning. For academics, e-portfolio is an assessment tool for both summative and formative assessment methods, a means for assessing learner's achievements and progress. But, how an e-portfolio could be use in continuing professional education?

Continuing professional education differ from HE accredited programs in many ways, mainly in lacking formative quality control, its voluntary mode of use and lack of external accountability. CPE program tends to maintain and enhance the knowledge, expertise and competence of professionals throughout their careers, according to a plan formulated with regard to the needs of the professional, the employer and society. Recent radical changes that affect workforce security in 'job for life' expectation, have multiple outcomes on career planning. Professionals accept reality that there is no safe job, that they must be open to potentially multiple careers instead of deeper specialisation in one single field, and that they must plan their own portfolio careers with horizontal development with little opportunities for vertical hierarchical promotion (Middlehurst and Kennie, 1994). In that regard, continuing professional education, which links education and practice and aims to maintain competence to practice become essential for professional survival. Providers of CPD will take substantial steps toward creation of mandatory and structured CPD opportunities for professionals. CPD portfolios or PDP (professional development plan) portfolios are wildly recommended for professional in all range of sectors.

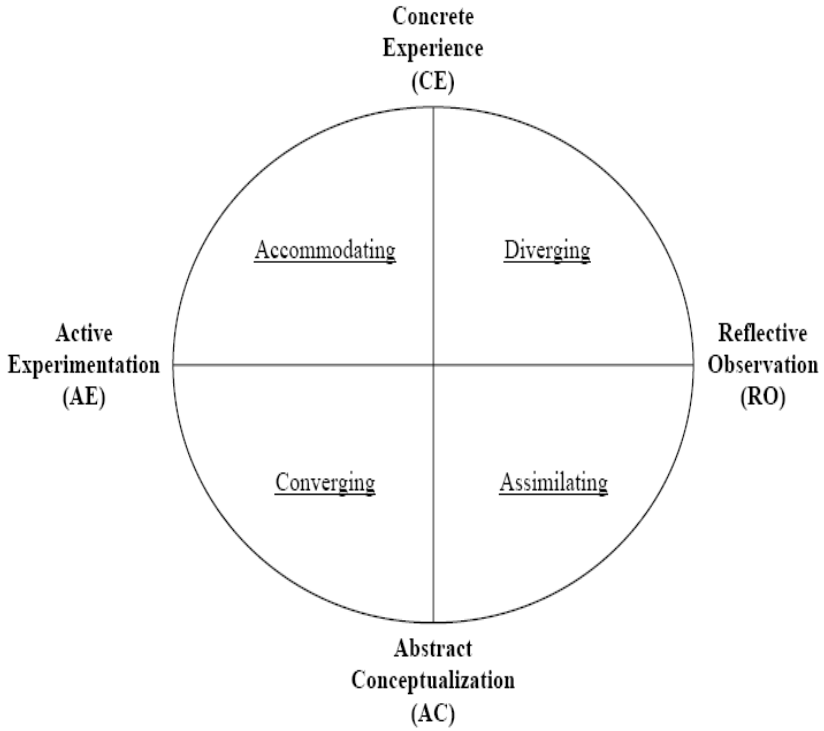
As a showcase of using portfolio for assessment of professional development in library and information science is a case of licensing membership for Chartered Information and library professionals (CILIP) in Great Britain. Watson (2008) indicates that portfolio, apart from being a means of presenting learner's professional competence, is a tool for personal development planning, concerning with current or future job roles, and should include additional activities that the individual undertakes outside the work-based environment. She discusses the portfolio assessment criteria which are in use in CILIP, stressing five objectives of good portfolio practice: reflective writing, curriculum vitae, professional development plans, professional statements and supporting evidence.

Reflective writing is the formal, or informal, recording of learner's thoughts. Most writers on reflective practice refer to Kolb's model of experiential learning (Figure 2).

Experiential learning is learning from direct experience. The idea of experiential learning is old as a Confucius sentence "tell me and I will forget, show me and I may remember, involve me and I will understand." Theory of experiential learning is heavily drawn on the work of John Dewey, Jean Piaget and D. A. Kolb. Learning from experience is individual process of a learner, process that requires or involves no teacher. But to transform an experience to knowledge,

some abilities are required: the learner must be willing to be actively involved in the experience; the learner must be able to reflect on the experience; the learner must possess and use analytical skills to conceptualize the experience; and the learner must possess decision making and problem solving skills in order to use the new ideas gained from the experience. Reflective writing about learning experience must go beyond descriptive writing, and must be evidence of learner's reflective thinking, his ability to analyse and synthesize.

Figure 2. The Experiential Learning Cycle and Basic Learning Styles (Kolb, 1984, 141).




Source: David A. Kolbe et al. (2000, 39)

Curriculum vitae (CV) is also an important statement about how we see ourselves and what image of ourselves we project to others. A good thought-through CV enables to identify key moments in career and in development.
CPD is not something that just happens in an unstructured way, but that we should be pro-active in determining our goals and how we are going to reach them. CPD should be planned in such a way that our knowledge and skills are enhanced and improved by a program of varied developmental activities. Pro-

fessional development plans provides direction, sets out objectives, identifies potential areas of development.
Personal statements are the most important element in CPD portfolio. It is like an executive summary in a report. By reading personal statement it must be immediately understood what portfolio presents. With personal statements we show evidence of analysis, evaluation and review of our knowledge and experience.
Supporting evidence is the largest part of CPD portfolio. Over the qualifying period we will collect a great deal of evidence of our professional developmental activities. There could be lot of types of evidence. The common items include: reports, published articles, presentations, performance reviews, photograph of exhibitions, minutes of meetings, notes from conferences or visits, web pages, diaries, blogs, letters, etc.
Based on developed practice of using portfolio assessment of prior learning, CILIP is heading to implement e-portfolio in near future.

## Constructing a model

Library and information professionals in Croatia are aware of great importance of continuing education. As Herbert White (1986) has noted, academic degree is not so much a qualification for a particular position, as it is a qualification for entry into the profession. Affected by dynamic environment of information technology, scientific innovations and mass production of information, library professionals change their professional routines and challenge daily their knowledge, skills and competencies. New knowledge and new skills is being learnt by evidence-based and work-based learning, non-formal and informal learning.
National program of continuing education for librarians and information professionals is Croatia is provided by The Training Centre for Continuing Education for Librarians in Croatia, founded in 2002 in National and University Library. Centre has been financially supported by Croatian Ministry of Culture. Program board, responsible for annual program scheme, consists of representatives from cofounder's institutions: National and University Library, Information Science Department of University of Zagreb Philosophy Faculty, Zagreb Public Libraries and Croatian Library Association. Short, one-day courses tend to refresh prior knowledge as to further new library or information skills. Program is developed for all types of library; academic, special, public and school libraries; and organized in Zagreb as in anther eighteen cities in Croatia. Last year, in 2008, Centre profound 500 hours of education for 1570 participants. At accomplishment of course, participants receive Certificate for participating in continuing education. Any kind of assessment or recognition process has not been in use. Library schools in Croatia also provide continuing education programs, and several major conferences are organized every year around LIS topics.

By one-year research project, "Lifelong learning for librarians", granted by The National Foundation of Science, Training Centre will integrate outcomes-based education in defining competency-based learning outcomes at unit and program level. Learning outcomes are statements about what learner will know, understand or be able to do after accomplishment of a learning program. Outcomes-based CPD will ensure high and sustainable competence standards for the library and information profession. Competence standards will help librarians to plan their professional development, to take an active part in continuing learning. On the other hand, competence standards will serve as a reference base for assessment and self-assessment process.

In summer 2009 University computer centre – SRCE implemented Moodle Community learning management system with integrated Mahara e-portfolio module. While SRCE maintains Moodle for academic users from University of Zagreb, Moodle Community was designed to serve large public community. One of stated aims declared by its E-learning Strategy (2007) is to establish and maintain an e-portfolio system - "a system of unified (interoperable) recording of the qualifications and experiences obtained in the course of education should help students not only in achieving mobility during their studies, but also in getting adequate jobs and starting professional career." Moodle Community LMS would be most suitable for librarians to participate in professional development e-learning program regardless of type of library he/she is working in. SRCE maintained Moodle learning management system for academic community, while Moodle Community would be suitable for community in general. Similarly, Minnesota's *eFolio* allows citizens of Minnesota to create a 'living showcase' of their education, career and personal achievements.

Maintaining a CPD Portfolio based on pro-forma e-portfolio system Moodle Community would be recommended for all participants in Training Centre who wish that theirs learnt knowledge in CPD will be assessed. Assessment criteria will be in use to assemble and display, for verification, the annual evidence for CPD of librarians. Self-presenting with a CPD e-portfolio is evidencing rather yours professional competencies than just your qualifications.

## Conclusion

Last ten years e-portfolios have been used in Higher Education as an alternative assessment tool. Despite of diversity of theirs structure and use, e-portfolios are means of presenting learner's professional competence, and a tool for personal development planning. Using an e-portfolio to maintain a record of achievements enables learner to reflect upon experiences and plot a development path for skills that will also help support learner in planning further career path. In the library and information profession the range of skills acquired by professionals can be extremely broad and may include aspects of building management, finance, personnel management, computing, teaching as well as some of the more traditional skills such as cataloguing and information retrieval. This

diverse set of skills makes the use of an e-portfolio by each individual imperative, to enable one to keep track of one's development in all areas. E-portfolio enables deep learning, evidence-based reflective writing, and summative and formative assessment. Like lifelong learning tool, e-portfolios gather evidence of learning experiences in formal, informal and non-formal learning. Thus, it is a most suitable for creating professional development plan and making records of continuing education.

National CPD program for librarians will integrate outcomes-based education in defining competency-based learning outcomes at unit and program level. Outcomes-based CPD will ensure high and sustainable competence standards for the library and information professionals in Croatia. Application of quality and competency standard will serves as a reference base for assessment of prior learning, which for an e-portfolio is its most appropriate tool. Recommendation of theoretical findings in aspect of assessment of prior learning and e-portfolio provide valuable information how to implement assessment of prior learning practices and procedures in CPD program. Implementation of CPD e-portfolio for library and information professionals in Croatia would be turn to more structured, mandatory and competency-based continuing professional development opportunity.

## References

Barrett, Helen C. Researching Electronic Portfolios and Learner Engagement: The REFLECT Initiative. 2005. http://electronicportfolios.org/reflect/whitepaper.pdf (2009-08-07)

Barrett, Helen C. ; Wilkerson, J. Confliction Paradigms in Electronic Portfolio Approaches. (2004). http://electronicportfolios.org/systems/paradigms.html (2009-08-07)

Batson, Trent. (December 2002). The electronic portfolio boom: What's it all about? // *Syllabus Magazine*. http://www.syllabus.com/aricle.asp?id=6984 (2009-08-07)

Chang, Chi-Cheng. Enhancing self-perceived effects using Web-based portfolio assessment. // *Computers in Human Behaviour*. 24 (2008), 1753-1771.

E-learning Strategy : 2007-2010. 2007. http://www.unizg.hr/fileadmin/rektorat/dokumenti/eucenje_strategija/University_of_Zagreb-E-learning_strategy.pdf (2009-08-07)

Endacott, R. et al. Using portfolios in the assessment of learning and competence: the impact of four models. // *Nurse Education in Practice*. (2004), 4; 250-257.

Herman, J ; Winters, J. Portfolio research: A slim collection. // *Educational Leadership*. 52 (1994); 48-55.

Jacobson, Wayne ; Sleicher, Dana ; Maureen, Burke. Portfolio assessment of intercultural competence. // *International Journal of Intercultural Relations*. 23 (1999), 3; 467-492.

Johnson et al. Developing portfolios in education: A guide to reflection, inquiry, and assessment. San Francisco : SAGE Publications, 2006.

Kimball, Miles. Database e-portfolio systems: A critical appraisal. // *Computers and Composition*. 22 (2005); 434-458.

Kolb, David A. ; Boyatzis, R. E. ; Mainemelis, C. Experiential Learning Theory: Previous Research and New Directions. / R. J. Sternberg and F. F. Zhang (Eds.), Perspectives on cognitive, learning, and thinking styles. New York : Lawrence Erlbaum, 2000. URL: http://www.learningfromexperience.com/images/uploads/experiential-learning-theory.pdf (2009-08-07)

Lin, Qiuyun. Preservice teachers' learning experiences of constructing e-portfolios online. // *The Internet and Higher Education*. 11 (2008), 3-4; 194-200.

Middlehurst, R. and Kennie, T. (1997) Leading professionals towards a new concept of professionalism, in J. Broadbent, M. Dietrich, and J. Roberts, (eds), The End of the Professions? The restructuring of professional work (London : Routledge), pp. 50- 68. Cited by Watkins, Jeff. UK professional associations and continuing professionl development: a new diection? // *International journal of lifelong education*. 18 (1999), 1; 61-75.

Watson, Margaret. Building your portfolio : the CILIP guide. London : Facet Publishing, 2008.

White, Herbert S. The Future of Library and Information Science Education. // *Journal of Education for Library and Information Science*. 26 (1986); 174-181.

Original scientific paper

# Marvin – A Conversational Agent Based Interface for the Study of Information Sciences

Nives Mikelić Preradović
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
nmikelic@ffzg.hr

Damir Boras
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
dboras@ffzg.hr

Sanja Kišiček
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
smatic@ffzg.hr

**Summary**

*In this paper we present initial results of the ongoing project, building a Conversational Agent (chatterbot) - Marvin. Marvin is designed to simulate intelligent conversation with students of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb. It is capable of providing basic feedback to students via textual methods.*
*The primary goal of this work is to inform the user of points of interest, to provide support, capture data from the user and promote study of information sciences. Furthermore, the goal is to enhance the presentation of information to students, especially information regarding the undergraduate study, obligatory and elective courses, ECTS (European Credit Transfer System) points and exams. Finally, the objective is to teach the students how to improve the quality of user experience using human-like conversational agents.*

**Key words**: Conversational agents, chatterbots, AIML, Information Sciences curriculum, ECTS points

## Introduction

Chatterbots are virtual characters capable of engaging a human counterpart in a meaningful conversation, often encountered as interfaces to help systems and

web-based search engines. They are systems that have the ability to parse natural language questions and, by referring to a knowledge base, generate natural language answers.

One of the most popular uses of the chatterbots is in the online tutoring process, by the application of online tutoring system based on dialogue [4], [5]. These chatterbots offer the possibility of control of the tutoring process itself, by monitoring the course of the learning process, recording the major characteristics of the process, and finally correcting it.

Although current systems lack the communication capabilities of the real human being, it has been shown [6] that even fairly simple chatterbots can increase the quality of experience for the user of interactive services and applications. In this work we explore the possibility of enhancing the quality of experience for the student using a virtual tutor Marvin.

The first version of Marvin is available at the following web address[1]: http://pandorabots.com/pandora/talk?botid=f98d56804e374cba.

Marvin was not designed to be a tutor that should actually teach particular subject matter, but to give an impression of a personal and human service that would increase the students' interest in the Study of Information Sciences, ease the search for particular information and boost satisfaction with the presentation of information.

We begin with a brief summary of related work and continue with an overview of AIML programming language, followed by the description of the system and its behavior. Finally, we discuss our future work and planned improvements to the system.

## Related work

For the last decades, computer users have been witnessing new paradigms in human computer interfaces technology. The arrival of new gadgets and technologies is boosting the realization of friendly and easy-to-use interfaces.

Many systems for the English language have already been developed in this research area, but none for the Croatian language. A few of the chatterbots for English are presented in this chapter.

ELIZA [7] was the first chatterbot written by Joseph Weizenbaum between 1964 and 1966. It was a simulation of a Rogerian psychotherapist operated by processing users' responses to scripts. It rephrased the user's statements as questions and posed those to the "patient".

PARRY [2] was another famous early chatterbot who attempted to simulate a paranoid schizophrenic. It was designed in 1972 by psychiatrist Kenneth Colby.

[1] http://www.pandorabots.com is a software robot hosting service that allows to create and publish robots on the web from any browser. Together with Oddcast Inc.'s VHost™ platform, it allows publishing of flash-based interactive characters onto web sites, Intranets and mobile devices.

Although today one can find chatterbots that act as very advanced tutoring systems with sophisticated knowledge bases [1], most of the existing systems are focused on the quality and accuracy of information and the way it is presented to users, rather than the production of the knowledge base.
AIML is the XML dialect developed by Richard Wallace and a worldwide free software community between 1995 and 2002. It formed the basis for what was initially a highly extended Eliza called "A.L.I.C.E." (Artificial Linguistic Internet Computer Entity), which won the annual Loebner Prize Contest[2] for Most Human Computer three times and became the Chatterbox Challenge Champion in 2004.
The most popular online chatterbots based on AIML are Eliza[3] (both English and German version), Cypher[4](first Persian AIML bot), iGod[5], Kyle[6] (artificial intelligence bot which employs contextual learning algorithms), Shakespeare[7] bot, Ailis[8] (Italian bot) and Prelude[9] (a self learning bot with AIML support).

## System description

This paper introduces the concept of using virtual human characters to provide support, capture data from the user and promote the Study of Information Sciences. To realize the concept, a virtual chatterbot, named Marvin, is developed for the students of Information Sciences at the Faculty of Humanities and Social Sciences. Marvin offers a simple and user friendly interface.
Marvin borrowed its name and personality from the paranoid android, a fictional character in *The Hitchhiker's Guide to the Galaxy* novel by Douglas Adams.
Since the basic idea was to build a chatterbot that will serve as an information source for the undergraduate students at the Department of Information Sciences, the biggest part of its knowledge base is information on obligatory and elective departmental courses from all six semesters of study.

[2] The scientific point of the Loebner Prize Competition is not to fool the judges (as it is usually stated in the literature), but to design a candidate that has indistinguishable performance indistinguishable to any judge [3].

[3] http://www.denkwerkzeuge.ch/

[4] http://www.syavash.com/portal/projects/cypher-yahoo-messenger-bot-project

[5] http://www.titane.ca/concordia/dfar251/igod/main.html

[6] http://www.leeds-city-guide.com/kyle

[7] http://www.shakespearebot.com/

[8] http://ai-tech.com/showcase/

[9] http://prelude.lennart-lopin.de/

**AIML (Artificial Intelligence Markup Language) syntax**

Firstly, the AIML template was built and served as a basic structure for all the courses and was used and adjusted by all students working on this project.

The basic unit of knowledge in AIML is called a category. Each category consists of an input question, an output answer, and an optional context. The question, or stimulus, is called the pattern. The optional context portion of the category consists of two variants, called `<that>` and `<topic>`. The tag `<that>` appears inside the category and its pattern has to match Marvin's last utterance. Remembering one last utterance is important if Marvin asks a question. This tag was implemented to use the user's reply to point the conversation in the specific direction.

The `<topic>` tag appears outside the category, and collects a group of categories together. The topic may be set inside any template.

The AIML pattern language consists of words, spaces and the wildcard symbols _ and *. The words may consist of letters and numerals, but no other characters. The pattern language is case invariant. Words are separated by a single space, and the wildcard characters function like words.

AIML tags transform the reply into a mini computer program which can save data, activate other programs, give conditional responses, and recursively call the pattern matcher to insert the responses from other categories.

All categories with information regarding the specific course are grouped in `<topics>` that are reached by input string "*kolegij*" followed by the acronym of the specific course, e.g. *kolegij OZ* for the course *Organizacija znanja.* If the user needs the access to the list of acronyms, he has to ask Marvin for the specific semester of study in order to get the acronyms for that semester.

If the user wants to abort the conversation about the specific course, he exits the topic with the command *nova tema.*

In order to anticipate as many user's answers as possible in the input patterns, the synonyms (such as nastavnik, profesor, predavač) in all 7 cases of singular and plural were often used as keywords, as well as wildcards * and _.

**Recursion in AIML**

Marvin produces the same reply to many different query formulations that share the same or similar meaning. This was achieved by recursion technique.

When building Marvin, we decided to reduce many ways of saying the same thing to one category, which contains the reply:

```
<category>
  <pattern>NASTAVNIK</pattern>
  <template>OIT predaje Hrvoje Stančić. Zanima te što još
on predaje? </template>
</category>
```

```
<category>
  <pattern>_ PREDAJE </pattern>
  <template> <srai>NASTAVNIK</srai> </template>
</category>
<category>
  <pattern>PREDAVAČ</pattern>
  <template> <srai>NASTAVNIK</srai> </template>
</category>
<category>
  <pattern>_ DRŽI </pattern>
  <template> <srai>NASTAVNIK</srai> </template>
</category>
<category>
  <pattern>PROFESOR</pattern>
  <template> <srai>NASTAVNIK</srai> </template>
</category>
```

AIML implements recursion with the <srai> operator, so that the output depends not only on one matched category, but also any other recursively reached through <srai>.
There is a variety of applications for <srai> in AIML, since it is used to reduce complex grammar forms to simpler ones, to split the input into two or more subparts and combine the responses to each or to map different ways of saying the same thing to the same reply (synonyms). Also, this operator is used for spelling or grammar corrections and for detecting keywords anywhere in the input.

**Marvin's knowledge base and reply structure**

The course knowledge base covers the name of the lecturer, the course prerequisites (if any), the number of ECTS points for the specific course, the hours of lectures / seminars / labs per course, the status of the course (obligatory or elective) and the description of the exam. Also, Marvin gives information if a specific course is prerequisite to some other course or prerequisite to the graduate study. It also has built-in information regarding the profile of undergraduate study (such as the required number of ECTS points for each semester of the study, etc).
Apart from that, Marvin has built-in information regarding the Department profile, history and staff (names of the head of the Department, deputy head of the Department, Department administrator, Department librarian, working hours of the Department administration and library).
While coding this information, we did not use <topics> in AIML, but only <categories> that contain the corresponding keywords from the most probable user's queries in their entry patterns.

When providing the reply about the name of the teacher/lecturer, Marvin tries to boost communication with the user posing the question to the user such as: *are you interested in other courses taught by this lecturer?*
The communication flow from this point onwards depends on the user's answer (this part of communication is coded with the tag `<that>`).
Since the same question can be formulated in many different ways and since we cannot predict that user will pose the question in most frequent or in the shortest way, we had to use the wildcards * and _ with extreme caution. Therefore, for some keywords we decided to introduce several combinations with * and _, depending on the words preceding and/or following the keyword.
The exact content and the structure of Marvin's reply are different in two levels:
1) Every topic change triggers the content of the reply in a way that it attempts to follow the normal flow and nature of the communication (e.g. if user changes the name of the course in his utterance, Marvin will immediately change the content of the reply, providing the user only with the necessary information)
2) Most of Marvin's replies are enhanced with our personal comments, which are usually slightly humorous. The comments are designed with a goal to convince the user that the chatterbot is a student himself, who took all or at least some of the courses. The example of the comment is: *"Oh, in my time it was different…", "When I was taking this exam, it was 12 pages long, can you believe it?"*

**Implementing keywords**
Our goal was to write an AIML template which can be activated by the appearance of a keyword anywhere in the input sentence. The AIML categories that are built into Marvin are illustrated by the following example:

```
<category>
<pattern>TKO _ TAJNICA</pattern>
<template>Tajnica odsjeka je Nevenka Petak.</template>
<category>
<pattern>_ TKO _ TAJNICA </pattern>
<template><srai> TKO _ TAJNICA </srai></template>
</category>
<category>
<pattern> TKO _ TAJNICA _</pattern>
<template><srai> TKO _ TAJNICA </srai></template>
</category>
<category>
<pattern>_ TKO _ TAJNICA *</pattern>
<template><srai> TKO _ TAJNICA </srai></template>
</category>
```

The first category both detects the keyword when it appears by itself and provides the generic response. The second category detects the keyword as the suffix of a sentence. The third detects it as the prefix of an input sentence, and finally the last category detects the keyword as an infix. Each of the last three categories uses `<srai>` to link to the first, so that all four cases produce the same reply, but it needs to be written and stored only once.

**Conversation context**

In AIML syntax `<that>...</that>` encloses a pattern that refers to the bot's previous utterance. Specifically, if the chatterbot responds with a multiple sentence paragraph, the value of `<that>` is set to the last sentence in the sequence. A common application of `<that>` is found in yes-no questions:

```
<category>
  <pattern>DA</pattern>
    <that>ZANIMA TE ŠTO JOŠ PROF. TUDMAN PREDAJE</that>
  <template>Na preddiplomskom studiju prof. dr. sc. Tudman
predaje osim OZ-a i TIZ.
  </template>
</category>
```

This category is activated when the client says YES. The chatterbot must find out what he is saying "*yes*" to. If the bot asked, *"Are you interested in what prof. Stancic is teaching apart from this course?*" this category matches, and the response, *"Apart from OIT, he teaches the part of the ODOTIS course, as well as Informatics for Archaeologists",* continues the conversation along the same lines.

The AIML interpreter stores the input pattern, `<that>` pattern and `<topic>` pattern along a single path (e.g. `INPUT <that> THAT <topic> TOPIC`). When the values of `<that>` or `<topic>` are not specified, it implicitly sets the values of the corresponding of `<that>` or `<topic>` pattern to the wildcard *.

The first part of the path to match is the input. If more than one category have the same input pattern, the system will distinguish between them depending on the value of `<that>`. If two or more categories have the same `<pattern>` and `<that>`, system will choose the reply based on the `<topic>`.

**Symbolic reduction**

Symbolic reduction refers to the process of simplifying complex grammatical forms into simpler ones. Usually, the atomic patterns in categories storing Marvin's knowledge are stated in the simplest possible terms. In other words, we tend to prefer patterns like "*WHO IS THE HEAD OF THE DEPARTMENT*" to ones like *"DO YOU KNOW WHO THE HEAD OF THE DEPARTMENT IS*" when storing information about the head of the Department.

Many of complex forms were reduced to simpler forms using symbolic reduction:

```
<category>
<pattern>DO YOU KNOW WHO * IS</pattern>
<template><srai>WHO IS <star/></srai></template>
</category>
```

Whatever input matched this pattern, the portion bound to the wildcard * may be inserted into the reply with the markup `<star/>`. This category reduced any input of the form "*Do you know who X is?*" to "*Who is X?*"

One of the most useful applications of `<topic>` are subject-dependent "pickup lines", like:

```
<topic name="mediji">
<category>
<pattern>*</pattern>
<template>
<random>
<li>Gledas li televiziju?</li>
<li>Koje novine citas?</li>
<li>Koji radio slusas?</li>
<li>Koje portale pratis?</li>
</random>
</template>
</category>
```

Finally, apart from the above mentioned functions, Marvin also acts as a language tutor and corrector of spelling and grammar mistakes that students make. Here is the example of that function:

```
<category>
<pattern>UVIJET *</pattern>
<template>Mislim da si htio reći uvjet, zar ne?
</template>
</category>
```

## Evaluation

Marvin's evaluators were four most successful undergraduate trainee teachers in the final year of the undergraduate study at the Department of Information Sciences who tested out the knowledge base through dialogue with the chatterbot. They checked the database for accuracy, interpretation and relevance to the types and the level of questions being asked.

The discussion consisted of 50 inputs and answers and each evaluator spent 30 minutes on average chatting with the chatterbot. A different set of input sentences was used by each evaluator.

The evaluators chatted with Marvin from two perspectives: as the competent information scientists that they were, as well as pretend bachelor students in order to see how well the chatterbot interacted in the specific scientific field and overall knowledge.

Analyses were conducted post-hoc using the transcripts saved by the evaluators from all of their chat sessions.

From the pretend bachelor students' perspective, trainee teachers concluded that the chatterbot works best with single clause utterances, each exchange being treated virtually independently. When they moved to multi clause units, or look at exchanges which range over more than one turn, Marvin's limitations become much more apparent. In other words, although Marvin's database contains some very basic factual world knowledge, the depth of knowledge is too narrow to cope with open-ended conversations with humans.

From the competent information scientists' perspective, trainee teachers evaluated Marvin's answers based on the following categories: good answers, reasonable answers, and off topic answers, that seem to have little or nothing to do with the input clause.

Although 47% of the answers were classified as good answers and although 41% of the answers were classified as reasonable answers, evaluators concluded that clusters of knowledge about different topics that have been input into Marvin's knowledge database are at a somewhat superficial level, since Marvin has no actual knowledge of what it is talking about and cannot discuss a topic.

Since Marvin draws from his general knowledge database in terms of heuristics to avoid answering a question it in fact has no answer for, and suggests a new topic, giving the illusion that it actually has something to say about the new topic, the evaluators suggested improving Marvin's database with more factual world knowledge.

## Conclusion and future work

As part of our future work, we plan to modify Marvin to take on the personality and knowledge base of named individuals. If a sufficiently high-quality knowledge base is constructed, then there is scope in higher education for using such a chatterbot as a substitute for the expert academic.

Many students of information sciences experience problems in learning theory for the specific courses in both undergraduate and graduate studies. They usually find it very hard to get started with reading, need cartoon level introduction, leading on to more complex material and they find it hard to know what is relevant discussion, so seminar discussion often wanders off topic.

In this context, an accessible chatterbot with a knowledge base reflecting key areas of information science could make an important educational contribution.

In our future work we plan to develop a small number of knowledge bases for use in information science study, particularly in Knowledge Organization course and Theory of Information Science course. These will consist primarily of personality and knowledge linked with key information theorists. The knowledge bases will be designed to be sensitive to the knowledge levels of potential student users and to be open to explanatory questions. Such a chatterbot can be used as a general advisor or expert. Students who have a research question or

essay title that needs to be researched could use the bot to generate content that can be included in their assessed work. Conversation with the bot will be recordable and can then be cut and pasted into an essay. Students will be required to edit this information into a coherent essay, just as they would with information collected from texts.
Such chatterbot would use Wikipedia information to build its conversations and would offer links to Wikipedia articles in the field of information science that students wrote and submitted to Wikipedia. The result would be the development of AIML chatterbot that would use key information science knowledge bases and provide extra services and linked information website. This should produce a better understanding both of the form and type of content that best matches student needs and also information about the best ways in which the bots can be used educationally.

## Acknowledgement

This work was partly carried out within the research projects “Croatian Dictionary Heritage and Croatian European Identity” and “Public Knowledge Design and Management in Information Space” supported by the Ministry of Science, Education and Sports of the Republic of Croatia. Much of the work was done as a student project at the Faculty of Humanities and Social Sciences in Zagreb, Croatia.
The implementation and modeling was done by undergraduate students: Jasna Turković, Mirjana Horvacki, Eva Cukor, Bojan Kopitar, Tomo Šala and Julija Maksimović.

## References

Burleson, W. Affective Learning Companions: strategies for empathetic agents with real-time multimodal affective sensing to foster metacognitive and meta-affective approaches to learning, motivation, and perseverance : PhD thesis. Massachusetts Institute of Technologies, 2006

Guzeldere, G.; Franchi, S. Dialogues with colorful personalities of early AI. // *Constructions of the Mind: Artificial Intelligence and the Humanities*. volume 4 (1995), issue 2

Harnad, S. The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. // *SIGART Bulletin*. volume 3 (October, 1992), number 4; pp. 9-10

Kovacic, B.; Skocir, Z. Development of Distance Learning System Based on Dialogue. // *Proceedings of the IEEE Region 8 Conference EUROCON 2003 : Computer as a tool.* / Zajc, B.; Tklacic, M. Ljubljana, Slovenia : Faculty of Electrical Engineering, University of Ljubljana. volume 1 (22-24 Sept. 2003); pp. 224-228

Kovacic, B.; Skocir, Z. Formal Model for Distance Learning System Based on Dialogue. // *Proceedings of the International Conference ICT2001*. Bukurešt. volume 1 (July 2001); pp. 231-236

Massaro, D. W., In Spencer, P.E.; Marshark, M. (Eds.). A computer-animated tutor for language learning: Research and applications. Advances in the spoken language development of deaf and hard-of-hearing children (pp. 212-243). New York, NY: Oxford University Press, 2006

Weizenbaum, J., ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. // *Communications of the ACM*. volume 9 (1965), issue 1; pp. 36-45

Original scientific paper

# Virtual Cultures and Races in RPG as Educational Means of Multicultural and Multiracial Social Relations

Winton Afrić
Faculty of Teacher Education, University of Zagreb
Savska Cesta 77, 10 000 Zagreb, Croatia
winton.afric@ufzg.hr

**Summary**

*In this article the focus lies on the educational character of RPG treatment in different virtual cultures and races. In any RPG system the structure is being made out of elements none of which can undergo a change without effecting changes in other elements.*

*This is of great importance for this study as it represents the way cultures and races behave when interacting in social and cultural relations. We can safely say that when facing any virtual culture or race we will be facing a sheet of unique characteristics consisting of cultural and racial attributes and traits commonly ranging from physical, mental, social and cultural groups. In the concept of RPG so far, the difference between race and culture is in fact blurred because virtual cultures are commonly created as virtual races of a specific cultural type.*

*The form on which virtual cultures and races are most commonly made seems to be stemming from a widely accepted view that culture presupposes society. In this paper the" four level approach" is used to the study of human beings based on body, psyche, society and culture. The creation of virtual races and cultures is therefore directly reliant on the idea that the biological and psychological are setting constraints or limits on culture, as well as culture being understood as humanity's unique form of adaptation to meet needs that are simultaneously social and biological. All the changes that make any virtual race or culture into a truly unique element are in fact based on changing the human potential norm. Changes can be in any of the aforementioned four level fields or all of them. The starting point of creating a virtual culture is therefore most commonly a modification of human standards of physical, mental and social attributes and traits.*

*From any specific cultural or racial point of view any other type of behaviour may well seem abnormal but as elements of a system they are all in fact behaving in the same manner .This type of creativity, regardless of being the creator or merely the student of a virtual structure of cultures and races educates all agents about the characteristics of cultures and races in general. Trough the in-*

*sight given by the transparency of the model the main focus of the education naturally becomes the understanding of racial and cultural differences, but in this case it becomes uniquely transparent that despite their differences in attributes traits and constellations of characteristics, they are all in fact the same, seeing as they exhibit the same kinds of behaviour.*
*Dealing with RPG systems at any level of immersion from creator to participant all agents are therefore educated in understanding and dealing with racial and cultural differences as characteristics of equal value but different apparel, and in such understanding given the chance to understand cultural and racial respect and equality learning the nature of their differences resulting in acquiring racial and cultural tolerance and understanding.*

**Key words:** Role Playing Game, Race, Culture, Identity, Multiculturalism, Education

In modern history games have wrongfully fallen into the category of violence instillers and generally tools of bad influence. As it is with all games RPG (pencil and paper and computer ones alike) had also been branded in the same bad manner. This is however due to general misinformation and prejudice towards gaming as a way of spending time. Games however have from the early human days had a vital educational role, firstly to note in non formal education and later on in all forms of education. RPG being in its base a simulation model is a powerful educational tool but sadly it has not been harnessed in that manner, not even close to its full potential. As a multi agent virtual reality system RPG in fact develops social communication and intelligence, as well as the competence of living together. Becoming a part of a virtual environment and learning to live there participants learn to understand and accept the new and strange trough various forms of social interaction. Thus RPG is a potent tool which educates its participants in accepting multiculturalism and racial tolerance. One of the strong components of RPG as an educational tool in learning tolerance and virtue of living in general is the literary "story element". People used to listen to stories as hot media and learn from them how to lead a virtuous life, how to make good, socially acceptable decisions and generally behave in an honoured manner. Today, playing RPG puts us in the same educational environment where people where when listening to those stories but from a much more immersed cold media point of view. It is safe to say that the best education on how to live can be given trough a simulation of life, and RPG is a tool which does just that. It enables us to transmit instructions, using the system and its rules on how to live, behave as a part of a social community and in giving its participants the freedom of choice of behaviour gives first hand experience on social interaction in specific situations. The very obvious difference between interaction in an RPG system and real life comes however from the Game part of the name. In understanding that as it is with every game we know the rules of

the game to start with, but in real life we have to discover the rules as we grow as members of a society and culture. Being modelled as a simulation or rather virtualisation of the real world much of these rules and roles in social interaction remain the same, thus it is very transparent and easy to apply the "preparation for life" we gain from RPG into our everyday living. In any standard RPG environment agents are bound to come across many different virtual cultures and races. All of those regardless of how well or poorly known share basic attributes and traits as they are all made based on different types of cultural forms depending on the system they are a part of. Seeing as they are a part of a virtual environment (world) the logic they must abide by as a part of a network of corresponding elements is always drawn according to rules and standards acquired from real cultures and races and their cultural and social relations. Any virtual culture we can encounter is therefore as an element a part of the structure of the RPG system. When looking at social and cultural relations between virtual cultures we are drawn to the question, what kind of requirements should the model meet for it to be called a structure? As a question of methodology of science in general we can refer to Claude Levi-Strauss when saying that in order for virtual cultures and races to be applicable in an RPG system we are in fact looking at a structure that consists of a model that must exhibit the characteristics of a system. The structure is therefore being made out of elements none of which can undergo a change without effecting changes in other elements. This is of great importance for this study as it represents the way cultures and races behave when interacting in social and cultural relations.

When dealing with virtual worlds built using a RPG system any virtual culture, race, ethnic group or nationality will inside the system be an element of the model structure we are facing. Each of those elements will likewise be comprised of elements we see as specific attributes traits and characteristics. Each of those micro elements will stand out as a unique part of a whole due to its specific constellation of characteristics which will define it corresponding to any other element within the model structure. The first question that comes to mind is how to define virtual cultures in relations to virtual races, virtual ethnie and virtual nations. If we want to define any of the aforementioned elements we need in turn to look at their real life counterparts and how they are defined and understood over the course of history. "Cultures have been traditionally conceived as encompassing and firm spiritual boundaries, which define how their members view the world and other cultures. They have been introduced as social formations with unique structures (constellations of elements) and specific beliefs. They have been first and foremost defined by the way specific communities claim certain territory and how they follow the forms of social communication in their everyday life. The constitutive elements of culture are in tradition social forms like, language, myth, tradition, ceremonies, customs and self comprehension of a community. Members of a culture do not view their duty solely in terms of preservation of specific practices and symbols. They also feel inter-

connected and are guided by solidarity. The uniqueness of a culture is best shown when comparing it to a different culture“(Mesić 2006). Such understanding of a culture comes from looking at small communities. That is of vital importance when looking at virtual cultures because of the practice that a lot of the cultures we commonly meet in RPG systems are in fact either small communities or are modelled after them. We can safely say that when facing any virtual culture or race we will be facing a sheet of unique characteristics consisting of cultural and racial attributes and traits commonly ranging from physical, mental, social and cultural groups. In the concept of RPG so far, the difference between race and culture is in fact blurred because virtual cultures are commonly created as virtual races of a specific cultural type. Examples of the Dwarven and Elven race/culture come to mind as most common ones. The form on which virtual cultures and races are most commonly made seems to be stemming from a widely accepted view that culture presupposes society, society is based on individuals and individuals have both minds and bodies (Kroeber, 2006, 36).

If we look at Tylors (1909) definition of culture specific interest falls on the two different understandings of his definitions. Firstly there is a definition of culture as one of the defining attributes of any ethnic group or ethnic collective. It is viewed upon as a union of common belief, custom, value, and constructions of meaning, as well as the way of being shared by the members of such a collective. Therefore culture is understood as a vital ingredient of ethnic identification. A different approach is given by Clifford Geertz (2006, 236). According to him culture is a union of common knowledge, beliefs and values which form the basis for social, economic, political and religious institutions. By Geertz “culture is a historically transmitted form of meaning embodied in symbols, a system of inherited conceptions realised in symbolic forms used for communication, renewal and development of knowledge of living as well as their view of life” (Geertz, 2006, 236). This approach is of specific interest to us because it points out the unifying elements of culture and virtual culture alike instead of what sets them apart. It is of course an inherited human ability to exist as a cultural being. Every human being has a culture as well as a language and that connects them in their essence as humans. Likewise every virtual being following the same pattern will have a culture as well as a language regardless of how it is determined biologically. The “four level approach” to the study of human beings based on body, psyche, society and culture is directly applicable to the manner that races and cultures are generated and treated in any RPG system. The creation of virtual races and cultures is therefore directly reliant on the idea that the biological and psychological are setting constraints or limits on culture (Steward, 2006, 100; White, 2006, 107), as well as culture being understood as humanity's unique form of adaptation to meet needs that are simultaneously social and biological (Kroeber, 2006, 36; Malinowski, 2006, 88). Therefore we can deduce that any virtual culture is in its basic potential or norm

starting out as human. All the changes that make any virtual race or culture into a truly unique element are in fact based on changing the human potential norm. Changes can be in any of the aforementioned four level fields or all of them. The starting point of creating a virtual culture is therefore most commonly a modification of human standards of physical, mental and social attributes and traits. The resulting constellation of characteristics gives us a unique virtual culture/race which in turn exhibits its own standards.

It is safe to say that when generating a virtual culture as a part of an RPG system we are in fact offered a choice between various points of view and cultural understandings. If we should want a virtual culture modelled after small communities it is justifiable to look at it stemming from that specific point of view. The basis of generating any virtual culture is to follow a form which is unique to all the cultures and sets that specific culture apart from any other by its specific constellation of elements, like attributes and traits. Therefore it is transparent that all cultures even though each of them is unique in their own right are always of equal value. When facing any kind of cultural or racial interaction between virtual cultures it is very transparent, because of the transparency of the racial / cultural model itself that each culture will act according to their own standards as a result of its unique constellation of characteristics. From any specific cultural or racial point of view any other type of behaviour may well seem abnormal but as elements of a system they are all in fact behaving in the same manner. This is also true for real cultures, where naturally Benedict (2006, 77) comes to mind in observing that when the Kwakuti exhibit a constellation of elements which appears abnormal by western standards, this judgment is in fact invalid since the behaviour is normal by Kwakuti standards.

When looking at the terms virtual race and virtual ethnicity it is vital to ask ourselves of their specific meanings and definitions. The term race when viewed from an anthropological or sociological view is always tied closely to biological determinism. The problem being biological determinism states that social, economical and behavioural differences in human groups are defined by race and only afterwards class or sex. Therefore it states that their differences are defined by their biological heritage. The term race is falling largely out of practice of use due to biological determinism, and the way that term has been polluted in modern history. Therefore the term ethnicity is being largely used in races stead. If we look at this problem from inside a RPG system virtual reality we stand at a different ground completely. When looking at virtual races, cultures and ethnicity in any virtual world we can safely say that biological determinism has a rightful place of its own when defining all the elements (virtual cultures, races, ethnie and nations) in question of the RPG system.

At this point I would like to refer to J. R. R. Tolkiens Middle Earth setting. If we look at some of its denizens in terms of virtual race, virtual culture, virtual ethnie and nation and single out the very apparent ones we may as well focus on Elves, Dwarves, Orcs and Humans. Each and every one of these elements

(races) has a specific constellation of traits defining them as unique when compared to others. Each of them has undisputed biological heritage, and as being biologically determined it is safe to dub them virtual races in their own right. The thing to note here however is, that biological determinism even though in service here does not imply superiority or inferiority of any race in comparison to any other race, but instead it clarifies their differences in their equality. This is such due to the nature of the RPG system. If we look at any RPG system as a structure of elements any virtual race will follow the same form of elements it is comprised of. RPG systems, being in their nature models, are all about balance in any specific state. It is therefore a common practice to make controls of value using point systems (or similar) to ensure none of the elements (in this case virtual races) fall out of scheme, thus violating the system. Differences between races can be vast but for instance if a virtual race claims physical superiority over others they will be “lacking” in another field keeping them in check as a balanced part of the system. It all in fact comes down to uniqueness in difference and preference in liking. Although some may say that relying on biological determinism is in fact an act of approval of racism in its core, the approach we take at using biological differences as a means of diversification in fact teaches us the very opposite. We have to be able to accept a virtual setting where racism as a term is being purified from its polluted historical “real life use” and is used in an entirely different context: the context of acknowledging difference, and at the same time, the understanding of equal value. I would like to turn our attention to the racist myth of giving value to individuals or groups by measurement of intelligence as a value. In every RPG system to date in one form or another, Intelligence is represented as an attribute of every being. Intelligence is measured within the system and is in fact separating the daft from the acute the more capable from less capable. There are systems and books where we may encounter suggestions or examples of virtual races of inferior or superior intelligence in comparison to the “human” standard. As stated before though, the difference will always be kept in check balancing it out by adding another superior or inferior trait or value in accordance to any specific case. I have to pause here to notice that in most such systems like GURPS, DnD and similar Intelligence is most commonly treated as an equal attribute with many others. In this case I beg to differ in saying the RPG needs to reconstitute the value of intelligence as an attribute and model it accordingly for it to be kept in check and be proportionate to other attributes. In this manner I believe that intelligence as an attribute in RPG is mostly underrated and should be of greater value in comparison to other attributes. Some good examples have been made not to treat Intelligence as a single attribute in value but in fact dismember it into several ones. For example the D20 system institutes Wisdom as a separate attribute and others like White Wolf systems deal with Perception, Intelligence and Wits as mental attributes of a being. The reason why I take specific note of intelligence in regard to the racist question is not only due to the fact of physical anthropology and its

craniometry but also the psychological tests used to prove inferiority of other races or sex in comparison to the "great white male". Intelligence when measured (as value) for example in animals constitutes a boundary of understanding, a checkpoint for action and ability. It is vital to use such a scale in RPG for instance, for the sake of programming behavioural patterns in artificial intelligence. When looking at virtual race or culture intelligence however an important attribute is nothing more than an attribute, and if kept in proportion with other attributes and traits it by no means defines any superior or inferior choices but is again used solely as an element in generating a unique structure of elements. Therefore biological determinism in RPG is a school of taught which teaches us to understand and respect differences in a system of equal value. When talking about biological determinism we should mention the two main currents of scientific racism: the one which follows the idea of mono-genesis, and the other which follows the poly-genesis idea. Mono-genesis as such is very rarely present in RPGs as systems mostly due to the fact most virtual worlds are generated with a Poly-genesis structure in mind. When looking at Tolkiens Middle Earth, Dwarves and Elves are biologically completely different species. The fact remains that each school of taught is nothing more that a choice in an artistic approach when conceiving a virtual reality setting, and that each of them will work. In any case the system itself could support generating superior and inferior races (under the idea of artistic freedom, as well as what it can offer in terms of choices that exist), as such a model could suggest, but then it would no longer be a valid RPG structure or a model, and would loose its use as a system being more of a literary concept based around guidelines, than a true RPG.

From an educational point of view we are facing a situation where when making choices in creating a virtual reality we can conceive any possible setting but the truth remains, we learn to understand the concept of diverse elements of equal value, thus generating a valid RPG system. In the same manner a concept of diverse elements which can be graded into superior and inferior ones could exist in theory, then however it would loose one of RPGs defining characteristics. One of the main defining characteristics of RPG is the fact that it from its very beginning instituted measure and scale to represent power, ability, superiority and inferiority. The idea behind this system of measure is such that in order to advance in the system gaining power one must be facing the challenges suited for their own level. Since the first appearance of levels, they represented a measure in growth. In early games levels grew harder as the agent progressed, to measure the agent's skill at playing a game as it developed. RPG took that concept and turned it into a system measuring agents characters as well as every other aspect of the virtual reality surrounding them, doing so in proportion. So the very nature of RPG is to understand superiority and inferiority but to value equality and diversity, as there is no progress in personal (character / avatar) growth when dealing with only inferior elements. Progress in RPG is achieved most commonly trough the agent gathering experience by dealing with other

elements of the system (surpassing challenges) which are most commonly of proportional power to that of the agent. With this in mind I have to reflect on my mentioning Tolkiens Middle Earth, that is literary work and not an RPG system. It does however share the same basic principle as RPG. That principle is diversity equality, measure, proportion and it in fact behaves as a complex system of corresponding elements. For instance, the race of Hobbits although not as strong as Humans or Dwarves, or as wise as the Elves has the greatest strength of character, and thus a Hobbit is chosen to be the ring bearer. Another aspect of note is the fact that RPG systems establish a scale of proportions between elements according to the "human model". In a virtual reality looking firstly from a biological point of view, a calculated model of the most average human judging by its attributes, characteristics and traits is given as a starting point according to which any other being is measured. Likewise when generating rules for virtual race creation we are in general speaking of the most average representative of that race. To look back at the problem of what is virtual race in comparison to virtual ethnicity or virtual ethnic group, the answer is simple. If we look again at Tolkiens Middle Earth, Orcs are said to have been Elves who abandoned virtue (actually abandoning Elven culture) and as a result have "fallen" (changed the way they looked and acted according to their surroundings) and given artistic freedom have changed physically as well. So they were of same ethnic background but over the course of history they have biologically and culturally (de)evolved into a different virtual race.

So when defining virtual ethnicity the best approach would be to look at Kivisto (2002) when saying it is an "umbrella term". In that respect virtual ethnicity describes social boundaries which are constructed under the assumption of a common genealogy, cultural forms such as language (tongue of Mordor being described as some form of elvish), religion, customs, tradition, common history, folklore and common geographical history (according to Mesić,2006). The only problem in this understanding lies that we have to allow for biological diversity to exist under ethnicity as an "umbrella term" for this definition to be valid. Therefore race as a marker is understood as one of the attributes of ethnicity, such as religion. Virtual culture in that manner is a different thing.

This type of creativity, regardless of being the creator or merely the student of a virtual structure of cultures and races educates all agents about the characteristics of cultures and races in general. Trough the insight given by the transparency of the model the main focus of the education naturally becomes the understanding of racial and cultural differences, but in this case it becomes uniquely transparent that despite their differences in attributes traits and constellations of characteristics, they are all in fact the same, seeing as they exhibit the same kinds of behaviour. Dealing with RPG systems at any level of immersion from creator to participant all agents are therefore educated in understanding and dealing with racial and cultural differences as characteristics of equal value but different apparel, and in such understanding given the chance to understand

cultural and racial respect and equality learning the nature of their differences resulting in acquiring racial and cultural tolerance and understanding.

## Conclusion

RPG as any game has been viewed from many different standpoints as a harmful and generally bad companion to spend people's time on. This is however almost exclusively due to prejudice connected with games and violence in the first place. As a simulation model RPG is an excellent tool to be used in education. In both non formal and formal way RPG teaches that as a simulation of life RPG is preparing its participants educating them how to live. RPG develops social intelligence and ones competence of living together. Trough going trough countless social interaction situations and due to the nature of RPG being a model structure it teaches us multiculturalism, and racial tolerance. By participating in an RPG on any level we gain knowledge's on social structures, cultures, races, ethnie and nations and their interaction. This unique insight which we act within as participating in a cold medium situation, is teaching us the understanding of social and cultural diversification in viewing the different as specific constellations that form unique elements of equal value. Therefore RPG is a modern educational technology for promoting multiculturalism and racial tolerance.

This research was a part of main Scientific research named "Analytical Model for Monitoring of New Education Technologies for Long life Learning" conducted by Ministry of Science, Education and Sports of the Republic of Croatia (Registered Number 227-2271694-1699).

## References

Benedict, Ruth. The Individual and the Pattern of Culture, edited in Anthropology in theory, issues in epistemology by Moore, Henrietta L. and Sanders Todd, Blackwell Publishing, 2006

Geertz, Clifford. Thick description: Toward an interpretative theory of culture, edited in Anthropology in theory, issues in epistemology by Moore, Henrietta L. and Sanders Todd, Blackwell Publishing, 2006

Goffman, Erving. Frame Analysis: An Essay on Organization of Experience, 1974

Hakkarainen, Henri. Stenros, Jaako, Thoughts on Role Playing, 2003

Kivisto, Peter. Multiculturalism in a Global Society, 2002

Kroeber, Alfred L. The Concept of Culture in Science, edited in Anthropology in theory, issues in epistemology by Moore, Henrietta L. and Sanders Todd, Blackwell Publishing, 2006

Lortz, Stephen L. Role Playing, Different Worlds, 1979

Malinowski, Bronislav. The Group and the Individual in Function Analysis, edited in Anthropology in theory, issues in epistemology by Moore, Henrietta L. and Sanders Todd, Blackwell Publishing, 2006

Mesić, Milan. Multi Kulturalizam, Školska Knjiga Zagreb, 2006

Montola, Markus. Role Playing as Interactive Construction of Subjective Diegeses, 2003

Padol, Lisa. Playing Stories, Telling Games. Collaborative Storytelling in Role Playing Games, 1996

Park, Robert. Race and culture, 1950

Stenros, Jaako. Genre, Style, Method and Focus. Typologies for Role Playing games, 2006

Steward, Julian H. The Concept and Method of cultural Ecology, edited in Anthropology in theory, issues in epistemology by Moore, Henrietta L. and Sanders Todd, Blackwell Publishing, 2006

Tolkien, J. R. R. History of Middle Earth, volumes 1-5, 1992-2003

Tylor, Edward B. Anthropology: An Introduction to the study of Man and Civilization. New York: D Appleton

White, Leslie A. Energy and the Evolution of Culture, edited in Anthropology in theory, issues in epistemology by Moore, Henrietta L. and Sanders Todd, Blackwell Publishing, 2006

Professional paper

# Recommendation for a World Virtual School Project

Neven Sorić
American International School of Zagreb
Voćarska 106, 10 000 Zagreb, Croatia
neven.soric@aisz.hr

Sanja Kišiček
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
smatic@ffzg.hr

Damir Boras
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
dboras@ffzg.hr

**Summary**

*World Virtual School (WVS) is a project sponsored by the U.S. Department of State Office of Overseas Schools. It has gathered eight representatives from eight major international school regions to form a WVS Advisory Group. The group's objective is to conduct research in virtual learning environments (VLE) of overseas schools, to analyze and present the results in order to offer a WVS structure plan. The goal is to increase the usage of VLE systems throughout the world to make this project global having in mind its relevancy, accessibility, durability and affordability. Therefore, we bring a possible WVS project outcome, recommending a shared global VLE.*

**Key words**: WVS (World Virtual School), VLE (Virtual Learning Environment), Moodle

## Introduction

The usage of virtual learning environment in schools (VLE) has a great impact on education. While originally restricted to physical models,[1] nowadays, as the quantity and modality of different VLE activities and resources grows, the broader is the content for VLE management. There are many schools partici-

[1] Dillenbourg, P. Virtual Learning Environments, p.8.

pating in VLE interested in sharing. World Virtual School (WVS) is a project sponsored by the U.S. Department of State Office of Overseas Schools[2] to assist participating schools and regions in terms of curriculum quality and continuity, opportunities for collaboration, progressive professional development, and resiliency in the face of natural or man-made emergencies. This paper's objective is to determine how the World Virtual School Project could be realized. The World Virtual School initiative has gathered representatives from eight major international school regions, Central and Eastern European Schools Association (CEESA), Association of International Schools in Africa (AISA), Near East South Asia Council of Overseas Schools (NESA), The East Asia Regional Council of Overseas Schools (EARCOS), Association of American Schools in South America (AASSA), The Association of American Schools of Central America, Colombia, Caribbean and Mexico (TRI-A), The European Council of International Schools (ECIS) and The Mediterranean Association of International Schools (MAIS).

This unique project is based on the following assumptions:

- schools value the integration of relevant and effective online resources for their varied learning communities;
- schools value working regionally to collaborate and to share perspectives and methodologies and to consider and nurture best practices;
- schools value their integrity in the face of adversity and seek cost-effective and reliable means of assuring continuity of operations.

Building upon the success of the NESA[3] Virtual School[4], a large consortium of schools that are cost-sharing and co-managing an enterprise level Blackboard service, this World Virtual School initiative intends to gather representatives from eight major international school regions to focus on general principles and practices related to standalone and shared virtual learning environments, aiming towards the possibility of further developing regional consortia using online vehicles such as Moodle, Blackboard, or newly evolving Web 2.0 tools.

## The World Virtual School project

Idea for this project started four years ago while brainstorming after attending a successful NESA Virtual Science Fair[5] and seeing technology being accessible to most students in international schools worldwide. State department has or-

[2] http://www.lincoln.edu.ar/aassa/booklet/World%20Virtual%20School%20Project.pdf, by Ken Paynter, the WVS project facilitator.

[3] Near East South Asia Council of Overseas Schools (NESA)

[4] The NESA Virtual School (NVS) is a consortium of currently 19 NESA member schools cost sharing a single ASP installation of enterprise level Blackboard, augmented with Learning Objects building blocks.

[5] http://wvsgeo.org/drupal/node/4

ganized a meeting where eight representatives met and discussed possibilities of creating network of virtual schools all connected into one place, called World Virtual School.
We bring the conclusions from the meeting and recommendations for the WVS project

**Conclusions from the meeting**

1. The evolution of web communication tools is very rapid, therefore one can reasonably assume that within two years many current precepts may have changed. Accordingly, we should think and act on goal based principles and practices and we should assume that functional convergence (platforms, operating systems, browsers) is inevitable and will work to our advantage to eliminate some of what seem to be present day inhibitors.
2. Many international schools are eager to engage with virtual learning environments on some level. Although some schools have begun on their own, and some have started to act together, there may also be a large number of schools ready to take some sort of action, schools that would appreciate knowledge and advice that our group could generate.
3. Independently hosted consortia can allow organized and sustainable cost-sharing, transparent mutual access for collaboration, and improved resiliency. Some startup schools that might naturally tend towards an economic "stovepipe" (vertical only) installation, might ultimately benefit by building a philosophy and practice with horizontal collaborative components planned for in advance. Perhaps our group can help engender this understanding and establish functional knowledge from which to act accordingly.
4. Courseware, although perhaps the most obvious and transparent emulation of the overall school environment to its own community, seems to not necessarily be the best platform for strictly collaborative and flexible projects. We recognize also, that as student information systems move towards web platforms, there is an increasing competition within schools for various products to host and deliver a variety of web-based services (such as calendaring, grade book, posting homework, discussions, etc.), for example Power School[6] However, with rapid convergence and integration of an increasing variety of web 2.0. type tools, all of this will likely continue to change, and cannot be categorized or determined at this or probably any phase of this project.

**Recommendations for the WVS project**

1. WVS group should act as a knowledge building and advisory group, working as closely as possible with the Directors of eight international regions.

[6] http://www.powerschool.com

As knowledge is gained and as initiatives related to the scope of our overall objectives are developed, we will keep each other informed. The group of eight representatives will be henceforth referred as the WVS Advisory Group.

2. The members of the WVS Advisory Group (including the Directors of all eight regions) will be in contact, using the Blackboard WVS course at the moment (hopefully with improved wiki/blog functionality in the near future). We plan to meet yearly at JOSTI[7], although, especially as we are dealing with rapidly changing circumstances, the WVS Advisory Group will start to meet twice a year.
3. To gather baseline and trends data about the usage and needs for virtual learning environments within schools throughout the eight regions, the WVS Advisory Group has developed a survey. The WVS Advisory Group will work together to analyze data and to advise the regional directors regarding significant circumstances and/or trends.
4. The WVS Advisory Group will be refining this summary statement with recommendations for presentation to the Regional Directors on the yearly basis at their directors' meeting by the appointed representative of the WVS project.

## Survey on the usage of VLE in schools

The WVS VLE[8] Survey was one of the several outcomes from the WVS meetings at JOSTI 2007 conference. The primary goal was to gather baseline information that would help the WVS Group establish VLE practices in place, issues in implementation, and potential needs of schools in these regards. The survey was developed by the members of the WVS Advisory Group so as to be as clear and as inclusive as possible with regards to our objectives.

The survey was conducted via Free Online Surveys[9] from mid-September to early November 2007. 114 schools altogether from eight major international school regions conducted the survey (28 small, 53 medium and 33 large schools). The charts below illustrate some segments of the survey analysis regarding the VLE usage in small (enrolling up to 250 students), medium (enrolling 251-850 students) and large schools (enrolling more than 851 students).

Chart 1 illustrates the usage of the two most popular VLEs, Moodle and Blackborad, in small, medium and large schools, also indicating in what rate the schools use some other VLE or do not use any. Moodle is most frequently used in all schools, while Blackboard has the most significant rate of usage in medium size schools. Small schools basically use Moodle or do not use VLE at all, while medium schools mostly do not use any VLE as opposite to large schools mostly using VLE.

[7] JOSTI - Jefferson Overseas Technology Institute for American Sponsored Overseas Schools

[8] Virtual Learning Environment (VLE)

[9] http://freeonlinesurveys.com/

Chart 2 illustrates the frequency of usage of VLE in small, medium and large schools indicating that in medium size schools there is the highest rate of VLE usage, then in large and small schools respectively.

Chart 1: The usage of VLE platforms in small, medium and large schools

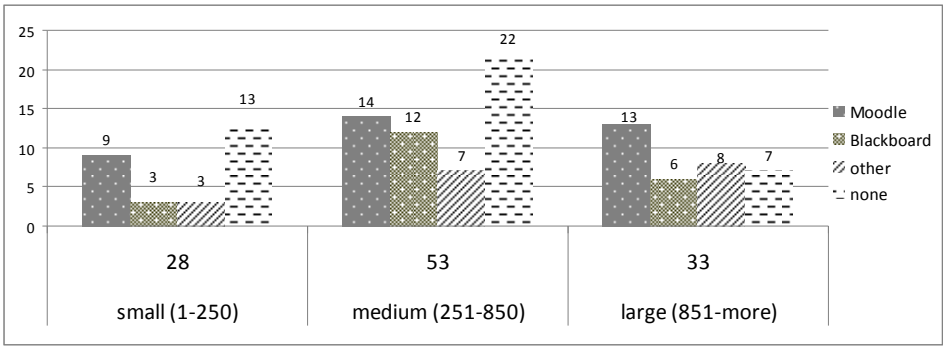
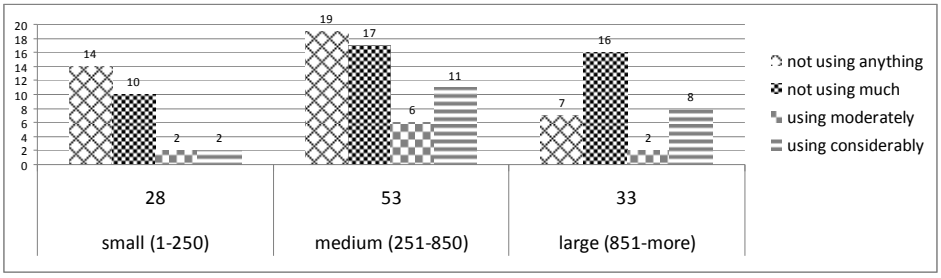



Chart 2: The frequency of VLE usage in small, medium and large schools




## Discussion

Considering foundational concepts, we have to make this project global. Way to do this is to increase the usage of VLE systems throughout the world having in mind that a VLE system has to be relevant, accessible, durable and affordable. For example, one of the topics we would like to explore is the idea of Moodle ASP (application service provider). The concept refers to a possibility to contract a Moodle integrator to set up a server, off-site, with Moodle service(s). An important issue is site security. If an online Learning Management System is a part of a security plan, for evacuations, or quarantines, for instance if a school has to be evacuated due to either internal or external reasons, its continuous operation is critical, therefore teaching and the entire process of education should be geographically independent. Having the system offsite, one does not have to approach the server physically to do system administration or solve problems.

Table 1: The big four VLE issues

| The Big Four VLE issues | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| RELEVANT | + | ACCESSIBLE | + | DURABLE | + | AFFORDABLE |
| ***Sustainable engagement at each school*** | | ***Ready authentication for all regional users*** | | ***ASP[10] / offsite / "neutral" hosting*** | | ***Best value/ ROI[11] short and long term*** |
| (programmatic integration at various institutional levels) | | (collaboration & professional development) | | (true resiliency and equitable costing and management) | | (leveraging group pricing and symbiotic opportunities) |

Also, having the system offsite guarantees continuity of education, because the system is not site-dependent in case of any physical damage. Offsite resources require only Internet connectivity for users to sustain activity. On the other hand, having the system onsite can sometimes guarantee liability. For instance, if a problem occurs, the school's support personnel do not have access to the system if the system is offsite, due to its connection with an outer server. If the system is onsite, users do not have to wait for the problem to be solved but they can solve the problem by themselves due to system being connected to a local server.

With Moodle being open source, one might ask why a school would pay for something that can be free? Firstly, all online services require an infrastructure and support which the end user does not see or feel, but the institution does. It costs time and money to provide reliable equipment, connectivity, network security, backup and timely and intelligent upgrades of security and operating system application(s). All of this comes into focus as the service itself becomes more and more mission critical with one of the Department's missions being continuity of education in schools, with daily teaching functioning on a regular basis.

We have calculated that if 15 schools were to share a Moodle ASP service (on which each of their schools could have their own Moodle site), the cost can be as little as about $1,000 year. Even more schools involved would bring the costs down.

Furthermore, there is an additional benefit to collaborating that we have not yet rolled into this discussion - course sharing and professional development opportunities. If schools are working closely together on their Moodle management, it is possible that users can be shared. Some of the technology to facilitate

[10] Application Service Provider

[11] Return On Investment

this is underway as we speak, but the concept refers to collaborating institutions that share their user and course base to some degree - allowing shared courses and professional growth opportunities. This kind of combined and secure ASP environment is the one we were able to create with the NESA Virtual School (using Blackboard) currently involving 16 schools in the NESA region offering stability to participating schools.

## Ongoing projects and plans for the future

CEESA Moodle Server is a pilot project that has started three years ago to provide smaller schools with VLE. The coordinator of the project is Neven Sorić, the coauthor of this paper, also a member of the WVS Advisory group representing the CEESA region. This project has become a role model for other regions, and an example how all international schools in the world can work in VLE, this being one of the main requirements for the WVS realization. The server on which eight virtual servers with separate Moodle instances are running each of them supporting different schools' needs, is stationed in Zagreb. An administrator takes care of backup, security, updates and all other requirements for Moodle operating for all eight schools sharing the cost of administration and server expenses. Three years from now, the schools have integrated VLE into their everyday school life where teachers and students all benefit from it. It was initially fully subsidized by the US State Department Office for Overseas Schools., and now it is financed only by the schools using it. We hope to be able to conduct some collaborative experiments soon, even without the ASP environment available to all schools around the world.

This is the 3rd year of the WVS project with three more ahead. We plan to meet again, with regional education directors joining us to hear more about our plans, and to contribute in realization of the WVS project. We also plan to develop and conduct more surveys to refresh the data, to see if any changes occurred, etc. Having all the theoretical ideas we needed, we have to move into practical realization of WVS. There is a VLE project created by CEESA which was able to purchase and set-up a server for small schools that use Moodle. The user schools are responsible for maintenance and provider costs. With one server running and one server administrator there are eight separate VLE running. That model is accepted, but in order for it to succeed, we have to prove that the service is relevant, accessible, durable and affordable. Therefore, a possible situation could be an IT company collaborating with the WVS. A chosen IT company would be responsible for installation and maintenance of all Moodle sites. It would also be important that the chosen IT company has some kind of credentials, for example being a Moodle partner[12] and providing 24 hour support that is essential as schools are situated over all time zones in the world.

[12] http://moodle.com/partners/

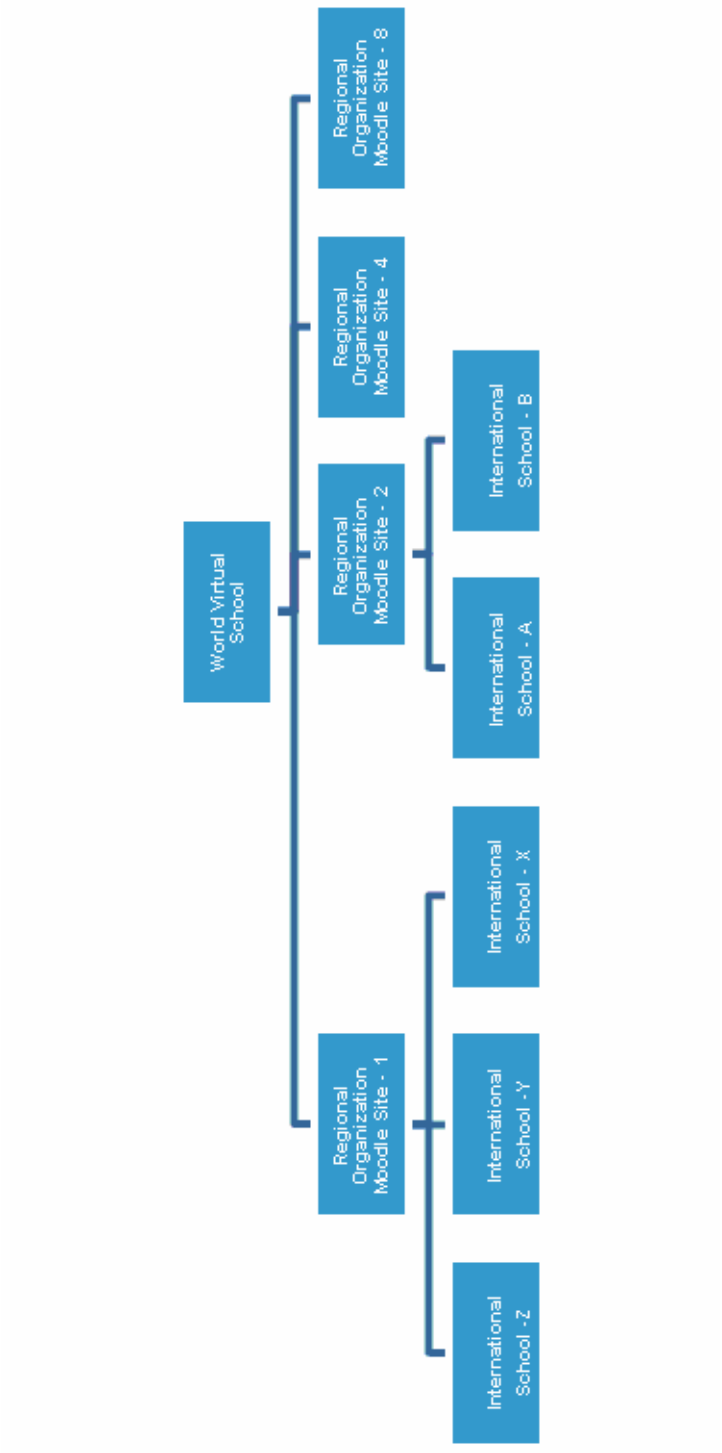



Chart 3: Hierarchical WVS structure plan

Structure of a WVS network, as shown in Chart 3, would be a hierarchy with WVS on top, regional VLE following, where courses of regional interest would be stationed, and in the end, separate VLE schools where curriculum of each single school would be running. WVS and regional environments would be administered by a regional point person, and each school would have their own VLE administrator. Advantage of creating a hierarchy model is the usage of users (teachers and students) from each school in all higher positioned VLEs. More students and teachers would start using VLE on a daily basis.

**Conclusion**

If we consider greater issues of access and reliability, we have to lean towards collaborating in a cost-share and offsite service with an IT company administering the system. This way we would provide the schools with a stable VLE system that requires as little administration for current school staff. Furthermore, having an IT company responsible for the functioning of the system would ensure stability and durability, and would eliminate the possibility of one or two enthusiasts running the project and leaving it behind.

However, the system has to be entrusted with a reliable company in order to preserve the principle of education continuation, suffering as less as possible system breakdowns, if any. We strongly recommend introducing an ASP, especially as these services become increasingly mission critical meaning that the continuation of VLE education in case of emergency is challenged. The system would be onsite, connected to an outer server requiring only Internet connectivity for users' activity. Realization of the WVS structure plan would significantly reduce support and administration expenses, since all the schools participating would share costs of a shared VLE.

## References

A Global Paradigm Shift in Science Fairs. http://www.nais.org/resources/index.cfm? ItemNumber =149604 (15th August 2009)

Dillenbourg, P. Virtual Learning Environments. // *EUN Conference 2000: Learning in the new millennium: Building new education strategies for schools.* University of Geneva, 2000, p.8

Moodle Partner. http://moodle.com/partners/ (21st August 2009)

NESA Science Fair. http://wvsgeo.org/drupal/node/4 (11th August 2009)

# E-SERVICES, E-GOVERNMENT AND BUSINESS APPLICATIONS

Original scientific paper

# Two Statistical Models on European and Croatian Information Society

Božidar Tepeš
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
btepes@ffzg.hr

Ivan Mijić
Diplomatic Academy, Ministry of Foreign Affairs and European Integration
Zagreb, Croatia
Ivan.Mijic@mvpei.hr

Krunoslav Tepeš
City Office for Transport Planning, City of Zagreb
Zagreb, Croatia
krunoslav.tepes@zagreb.hr

**Summary**

*The Council of the European Union (EU) defined information society (IS) with new Strategy i2010 for information and communication technology (ICT). Eurostat has European statistics for EU countries. Central bureau of statistics of Croatia and IDC Adriatics have numerical data benchmark indicators for information society for Croatia as candidate country. In our paper we use eight numerical data variables in two statistical models. Factor analysis model looks for the most important variables for information society. Causal structure model defines causal relation between these variables for development of information society in EU and Croatia.*

**Key words:** Information society, benchmark indicators, factor analysis and causal model.

## Information society

Among numerous attempts to define Information society (IS), we will use one which defines it as a society in which the creation, distribution, integration and manipulation of information is a significant economic, political, and cultural activity.[1] The knowledge economy represents its economic counterpart, where

[1] http://en.wikipedia.org/wiki/Information_society

wealth is created through the economic exploitation of knowledge.[2] Specific to this kind of society is the central position ICT has for production and economy.[3] Information society can be seen as the successor of industrial society.[4] The aim of the information society is to ensure the access to information to everyone. Bearing in mind the ever growing amount of information accessible in digital shape, it is important that each and every citizen is provided with an equal access to information, i.e. create the conditions in which the knowledge society benefits all, regardless of geographic, social, age or any other factor. The key activities to meet this goal are enabling equal access to the information society technologies to all social groups and continuous efforts of business sector and academic community to foster information knowledge throughout the entire population. From statistical point of view, IS is connected with ICT. In Guide to measuring IS, 2009. Organisation for economic and co-operation and development (OECD) presented possible conceptual model of IS.[5] Main elements of this model are ICT supply, ICT demand, ICT infrastructure, ICT products and content. ICT supplies are producers and production ICT goods and services. ICT demands are users and uses, households, individuals, and businesses. ICT infrastructures are investments and services. ICT products are price and quality. Content or flows of information are production, publishing, and electronic distributions. ICT demand and ICT infrastructure as part of IS are electronic commerce (e-commerce).

In EU the i2010 strategy is the EU policy framework for the IS and media. Presented it in June 2005 by the European Commission as the new initiative, it promotes the positive contribution that ICT can make to the economy and society,[6] The i2010 strategy has three main aims: to create a single European IS which promotes internal market for IS and media services, to strengthen investment and innovation in ICT research, and to support better public services and quality of life through the use of ICT.[7]

## Benchmark indicators

The eEurope is a political initiative that emerged to ensure that future EU generations maximize the changes the IS brings. These changes affect a vast array of factors and agents, create prosperity and share knowledge. That is why they have an enormous potential for enrichment. The good management of this trans-

[2] http://Ibid.

[3] http://www.it.iitb.ac.in/~prathabk/pages/tech_archives/global/post_industrial_society.pdf

[4] Ibid.

[5] http://www.oecd.org/dataoecd/25/52/43281062.pdf

[6] http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm

[7] http://ec.europa.eu/information_society/eeurope/i2010/strategy/index_en.htm

formation represents the main economic and social challenge, since it also involves serious repercussions for employment, growth and the greater integration of EU.8

In May 2002, the Commission (in view of the Seville European Council) presented the eEurope 2005 Action Plan. This provides policy actions for European institutions and member states to accelerate the development of the IS in Europe. In order to monitor progress of the Action Plan, it also contained proposals for a benchmarking exercise. They were based on a set of indicators which would be proposed by the Commission and endorsed by the Council.9

The Eurobarometer surveys used for several eEurope 2002 indicators provided rapid results (within 6 weeks of survey) and used a single methodology for all member states of the European Union. Greater use should be made of surveys undertaken by National Statistical Institutes (NSIs) and Eurostat, and additional ad hoc surveys. It should be stressed that candidate countries were invited to participate in Eurostat surveys from 2003 and additional surveys run by the Commission will be extended to candidate countries as soon as possible.10

The Commission proposes fourteen policy indicators and twenty two supplementary indicators (with their sources and frequency of collection).11 In addition, one policy and tree supplementary indicators are proposed for which pilot studies need to be carried out.

The following eight basic benchmarks were used in this article:

Percentage of households or individuals having access to the Internet at home (A1)

Percentage of individuals regularly using the Internet (A2)

Percentage of persons employed using computers connected to the Internet (B1)

Number of basic public services fully available6 on-line (D1)

Percentage of population using Internet to seek health (F1)

Percentage of enterprises total turnover from e-commerce (G1)

Percentage of enterprises with broadband access (J1)

Percentage of households or individuals with broadband access (J2)

## Factor analysis

Main idea of factor analysis method [1], [6] is to find smaller dimension space to analyze multi dimensional space of measured stochastic variables [6]. In IS we are looking at eight measured variables for EU countries from 2003 to 2008. Our measured variables are eight basic benchmarks (A1, A2, B1, D1, F1, G1,

[8] http://www.ine.es/en/docutrab/tic/inventario_in05_en.pdf

[9] http://ec.europa.eu/information_society/eeurope/2002/news_library/documents/benchmarking05_en.pdf

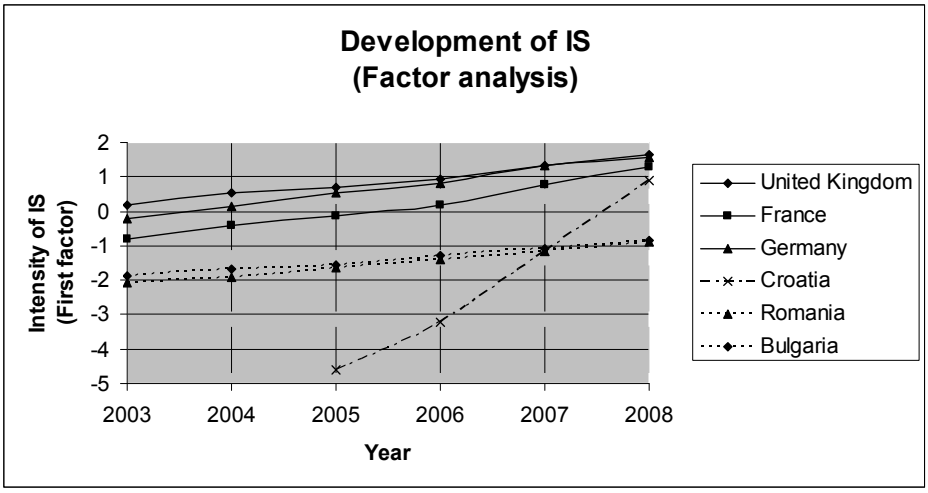
[10] Ibid.

[11] Ibid.

J1, J2). From Eurostat statistics, we are observing this eight measured variables for twenty seven EU countries for six years. We are using principle component method [1] for finding factors for describing IS. Using statistical software Statistica first factor or principle component describe 75%,of total variance of all measured variables. We named this first factor Intensity of IS (Y). The relation between first factor and measured variables is:

$$Y = 0{,}947A1 + 0{,}933A2 + 0{,}886B1 + 0{,}756D1 + 0{,}871F1 + 0{,}879G1 + 0{,}834J1 + 0{,}916J2$$

Software Statistica found factor score for every EU country in years we were observing. Scores for United Kingdom, France, Germany, Romania and Bulgaria are in Picture 1. Central bureau of statistics of Croatia and IDC Adriatics have numerical data benchmark indicators for IS for Croatia. We used this data for 2005, 2006, 2007 and 2008 with normalization and relation first factor and measured variables as before, and we found score for Croatia. You can see the results in Picture 1.




Picture 1. Development of IS (Factor analysis)

From factor analysis we can say that in 2005 IS in Croatia was far from IS in EU countries, in 2007 we were at level of negative development of IS countries Romania and Bulgaria, but in 2008 we were near development of IS countries United Kingdom, France and Germany.

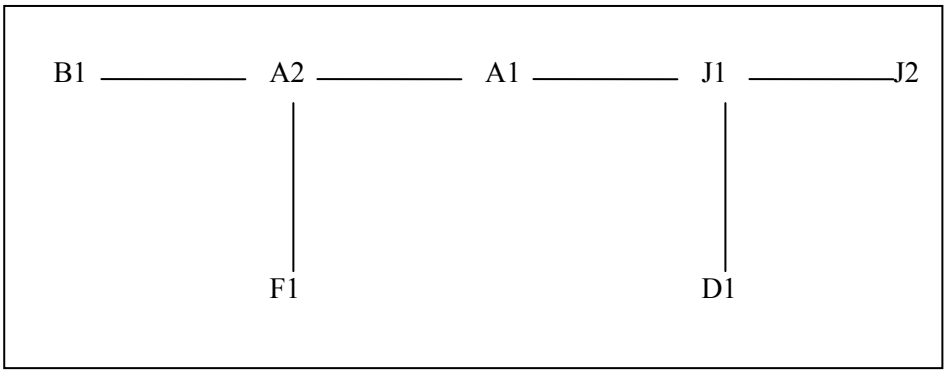
## Causal model

Causal statistical model [2], [3], [4], [5] use same measured data or eight benchmarks. In analyzing causal structure we are looking for partial correlation coefficients between measured variables. In first step we are looking for partial correlation coefficients between two measured variables and all others or $\rho_{ij\cdot kl\ldots s}$ where two variables are $ij$ and all other variables are $kl\ldots s \neq ij$ Results are in the following Table 1.

Table 1. Partial correlation coefficients

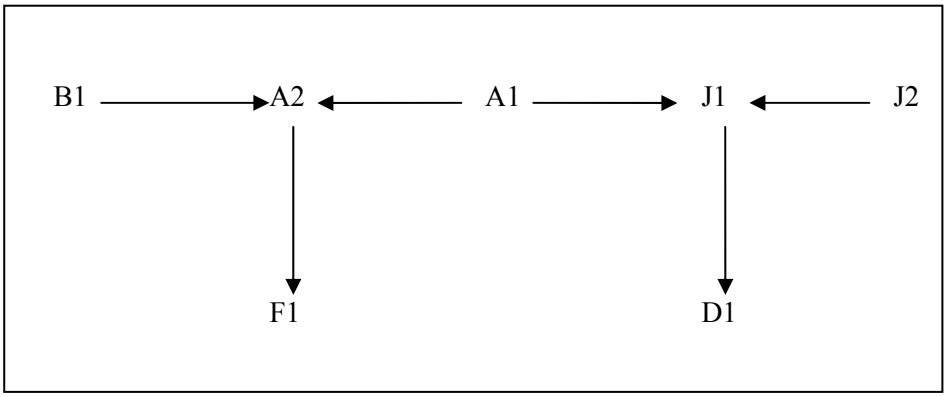
| Variable | A1 | A2 | B1 | D1 | F1 | G1 | J1 | J2 |
|---|---|---|---|---|---|---|---|---|
| A1 | - | 0,510 | 0,213 | 0,126 | 0,112 | 0,283 | -0,800 | 0,160 |
| A2 | | - | 0,312 | -0,184 | 0,345 | -0,159 | 0,067 | 0,275 |
| B1 | | | - | 0,154 | 0,075 | 0,237 | 0,152 | -0,288 |
| D1 | | | | - | -0,077 | 0,231 | 0,401 | 0,110 |
| F1 | | | | | - | -0,028 | -0,077 | 0,212 |
| G1 | | | | | | - | -0,121 | -0,148 |
| J1 | | | | | | | - | 0,467 |
| J2 | | | | | | | | - |

We are looking for six bold partial coefficients in Table 1. absolutely greater than $0{,}3$ or $\left|\rho_{ij\cdot kl\ldots s}\right| \geq 0{,}3$ and find causal structure [7], [8], [9] in Picture 2.




Picture 2. Causal structure IS

With more analysis [2], [3], [4], [5] we can find causal V structures $A1 \rightarrow A2 \leftarrow B1$ and $A1 \rightarrow J1 \leftarrow J2$ together with causal connections $A2 \rightarrow F1$, and $J1 \rightarrow D1$. Final causal model of IS for EU countries is in Picture 3.




Picture 3. Causal model IS

From causal model of IS we can see three most important causal measure variables for development of IS. These three variables are percentage of households or individuals having access to the Internet at home (A1), percentage of persons employed using computers connected to the Internet (B1) and percentage of households or individuals with broadband access (J2). It means that Internet at home and working place, together with broadband access at home, are the most important for IS. Second level for IS developments are two variables: percentage of individuals regularly using the Internet (A2), and percentage of enterprises with broadband access (J1). It means that second level is Internet for individuals research and enterprise with broadband access. At the third level there are two variables: IS percentage of population using Internet to seek health (F1), and number of basic public services fully available on-line (D1). They relate to public services and health, as one of the most important goals of public services. As we can see variable percentage of enterprises’ total turnover from e-commerce (G1) is not in causal model because e-commerce is ICT demand and ICT infrastructure as part of IS.

## Conclusion

Our research of IS form statistical point of view or statistical modeling of IS. Model show causal structure of IS and the most important goals for development of IS. Also, we can see position of Croatia in development of IS or knowledge society. Our model have only eight measured variables, and more realistic model of IS must have more variables. At the same time, period of research was only six year. Development of IS is connected with globalization and model of IS must include more countries than EU countries.

## References

[1] Hill, T., Lewicki, P., STATISTICS: Methods and Applications, StatSoft, 2005
[2] Pearl, J., Causal Diagrams for Empirical Research, Biometricka, 82(4), 669-710, 1995
[3] Pearl, J., The logic of counterfactuals in causal inference, Journal of American Statistical Associations, 95(450), 428-435, 2000
[4] Pearl, J., Causality, CAUSALITY: Models,Reasoning, and Inference, Cambridge University Press, 2001
[5] Pearl, J., Statistics and Causal Inference: A review, Test Journal, 12(2), 281-345, 2003
[6] Tepeš, B., Skripte iz statistike, Zagreb, 2009 (http://www.ffzg.hr)
[7] Tepeš, B., Predavanja iz statističkih modela na grafovima, Zagreb, 2009 (http://www.ffzg.hr)
[8] Tepeš, B., Predavanja i vježbe iz kombinatorike i grafova, Zagreb, 2009 (http://www.ffzg.hr)
[9] Veljan, D., Kombinatorika I diskretna matematika, Algoritam, Zagreb, 2001

Professional paper

# Compiling, Processing and Accessing the Collection of Legal Regulations of the Republic of Croatia

Tanja Didak Prekpalaj
Tamara Horvat
Diana Miletić
Dubravka Mokriš
Hrvatska informacijsko-dokumentacijska referalna agencija - HIDRA
Siget 18C, Zagreb, Hrvatska
Didak@hidra.hr, THorvat@hidra.hr, Diana.Miletic@hidra.hr, Mokris@hidra.hr

**Summary**

*Accelerated Internet development and expansion contributed to the growing number of regulations published in an electronic form. Due to that fact, one of the latest research objectives is development of new programs for searching through such documentation and development of digital collections. Croatian Information Documentation Referral Agency (in further text: HIDRA) systematically and regularly gathers publicly available official documentation in print and electronic form, and its goal is to gather documentation to enable simple user access to information from a single point. On HIDRA's web site there are fully available legal documents of the Republic of Croatia with all changes and amendments, bylaws and other data and relations accessible through hyperlinks, relevant to a regulation in question. Thanks to electronic documents and good information-documentation practice, complete Croatian legislation, including that enacted prior to 1990, is for the first time available on-line in a unique way.*
*This Paper will show how a simple access to a complex content (such as legislation) is developed, and which knowledge, technologies and institutions must be included in its realization.*

**Key words**: electronic document, regulation, document retrieval, official document

## Introduction

Being familiar with legislation as well as its accessibility are the key factors which enable a society to function. Not only is it important to individuals and native legal entities, but also to all international factors which are interested in legal regulations and business dealings of a particular country.
Contemporary users can access legislation in a number of different ways, due to the fact that all documents are presented and published electronically. However,

such presentations vary in quality, they provide more or less updated and correct information, and in most cases in order to access them one has to pay a fee. The government should aim, within a variety of private initiatives, to offer its citizens and other interested parties high service quality, free of charge. This way the entire legislation, which is a foundation of its judicial system, will be presented in a coherent way and made easily accessible to everyone.

## The importance of accessibility of legal regulations

Well-designed display of required information would allow a user to have a clear and simple access to a system of legal regulations. This also shows the level of a country's democratic development, and another very important aspect which is known as e-democracy [1]. The term e-democracy implies, among other things, that users are allowed to use all tools which they need in order to actively benefit from information systems provided by government administration. Such system should be a source of comprehensive data, which will be easy to process, simple to combine into one coherent whole and exchanged with other users. Therefore, it will stimulate the environment where a citizen, once familiar with their rights, will be able to actively take part in decision making.

When it comes to tools which are necessary for searching and locating such system data, their most important feature must be clarity and simple use. Whether a user is searching for a legal regulation or trying to locate some other document, it is crucial that simple, fast, coherent and long-term formats are applied. As a result, legislation would become more user-friendly and the range of users would expand, thus enhancing better understanding between citizens and decision makers and strengthening confidence in democratic processes.

A part of collection of official documentation in HIDRA refers to legal regulations that have been processed in this very organization and offered to users in a unique way. Based on experience in dealing with users, a significant number of inquiries were noted in the area of legal regulations and legislation search, which made us, conclude that finding and knowing legal regulations is very important to citizens both in business and private aspects of their lives. Therefore it is crucial to ensure that a required regulation is found easily and as fast as possible.

## Creating a collection of legal regulations of the Republic of Croatia

The main source for such collection of legal regulations of the Republic of Croatia is The Official Gazette – the official newspaper of the Republic of Croatia where all legal regulations are published before they come into force.

Considering the fact that in most cases there is a short time period (*vacatio legis*) between the date legal regulations are published and the date they come into force, and taking into account a well-known principle both in legal theory and practice, that not being familiar with the regulation may cause you harm

(*ignorantia iuris nocet*), a team of experts from various fields is gathered and starts work immediately upon an e-version of Official Gazette is published.
In HIDRA, the first step in creating a collection of legal regulations of the Republic of Croatia is to select the ones which will find their place in it. The primary criterion is the one of addressee*, which implies that the first acts to be processed are predominantly the bylaws which are enforceable against all citizens or a particular group of citizens of the Republic of Croatia. Singular acts which are enforceable against an individual or a group of individuals are left out as they usually refer to an individual case and therefore are unique.
Bylaws passed by bodies of local and regional self-government (ordinances, statutes, etc.) are not incorporated in this collection as the similar collection is being created on the county, city and municipality level. It is important to note that international agreements with the Republic of Croatia as a party are incorporated in its national judicial system and according to their legal effect they are above the law. As such, these agreements are found in a separate collection known as International agreements.
HIDRA's collection contains complete texts of legal regulations of the Republic of Croatia in its electronic form. It has to be said that it also contains electronic versions of those regulations by preceding states, which were enacted before 1990 and were still in force at the beginning of 2007, the year when the project of their digitization began. These regulations had to be searched for in old volumes of former country's official gazettes, some being over 60 years old (The Official Gazette of FNRJ, The Official Gazette of SFRJ, The Official Gazette of SRH) and together with their amendments they were scanned and saved in PDF format in digital archives of HIDRA.
This makes HIDRA the only institution, which on its web pages, offers the possibility of viewing those regulations that were enacted before 1990 and are still in force. For the first time, the entire enforceable legislation of the Republic of Croatia has been integrated and made accessible online [2].

## Processing legal regulations of the Republic of Croatia

### Documentation processing

Regardless of its means of publication, all collected documents are processed, formally and substantially in documentation databases. The entire compiled material is processed in the same place and with one, uniform procedure. Data processing is performed applying international recommendations and standards for bibliographic description of publications and documents and HIDRA's normative infrastructure:

- ISBD standards (for bibliographic description of material)
- ISO standards (for data exchange) [3]
- UNIMARC format (for storing data in machine-readable medium)

* Persons for whom the legal standard is intended.

- Pravilnik i priručnik za izradbu abecednih kataloga / Eva Verona (for bibliographic processing)
- EUROVOC Thesaurus with Croatian appendix [4] (for selection of document content determinant as well as standardized names of bodies/institutions in charge of drawing and publishing legal regulations)

The processing involves determining unique document contents according to a word or expression listed in a controlled dictionary. The document is further processed by the method of intellectual analysis, which implies a complete understanding of its contents. Descriptors which outline the contents of the document in short, will allow the end user to access the data via search engine.

**Eurovoc Thesaurus**

Eurovoc Thesaurus (in further text: Eurovoc) [5] is a controlled dictionary which has been used in this type of official documentation content processing in EU for the last 20 years. It is a multi-disciplinary, multi-language thesaurus, primarily used as a documentation tool which enables the contents to be presented in a unique and unambiguous way. Apart from this, by using Eurovoc, document search is made easier and much faster. It was published in seven languages in 1984 by The Publications Office of the European Union and till today a total of 23 language versions have come out, including Croatian.

Eurovoc consists of 21 fields and 127 microthesauri in the form of a structured and controlled list of names used for all important terms in different official documents. The terms in Eurovoc are called descriptors, because they help us to describe the documents. But what makes it so specific is the great number of non-descriptors; i.e. the names that have not been recommended and have the same meaning in everyday language. This feature makes the document search a great deal easier. Currently, Eurovoc consists of more than 7000 descriptors and approximately the same number of non-descriptors.

For a precise way of processing official documentation of the Republic of Croatia, HIDRA has designed a National appendix to Eurovoc. It contains a structured directory of Croatian state bodies and bodies of local and regional self-governments, the Republic of Croatia diplomatic missions as well as foreign diplomatic missions in the Republic of Croatia, the list of political parties and geographic features, all together in 3500 normative records. Some 200 descriptors for the need of specific Croatian contents have been built in the structure of the original Eurovoc. All items in Croatian appendix are marked as Crovoc.

Using this unique thesaurus while processing official documents is a guarantee that the required document will be discovered much faster in an ever-rising pool of information published daily by the bodies of public authorities.

## AIDE – Automatic Indexing using Eurovoc Descriptors

Apart from consistent application of standards and being well-familiarized with the thesaurus which is used in official documentation content processing, another important condition for a good quality processing is competence in various professional areas. Even in cases where these conditions are well-met, the marking procedure depends to a great extent upon a person who is performing indexing, therefore the results may vary. This is highly undesirable in any content processing, even more so in the processing of legal regulations. For this reason, one should strive to reduce such disparity to a minimum and this can be achieved by automatic indexing.

AIDE project – *Automatic Indexing using Eurovoc Descriptors* [6] was launched by HIDRA in 2004 after Croatian experts participated in the workshop organized by Joint Research Centre of The European Commission, entitled *Addressing the Language Barrier Problem in the Enlarged EU – Automating Eurovoc Descriptor Assignment*. At the workshop, JRC experts presented their achievements in the field of automatic indexing by Eurovoc descriptors, the new automatic system which was imitating the Eurovoc descriptor assignment performed by humans. At the end of 2007, the automatic indexing programme was developed, based on the above experience and the work of experts from HIDRA, Department of Electronics, Microelectronics, Computer and Intelligent Systems with the Faculty of Electrical Engineering and Computing, University of Zagreb (ZEMRIS), and experts from Department of Linguistics with the Faculty of Philosophy, University of Zagreb (ZZL). This was a software system for indexing official texts in Croatian language by using Eurovoc descriptors. The entire AIDE project is considered to be the result of Croatian know-how.

AIDE is intelligent software which learns from examples and is based on the principles of machine learning [7]. The development of algorithms for analysing and indexing legal texts is, according to documentalists, based on two premises: the first one is a great number of e-documents in learning corpus and the second one is high-quality indexing model which has been designed by relevant experts from various professional fields. Today AIDE software efficiently proposes ways of indexing texts of legal regulations which are still unindexed, while the team of experts revises the accuracy and completeness of recommended solutions. Thus, further increase in the number of processed documents as well as regular revision of proposals for automatic indexing software make AIDE more and more successful, which directly affects the quality of search results.

Apart from being increasingly successful in indexing new, unindexed documents, the software offers the possibility of selecting statistical associations in the texts of legal regulations. These are first terminologically checked and then incorporated into Eurovoc, this way enriching the range of its non-descriptors. As a result, the further development of Eurovoc is ensured in the best possible way, directly from the texts of official documentation.

WinAIDE station, built for computer assisted indexing by applying automatic indexing method has become a routine tool which is regularly used in HIDRA when processing a collection of legal regulations. It brought about the development of publicly available web service automatic indexing programme called WebAIDE, which can meet the requirements of IT infrastructure developments in all public authority bodies in the Republic of Croatia.

## Data search and display

### Searching for legal regulations through an *e-catalogue*

HIDRA web pages offer an *E-catalogue of official documentation of the Republic of Croatia* [8], where one can access bibliographic description of legal regulations. Apart from the title, Official Gazette's issue number and year, the following data is offered to the user:

- complete texts of legal regulations
- data on a regulation and all amendments, revisions and links to the former
- date of coming into force
- data on the body which passed the legal regulation
- data and link to legal grounds for passing a particular regulation
- data and link to any regulation that has in any way intervened in the contents of a regulation, together with a short note on the nature of intervention
- data on invalidity – a regulation which is not in force any more is marked as "not in force", with the quoted date of invalidity and link to a regulation whose provisions repealed it
- link to all supporting legislation which was passed based on this regulation
- data on the harmonization of the regulation with EU legislation and a link to complete texts (directives, ordinances, Treaty establishing the European Community etc.) of relevant EU legal acts
- data on negotiation chapter pertaining to the regulation, in line with the harmonization of legislation of the Republic of Croatia with the EU acquis (35 chapters)
- data on the field of competence of the body of public authority which has drawn up the regulation (27 fields)
- data on the contents of a regulation (using Eurovoc descriptors)

Various ways of data marking in *E-catalogue* give the possibility of selective data search. When it comes to legal regulations of the Republic of Croatia, this enables not only the search for all legal regulations, but also the access to:

- only those regulations which are in force
- those regulations that have been harmonized with EU legislation
- only those regulations in force which have been harmonized with EU legislation

What is more, the *E-catalogue* also has the option of accessing the above quoted categories in terms of the field of competence of the bodies of public authorities, and also in terms of negotiation chapters pertaining to regulations which have passed the process of harmonization with the EU legislation.

**Zakon o genetski modificiranim organizmima [online e-arhiv] / izrada nacrta Ministarstvo kulture ; donositelj Hrvatski sabor. - (NN 070/2005).**

| | |
|---|---|
| Tip dokumentacije: | PRAVNI PROPISI RH NA SNAZI usklađeni s EU zakonodavstvom |
| Tijela javne vlasti: | Ministarstvo kulture |
| | Hrvatski sabor |
| Napomena: | Prijedlog zakona: P.Z. br. 212 |
| | Datum stupanja na snagu: 2005-06-16 |
| | Stupanjem na snagu Zakona o hrani (NN 46/07) prestaju važiti odredbe koje su u suprotnosti s odredbama Zakona o hrani |
| Vidi: | * UTJECAJNI PROPIS: Zakon o hrani [online e-arhiv] / izrada nacrta Ministarstvo poljoprivrede, šumarstva i vodnoga gospodarstva ; donositelj Hrvatski sabor. |
| Poglavlje pregovora: | Okoliš |
| Obuhvaća: | * Podzakonski akti |
| Područje djelatnosti: | Okoliš |
| URL: | DOKUMENT: http://hidra.srce.hr/arhiva/263/18315/ww... |
| DOKUMENTI EU: | Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC - Commission Declaration, http://eur-lex.europa.eu/LexUriServ/LexU... |
| | Council Directive 90/219/EEC of 23 April 1990 on the contained use of genetically modified micro-organisms, http://eur-lex.europa.eu/LexUriServ/LexU... |
| | Council Directive 98/81/EC of 26 October 1998 amending Directive 90/219/EEC on the contained use of genetically modified micro-organisms, http://eur-lex.europa.eu/LexUriServ/LexU... |
| SKUPNA RAZINA: | NN 070/2005 |
| Sadržaj: | genetički promijenjen organizam |
| | zdravstvena politika |
| | zaštita okoliša |
| | mikroorganizam |
| | utjecaj na okoliš |
| | sprečavanje rizika |
| | bioindustrija |
| | genetički inženjering |
| | sigurnost na radu |
| | priprema za tržište |
| | javno zdravstvo |
| Stanje podataka | 2009-07-08 |
| baza | PPRH |

Picture 1. Data display on chosen regulation (E-catalogue of official documentation of the Republic of Croatia)

**Direct full text search through CADIAL search engine**

Apart from classical e-catalogue search, HIDRA offers the possibility of direct search through complete texts of legal regulations with the assistance of CADIAL search engine, which is the result of Croatian-Flemish project CADIAL (Computer Aided Document Indexing for Accessing Legislation) [9][10]. It is a continued cooperation among the three parties: the University of Zagreb experts, researchers from Katholieke Universiteit Leuven Interdisciplinair Centrum voor Recht en Informatica (ICRI) in Belgium and Joined Research

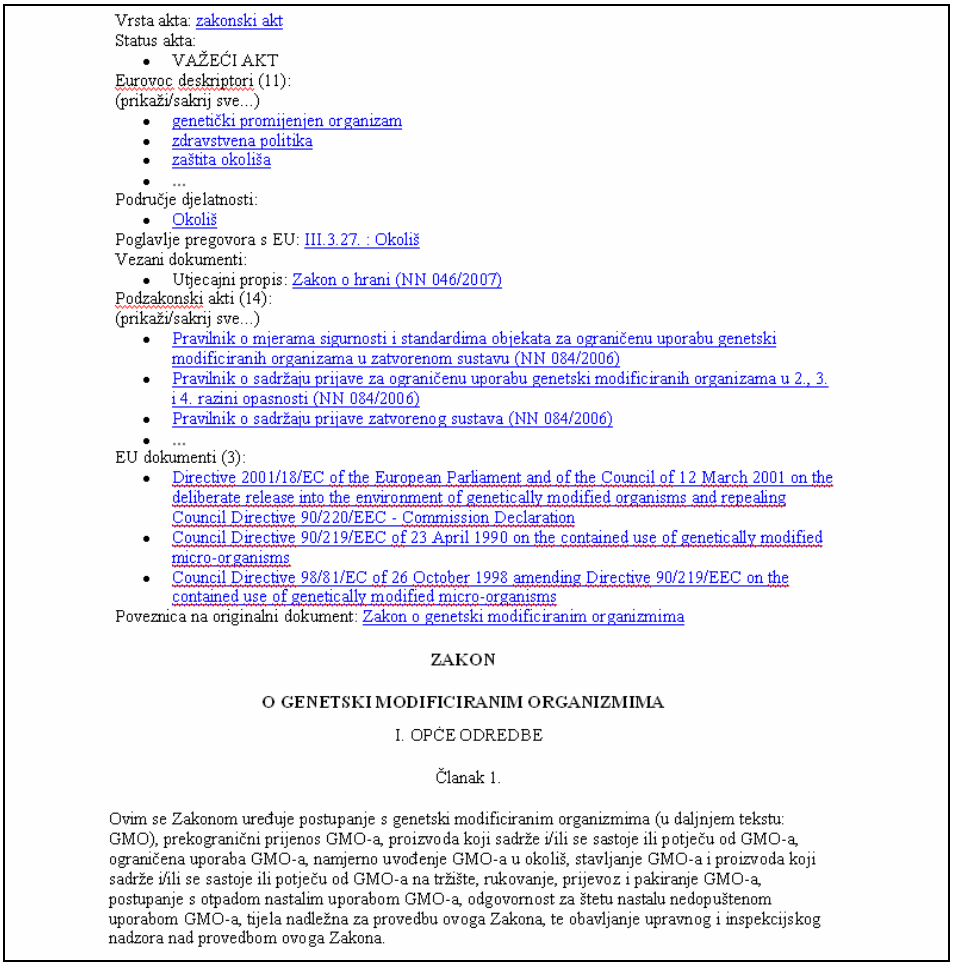
Centre. The project was based on the results of AIDE, and its objective was further development of algorithms for analysing and indexing legal texts in Croatian language and their intelligent search. Search engine CADIAL has been publicly available since the end of 2008 for searching HIDRA’s collection of legal regulations of the Republic of Croatia.

Picture 2. CADIAL search engine user interface display

The search starts by selecting one or more of the three possible options: searching the complete text of the document, searching the document title and searching Eurovoc thesaurus descriptors. Another possibility is to search for only those legal acts which are in force, rather than searching for all regulations. By selecting advanced options, the search becomes more focused according to the type of act, negotiation chapter and area of competence.

With CADIAL, the search is carried out on the complete text of the legal regulation. Just one click on the title of the legal regulation which is selected among the search results, leads the user onto the display of its complete text. At the very beginning of this display, there is data which describes the regulation and offers direct further links. This way the user has the opportunity to directly access groups of linked regulations, what is more, the user can have an insight into the complete text of each of these. Therefore, all linked regulations can be viewed directly, without referring to bibliographic description. Apart from links to complete texts of all supporting legislation and links to the underlying regulation as well as to all EU documents the regulation has been harmonized with, the display also contains the link to the authentic regulation document.

What is of the utmost importance is that CADIAL has been using the software for morphological normalization [11] of Croatian words, and therefore has solved the problem of locating a word in all its forms, which made the search a great deal easier for each user. The search engine has the option of bilingual search, in Croatian and English, and as such enables the user to search for Croatian legal regulations by using Eurovoc descriptors in English.



Vrsta akta: zakonski akt
Status akta:
- VAŽEĆI AKT

Eurovoc deskriptori (11):
(prikaži/sakrij sve...)
- genetički promijenjen organizam
- zdravstvena politika
- zaštita okoliša
- ...

Područje djelatnosti:
- Okoliš

Poglavlje pregovora s EU: III.3.27. : Okoliš
Vezani dokumenti:
- Utjecajni propis: Zakon o hrani (NN 046/2007)

Podzakonski akti (14):
(prikaži/sakrij sve...)
- Pravilnik o mjerama sigurnosti i standardima objekata za ograničenu uporabu genetski modificiranih organizama u zatvorenom sustavu (NN 084/2006)
- Pravilnik o sadržaju prijave za ograničenu uporabu genetski modificiranih organizama u 2., 3. i 4. razini opasnosti (NN 084/2006)
- Pravilnik o sadržaju prijave zatvorenog sustava (NN 084/2006)
- ...

EU dokumenti (3):
- Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC - Commission Declaration
- Council Directive 90/219/EEC of 23 April 1990 on the contained use of genetically modified micro-organisms
- Council Directive 98/81/EC of 26 October 1998 amending Directive 90/219/EEC on the contained use of genetically modified micro-organisms

Poveznica na originalni dokument: Zakon o genetski modificiranim organizmima

ZAKON

O GENETSKI MODIFICIRANIM ORGANIZMIMA

I. OPĆE ODREDBE

Članak 1.

Ovim se Zakonom uređuje postupanje s genetski modificiranim organizmima (u daljnjem tekstu: GMO), prekogranični prijenos GMO-a, proizvoda koji sadrže i/ili se sastoje ili potječu od GMO-a, ograničena uporaba GMO-a, namjerno uvođenje GMO-a u okoliš, stavljanje GMO-a i proizvoda koji sadrže i/ili se sastoje ili potječu od GMO-a na tržište, rukovanje, prijevoz i pakiranje GMO-a, postupanje s otpadom nastalim uporabom GMO-a, odgovornost za štetu nastalu nedopuštenom uporabom GMO-a, tijela nadležna za provedbu ovoga Zakona, te obavljanje upravnog i inspekcijskog nadzora nad provedbom ovoga Zakona.

Picture 3. Display of complete text of legal regulation and supporting information

Regarding a very high degree of data integration offered by CADIAL, it is possible to say that at the moment it is the easiest and the fastest search for legal regulations of the Republic of Croatia.

## Conclusion

The main objective of any democratic society should be to make sure that simple and high-quality search for legal regulations is made possible. As one of the agents in designing IT infrastructure between the Government and state bodies of the Republic of Croatia, HIDRA has been monitoring the needs of its users. It has therefore designed the collection of legal regulations that offers high-quality information to the user and makes it easily and quickly accessible. The presented collection was created and is further being developed as a combina-

tion of experts’ knowledge from various professional fields and applying tools which are the result of new technological advances. CADIAL search engine shortens the time users need to find the regulation, at the same time making the search more simple. This way the user has an easy access to comprehensive contents of national legislation. Finally, if we take into consideration the advantage offered by a bilingual Eurovoc, which is overcoming language barriers while searching Croatian as well as EU member states legislations, it is obvious that the intelligent CADIAL search engine truly improves the search by making legal regulations easily accessible to everyone.

## Literature

[1] CM/Rec(2009)1E / 18 February 2009. Recommendation of the Committee of Ministers to member states on electronic democracy (e-democracy). http://www.coe.int/t/e/integrated_projects/democracy/02_Activities/002_e-democracy/Recommendation%20CM_Rec_2009_1E_FINAL_PDF.pdf

[2] HIDRA - Hrvatska informacijsko-dokumentacijska referalna agencija. http://www.hidra.hr

[3] ISO 5963-1985. Documentation - Methods for examining documents, determining their subjects, and selecting indexing terms. Geneva : International Standards Organization, 1985.

[4] Hrvatska informacijsko-dokumentacijska referalna agencija – HIDRA. Pojmovnik Eurovoc. Verzija 4.3. Zagreb : HIDRA, 2009. http://www.hidra.hr/eurovoc/eurovoc1.HTM.

[5] Eurovoc Thesaurus. http://europa.eu/eurovoc/

[6] Hrvatska informacijsko-dokumentacijska referalna agencija – HIDRA. Projekti. AIDE. Zagreb: HIDRA, 2009. http://www.hidra.hr/hidra/hidraproj-c.htm

[7] Kolar, Mladen; Vukmirović, Igor; Dalbelo Bašić, Bojana; Šnajder, Jan. Computer-Aided document Indexing Systems. // Journal of Computing and Information Technology - CIT. 13 (2005), 4; 299-305

[8] Horvat, T.; Pekorari, R. Pitajte HIDRU : kako najjednostavnije i najbrže doći do potrebne službene informacije ili dokumenta? // Infotrend, Prilog eGovernment. 169 (3/2009), str. 1-5.

[9] CADIAL – Computer Aided Document Indexing for Accessing Legislation. http://www.cadial.org/

[10] Mijić, Jure; Moens, Marie-Francine; Dalbelo Bašić, Bojana. CADIAL Search Engine at INEX. // Lecture Notes in Computer Science, Advances in Focused Retrieval (INEX 2008). 5631 (2009)

[11] Šnajder, Jan; Dalbelo Bašić, Bojana; Tadić, Marko. Automatic Acquisition of Inflectional Lexica for Morphological Normalisation. // Information Processing & Management. 44 (2008), 5; 1720-1731.

# Internet Voting: State in EU and Croatia

Neven Pintarić
Department of Economics, University of Zadar
Trg Kneza Višeslava 9, 23000 Zadar, Croatia
neven.pintaric@unizd.hr

Ante Panjkota
Department of Economics, University of Zadar
Trg Kneza Višeslava 9, 23000 Zadar, Croatia
ante.panjkota@unizd.hr

Josipa Perkov
Department of Economics, University of Zadar
Trg Kneza Višeslava 9, 23000 Zadar, Croatia
josipa.perkov@unizd.hr

**Summary**

*Voting through the internet (i-voting) should be a service through which governments enable their citizens to participate in decision making and establish greater democracy in society. This paper investigates the i-voting situation in EU and Croatia, as well as the attitude of student population towards i-voting. The first part of the paper relates to the application of i-voting in EU and Croatia, according to the implementation of i-voting, its problems and legislation. In the second part we present results from the initial research of attitudes among the student population at the Department of Economics, University of Zadar. The research was done with a questionnaire in respect to: participating in the election, advantages and disadvantages of i-voting, problems with ICT related to i-voting, reasons of the absence of i-voting in Croatia. Estonia is the first country which has implemented i-voting in the elections, whereas other countries are still testing it and resolving problems present in the implementation. Croatia still does not plan to implement i-voting and needs to consider strategic steps in this process. The research indicates insufficient information about i-voting among the student population. It is necessary to include topics from e-government in the education curriculum, i.e. in the subjects covering the field of Information science, in order to create education preconditions for the implementation of i-voting.*

**Key words***:* i-voting, EU, Croatia, IKT, SWOT Analaysis, attitude of student population

## Introduction

The development of information and communication technology (ICT), especially the development of the Internet from 1990-ies have changed the way we work, learn and live. The Internet has become the most important infrastructure for data transmission and services.

Internet voting (i-voting) is a method of voting in elections which use internet, where a voter accesses the web pages (services), identifies himself/herself and votes (Kovačić, Škrablin, 2009). Voting is possible from any location, with general principles of classical election. I-voting is one of possible ways of electronic voting (e-voting). E-voting is the aggregate name for voting with electronic devices, which enable voting at the election place or other, remote location. E-voting and i-voting are often used in the same meaning in literature. The framework of i-voting changes from previously fully controlled conditions to the new framework, which is not fully controlled. All procedures and controls are realized with ICT and through the acts of the voters.

ICT is requested to be closed and secured. For this reason, different procedures of control and monitoring have been applied (e.g. logging, every packet of data which is transmitted over the internet has its original address etc.). The elections must provide the confidentiality of vote as well as the transparency of the process.

According to Gibson (Gibson, 2005, 31) it is necessary to research the impact of new technologies related to voting at elections. It is also necessary to perceive all effects which can appear.

Croatia does not have i-voting, there are few data and information about it. The tendency is to create the Information Society and greater democracy. The assumption is that in future Croatia will prepare and test i-voting in order to increase the participation of citizens. There is a need to analyze which factors would appear with i-voting; what and how they would affect. I-voting in Croatia could be mostly used by younger population. According to the research in the USA (Gibson, 2005, 29) there is a low turnout at the election of population between 18 and 24 years. Austria aims at improving students’ participation in the student elections by i-voting.

The following research questions are asked in this paper:

- What is the status of i-voting in the EU and Croatia?
- How well is the student population informed about i-voting?
- What are their attitudes towards:
  - what i-voting enables and disables
  - problems related with ICT, which can be present in i-voting
  - reasons that Croatia does not have i-voting?

## Method

The state of i-voting in the EU and Croatia will be investigated based on the qualitative analysis of available literature in order to understand the factors that are related to enforcement, legislation and problems present in the i-voting.
Factors that are identified will be analyzed with the SWOT analysis.
The investigation of students' attitudes related to i-voting was done with a questionnaire. This population was chosen because it is expected that students would use the i-voting when it would be available as a service of the state administration, and because in some states i-voting should increase their participation in elections. It is assumed that students could have obtained the information about i-voting through the public media, the Internet, and about ICT through previous education and experience with the use of ICT.
We prepared 8 questions to determine the attitudes of the student population concerning the following topics (Q): 1.) voting in elections; 2.) the use of i-voting; 3.) familiarity with the topic of i-voting; 4.) in which EU country is i-voting organized at the state level (answers: Finland, Austria, Switzerland, Estonia, I do not know); 5.) what i-voting provides (confidentiality, equal access / participation, transparency, security); 6.) what i-voting does not provide (confidentiality, equal access / participation, transparency, security); 7.) which problems are related with ICT, and are connected with i-voting (theft and replacement of identity, multiple voting, manipulation of records in the database, problems with the PC, problems with the Internet, problems with the main server, problems with the software); 8.) which are the reasons that there is no i-voting in Croatia (laws, internet connections, the number of potential users, investment in the process of computerization of state administration, electronic identity and signature, security problems).
Questions and answers are attached to values. Replies in questions 1, 3, 5, 6 and 8 have a value scale from 1 to 5, where 1 denotes negation, and 5 full agreement. The second question was answered with yes and no. Questions 4 and 7 were answered by choosing one of the listed claims. The analysis of statistical results of the survey was done with Mystat 12 tools.

## Results

### The situation in the EU

The right to participate in the elections and vote is the foundation of democracy. Together with the development of ICT technologies, with political changes that have happened in the EU (integrations) and in the world in the last 10 years, it has been investigated how the ICT technology affects society, democracy, and in what way it can help in these processes. Given the increasing use and impact of ICT technologies (especially the Internet) in the realization of democracy we can talk about electronic democracy (e-democracy). The participation in democracy according to the Hague and Loader and Karmacka and Nya is realized through (Kersting, Baldersheim, 2004, 4): information (www, e-mail), commu-

nication (chat, forum, ..) and transactions (e-voting, services). At the end of the 1990-ies and the beginning of 2000 various projects and research have been done in the area of i-voting. Buchsbaum T. M says there are following categories in the introduction and use of e-voting (Buchsbaum, 2004, 39):

- early (private) pilot projects financed with the EU funds
- states that seek to introduce e-voting
- e-voting and testing in the academic community
- advanced pilot projects and elections (e-voting) that are at the level of local self-government.

The EU Commission in the framework of FP5 launched the project CyberVote associated with i-voting in 2000. European countries have been considering the introduction of i-voting and running various projects in order to research all forms of e-voting. In Switzerland pilot projects on the use of i-voting for the canton Neuchâtel, Geneva and Zurich were launched in 2002; Germany had a research project Wahlen in elektronischen Netzwerken (W.I.E.N.) associated with i-voting, and the United Kingdom (UK) has launched 17 pilot projects related to e-voting (i-voting, e-voting at polling stations, telephone, SMS, digital TV).

The main topics related to i-voting, which are present within the research are: the basic assumptions of i-voting (F. Mendez, A. Treschsel, R. Gibson), legislation (P. Garrone, A. Aurer and M . Mendez), security (L. Pratchett, M. Wingfield, B. Fairweather, S. Rogerson, CARNet CERT and LS & S) digital divide (Grönlund, P. Norris), experience with i-voting H. Gaser (Switzerland), W. Dreschsler and U. Madise (Estonia), R. Krimmer (Austria), A. Karger (Germany) and Lawrence Pratchett and Melvin Wingfield (UK).

For the purposes of the introduction of i-voting, it is necessary to pass the laws. The Committee of Ministers of the Council of Europe issued a recommendation on e-voting in elections and referendums on 30th October 2004, on how to create, implement, monitor system of e-voting in order to ensure that voices are real like those obtained in the classic way.

Recommendations are related to (Council of Europe, 2004):

- legislation – general frame, equality, freedom and secrecy of votes; transparency, checking the proper operation of the system, the possibility of recounting the voices; the system must be reliable and free of possible threats.
- operational standards – informing the voters of elections; the necessity of registration of voters; candidacies; the procedure, way and duration of voting; the procedure of generating the results; the system should enable checks.
- technical requirements – the necessity of determining and assessing the risks; all voters should be able to access and use the service; using open standards to enable interoperability; the list of equipment and program

support, procedures related to unaccepted situations; security database; secure location, security procedures (before, during and after voting), checking the system; monitoring; checking and confirming coordination with laws, certifying the system.

Among teh EU countries Estonia and Austria have done most in the application of i-voting. Estonia introduced i-voting in 2005 for local elections, and in 2007 for parliamentary elections.

According to the report (Trechsel, Shuman, 2009) for the Council of Europe, the main reasons for implementing the i-voting were the procedures that lead to the creation of an information society (e.g. in these countries the right to access the Internet is the social right; however, Estonia also leads in the amount of investment in ICT in relation to GDP). Requirements directly related to the i-voting were: legal framework, the introduction of identification cards (the introduction in 1997) which contains a certificate for authentication and certification of digital signatures, and system-voting based on "double envelope" which made it similar to the classic voting . In 2005 from a total of 1,059,292 voters, 9681 voters used the possibility of e-voting, which is 1.92% compared to the voters who came out to the polls. In the parliamentary elections in 2007 the i-voting was used by 30,275 citizens, which is 5.4% of the total number of voters who came out to the polls. In traditional elections voting is allowed only once. The Estonian election law related to i-voting allows the voters the possibility of multiple voting (voting repetition), which counts only the last vote. This was introduced in order to prevent any possibility of buying votes and influencing the voters.

According to the research (Madise, Martens, 2006, 23) conducted after the local elections, the reason for the reduced use of i-voting is the lack of Internet access, lack of knowledge about computers and the adequacy of classical voting by ballots. It was also found that the age of voters plays an important role in participating in i-voting.

Austria started the project at the beginning of 2000 in order to increase the participation in the student elections. I-voting was conducted in May 2009 in the official elections for student representatives. The prerequisite for i-voting for citizens (students) was the possession of the identification card (electronic card e-Card, Bürgerkarte) and certificates.

On the web site of the Ministry it is indicated that Austria still has no legislation for the elections at the state level.

**The situation in Croatia**

In Croatia, the elections and voting are possible according to the following legislation: the Constitution of the Republic of Croatia, the Law on Elections for the Croatian Parliament, the Law on the Election of the President of the Republic of Croatia, the Law on the Election of members of representative bodies of local and regional (regional) governments, Law on election of municipal may-

ors, mayors, prefects and mayors of Zagreb, the Constitutional Law on National Minorities, the Law on Election of Members of European Parliament (in the process of adoption), the Law on the constituencies, the Law on the Election Commission and the Law on voter registration. Through this legislation the following is defined:

- the right to participate in the elections in Croatia
- the freedom of affiliation of voters and the secrecy of their vote
- the right and obligation to vote only once in the elections
- vote on the basis of residence of voters
- voter gets an extract from the list of voters which refers to him
- secrecy of voting is carried out and allows the voter Board
- Voters Board verifies enrollment of voters in the voters' list
- voting is done in person, via the ballot
- the ballot has a serial number.

Current legislation relating to elections does not provide the possibility of organizing and carrying out the i-voting.

The Central State Office for Administration (CSOA) allows verification of voter registration lists by the single parent and an SMS message.

In the strategic document of the state government "*Public administration reform strategy for the period of 2008 – 2011*" it is stated that one of the goals of the reform is the application of modern information-communication technologies (CSOA, 2009, 1) whereas one of the five main directions is the realization of e-administration (CSOA, 2009, 2). Improving the application of the ICT system of the state administration is seen through (CSOA, 2009, 7): the supply of equipment, development and procurement of appropriate computer software, networking, IT education and training of civil servants. The application of ICT is ensured through the implementation of public administration reform strategy, and is realized through the "*disclosure of the electronic address for communication with the citizens on the web site of all public administration bodies and determining at least one civil servant in each of those bodies in charge of replying to electronic messages of citizens; ensuring full communication with citizens electronically, including applications, requirements and other patterns in all government agencies, except in proceedings relating to the status issues of citizens; courses for beginners and advanced IT training for civil servants*".

The main holder of the ICT activities in the state administration is the Central Office for e-Croatia.

The starting point for the definition of the ICT strategy within the framework of this strategic document was the "*strategy of the Information and Communication Technology – Croatia in the 21st century*" from 2002. The Central State Administration Office in the framework of this strategy has defined the activities in collaboration with the Central office for e-Croatia, but there are no activities related to the conduct of the elections by i-voting.

In the framework of the Strategy for developing the electronic government in the Republic of Croatia (Croatia Government, 2009) for the period between 2009 and 2012 i-voting is not specified as one of the goals or services that are planned for that period.
The goals which are planned within the framework of the strategy, and would be related to i-voting are: the establishment of a central authentication and authorization system (e-identity) (Croatia Government, 2009, 17) and the legal source, base for electronic services that are provided (Croatia Governmet, 2009, 20).
Besides the already mentioned legislation for elections it will be necessary to consider or finalize the legislation when implementing the i-voting related to: information security, electronic signature, electronic document and electronic communication.
According to the above strategies Croatia is on the level of informing and communication with its citizens within participation in the e-democracy.

**SWOT analysis**

In the available research literature we find both positive and negative factors that may affect the i-voting. These factors were analyzed using the SWOT analysis, which consists of internal factors (strength, weakness) and external factors (opportunities and threats).
Strength – voting regardless of the location, getting results faster, reducing costs (e.g. persons who are involved at the polling places, in printing, distribution), the implementation of the referendum, a longer period and re-voting (the possibility of reflection), reducing the number of errors in voting.
Weakness – poor turnout of voters, privacy, dependence on other technologies, Internet access, possible technical problems in the implementation of voting, the possible social engineering, a more complex procedure of voting (a problem for people who do not use a computer every day), very hard to establish the control of i-voting (down transparency), inaccessibility to a greater number of people, the secrecy of voting (ballot), identifying and correcting mistakes, misuse of electronic identity, possible dependence on the software manufacturer.
Opportunities – increasing the efficiency of administration, increasing the degree of democracy, the conduct of voting at the EU level, increasing participation (exit) of voters in elections.
Threat – a negative effect due to public perception, influence on political parties, vote trading, the destruction of the ritual of classical voting, questionable transparency of the voting process, technological development and new technologies (such as voting by mobile phone), changing legislation, changing the parliamentary democratic system (direct democracy).
As there is a great number of weaknesses and threats it is not possible to implement i-voting fast.

**Analysis of students' attitudes**

The questionnaire was carried out on the student population at the 3rd year of undergraduate study of management at the Department of Economics, University of Zadar. The respondents belong to the age group from 21 to 23 years. Tha analysisi comprises N = 56 questionnaires.

Based on the analysis of responses for Q1, which is related to voting in the elections, $\bar{x}$ = 3.625, Mo = 4 (mostly yes), and coefficient of variation is V = 33%. As coefficient of variation refers to insufficient homogeneity, according to the mode we can conclude that the student population generally votes in the classic elections.

In Q2 $f$ = 47 (83.93%) respondents expressed willingness to use i-voting, whereas $f$ = 9 (16.07%) respondents did not express they would vote through the Internet. According to a study from Austria (Kersting, Baldersheim, 2004, 116) conducted among the student population, 84% of respondents expressed willingness to use the i-voting.

Familiarity with the i-voting is associated with Q3 and Q4. Answers related to Q3 are $\bar{x}$ = 1.5 and Mo = mode 1 (not general). We can conclude that the student population is not familiar with the topic of i-voting.

In Q4 – in which EU country i-voting has been carried out, $f$ = 54 (96.43%) respondents indicated that they did not know, $f$ = 2 stated Switzerland.

Answers to Q5 relate to what i-voting allows. According to the opinion of the students voting via the Internet mainly provides secrecy of voting (Mo = 4), generally provides equal access and participation in voting in elections (Mo = 4) and transparency (Mo = 4). There is a possibility of expression of doubt in the safety procedures (Mo = 3).

In Q6 Mo was taken into account. According to it the attitude of respondents related to what i-voting does not allow is the following: sometimes it does not provide secrecy of the vote (Mo = 3), respondents express confidence in equal access / participation in voting (Mo = 2), according to them transparency of voting (Mo = 3) as well as safety in the process of elections (Mo = 3) are sometimes made impossible.

Having established the correlation coefficients between Q5 and Q6 a significant negative relationship was obtained.

In Q7, which problems are related to ICT and can connect to or influence the i-voting the respondents list: theft and replacement of identity $f$ = 37, the problems with the network (Internet) $f$ = 32, possible manipulation in the database records $f$ = 30, the possibility of multiple voting $f$ = 29, the main problems with servers $f$ = 21, problems with PC $f$ = 18, and problems with the programs i-voting $f$ = 11.

In Q8, relating to relevant reasons why Croatia has no i-voting in elections, the answers of respondents who answered affirmatively to the exit-polling (Q2) were taken into account. The referred question was put in correlation with Q8.

The result is a correlation: if the respondent is ready to vote via the Internet, then he/she believes that the reason why there is no i-voting in Croatia lies in legislation and investment in the process of computerization of the state administration.

## Conclusion

I-voting in the elections in the EU is still in a testing phase. There are different problems that need to be resolved before the i-voting can be widely applied. It is necessary to ensure the credibility of the ICT technology as well as to be aware of all the changes that can occur due to i-voting. For now, the i-voting is only allowed in Estonia, on the national and local level. A small percentage of voters participated in the elections through the Internet, 2005 – 1.92% 2007 – 5.4% compared to the total number of voters. Such a small number of voters was due to insufficient access to the Internet and to the ignorance of voters in the use of computers.

At this point Croatia does not have i-voting in the legislation or the strategies. The latter is a disadvantage, because it can be a limiting factor in possible projects related to the i-voting and seeking funds from the EU.

According to the initial research conducted among the student population, respondents would participate in i-voting, but at the same time they have no knowledge about i-voting. Problems that are related to the ICT and can appear in the i-voting are: the theft and replacement of identity, problems with the network (the Internet), the possible manipulation of records in the database. As reasons for the lack of i-voting in Croatia, students state the lack of legislation and investment in the process of computerization of the state administration. It would be necessary to introduce topics that are related to e-governance in the educational process in order to create the environment for the application of this service.

Since i-voting is connected with a lot of problems and carries with it a series of changes, additional research in the area of i-voting is required (e.g. area security, influence of i-voting on the society, the analysis of justification, the use of open standards (codes), defining metrics, models monitoring, ISSX – students e-cards), in order to ensure the application of ICT technologies in the social areas.

## References

Buchsbaum T. M.: E-Voting: International Developments and Lessons Learnt članak u Prosser A., Krimmer R.: Electronic Voting in Europe Technology, Law, Politics and Society, Köllen Druck + Verlag GmbH, Bonn, 2004

Council of Europe: Recommendation Rec (2004) 11 of the Committe of Ministers to member states on legal, operational and tehnical standards for e-voting:, available on: https://wcd.coe.int/ViewDoc.jsp?id=778189 accessed at: 5. March 2009.

Croatia, Central State Office for Administration Strategija reforme državne uprave za razdoblje 2008.-2011., available on: http://www.uprava.hr/strat-hr.pdf; accessed at: 12. March 2009.

Croatia, Governmet: Strategija razvoja elektroničke uprave u Republici Hrvatskoj za razdoblje 2009.-2012. godine, available on: http://www.e-hrvatska.hr/sdu/hr/Dokumenti/StrategijeIPrograma/categoryParagraph/01116/document/Strategija_razvoja_elektronicke_uprave_u_Republici_Hrvatskoj_za_razdoblje_od_2009_do_2012_godine.pdf; accessed at: 10. March 2009.

Gibson K. R.: Internet Voting and the European Parliament elections in Trechsel H. A; Mendez F.: The European Union and e-Voting, Addressing the European Parliament’s internet voting challenge, Routledge, New York, 2005

Kersting N.. Baldersheim N.: Electronic Voting and Democrtaic Issues: An Introduction in Kersting N.. Baldersheim N. Electronic Voting and Democracy A Comparative Analaysis, Palgrave Macmilla, New York, 2004

Kovačič M., Škrablin J.: Internet volitve v Sloveniji?; available on: http://www.elektronske-volitve.si/i-volitve.pdf; accessed at: 10. Juni 2009.

Kripp M. Krimmer R..: Information Technology in The Electoral Processs: Electronic Voting state of the Art in Europe, available on: http://www.a-i-c.at/upload/08-10-27%20Pr_sentation_Kripp_evoting.pdf, accessed at: 10. March 2009.

Madise Ü.,: e-voting in Estonia experience, available on: http://static.twoday.net/evoting/files/First_Experience_with_E-Voting_in_Estonia.pdf, accessed on: 20. March 2009.

Madise Ü., Martens T.: E-voting in Estonia 2005. The first practice of country-wide binding Internet voting in the world, in Krimmer A.: Electronic Voting 2006, Köllen Druck + Verlag GmbH, Bonn, 2006

Trechsel A. H., Schuman R.:, Internet voting in the March 2007 Parliamentary Elections in Estonia, available on: http://hdl.handle.net/1814/7549 accessed at: 10. ožujka 2009.

ÖeH, Österreichische Hochschüle Wahl, available on https://oeh-wahl.gv.at/Content.Node/33092_7.html, acceesed at: 17. ožujka 2009.

# User Experience with Advertising over Mobile Phone: A Pilot Study

Neven Bosilj
T-mobile, Croatia
Ulica grada Vukovara 23, Zagreb, Croatia
E-mail: neven.bosilj@t-mobile.hr

Goran Bubaš
University of Zagreb, Faculty of Organisation and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
goran.bubas@foi.hr

Neven Vrček
University of Zagreb, Faculty of Organisation and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
neven.vrcek@foi.hr

**Summary**

*This paper discusses potential uses of Short Message Services (SMS) as a means for mobile advertising. The adoption and user experience of mobile advertising was investigated in a pilot study. A convenience sample of 62 informatics students was exposed to a two week mobile advertising campaign in which they received and responded to 1-4 different textual messages per day. A survey was performed before and after the campaign to investigate the factors that could influence the adoption of mobile advertising as well as its effects on students' experience as participants in the campaign. Conclusions are drawn regarding the factors which have to be taken into account to facilitate consumer participation in mobile advertising campaigns.*

**Key words:** mobile advertising, short message services, user experience, survey

## Introduction

In comparison to land-line telephones, television, radio, and most other electronic communication media, mobile phones are much more personal devices. A mobile phone usually only has a single user and this attribute makes mobile phones suitable for high-precision targeting if they are used as a communication channel in marketing campaigns. According to Ahonen (2007) mobile devices are the most dominant devices worldwide (in comparison to other technological devices), with as many as 2.7 billion users in 2007 (see Table 1). E-marketing

potential is illustrated in the fact that in 2007 there were about 1.5 billion e-mail addresses on the Internet that were used by slightly less than 800 million people. On the other side, in 2007 almost 1.9 billion out of 2.7 billion mobile device users were actively using the SMS service, which indicates a comparable m-marketing perspective.
Ahonen (2009) recently claimed that there are 4 billion subscribers of mobile devices and that about 1.1 billion new mobile phones were sold in 2008. It must be noted that in many developed countries the penetration of mobile devices has exceeded 100%, i.e. there is a substantial percentage of users who are using not only one, but two or three mobile devices. According to Ahonen, in 2009 there have been about 2.7 billion active SMS users.

Table 1. Comparison of the number of technological devices worldwide

| Device/technology | Years of existence | Number |
|---|---|---|
| car | 100 | 800.000.000 |
| television | 60 | 1.500.000.000 |
| personal computer | 30 | 850.000.000 |
| regular phone | 110 | 1.300.000.000 |
| Internet | 15 | *1.100.000.000 |
| digital camera | 20 | 200.000.000 |
| mobile phone | 35 | 2.700.000.000 |

Source: Ahonen (2007); * number of users

Mobile messaging is a joint name for SMS (Short Message Service) and MMS (Multimedia Message Service). SMS advertising messages are limited to text with a maximum length of 168 characters, but they are ideal for access to all mobile users regardless of the type of a mobile device, available data space on a device or similar limitations that can present an obstacle to the success of an MMS advertising campaign. An MMS can contain a picture, audio content or video content. However, problems in MMS delivery can arise because of the inability of certain mobile devices to open specific types of formats (mpeg, avi, mp3 and the like). It must be emphasized that SMS is the most widely utilized data application in the world that is used by more than 75% of all mobile device users (Gopal and Tripathi, 2006).

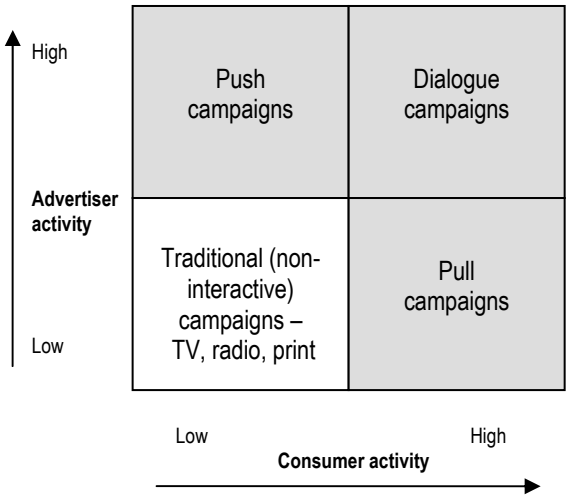
## Mobile marketing

### Push, pull, and dialogue marketing campaigns

The concept of sending advertising and promotional messages directly to mobile phone users has changed the previous concept of advertising and has opened up new possibilities to companies for creating innovative marketing campaign forms. However, mobile phones are still an insufficiently utilized advertising medium (O'Shea, 2007).
Mobile campaigns can be divided into three main groups (Jelassi and Enders, 2004): (a) pull, (b) push, and (c) dialogue. The previous division, depicted in

Figure 1, takes into consideration the activity level of the advertiser, as well as the level of user activity. *Push* mobile advertising is represented by sending unsolicited messages, usually via an SMS alert. In *pull* advertising the messages (usually promoting free information such as traffic reports or weather forecasts) are added to the browsed content requested by the customer (Schreiber, 2000; Dickinger et al., 2004). As can be observed in Figure 1, the highest degree of interactivity is present in a *dialogue* campaign between the advertiser and the user. On the other side, *traditional* campaigns (TV, radio, print, etc.) have the lowest level of advertiser and consumer activity.

Figure 1. Interactivity and mobile campaign types




Source: Adapted from Jelassi and Enders (2004)

**Personalization in mobile marketing**

The two main trends in mobile marketing are (1) customization, and (2) customerization. *Customization* denotes the dimension of customer service that represents personalization that is oriented toward the use of specific user profiles. According to Barnes and Scornavacca (2004), mobile users want to receive highly personalized information. On the other hand, *customerisation* represents a new/higher level of personalization and individualization that is made possible by mobile devices. Campaigns with a higher level of adaptability to users can have greater marketing effects and contribute to the design of mobile customer management systems (mCRM) for business (McManus and Scornavacca, 2005).

**Viral marketing**
Viral marketing is a special marketing strategy that encourages a client to forward a received message to third persons, through which a multiplication effect (avalanche effect) is achieved very quickly. The aim of the message is to arouse the user's interest in different ways (fun message, humor, surprise, exceptionally useful message, etc.) or to offer some added value (e.g. monetary stimulation) so that the user forwards it to his/her friends and colleagues (Pousttchi and Wiedemann, 2006). In this way, the client carries out marketing activities instead of the company. Those who receive a mobile advertising message from a friend are more likely to participate in the campaign (Salo and Tahtinen, 2005). This form of marketing deserves special attention in the development of mobile SMS advertising.

**Privacy**
Mobile advertising may have an adverse effect in raising users' fears in terms of privacy. Privacy is defined as "the right of any individual to control the information held about them by third parties" (Chaffey, 2003). It must be noted that Dickinger et al. (2004) observed that *"The mobile phone cannot distinguish between spam and genuine communication automatically"*. They also found out that consumers fear registration to SMS-based information services for privacy concerns. Permission-based mobile advertising (PBMA) is considered to be the easiest way to tackle the privacy issue (Godin, 1999; Cleff, 2007).
Ahonen's (2007) comparison of e-mail messages and SMS messages indicated that an e-mail message is opened, on average, within 24 hours, while a reply occurs within 48 hours. On the other hand, SMS messages are read within 15 minutes of their arrival and the average response time amounts to less than 60 minutes. However, while as much as 65% of e-mails are spam, the percentage of spam among SMS messages is less than 10%. Therefore, protection of the mobile world from being polluted by spam is an important issue. The problem of unwanted messages that swamp e-mail inboxes is crucial and the answer to this problem will influence the development of mobile advertising.

**Adoption problems**
Even though mobile marketing adoption is on the rise, the marketers should have a clear understanding of the factors which drive consumer acceptance to attain and preserve the ability to consistently generate profits (Spurgeon, 2008, 95-100; Merisavo, et al., 2007). One of the basic problems is that advertisers do not know what users are doing when they see an advertisement on their mobile device, which means that they do not have direct and immediate feedback in relation to the success of a mobile advertising campaign. Until this problem is resolved, mobile advertising may not attract significant investment. However, in the online world it is fairly easy to determine the success of a marketing campaign. Websites can record what users do after they see an advertisement on a

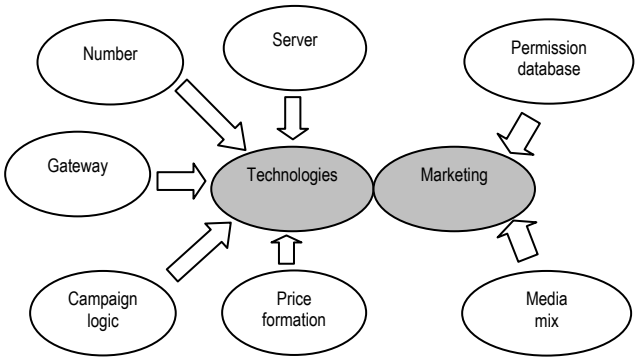
specific webpage, for instance with the help of cookies and software for monitoring user habits. The data collected about user behaviour makes it possible to personalize online advertising messages and introduce targeted display of advertisements which are most likely to attract user attention and response.
The mobile phone industry has rejected the approach which would monitor user behaviour with cookies or similar technology solutions. Most mobile operators block cookies before they reach mobile phones, because operators claim that they pose a security threat to servers. An additional argument for operators is a potential increase in traffic resulting from the use of these additional programs, which could impede the normal use of mobile devices and perhaps even make calls impossible.
Potential problems may also arise from the lack of standards regarding the measurement of the effects of mobile advertising. While one advertising agency may want to know how many advertisements were shown to users included in the campaign, another agency may want to know whether the airing of advertisements led to actual purchase. The research that is presented in this paper attempts to provide insight into some of the ways the users perceive mobile advertising before, during, and after they were exposed to a campaign. In addition, the intention of the preliminary study that is presented in this paper was to measure the discomfort of users of mobile devices as a consequence of their exposure to SMS advertising.

## Mobile marketing campaign

Sinisalo et al. (2006) offer a list of key problems which they identified during the process of creating a mobile campaign. The concrete campaign in their study was related to the implementation of an mCRM system in one of the largest Finnish companies in 2004 and 2005. The key problems of the mobile marketing campaign which were observed (see Figure 2) can be divided into two basic groups: (1) technological and (2) marketing problems.
On the technological level, five types of mobile campaign problems were identified: server, number, gateway, campaign logic, and price formation. On the marketing level, two basic campaign problems were related to permission database and media combination.
To provide the technological basis for our pilot study of mobile marketing we had to solve the *gateway* problem. A gateway is a communication device that links two independent systems that use different protocols and is dependent on the number of mobile operators that do business in the country in which the mobile marketing campaign is carried out. In Croatia, three operators are currently acting as mobile telecommunications providers (T-Mobile, Tele2, and Vipnet) and from the technological aspect this does not pose a big problem or create high costs, as opposed to, for example, Finland with more than 17 active operators (Sinisalo et al., 2006). If a mobile marketing campaign is carried out, users from all nationally available networks have to be able to access the cam-

paign server without complication. In principle, operators have to provide a service for connecting to their networks for SMS/MMS messages. This service is often called *messaging interface*. It enables the sending of SMS and MMS messages among mobile devices which are registered with various operators, on the one side, and the information system of the company that implements the mCRM, on the other side. This service basically includes three different links: content gateway, short phone number/SMS number and a price formation service. Once the service is established, all mobile subscribers can initiate an SMS dialogue with the company.

Figure 2. Key points in creation of mCRM [15]




Source: Sinisalo et al., 2006

For the purpose of this research, we developed a Java application which allowed the personal computer to which a mobile device with a number was connected to behave like an SMS server that sends messages to users who are in the data base. The return messages of users were saved in a special database suitable for further analyses. In this way we created a simple, effective and low cost technological solution to support the planned methodology for our pilot study and also avoided the obstacles that Sinisalo et al. (2006) recognized in their case analyses of a mobile campaign.

## Problem and hypotheses

The main goal of our study was to investigate how potential clients would react to a mobile advertising campaign. We decided to perform a pilot study on a small group of students who would receive 1-4 marketing SMS messages per day for a period of two weeks and evaluate the usefulness of each of the incoming messages. Therefore, the main problem of our study was to determine the potential acceptance rate of SMS marketing. For this we combined quanti-

tative and qualitative data on students' responses to mobile marketing messages and interpreted our campaign as a case study. However, we were also interested in determining the factors which contribute to the acceptance of SMS marketing. Several self-assessment scales were used for this purpose, which were correlated with measures of acceptance of mobile marketing.

The following hypotheses were formulated in relation to the problems of our pilot study:

- H1: Students in our convenience sample will on average demonstrate a high level of acceptance of SMS marketing.
- H2: The level of engagement of students in our convenience sample during the marketing campaign will not decrease over time.
- H3: Message attributes (related to brand, information value, entertainment, personalization, perceived usefulness) have an influence on the acceptance of mobile marketing.

## Methodology

Our convenience sample consisted of 62 students of the fourth year of study of informatics at the University of Zagreb, Croatia. The subjects were aged between 21 and 24, 66% of them were male and 34% female. All of the subjects voluntarily participated in the study.

Before the students provided the researchers with their mobile phone number they were given a written statement regarding the privacy and confidentiality of data collected in our pilot study. Also, prior to their participation in the mobile marketing campaign the subjects responded to a survey regarding their actual use of mobile technology (phone, Internet etc.). In addition, the survey consisted of self-assessment scales that measured various constructs (user attributes) that could be related to the acceptance of mobile marketing. These constructs were associated with attitude about mobile advertising (Shimp and Kavas, 1984; Pollay and Mittal, 1993), perceived usefulness (Venkatesh et al., 2003), message attributes (brand, information value, entertainment, personalization), and use of services related to mobile advertising (Merisavo et al., 2007). Most of these self-assessment measures demonstrated an internal reliability in the range from satisfactory to very good (Cronbach alpha from 0.70 to 0.90).

The mobile marketing campaign in our pilot study lasted two consecutive weeks. During this campaign the subjects received SMS messages that were related to their college, program of the local cinema and theatre, city swimming pool, student restaurants and town bars/pubs, or included diverse advertisements of products and services. The subjects were asked to respond to each of the received SMS messages regarding its usefulness on a 1-5 Likert-type scale. After a week of participation in the mobile marketing campaign they also completed a brief survey regarding the received messages and their effects.

## Results

After a week of exposure to the mobile marketing campaign the subjects in our convenience sample responded to the survey question "Mobile advertising is a useful concept which I plan to use in the future" with an average response of 4.14 on a Likert-type scale ranging from *"1 – I do not agree at all"* to *"5 – I completely agree"*. Typical positive verbal responses of subjects to the campaign after its first week were: *"Many messages were very useful to me (menu at the student restaurant, cinema, discount at the city swimming pool, college information) and I will miss them when the campaign is over."*; *"Interesting, I approve of it and support this kind of advertising as long as the user is capable of choosing and controlling the content that he/she receives and its amount."*; *"I support this and hope that it will benefit the students"*. In total there were 19 positive verbal responses to the mobile marketing campaign in our pilot study, three neutral verbal responses which were predominantly related to the need for personalization of the campaign form and content of messages. There was only one negative verbal response which indicated that this student considered mobile marketing as a form of spam. Both the quantitative and qualitative data colleted in our pilot study confirm the first hypothesis (H1) and it can be concluded that the acceptance rate of mobile advertising was rather high and that most of the reactions of the subjects were positive.

Since the subjects in our pilot study were asked to respond to each of the received SMS messages on a 1-5 Likert-type scale regarding how useful the message was for them we were able to indirectly measure the effect of each message and of the campaign as a whole. It must be noted that most of the SMS's (85% of messages) during the *first week* of the mobile marketing campaign were highly personalized and received an average rating for usefulness of 3.0 or above from the subjects in the study. However, the messages during the *second week* of the campaign were more oriented toward specific products/services and received a lower average usefulness rating (only 62% of those SMSs received an average rating of 3.0 or above). The interactivity of the campaign in our pilot study was also higher in the first week, with about 70% or more subjects responding with their ratings or comments to most of the incoming SMSs. In Figure 3 the percentage of responses to the marketing SMSs in our pilot study is displayed. A lower percentage of responses is evident throughout the second week of the campaign. The percentage of students' responses with an evaluation of the received marketing SMS could be considered an indirect measure of their interest in the campaign. Therefore the second hypothesis (H2) was not confirmed since the level of engagement of students in the marketing campaign was not stable but decreased over time.

In our initial survey we used several self-assessment scales to investigate the potential acceptance factors of mobile marketing. As a measure of *acceptance of mobile marketing* we used a self-assessment scale with four items and with internal consistency (Cronbach alpha) of 0.83. The items of this scale were predominantly related to the intention of future use of mobile marketing. To inves-

tigate the relations of marketing message attributes with acceptance of mobile marketing this variable was correlated with message related constructs like brand information in message content, information and entertainment value of message, personalization of message content, and expected usefulness of participation in a mobile marketing campaign. Most of the self-assessment scales which were designed to measure those constructs had a satisfactory internal consistency (see Table 2).

The data presented in Table 2 confirms that message attributes have an important effect on the acceptance of mobile marketing. The *expected usefulness of participation* in a mobile marketing campaign (which is related to the content of the messages the subject is exposed to), the *personalization of the content* of messages (receiving predominantly those messages which are of personal interest to the subject), and the *information value* of the received messages had the strongest association with the acceptance of mobile marketing in our pilot study. The brand of product/service and the entertainment value of messages were also associated with the acceptance of mobile marketing although to a lesser degree. These results confirm our third hypothesis (H3) – that the acceptance of mobile marketing is related to the messages the user will be exposed to during a marketing campaign. However, the influence of the initial interest should not be disregarded since our pilot study indicated that after a period of exposure to mobile marketing messages the interest of the recipients may decrease (see Figure 3).

Figure 3. Percentage of subjects who responded with an evaluation of the received marketing SMS during the first and second week of the campaign

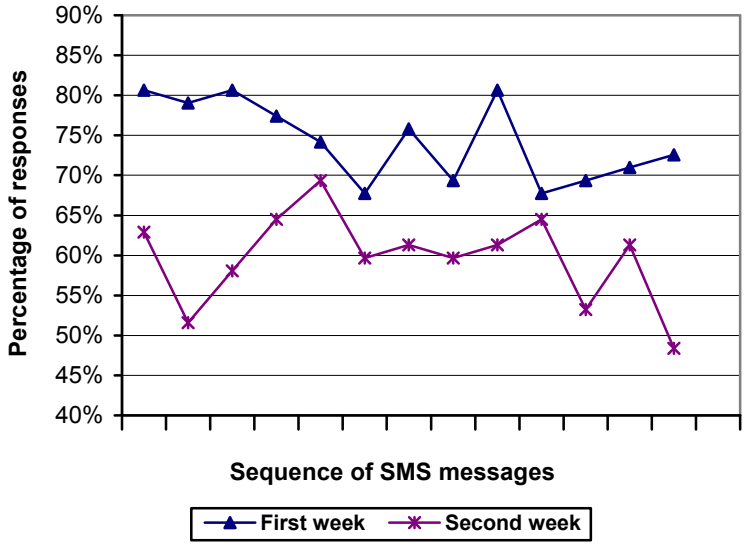


Table 2. Correlation of acceptance of mobile marketing with the constructs related to the attributes of mobile marketing messages and internal consistency of scales used to measure these constructs (N=62)

| MESSAGE RELATED CONSTRUCTS | CORRELATION WITH ACCEPTANCE OF MOBILE MARKETING | INTERNAL CONSISTENCY OF SCALE *(CRONBACH ALPHA)* |
|---|---|---|
| Attractive brand in messages | 0.47 | 0.90 |
| Information value of messages | 0.59 | 0.81 |
| Entertainment value of messages | 0.33 | 0.65 |
| Personalization of message content | 0.67 | 0.70 |

Most of our data indicate that for the subjects in our convenience sample the participation in the mobile marketing campaign was generally positively accepted. Furthermore, after the first week of the mobile marketing campaign the subjects were asked to rate how uncomfortable they felt because of the participation in our study about acceptance of mobile marketing. It must be noted that none of the participants indicated that they felt uncomfortable because of their participation in our pilot study. Furthermore, the verbal content of their SMS responses to received mobile advertising messages was usually positive and did not indicate annoyance or other kind of unpleasant emotion, except for the advertising messages received on Sunday, which were not approved of by some of the subjects in our study.

## Conclusion

The mobile phone is one of the most widely used technological devices and SMS is one of the most individualized channels of electronic marketing communication. However, marketing over a mobile phone may not have been widely accepted because of privacy concerns and technological complexity. In our pilot study we first resolved the technological problem with a simple Java application and a low cost solution to send marketing SMSs to the subjects. Then, by conducting a survey on a convenience sample of 62 informatics students we found out that mobile marketing could be accepted at least by a younger generation of computer literate mobile phone users. Most of the subjects positively evaluated their experience with the mobile marketing campaign in our pilot study. However, our findings also indicate that user/consumer interest in the participation in mobile marketing campaigns may decrease over time and that the mobile advertising messages need to be personalized and contain adequate information value to attract attention, receive greater interest, and avoid being perceived as spam.

## References

Ahonen, T. Putting 2.7 Billion in Context: Mobile Phone Users. // January 2007, the blog of the book "Communities dominate brands: Business and marketing challenges for the 21st century", by Tomi T. Ahonen and Alan Moore. 2007. http://communities-dominate.blogs.com/brands/2007/01/putting_27_bill.html (10 July 2009)

Ahonen, T. Bigger than TV, bigger than the internet: Understand mobile of 4 billion users. // February 2009, the blog of the book "Communities dominate brands: Business and marketing challenges for the 21st century", by Tomi T. Ahonen and Alan Moore. 2009. http://communities-dominate.blogs.com/brands/2009/02/bigger-than-tv-bigger-than-the-internet-understand-mobile-of-4-billion-users.html (10 July 2009)

Barnes, S.J.; Scornavacca, E. Mobile marketing: The role of permission and acceptance. // International Journal of Mobile Communication, 2 (2004), 2, 128-139.

Chaffey, D. E-Business and E-Commerce Management, London: Prentice Hall, 2003.

Cleff, E. B. Implementing the legal criteria of meaningful consent in the concept of mobile advertising. // Computer Law and Security Report, 23 (2007), 262- 269.

Dickinger, A.; Haghirian, P.; Murphy, J.; Scharl, A. An investigation and conceptual model of SMS marketing. // Proceedings of 37th Hawaii International Conference on System Sciences, (HICSS'04). Hawaii, USA. Track 1, vol. 1, 2004, 31-41.

Godin, S. Permission Marketing: Turning Strangers into Friends, and Friends into Customers. New York: Simon and Schuster. 1999.

Gopal, R. D.; Tripathi, A. K. Advertising via wireless networks. // International Journal of Mobile Communications, 4 (2006), 1, 1-16.

Jelassi, T.; Enders, A. Leveraging wireless technology for mobile advertising. // ECIS 2004 Proceedings. Paper 50. 2004. http://aisel.aisnet.org/ecis2004/50 (10 July 2009)

McManus, P.; Scornavacca, E. Mobile marketing: Killer application or new hype? // Proceedings of the International Conference on Mobile Business (ICMB'05), Sydney, Australia. 2005.

Merisavo, M.; Kajalo, S.; Karjaluoto, H.; Virtanen, V.; Salmenkivi, S.; Raulas, M.; Leppäniemi, M. An empirical study of the drivers of consumer acceptance of mobile advertising. // Journal of Interactive Advertising, 7 (2007), 2. http://www.jiad.org/article92 (10 July 2009)

O'Shea, D. Small screen for rent. // Telephony.Online, 34 (2007). http://telephonyonline.com/wireless/news/telecom_small_screen_rent/ (10 July 2009)

Pousttchi, K.; Wiedemann, D.G. A contribution to theory building for mobile marketing: Categorizing mobile marketing campaigns through case study research. // Proceedings of the International Conference on Mobile Business – ICMB 2006, Copenhagen, Denmark. June 2006, 26 - 27.

Salo, J.; Tahtinen, J. Retailer use of permission-based mobile advertising. In Irvine, III Clarke, Theresa B. Flahertypp (Eds.), Advances in Electronic Marketing, Hersshey, PA: Idea Group Publishing. 2005, 139-155.

Schreiber, G.A. Schluesseltechnologie Mobilkommunikation. Koeln: Deutscher Wirtschaftsdienst. 2000.

Sinisalo, J.; Salo, J.; Karjaluoto, H.; Leppäniemi, M. Managing customer relationships through mobile medium — underlying issues and opportunities. // Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06). Track 6, vol. 6, 2006, 112b.

Spurgeon, C. Advertising and new media. London: Routledge. 2008, 95-100.

# The eOffice Project by Ericsson Nikola Tesla

Marko Lukičić
Ericsson Nikola Tesla
Krapinska 45, Zagreb, Croatia
marko.lukicic@ericsson.com

**Summary**

*In the last ten years separate EU countries has brought specifications to support delivery of eGovernment services to citizens and businesses. However, at the national level, Croatia still lacks of such interoperability standards and frameworks.*
*This paper is presenting a work on the eOffice solution developed by Ericsson Nikola Tesla company. The aim of this project is to offer the standardized eOffice application core that fulfills both national legislative and regulative and general practice. Such solution will help government institutions to deliver new services to public in a context of eGovernment.*
*First, existing Croatian eGovernment legal foundation is presented. National strategies and novelties in the new version of Records Management Ordinance are depicted. Finally, the state of the work and the experience on eOffice project are presented.*

**Key words:** eOffice, eGovernment, writing office

## Introduction

In the last century societies has become complex as never before. There have never been more government services provided to citizens and businesses than now. There have never been more open cases than now and the cases have never been more complex. Furthermore, citizens and businesses has never been such demanding as now requesting new services and optimization of old ones.
To cope with these problems, governments start to transform into electronic governments or eGovernments. eGovernment is a general term for using information and communication technologies (ICT) to facilitate more accessible government services, allow greater public access to information, and make government more efficient and accountable to citizens[1]. It should facilitate opening of new services and optimization of existing ones to sustain growth of complex society.

[1] C. Jain Palvia, Shailendra; S. Sharma, Sushil. E-Government and E-Governance: Definitions/ Domain Framework and Status around the World. International Congress of e-Government. Foundations of e-Government. 2007.

However, before implementing eGovernment two critical tasks must be performed: one at the national level and another at the institutional level. At the national level, government must agree on national-wide technical and non-technical eGovernment frameworks, strategies, standards and principles. At the institutional level, every institution must reconsider its IT infrastructure and back-end systems and applications in respect to national eGovernment guidelines to support the eGovernment initiative.

Although the first task is crucial for eGovernment implementations, in practice the second task has shown as stumbling stone on the eGovernment way. The main reason lies down in inadequate ICT infrastructure, at the institutional level, not capable to follow national eGovernment guidelines. The infrastructure deficiency is caused by lack of sufficient enterprise architecture. This can be visible on all infrastructure layers such as: data layer, security layer, interoperability layer, legacy adapters, etc. Another problem is lack of functionalities (or non-compatible functionalities with eGovernment guidelines) in back-office systems. In most cases this lack is caused by:

- institutional regulations that doesn't comply with eGovernment framework, and
- back-end applications that only partly implements business processes leaving some of its parts un-automated.

If such ICT infrastructures remain unchained, implementation of eGovernment at the local institution level can lead to cost burden and administration efficiency degradation. As a classical front-office can not be dismissed, higher costs can be introduced by new Web based services that acts as Web based front-office. As this services still stays un-automated, additional labor must be performed at back-office to process the Web requests issued by citizens and businesses.

For that reasons re-evaluation of back-end systems should be a pre-task performed before implementing eGovernment at the institutional level.

One of crucial back-office applications is a writing office application. It is often considered as an application for nothing more than keeping tracks of all inbound and outbound communication (mainly mail) and case file documentation. However, with an appearance of new communication channels (such as E-mail) together with introduction of electronic documents and records into institutions, such applications are upgraded to provide functionalities such as: management and delivery of electronic files, cases and records, integration with case file management systems and archives, integration with collaboration portals, etc.

These applications are called eOffice applications, and offices that benefits of those applications are called *paperless offices*[2].
This paper introduces the legal foundation for records management and eGovernment. National strategies together with the new Records Management Ordinance are discussed. The overall Ericsson Nikola Tesla's eOffice project, an experience on all three implementation phases and current project state are presented.

## The legal foundation

In 2009, Central State Administrative Office for e-Croatia[3] has issued two strategic documents regarding eGovernment initiative in Croatia. Electronic Government Strategy of the Republic of Croatia for the Period from 2009 to 2012[4] was adopted by Croatian Government in January 2009. The document presents a foundation for building of modern, transparent, efficient and streamlined public services for citizens. The strategy introduces ICT as a fundamental tool for reforming the public administration in respect to Public Administration Reform Strategy issued by ex Central State Office for Administration (later transformed to Ministry of Public Administration)[5]. The document introduces two phases of the eGovernment implementation. In the first phase, the government will:

- assess the current information systems, communication networks and eGovernment services, and
- set a single methodology and standards for the functioning of the various segments of eGovernment.

In the second phase, the public authorities will have at their disposal a complete ICT infrastructure enabling them to communicate with each other in a unified environment.
Establishment of an unambiguous system for managing electronic documents is stated as one of the strategy targets. Development of project documentation for the eOffice reference model is stated as one of the target's activities. It is aimed

[2] Volarevic, Marijo; Strasberger, Vito; Pacelat, Elvis. A Philosophy of the Electronic Document Management. Proc. of the 22nd International Conference on Information Technology interfaces; 2000 Jun 13-16; Pula, Croatia. Zagreb: SRCE University Computing Centre, University of Zagreb; 2000. p. 141-146.;

[3] Središnji državni ured za e-Hrvatsku

[4] Central State Administrative Office for e-Croatia, Electronic Government Strategy of the Republic of Croatia for the Period from 2009 to 2012, The Government of Republic of Croatia, URL: http://e-hrvatska.hr/sdu/en/Dokumenti/StrategijeIPrograms.html, 2009.

[5] Central State Office for Administration, Public Administration Reform Strategy, The Government of Republic of Croatia, URL: http://www.uprava.hr/strat-hr.pdf, 2008.

to be used as a basis for the future implementations of eOffice applications for the public institutions.
The second document is The Action Plan for the Implementation of the One-Stop-Shop Program[6]. This operational plan, adopted by Croatian Government in June 2009, defines targets, tasks and services as well as implementation and audit mechanisms for a realization of the One-Stop-Shop program. Standardization of electronic systems for the management of electronic and non-electronic records in compliance with MoReq2[7] specification is stated as one of the measures needed for the One-Stop-Shop implementation. MoReq2 is a comprehensive catalogue of generic requirements for an Enterprise Records Management (ERM) system. It builds on the original MoReq specification, which was published in 2001. Specification is intended for use in public and private sector organizations which wish to use ERM systems. This standardization is a foundation for adjustment of back-office applications (such as eOffice) to the eGovernment architecture.

**Records Management Ordinance**
Records Management Ordinance[8] was adapted by Croatian Government on January 2009 with the purpose of modernizing writing office by automating its processes and introducing electronic documents. It is a cornerstone document for any eOffice application functional specification.
However, this ordinance is conceptually identical to the previous one[9] that originates from the year 1987. The document covers classical mechanisms for case prosecution audit trough evidencing and tracking the case file flow in the institution. Still, the ordinance does not cover the management of other non-case documentation regardless how important non-case documentation is for the operation of institutions.
Moreover, making it the central body for mail and file case documentation distribution the regulation is conceptually still oriented to writing office. On the other hand, implementation of eOffice applications removes this documentation flow bottleneck enabling direct distribution of electronic documentation from

[6] Central State Administrative Office for e-Croatia, The Action Plan for the Implementation of the One-Stop-Shop Program, The Government of Republic of Croatia, URL: http://e-hrvatska.hr/sdu/en/Dokumenti/StrategijeIPrograma.html, 2009.

[7] Cornwell Affiliates plc. Model Requirements for the management of electronic records. MoReq2 specification. Office for Official publications of the European Communities as INSAR supplement VIII. Bruxelles. Luxembourg. 2008.

[8] The Government of Republic of Croatia, Records Management Ordinance, Narodne novine, 07/09, January 2009.

[9] SR Croatia, Records Management Ordinance, Narodne novine, 38/87 and 42/88, 1987.

one referent directly to another. This documentation flow can be managed by previously defined business processes and well tracked in the overall system.
Another document's defiance is its primary concentration on typical document cycle aspect of records management that consists the phases: receive evidence, distribute, dispatch and archive. However, classification scheme configuration and management, rights management, management of retention periods and disposal schedules are not covered by the ordinance.

## The eOffice project

The purpose of eOffice project is to implement a core for the eOffice application which will:

- form a comprehensive basis for the eOffice application,
- comply with the legal foundation, corresponding standards and eGovernment trends,
- provide a basic eOffice functionalities (functionalities defined by Records Management Ordinance combined with the best practice)
- provide a broad and extensible data and security model, and
- allow easy and time non-challenging upgrade with specific client's functionalities to a final product.

eOffice project is split into three main phases. The first phase is an investigation phase. The investigation phase is consisting three components: economic, operational and technical. The operational component consists of methodology selection and setting up the project. Two most popular classes of methodologies have been evaluated: the classic and agile methodologies. Because of particular specificities of eOffice project, it was decided that a RUP (Rational Unified Process)[10] methodology, as a classic methodology, will be modified in some areas. This modifications were concerned mainly on the structure and content of the particular documents prescribed within RUP methodology.
The technical component is the crucial one. Its purpose is to identify key eOffice processes and functionalities and to help with selection of key technologies. In this phase Records Management technology was recognized as basis for the eOffice. The different Records Management platforms were evaluated against earlier identified project requirements and MoReq specification as well. Finally, the IBM FileNet was recognized as a platform that suits the eOffice solution most. As FileNet platform consist of a Content Management module, Business Process module and Records Manager module, the eOffice application can be easily upgraded to support specific case file management, automate specific business processes and implement an archive – what is recognized as crucial upgrades for achieving full back-office automation, Figure 1. The platform's

[10] IBM. Rational Unified Process; Best Practices for Software Development Teams. Rational Software Whitepaper. Rational Software. 1998.

richness is recognized as one of the key elements for reducing further infrastructure and system integration investments.

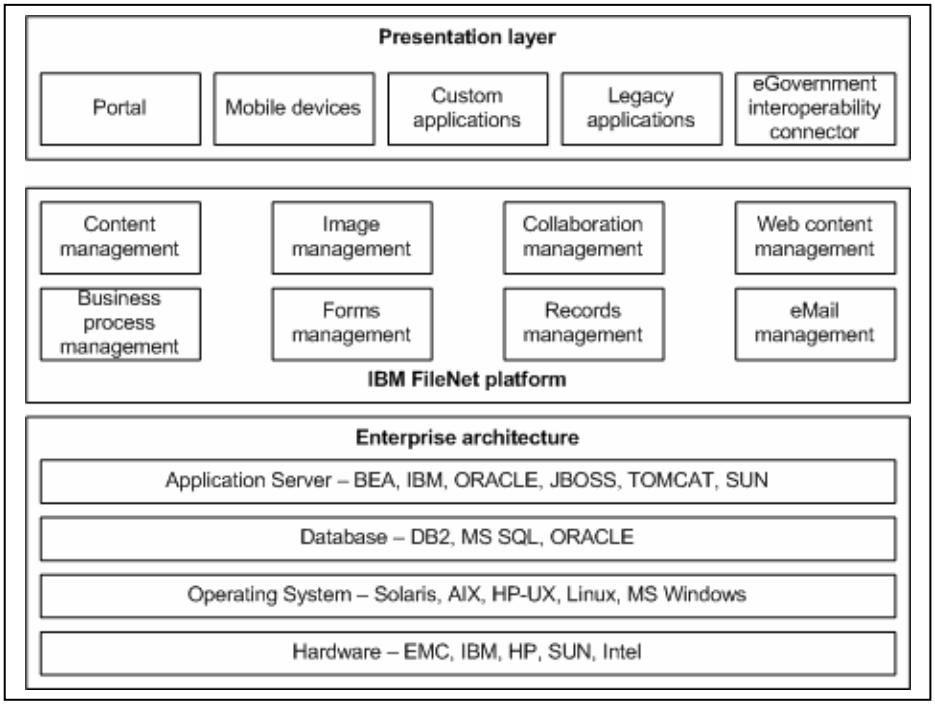



Figure 1. The FileNet platform in the enterprise architecture.

The first phase generated eOffice Conceptual Design document as an output. This document consists of:

- a basic description of the eOffice application,
- a description of organizational and security aspects,
- a preview of all business processes showed in BPM notation[11],
- use cases for every task in every process,
- an example of user forms, and
- a description of additional functional and non-functional requirements.

eOffice Conceptual Design was the input for the second phase: the preparation. In the preparation phase all documentation needed for eOffice application programming must be done. The basic technology foundation is extended with specific development technologies and sub-systems in accordance with functional and non-functional requirements from the conceptual design. As European Un-

[11] A. White, Stephen. Business Process Modeling Notation. The Business Process Management Initiative. May, 2004.

ion suggests the use of open standards[12], main technology criteria was support of: UML or RDF for data modeling, XSLT for data transformation, Dublin Core (possibly with national extensions) for metadata, Web Services for interoperability, etc. The Solution Design document is the end result of this phase. It contains all implementation details needed for programming the eOffice. It details the solution from deep technical perspective and elaborating each eOffice technical layer (ex. data layer, business layer, etc.) from the programmer's viewpoint (ex. detail specification of database definition, object model, program modules, etc.).
The last phase is the Development and Test phase. In this phase teams of developers maps solution design instructions into program modules, develops final eOffice program, and develops test scripts, Figure 2. Test scripts are performed on a testing environment to indicate on eventual program defects and weaknesses. Key users, such as domain experts, perform functional and ergonomic tests, and database specialists perform consistency tests on a data model trying to cause application malfunction by corrupting data. The development and test phase is considered as finished when all tests results meets predefined criteria.

**Current status**

The investigation and preparation phases are successfully completed with accepted eOffice Conceptual Design and eOffice Solution Design documents. The development and test phase was started and in progress at the time of writing the article. A FileNet team of specific domain experts (such as developers, database administrators, etc.) is allocated. This team will carry out the most tasks of design and test phase.
Until now, the FileNet team customized a FileNet platform to meet eOffice needs, and already developed the critical components of the program (such as data and object model, security, audit, etc.). Most of writing office processes are designed and implemented in FileNet platform using FileNet Process Designer. A basic concept of graphic user interface is designed and connected to FileNet processes. However, there still remains one large task to be performed: implementation of testing scripts and performing tests. This task will start with finalization of the eOffice development.

[12] Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens. European Interoperability Framework for Pan-European eGovernment Services. European Commission. Office for Official Publications of the European Communities. ISBN: 92-894-8389-X. 2004.

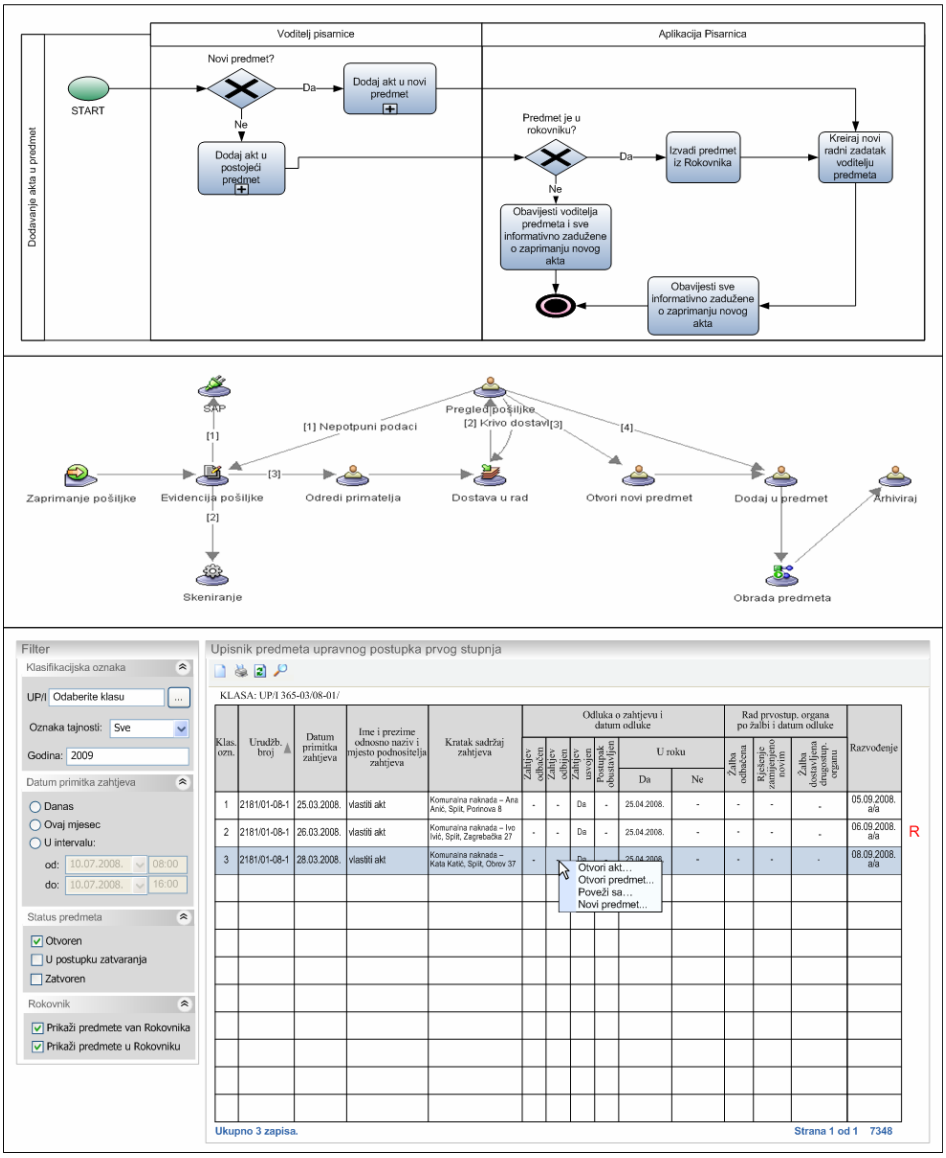



Figure 2. Top: An example of the process definition in BPM notation.
Middle: An example of a process implemented in FileNet platform.
Bottom: Graphical user interface of the eOffice application running FileNet processes in the background.

## Conclusion

This paper presents the work on the eOffice application. As this application is building with keeping in mind the legal foundation, national strategies, open standards and EU standards and guidelines regarding eGovernment, the eOffice will avoid incompatibility of back-office systems when implementing the eGovernment framework at the institutional level.

At the other hand, considering not only Records Management Ordinance but best practices as well, we believe that the eOffice application provides additional functionalities that will empower end users with all features of paperless office.

From the technical perspective, with the reliance on comprehensive platform, such as FileNet, IT departments can accomplish homogenous infrastructures, minimize the future investments and benefit from total cost reduction.

## References

A. White, Stephen. Business Process Modeling Notation. The Business Process Management Initiative. May, 2004.

C. Jain Palvia, Shailendra; S. Sharma, Sushil. E-Government and E-Governance: Definitions/ Domain Framework and Status around the World. International Congress of e-Government. Foundations of e-Government. 2007.

Central State Administrative Office for e-Croatia, Electronic Government Strategy of the Republic of Croatia for the Period from 2009 to 2012, The Government of Republic of Croatia, URL: http://e-hrvatska.hr/sdu/en/Dokumenti/StrategijeIPrograma.html, 2009.

Central State Administrative Office for e-Croatia, The Action Plan for the Implementation of the One-Stop-Shop Programme, The Government of Republic of Croatia, URL: http://e-hrvatska.hr/sdu/en/Dokumenti/StrategijeIPrograma.html, 2009.

Central State Office for Administration, Public Administration Reform Strategy, The Government of Republic of Croatia, URL: http://www.uprava.hr/strat-hr.pdf, 2008.

Cornwell Affiliates plc. Model Requirements for the management of electronic records. MoReq2 specification. Office for Official publications of the European Communities as INSAR supplement VIII. Bruxelles. Luxembourg. 2008.

IBM. Rational Unified Process; Best Practices for Software Development Teams. Rational Software Whitepaper. Rational Software. 1998.

Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens. European Interoperability Framework for Pan-European eGovernment Services. European Commission. Office for Official Publications of the European Communities. ISBN: 92-894-8389-X. 2004.

SR Croatia, Records Management Ordinance, Narodne novine, 38/87 and 42/88, 1987.

The Government of Republic of Croatia, Records Management Ordinance, Narodne novine, 07/09, January 2009.

Volarevic, Marijo; Strasberger, Vito; Pacelat, Elvis. A Philosophy of the Electronic Document Management. Proc. of the 22nd International Conference on Information Technology Interfaces; 2000 Jun 13-16; Pula, Croatia. Zagreb: SRCE University Computing Centre, University of Zagreb; 2000. p. 141-146.

Professional paper

# Implementation of Digital Repository at the Ruđer Bošković Institute: Organizational and Technical Issues

Alen Vodopijevec
Ruđer Bošković Institute, Library
Bijenička cesta 54, Zagreb, Croatia
alen.vodopijevec@irb.hr

Bojan Macan
Ruđer Bošković Institute, Library
Bijenička cesta 54, Zagreb, Croatia
bojan.macan@irb.hr

**Summary**

*This paper will focus on implementation of digital repository at the Ruđer Bošković Institute (RBI). Based on the RBI needs and considering consolidation of IT services, it was decided that RBI digital repository will include three different types of archived content: scientific output, documentary and press clipping materials. This is the main difference against other commonly implemented institutional repositories which archive mostly only scientific articles and thesis. Choosing of appropriate software, its implementation and customization will be discussed, as well as the methodology of gathering content for the repository and the method of its archiving. Archived material will be organized in collections based on the RBI needs and divisional structure. During the process of creation of metadata schemes for different types of archived content, especially those of scientific character, special attention was paid to OAI compliance with other digital/institutional repositories, as well as with Croatian Scientific Bibliography (CROSBI). The goal is that researchers should archive scientific materials primarily in RBI digital repository, while metadata will be automatically exported to CROSBI. Successful implementation of this data-exchange mechanism could also be useful to other institutions which are considering creating their own repository.*

**Key words:** institutional repository, digital repository, IR, CDS Invenio, Ruđer Bošković Institute, open access, OA

## Introduction

The Ruđer Bošković Institute is the largest Croatian scientific research institution in the fields of natural sciences. In the multi-disciplinary environment of

the Institute 530 academic staff (375 researchers and 155 Ph.D. students) [8] work on problems in experimental and theoretical physics, chemistry and physics of materials, organic and physical chemistry, biochemistry, molecular biology and medicine, environmental and marine research and computer science and electronics. Institute has 11 divisions, 3 centers, a library and sections for maintenance, technical service and administration. In the year 2008, the RBI have had 136 projects in basic research, which are funded by the Ministry of Science, Education and Sports (MSES), as well as 41 international projects, 67 applied and technological projects and 4 HITRA projects [8]. RBI staff is also active in teaching at universities and in 2008 they contributed 78 undergraduate courses and 245 graduate courses to the program of higher education in Croatia. The total number of research articles published by RBI scientists in 2008 was 446, whereof the majority was published in high ranking international journals [8].
Institutional repositories are "digital collections that capture and preserve the intellectual output of a single or multi-university community" [3]. Benefits from IR are many. Individual researchers gain better visibility of their papers which can, therefore, be cited earlier and more often than papers which are not in OA. Research community can find and access information more easily and institutions gain on their visibility and prestige by collecting all its scientific output in one place, rather than to be spread amongst hundreds of journals [10].
On July 17th, there were 4 active institutional repositories in Croatia (School of Medicine, Faculty of Philosophy, Faculty of Mechanical Engineering and Naval Architecture, all from University of Zagreb and Digital repository of the Information Sciences Department at the Faculty of Philosophy, University of Osijek) and Portal of scientific journals of Croatia – HRČAK. At the same time, there were 1429 repositories registered in Directory of Open Access Repositories - OpenDOAR (4 from Croatia) [13] and 1411 in Registry of Open Access Repositories - ROAR (3 from Croatia) [7].

## The idea about RBI digital repository

The idea of opening the science to the public was the main "spiritus movens" that led towards the RBI Digital Repository project proposal. As stated in the document entitled "Science and Technology Policy of the Republic of Croatia 2006-2010", scientific and research output produced within publicly funded projects should be freely available to the public [15]. Considering the size, status and importance of the RBI for the Croatian and International academic society, and new trends in scientific communication, the RBI Library came up with idea about implementation of institutional repository at the RBI which should be such platform for depositing and disseminating the results of publicly funded projects. The idea was initially born in 2006 and Library wrote a draft of the proposal project for an institutional repository, which was introduced to colleagues from Public Relations office (PR Office) and from Center for Informatics and Computing. They supported the idea and suggested that the Project should be expanded and incorporate

other digital content produced on RBI. These suggestions were adopted so the project proposal was rearranged with their help and its final version was created in July 2007 and presented to the RBI administration [4]. The Project was approved, as well as its financial construction for a period of first two years needed for acquisition of necessary hardware. Also, one person was designated to the library for working part-time on digitization and organization of RBI old documentary materials (photographs).

The main objectives of this Project were:

- archiving and preservation of digital content of the RBI
- gathering all scientific output of the Institute on the single site and offering it in OA to the community
- creating a digital archive for archiving of documental and press clipping materials about the RBI
- increasing of the RBI's visibility and it's scientific contribution to the science
- promoting of the OA initiative at the RBI and in Croatia
- helping RBI staff to publish their work on their personal web pages without fears of breaking copyright law

RBI digital repository will consist of three virtually separated parts:

- self-archiving online platform of RBI's **scientific output** based on OA principles. Such platform is commonly referred as "institutional repository";
- **documentary archive** - online digital storage of important historical and current multimedia content produced by RBI or with RBI as main theme of such contents;
- **press-clipping archive** - online digital storage of press-clipping content.

It was planed that the implementation of the RBI digital repository will take place in 7 phases:

1. Setting up hardware and software
2. Resolving copyright and licensing issues
3. Training the personnel for digitalization and administration of the repository
4. Digitalization of the documentary materials and initial data archiving
5. Presenting repository to RBI staff and OA advocacy
6. Depositing materials and regular maintenance of the system
7. Establishing institutional self archiving mandating policy and depositing license

## RBI digital repository

### Choosing appropriate software system

After deciding what kind of repository does RBI need, it was necessary to choose appropriate open source software for it. Software that suit RBI's needs would have to fulfill following requirements:

- open source software
- functional and extendible integrated search engine

- OAI-PMH compliance
- integrated standard internationalization an localization functions
- variety of standards and data formats for metadata representation
- group or role based access right privileges system
- fully customizable collection tree structure.

Therefore testing of several software options (CDS Invenio, EPrints and DSpace) was conducted (table 1). As a result of this testing, CDS Invenio (http://cdsware.cern.ch/) was recognized as the most promising solution for RBI’s multi-purpose repository and it was decided to do more detailed tests on it. Despite that, it is still possible to change it if better software appears on the market. Until now a significant amount of work was done on localization and analyzing administration interface, especially the submission process. CDS Invenio is very flexible system and despite the fact that it requires lots of work on its customization, as well as other tested products, it was concluded that with Invenio it is possible to do more in less time.

Table 1: Feature comparison of tested software products

| **Tested functionality** | **CDS Invenio** | **DSpace** | **Eprints** |
|---|---|---|---|
| Programming language | Python | Java | Perl |
| Database engine | MySQL | PostgreSQL | MySQL |
| Localization | Yes – standard .po files | Yes – Language packs based on Java Standard Tag Library | Yes – XML files |
| Customization of UI and depositing system | Web interface, extensive HTML and Python programming required. | Web interface, larger changes require Java coding and recompiling application. | Majority of configuration is handled by XML files and some HTML templates. Basic knowledge of respective technologies required. |
| Default metadata standard | MARC with possible export to other standards | DC | DC |
| User authentication | Local database, LDAP, Shibboleth | Local database, LDAP | Local database, LDAP |
| Access control | Role based | Groups as roles | 3 default groups |
| Search engine | Python custom | Java Lucene | Perl custom |

Complexity of installation and maintenance was not rated nor compared because it depends on the extent in which one would have to customize default features of an application and, of course, it depends on competence and availability of IT staff team members involved in the process of implementation.

**Type of archived materials and supported formats of files**

As already mentioned, Digital repository of the RBI will archive three different types of materials: RBI's scientific output, documentary materials and press clipping materials about RBI. Those materials will be archived in textual, video, audio and pictorial form and repository will support uploading of all available file formats, although a certain formats for specific data types of archived materials will be preferred. Table 2 brings the list of preferred file formats for different types of digital content. Mentioned formats are preferred because of the possibilities to represent them on web UI, for e.g., showing images, streaming audio and video. Furthermore, OGG (theora for video and Vorbis, Flac for audio) are open-source and patent free.

Table 2: Preferred file formats of archived materials for different types of digital content

| Type of digital content | Preferred file formats of archived materials |
|---|---|
| Textual materials | PDF, DOC, ODF, RTF, PPT |
| Video materials | OGG (Theora) |
| Audio materials | OGG (Vorbis, Flac, Speex) |
| Pictorial materials | TIFF, JPG, PNG, PPT, ODF |

*RBI scientific output*

This is the type of materials which is usually archived in IR. It is scientific material in a form of articles, book chapters, books, reports, posters, data sets etc. Many of mentioned items are copyright protected and it will be necessary to investigate the terms under which those materials can be archived and in which version. This will be elaborated later in this paper when talking about copyright issues.

*Documentary materials*

Under this type of materials photographic materials of the Institute, its building, staff, equipment and all kind of promo materials (brochures, posters etc.) will be included. For that reason, a project of digitizing of those materials was conducted. All photographic materials were gathered, organized and digitized. During this process, it was realized that there's a great number of material, whereof lots of pictures are duplicates or very similar to each other and that the process of selection of those materials will be needed. It was also realized that it would be very useful to identify people on the photographs and to include those information to its description. Therefore it was decided to have a inter phase in which photographs will be uploaded in lower resolution into the online gallery (which currently holds over 6000 photographs from 1950s until today) and they will be visible to RBI staff who will be able to tag people on them and describe wider context of taken photographs. Number of visits to uploaded photographs will be one of the criteria for choosing which photographs will be archived to

the RBI digital repository, but the most important criteria will be historical significance of a certain photograph. It will be possible to archive single photograph or store more thematically related photographs into an album along with their short descriptions. Photographs will be archived in their original resolution and quality with smaller images for quicker preview.

*Press clipping material about RBI*

The third type of materials included to the RBI digital repository is press cut materials of all published articles about RBI, recorded TV or radio shows, as well as archive of press release material from RBI's Public Relations (PR). The idea is that RBI's PR write press release, deposit it to RBI digital repository and all interested media can download it from the repository and publish it. Permission to deposit press cut and press release materials will be given only to PR Office.

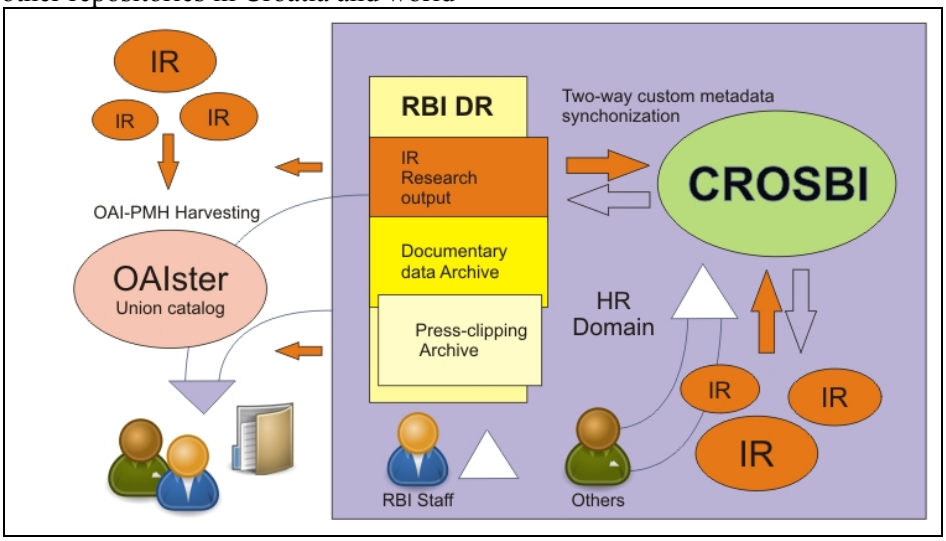
**Metadata structure and self archiving**

In 2002 the RBI Library started the project called Croatian Scientific Bibliographic Database (CROSBI) (http://bib.irb.hr/), a bibliographic database in which all Croatian scientists are "forced" by MSES to deposit metadata because they are obliged to attach the listing of their published works from it for their projects reviews, scientific promotion etc. Although it is primary bibliographic database, CROSBI also has an option for uploading full text documents. Until July 27$^{th}$ 2009, there were more than 243.000 bibliographic records archived into the database, whereof 13.000 full text documents [5].

Considering the fact that RBI staff has to deposit metadata about published papers in CROSBI database, and that the development of the new RBI digital repository is in progress, the plan is to make those two services compatible. Experience of institutions with similar situation were studied [1, 2] and it was decided that data synchronization between two systems should be enabled. That's why it was necessary to have metadata compatibility issue on mind when creating set of metadata needed for description of different types of materials expected to be hosted in the repository. While CROSBI is not based on any metadata standard, the metadata scheme used for describing items in the repository is based on MARC standard, but it could be mapped to other standard metadata formats as well (such as DC for OAI-PMH compliance). Therefore a mapping of CROSBI and RBI repository metadata fields will be done in order to enable this metadata exchange. Main goal regarding implementation of interoperability between existing services is to wipe out the need for multiple depositing of documents and/or multiple metadata submitting procedures. The idea is to insert metadata (and deposit document) only once, into researchers "home" repository and afterwards the system will do the replication process of certain types of content (scientific articles, posters etc.). Scientist would only have to supplement replicated record where applicable (e.g. add custom metadata regarding

publication details and category of an article in CROSBI). There will also be a possibility to harvest metadata from CROSBI to the RBI digital repository for initial import in Institutes repository. Successful implementation of this module could be later replicated by other institutional repositories in Croatia (Figure 1).

Figure 1: Interoperability of the Digital repository of the RBI with CROSBI and other repositories in Croatia and world




On the basis of metadata description, virtual collections will be formed. These virtual collections will be based on the RBI organizational structure and its specific needs, as well as on the fact that repository will archive three different types of materials which should be clearly visible at the repository homepage as separate collections.

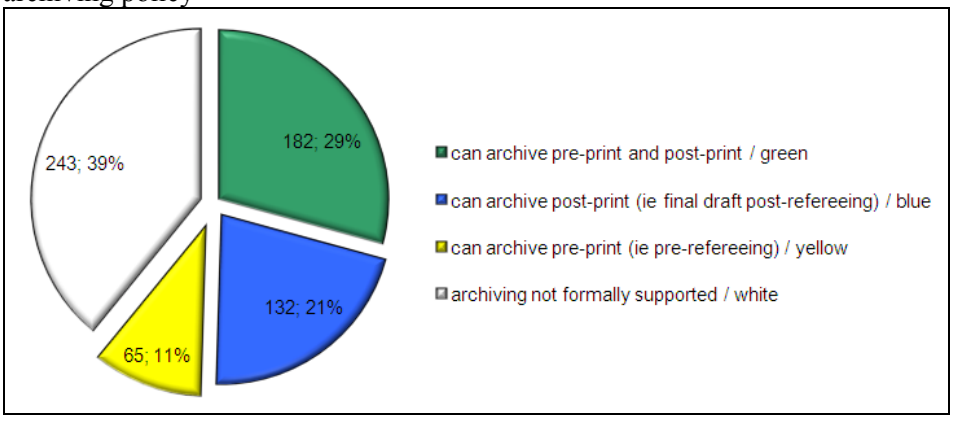
One of the often mentioned problems in literature about institutional repositories is problem about gathering content for repository [2, 6]. Some institutions decided that library staff will deposit items into the repository in behalf of their faculty staff [2], while others decided for self-archiving [11]. RBI scientists already have experiences with depositing metadata to CROSBI. To take advantage of this fact it was decided that our repository will be based on self-archiving, rather than on library staff depositing. Furthermore, it will be necessary to create institutional self-archiving mandating policy. This obligation shouldn't be a big problem for the RBI staff because they are already used to enter bibliographic data into CROSBI, and considering planned interoperability between two systems, it shouldn't take to much extra time for depositing.

**Copyright, licensing and access rights**

Copyright issues are crucial for every institution which is implementing an institutional repository and has to be considered very carefully. Copyright issues with documentary materials are in most cases very clear – RBI is a copyright holder and it can publish them in its repository. Press releases and press cut materials are, according to Croatian copyright law, free of copyright law and therefore, can be archived in our digital repository [14] with the exception of audio/video materials (TV and radio shows) which cannot be freely published.

However, the situation with scientific content is much more complicated because in majority of cases publications rights are transferred to the publisher. The SHERPA/RoMEO database of publishers is here out of great help. This database is used to determine the rights of authors to include papers published in scientific journals in the IR. At July 24th, SHERPA/RoMEO database included details of the policies of 662 publishers [9]. The publishers are divided into four groups according to archiving policy, and every group is represented with different color. Figure 2 brings data about publishers in RoMEO database according to their archiving policy. Most of publishers (61%) allow depositing of pre- or post-print version of paper in an IR.

Figure 2: Number and percentage of publishers in RoMEO according to their archiving policy




As mentioned before, archiving to the repository will be based on author's self-archiving and the final goal is to teach them how to check by themselves in SHERPA/RoMEO database which version of published paper can they archive to the repository. At first, the library staff will be there to help the authors to check the journal's policy about self archiving, but the library intervention is to be reduced to the minimum with time.

A suitable deposit license needs to be created and embedded into the submission form. This license will be based on the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported model (http://creativecom-

mons.org/licenses/by-nc-nd/3.0/) and the SHERPA model license and it will consist of clauses covering the institution's rights to store, disseminate and preserve deposited items, as well as providing assurance that copyright is owned by the author or permission has been given by the owner. Submitter of material will have to agree to the licensing terms and conditions or an item will be rejected.

Besides freely available contents, RBI digital repository will contain a great amount of digital data (documents, multimedia) that require access control mechanism, e.g., articles during embargo period, datasets that can be accessed only by few people, experimental data available to lab staff etc. CDS Invenio uses "role based access control" (RBAC) so it is possible to create fine grained rule sets for authorization purposes. User authentication and authorization mechanism relies on RBI's existing LDAP user database and further development is on the way to implement "Single Sign On" integration with Croatian Academic Authentication and Authorization infrastructure – AAI@EduHr. The same authentication and authorization mechanism users use for logging in to CROSBI and many other services developed by the RBI Library and other institutions in Croatia. Single sign-on principle is based on existence of central identity provider which guarantees that certain person is really who she pretends to be and it guarantees that she has certain rights when accessing service on service provider side. In Croatia every academic and research institution has its LDAP user database and there is central service maintained by SRCE (http://www.aaiedu.hr/) that acts as proxy between end user and service provider who wants to authenticate and/or authorize users through this mechanism. User can use one username and password but there is no interoperability so that remote independent services have the information that user is once logged in into system and can access any available resource. There comes Shibboleth (http://shibboleth.internet2.edu/) in place. Implementing of Sibboleth into CROSBI database is currently in the process and it was decided to use the same service for authenticating and authorizing users of RBI digital repository.

Benefits of such approach is that users have to log in only once and then they can access any available resource identified by Single sign-on system without the need for any further entering of login details. In that manner it is possible to restrict some content to much wider audience. For example, it is possible to configure group of Croatian students that have rights to download certain datasets, a group of users that belongs to some specific institution and, of course, RBI internal groups based on LDAP attributes such as Division, Labs, etc. that can access their "private" area within the repository.

## Future perspectives

To fulfill its goal and meaning IR must continuously grow and researchers must accept it as a way of scientific communication. A literature cites numbers of reasons why faculty participation rates are often so low and why are IR recruiting new content to slow: from the lack of awareness of the existence of institutional repositories [12], concerns about copyright and intellectual property is-

sues, concerns that depositing into the IR will be considered prior publication [6] to the fact that self-archiving is additional obligation for them and they are just not motivated to do it. There's no clear strategy that will lead for sure toward continuous and strong growth of the IR, but Mark and Shearer came to the conclusion that it is very important to promote repository on campus, because it raises awareness of the existence of the repository, and to establish institutional self archiving mandating policy [6]. That's why it is very important for RBI to complete institutional self archiving mandating policy as soon as possible to ensure better chances for successful growth of the repository. Furthermore, it is important for the RBI to stay on track with current achievements in exploring long-term data preservation strategies (backup solutions, possibility of access to data after longer period of time) and with ongoing activities concerning interoperability and disseminating of archived content e.g. OAI-ORE (http://www.openarchives.org/ore).

## References:


[1.] Afshari, Fereshteh; Jones, Richard. Developing an integrated institutional repository at Imperial College London. // Program: electronic library and information science. 41 (2007), 4; 338-352

[2.] Barwick, Joanna. Bulding an institutional repository at Loughborough University: some experiences. // Program: electronic library and information science. 41 (2007), 2;113-123

[3.] Crow, R. The Case for Institutional Repositories: A SPARC Position Paper, Scholarly Publishing and Academic Resources Coalition, Washington, DC. 2002. http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf (Access date: July 17th, 2009)

[4.] Digitalni repozitorij Instituta "Ruđer Bošković": prijedlog projekta. Knjižnica Instituta "Ruđer Bošković", Zagreb, 2007. (interni dokument)

[5.] Hrvatska znanstvena bibliografija. Skupni statistički podaci. http://bib.irb.hr/skupni_podaci (Access date: July 27th, 2009)

[6.] Mark, Timothy; Shearer, Kathleen. Institutional repositories: a review of content recruitment strategies. (2006). http://archive.ifla.org/IV/ifla72/papers/155-Mark_Shearer-en.pdf

[7.] Registry of Open Access Repositories (ROAR). http://roar.eprints.org/ (Access date: July 17th, 2009)

[8.] Ruđer Bošković Institute Annual Report 2008. Zagreb : Ruđer Bošković Institute, 2009

[9.] SHERPA/RoMEO. Publisher copyright policies & self-archiving. http://www.sherpa.ac.uk/romeo.php?stats=yes (Access date: July 24th, 2009)

[10.] Sparc Europe. Institutional Repositories: A Guide to Open Electronic Archive. http://www.sparceurope.org/resources/hot-topics/institutional-repositories (Access date: July 10th, 2009)

[11.] Sutradhar, B. Design and developement of an institutional repository at the Indian Institute of Technology Kharagpur. // Program: electronic library and information science. 43 (2006), 3; 244-255

[12.] Swan, Alma; Brown, Sheridan. Authors and open access publishing. // Learned Publishing, 17 (2004), 3; 219-226

[13.] The Directory of Open Access Repositories - OpenDOAR. http://www.opendoar.org/ (Access date: July 17th, 2009)

[14.] Zakon o autorskom pravu i srodnim pravima. Narodne novine. 79 (2003). http://narodne-novine.nn.hr/clanci/sluzbeni/2003_10_167_2399.html

[15.] Znanstvena i tehnologijska politika Republike Hrvatske: 2006.-2010. Zagreb : Ministarstvo znanosti, obrazovanja i športa, 2006.

# Clouds on IT Horizon

Sanja Mohorovičić
University of Rijeka, Faculty of Maritime Studies
Studentska 2, Rijeka 51000, Croatia
sanja.mohorovicic@gmail.com

**Summary**

*The aim of this paper is to present and elaborate a new IT trend – Cloud Computing. Cloud computing is rapidly evolving and is becoming increasingly popular. Cloud computing represents the move from desktop applications toward Web applications and services. The actual computing is moved into the cloud so the users are using third-party services and paying for them on pay-per-use model. The cloud (data and services) can be accessed from any device connected to the Internet. There are currently three distinct cloud service layers and three cloud computing types in existence. Main characteristics of cloud computing are presented and potential benefits are explained. Some concerns and things that need to be corrected are being discussed. Major cloud computing service providers and some of their services are mentioned, as well as some predictions for the future.*

**Key words:** cloud computing, computing, web applications, Internet

## Introduction

Computers are increasingly becoming an irreplaceable component of our everyday life. Information technology (IT) is evolving rapidly and new trends are persistently emerging. From its beginning up until today, the Internet gradually developed and received many new features. Subsequently this has changed the way we use computers, especially with the appearance of broadband connections and increased number of Internet users. Today, we spend more time in the Web browser than in the other (desktop) applications. A new trend in the IT field that is evolving rapidly, and is widely discussed approximately in the last two years is cloud computing.

The term cloud computing has become quite popular these days and many companies want to be part of it. But, what is cloud computing actually? Are we already using some form of cloud computing without even being aware of it? If you use a web-based e-mail account (e.g. Gmail) or a social networking web site (e.g. Facebook), then you have already experienced cloud computing. To use these services you don't have to install any software or save data on your computer – the software and storage for your account are in "the cloud".

Cloud computing uses the Internet as a platform to run applications and store data on servers. Data and services can be accessed from any connected device over the Internet thru Web browsers or specialized applications. The infrastructure behind the cloud is invisible to users which don't need to know how the cloud actually works. With this type of computing the users don't have to worry any more about the infrastructure. This means that organizations don't have to buy new and expensive computers or applications and invest in their expensive maintenance. Besides many benefits, cloud computing enables new forms of group collaboration at a distance.
What are advantages and disadvantages of cloud computing? What types of services exist? Can all desktop applications be moved to the cloud? Is cloud computing suitable for everyone? Is the data in the cloud safe? Some of these questions will be discussed in this paper.

## What is Cloud Computing?

Cloud computing is an emerging paradigm in IT industry which represents a major change in how we store data and run applications. It represents a move from desktop applications toward Web applications and services – computing is moved to a cloud of computers. The term first appeared in 2007.
To better understand the concept of cloud computing, the terms "cloud" and "computing" needs to be explained first.
The cloud is a metaphor for the Internet, which came from the cloud symbol that's often used in the networks diagrams.
Computing is the activity of using and developing computer technology. It includes different information operations performed on a computer.
When we join these two words together into one term, cloud computing, and try to define it, we encounter to a problem. Some providers present their services as cloud computing services but they aren't. Everybody seems to be talking about cloud computing but there's no unique definition and common agreement about what it really is so there are many different explanations. The question is which definition is the best? Therefore, a few definitions are mentioned.
Very simplified, cloud computing is using the Internet for computing needs.
Sam Johnston defines cloud computing as "the realisation of Internet ('Cloud') based development and use of computer technology ('Computing') delivered by an ecosystem of providers" (Johnston, 2008).
According to Kent Langley, cloud computing is "commercial extension of utility computing that enables scalable, elastic, highly available deployment of software applications while minimizing the level of detailed interaction with the underlying technology stack itself"(Langley, 2008).
Gartner (the world's leading information technology research and advisory company) defines cloud computing as "a style of computing where massively scalable IT-related capabilities are provided "as a service" using Internet technologies to multiple external customers" (Gartner Research, 2008).

National Institute of Standards and Technology (NIST) defines: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability…" (NIST, 2009).
As already mentioned, there is no strict definition of cloud computing so sometimes it's difficult to tell whether a service is truly a cloud computing service or simply a pre-existing offering that is labelled as one of them. Cloud computing is not something completely new because it uses some already established concepts.
Terms with which cloud computing can be confused are: Grid Computing, Utility Computing, Computing on demand, Software as a Service, Service-oriented architecture (SOA), outsourcing, etc. Cloud computing is closely related to most of these concepts (may include attributes previously associated with them) and often relies on virtualization, Web 2.0 and broadband networks. But, cloud computing goes beyond these concepts – it's a much more generalized, "umbrella" term.

## Traditional computing Vs. Cloud computing

With faster broadband speeds and increased internet access, some tasks (e.g. word-processing) can be done in web browser quickly as on a desktop. While traditional computing is desktop-centric and PC-centric, cloud computing is document-centric and has universal access, 24/7 reliability, and ubiquitous collaboration (Miller, 2009).
With traditional desktop computing, user runs copies of software programs on each computer he owns. Created documents are stored on the computer on which they were created and can't be accessed by computers outside the network.
With cloud computing, the applications that the user uses aren't being run from his computer. Computing resources, applications and data that is needed exist on a collection of servers in the cloud, somewhere at the Internet, and can be accessed from any computer connected over the Internet thru Web browsers or specialized applications. Users connect to them and use them as and when needed, and pay only for what they use.
Anyone with permission can access the documents and edit or collaborate on them in real time. If one user's computer crashes, or one of servers in the cloud, applications and created documents are still available for others to use.

## Main characteristics

Main cloud computing characteristics for end-users are scalable and elastic, on-demand, self-service, based on pay-per-use model, location independence and broad network access.

Scalability represents "the ability of a computing system to grow relatively easily in response to increased demand" (Langley, 2008). Computing system can dynamically get or release computing resources (elasticity) via on on-demand self-service, which reduces administrative intervention on the end-user side.
Some cloud computing services are billed on a pay-per-use utility model, while others on a subscription basis with little or no upfront cost. Utility computing is the combination of computing resources as a metered service similar to a traditional public utility such as gas or electricity. Consumers pay only for what they use.
Cloud computing is location independent. This characteristic refers to two things: computing resources location and users access. Computing resources are shared on multi-tenant model (serves a large pool of users) so the user doesn't know and doesn't need to know the location of the provided resources (e.g. storage). The other meaning of location independence is the users' ability to access their applications and data in the cloud from anywhere (not just from the office), at any time, and from any device (laptop, mobile phone…) connected to the Internet. Users are not confined to one computer or organisation's internal network.

**The main benefits of cloud computing**

Besides the advantages which derive from afore mentioned characteristics, there are many benefits for those who use cloud computing: reduced cost of ownership (infrastructure – hardware, software) and of maintenance; increased computing power; new forms of group collaboration; no need for physical space to store servers and databases (companies can store their data on someone else's hardware); virtually limitless storage capacity; independence of operating systems; better document format compatibility (all documents created by application in the cloud can be open by other users who have access to the same application), etc.
It is not necessary to have a high-powered computer (large hard drive, powerful processing power) in order to be able to use cloud computing. Lower priced model with smaller hard disks is adequate, which improves the performance of a computer because applications are not stored and run from it. Increasingly popular netbooks, designed for web browsing, often don't have a CD/DVD drive. They are ideal for the usage of cloud computing services.
By using cloud computing, users also have lower software costs. Organisations don't have to buy separate software packages and software licenses for each computer and they don't have to worry about software updates because updates are applied automatically, and each time the user access the cloud, he gets the latest version of the application. Also, many providers (e.g. Google) are offering their services for free.

As mentioned earlier, cloud computing users pay a provider only for what they use. On the providers' side (back-end), the cost is also reduced because providers can store infrastructure in locations with lower costs.
In the cloud, users can perform more complex calculations and tasks, and significantly speed up their completion, as opposed to what they can do on a single personal computer. They have the processing power of the entire cloud at their disposal.
One of the most important advantages of cloud computing is easier collaboration on the same documents and projects between users who are at distant locations, in real time.
As one of the advantages of cloud computing, increased data safety is often mentioned. If a computer, on which applications and data are stored, crashes, user can't access the data. If a computer in the cloud crashes, all the data is still in the cloud. Data stored within the cloud is safer than data on home computers or laptops (Preston, 2008) because data is not stored in just one place, it's duplicated, and providers invest more in security than any individual company would.

**The major drawbacks of cloud computing**

Cloud computing also involves some risks and has some disadvantages that need to be further improved.
Data safety is questionable, so it is often mentioned as one of the disadvantages because the data can be anywhere and users don't know how secure or insecure their data truly is.
The only real limitation is the Internet access – cloud computing requires a constant Internet connection because users access their applications and data via the Internet. If they don't have an Internet connection, they can't access anything. Also, low-speed connections impair the usage of cloud services.
Among the biggest concerns about cloud computing are the security and privacy of the data, lack of control, reliability, lock-in to cloud service vendor (an application built for one cloud service should be portable), regulatory compliance… "Some of these risks still don't have a industry-wide solution" (Spinola, 2009).
Once a company puts its data into the cloud, it loses control over that data because the data is outside the company's firewall and therefore might not be secure. The data security and control depends on third-party. For example, if computing service provider is experiencing problems, user may not be able to access his data at all. Regarding the possibility that user can access the cloud from any location, it's possible that the user's privacy could be compromised and sensitive data can fall into the wrong hands.
As one of the disadvantages, it can be mentioned that today's web-based applications might be limited (have less features) and can be slower than similar desktop applications.

## Cloud computing layers

Cloud computing service providers offer services on various layers. They provide Everything as s Service (XaaS), from raw hardware to end-user applications. These services can be grouped into three distinct architectural services layers (or service models or sub-areas) of cloud computing (Langley, 2008; Sun Microsystems, 2009; NIST, 2009; Spinola, 2009):

- Software as a Service (SaaS) – e.g. Salesforce.com, Gmail, Facebook applications
- Platform as a Service (PaaS) – e.g. Google App Engine, Microsoft Azure Platform, Facebook
- Infrastructure as a Service (IaaS) – e.g. Amazon Web Services, Joyent.

Figure 1 shows cloud computing service layers from system to the end-user. On the front-end, client side, is client's computer and application (web browser, like Internet Explorer and Firefox, or specialized application) with which accesses to the cloud. On the other side, back-end, there are servers that create the cloud.

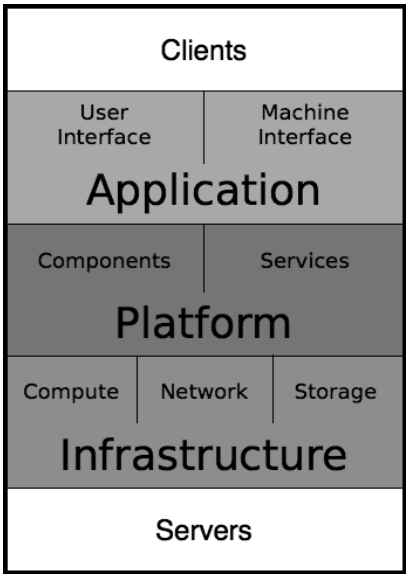



Figure 1: Sam Johnston's cloud computing architecture in stack with layers
Source: Johnston, Sam. The Cloud and Cloud Computing consensus definition?. 20.04.2009. http://samj.net/2009/04/introducing-cloud-computing-stack-2009.html

Software as a Service (SaaS) is the highest layer which features a complete application offered as a service, on-demand on multi-tenant principle, accessible over the network from various client devices (through a Web browser).

Platform as a Service (PaaS) is the middle layer which delivers development environments as a service. Consumers create their own applications using programming languages and tools supported by the provider on which's infrastructure are being run.
Infrastructure as a Service (IaaS) is the lowest layer which provides basic storage and compute capabilities as standardized services over the network, where the consumers are able to deploy and run arbitrary software (operating systems, applications).
At all layers, consumers do not manage or control the underlying cloud infrastructure. IaaS clouds are the underlying infrastructure of PaaS and SaaS clouds. Sometimes is hard to put some cloud service in one of these layers (e.g. PaaS itself can be cloud application and sometimes runs on the same IaaS that it manages).

## Cloud computing types

An organization may choose to use a service provider's cloud (public cloud), build its own cloud (private cloud) or use a hybrid cloud. These are cloud computing types (deployment models).
Private Cloud or Internal Cloud is a cloud of a single organization limited for their internal use. They own or lease infrastructure (e.g. server, network and disk) and control who can use it. This type of cloud allows greater control and customization, security and reliability.
Public Cloud or External Cloud is a cloud computing environment available to the public (individuals, organisations), offered by a third-party vendor. Many different customers use the same infrastructure within the cloud and each of them doesn't know who else is using the same servers as they do. Public clouds are more exposed to security threats and their services can be less flexible than in private clouds.
Private clouds usually cost more than public clouds but they are still cost-efficient. If either of these two types doesn't satisfy an organisation's needs completely, they can choose a hybrid cloud. Hybrid Cloud or Mixed Cloud combines two or more clouds (private and public). It is a cloud computing environment in which an organization possesses and manages some resources in-house and has others provided by third-party.

## Service providers and users

Moving to the cloud may or may not be the best choice for individual user or company. What types of users benefit the most from what cloud computing has to offer? Cloud computing is generally great for small and medium sized companies. However, it is not good for everyone, for now. There are discussions about the potential users of cloud computing, who appear to be good and bad "candidates" for cloud computing (Miller, 2009; Spinola, 2009).

Some cloud computing services are mentioned in this paper, but unfortunately due to limited space the matter is not investigated further.
The major cloud computing service providers (vendors) and some of their services are:

- Amazon: Amazon Web Services – several services including Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage System (S3)
- Google: Google App Engine, Google Apps (same of them are Gmail, Google Calendar, Google Docs, Google Sites)
- IBM
- Microsoft: Windows Azure Platform including Windows Azure, Live Mesh
- Yahoo
- Dell
- Vmware
- Sun Microsystems
- Rackspace
- Salesforce.com
- Ubuntu

Amazon is the pioneer in this field. Besides the mentioned vendors, there are many more, and their number grows every day, as also do the services that they provide.
Two interesting cloud computing services are announced to be released at the end of 2009: Google Wave (a new tool for communication and collaboration on the web) and Windows Azure (a cloud services operating system).
There are several cloud computing (online) operating systems which can be tested (although some of them are still beta versions): Cloudo (http://beta.cloudo.com/), eyeOS (http://www.eyeos.info/) and iCUBE Operating System (http://www.oos.cc/). This operating system is run in the web browser where a desktop is displayed. User can customize desktop, upload and store files, work with provided applications and services, etc. like on a regular desktop on a computer. The main difference is that user and desktop are not confined solely to one computer, but user can access his desktop from any computer that is connected to the internet.

**Future outlook**

Cloud computing is not the future – it is already happening. Many cloud services are already available. We're currently in the early days of the cloud computing evolution and it is difficult to predict its future. Although it is uncertain how the IT market will evolve in next few years, predictions exist which are the result of the conducted research.
"The projected shift to cloud computing will result in dramatic growth in IT products in some areas and in significant reductions in other areas" (Gartner

Research, 2008). From 4% of all IT spend on cloud services in 2008, in 2015 the proportion of IT spend for cloud services will be 17%, and a half of all IT spend by 2020 (Coda Research, 2009). "Over half (55%) of organisations plan to use cloud services extensively by 2015. The most popular services employed will be collaborative applications, followed by IT management applications, personal applications, business applications, storage, server capacity, and finally, application development" (Coda Research, 2009). These predictions seem to be very significant. If a software or hardware company wants to grab a piece of the IT market in upcoming years and to have a profit, it needs to think about cloud computing and develop (some) services toward that direction. On the other hand, potential cloud computing users (organisations) have to gradually change their businesses in order to benefit from the new business possibilities and advantages that come with cloud computing.

In July 2008, Yahoo, Hewlett Packard and Intel launched open cloud computing research test bed called Open Cirrus. It is created to "promote open collaboration among industry, academia and governments by removing the financial and logistical barriers to research in data-intensive, Internet-scale computing" (Yahoo! Research, 2009). There are more than 50 research projects currently in progress.

## Conclusion

The clouds are already over the IT horizon and future forecast of IT is cloudy. Trend of cloud computing will rapidly mature and transform the IT industry, but it is still unclear to which extent, because the development is still in the early stages. Research shows that more and more money is being invested in this field. The web is replacing the desktop and more and more desktop applications will turn into cloud services or at least become hybrid online/offline applications. With cloud computing, the ways the software is being used and the ways of programming are changing.

New opportunities and benefits of cloud computing are becoming interesting for all types of users – organizations and individuals, however (currently) cloud computing is not the best solution for some computing needs. More and more providers and services are emerging. Users can use only what they want, when they want and pay only for what they use. Some providers are even willing to provide their services free of charge, which is good. The most interesting novelties that cloud computing brings (compared to traditional computing) are the emergence of new types of business and new forms of collaboration.

It is of utmost importance that cloud computing enables fast, cost-efficient and secure services. The important prerequisite for cloud computing is high-speed Internet access. If it is not available, the potential users are deprived from cloud computing services. It is important to enable constant high speed Internet access, which will increase the need for cloud computing services.

In order to completely develop cloud computing and to spread its usage, many things need working on. Primarily, a clear and uniform definition of cloud computing is needed. Cloud computing providers need to find ways to increase protection of user privacy, improve the level of security, to work on portability of applications from one cloud service to another, etc.

## References

Armbrust, Michael; Fox, Armando; Griffith, Rean; Joseph, Anthony D.; Katz, Randy; Konwinski, Andy; Lee, Gunho; Patterson, David; Rabkin, Ariel; Stoica, Ion; Zaharia, Matei. Above the Clouds: A Berkeley View of Cloud Computing. 10.02.2009. http://d1smfj0g31qzek.cloudfront.net/abovetheclouds.pdf (15.08.2009.)

Coda Research. Cloud computing: An assessment. 2009. http://www.codarc.co.uk/cc2009/Cloud%20Computing%20An%20assessment%20-%20opening%20pages.pdf (16.08.2009.)

Demystifying the Cloud. http://www.infoworld.com/t/cloud-computing/rp/interactive-ebook-demystifying-cloud-140 (16.08.2009.)

Gartner Research. Gartner Says Cloud Computing Will Be As Influential As E-business. 26.06.2008. http://www.gartner.com/it/page.jsp?id=707508 (10.08.2009.)

Gartner Research. Gartner Says Worldwide IT Spending On Pace to Surpass $3.4 Trillion in 2008. 18.08.2008. http://www.gartner.com/it/page.jsp?id=742913 (10.08.2009.)

Johnston, Sam. The Cloud and Cloud Computing consensus definition?. 24.07.2008. http://samj.net/2008/07/cloud-and-cloud-computing-consensus.html (10.08.2009.)

Knorr, Eric; Gruman, Galen. What cloud computing really means. 07.04.2008. http://www.infoworld.com/d/cloud-computing/what-cloud-computing-really-means-031 (10.08.2009.)

Langley, Kent. Cloud Computing: Get Your Head in the Clouds. 24.04.2008. http://www.productionscale.com/home/2008/4/24/cloud-computing-get-your-head-in-the-clouds.html (10.08.2009.)

Martin, Richard; Hoover, J. Nicholas. Guide To Cloud Computing. 21.06.2008. http://www.informationweek.com/news/services/hosted_apps/showArticle.jhtml?articleID=208700713 (10.08.2009.)

Miller, Michael. Cloud Computing: Web-Based Applications That Change the Way You Work and Collaborate Online: Que Publishing, 2009.

NIST. Cloud computing. 2009. http://csrc.nist.gov/groups/SNS/cloud-computing/index.html (13.08.2009.)

Preston, Rob. Down To Business: Customers Fire A Few Shots At Cloud Computing. 14.06.2008. http://www.informationweek.com/news/services/data/showArticle.jhtml?articleID=208403766 (18.08.2009.)

Spinola, Maria. A Pragmatic, Effective and Hype-Free Approach for Strategic Enterprise Decision Making.2009. http://www.mariaspinola.com/whitepapers/An_Essential_Guide_to_Possibilities_and_Risks_of_Cloud_Computing-A_Pragmatic_Effective_and_Hype_Free_Approach_For_Strategic_Enterprise_Decision_ Making.pdf (12.08.2009.)

Sun Microsystems. Take Your Business to a Higher Level: Sun Cloud Computing. 2009. https://slx.sun.com/files/Cloud_Computing_Brochure_2009.pdf (14.08.2009.)

The Top Cloud Computing Solutions people are looking for in 2009 – Survey by onCloud Computing.com. 06.07.2009. http://www.oncloudcomputing.com/en/2009/07/the-top-cloud-computing-solutions-people-are-looking-for-in-2009–survey-by-oncloudcomputingcom/ (18.08.2009.)

What is CloudComputing?.http://www.univaud.com/about-cloud/what-is-cloud.php (16.08.2009.)

Wikipedia. Cloud computing. 2009. http://en.wikipedia.org/wiki/Cloud_computing (10.08.2009.)

Yahoo! Research. HP, Intel and Yahoo! Attract Leading Research Organizations to Collaborative Cloud Computing Test Bed. 08.06.2009. http://research.yahoo.com/news/2805 (20.08.2009.)

# Computer Technology in Insulin Based Therapy of Diabetes

Maja Baretić
University Hospital Rebro
Kišpatićeva 12, 10 000 Zagreb, Croatia
maja.simek@zg.t-com.hr

**Summary**

*Diabetes mellitus is a metabolic disorder characterised by chronic hyperglycaemia. Glucose regulation is one of the most refined ones in the body. Even slight elevation of glucose results in many hormonal and metabolic actions and it is able to produce devastating organ damages. While treating diabetes with insulin the key word is information. On the basis of information important decision considering diabetes therapy are done every day. Information technology gives possibility for rapid and easy exchange of information from patients to clinicians. In the paper the innovative use of computer technologies in health care of diabetes is presented: for education, collecting, viewing and interpreting home monitoring blood glucose data, for short and long-term glycemic control and also as a part of telemedicine techniques. At the end conclusions regarding computer technologies in insulin based therapy of diabetes mellitus improving health care utilisation in diabetes care is given.*

**Key words:** Information and communication technology, Diabetes mellitus, insulin based therapy

## Introduction

The incidence of diabetes mellitus has epidemic proportions both in the developing and developed world (http://www.who.int/diabetes/). This phenomenon has been attributed largely to westernised life style pattern and higher proportion of type 2 diabetes mellitus (Report of a WHO Consultation, 1999).

Diabetes mellitus is a metabolic disorder characterised by chronic hyperglycaemia (long term elevated blood glucose) with disturbances of carbohydrate, fat and protein metabolism. Those disturbances are resulting with defects in insulin secretion, insulin action, or both.

Sudden blood glucose elevation for a short time causes so-called "acute complications" that are life threatening and should be treated immediately. The most severe forms like ketoacidosis or a non-ketotic hyperosmolar state require immediate medical care. In conditions when blood glucose is elevated for a longer time, mostly without severe symptoms, many damages on different body sys-

tems occur. For example, eye damage from small changes to blindness, kidney damage from slight laboratory alterations to heamodialysis, vascular damage from subclinical ones to cerebral or myocardial stroke, poor peripheral circulation from undetectable one to gangrene and limbs amputation, neural damage etc. The two big famous studies Diabetes Control and Complications Trial (DCCT) and United Kingdom Prospective Diabetes Study (UKPDS) clearly illustrated that strict glicaemic control delays the onset and slow the progression of chronic complications (The DCCT research group, 1993 and UKPDS Group, 1998)

Glucose regulation is one of the most sophisticated ones in our system. Only slight elevation of glucose results in many hormonal and metabolic actions but that small elevation for a longer time produces described organ damages. Sudden drop in glucose level can result in hypoglycaemic coma and death.

Insulin, a key hormone in diabetes, is secreted from the pancreas. It regulates glucose uptake from the cell and controls many other metabolic processes. There are many types of diabetes mellitus, but most common ones are Type 1 and Type 2. In Type 1 diabetes mellitus (this type mostly occurs in childhood or in younger age before 30 yr.) humans one antibodies attack insulin secreting cells in pancreas. Destruction of such cells (called beta cells) results with immediate blood glucose elevation. Those people need insulin lifelong to survive. In type 2 (this type mostly occurs in older population) there is enough insulin-even more than normal. This insulin is not effective enough and those patients are "insulin resistant". Sometimes, not obligatory, they need insulin in therapy too.

In 1922 a 14-year-old boy Leonard Thompson, patient with diabetes type 1, was dying from ketoacidosis. He was the first patient who got insulin in therapy. Without insulin he would dye in a few days, but the fact that he lived 13 years more was considered a miracle. This historical event happened in Toronto General Hospital. Today we want that our diabetic patients all over the world live as long as non-diabetic people, we strive that they have as less complications possible and that they have good quality of life too.

So, what can we do to imitate nature? To replace complex mechanisms of glucose regulation without our own normal insulin function? It seems that some help of technology is required. Western civilisation develops every day more sophisticated, faster, reliable equipment supported by information science and technology that could be applied also in insulin treatment.

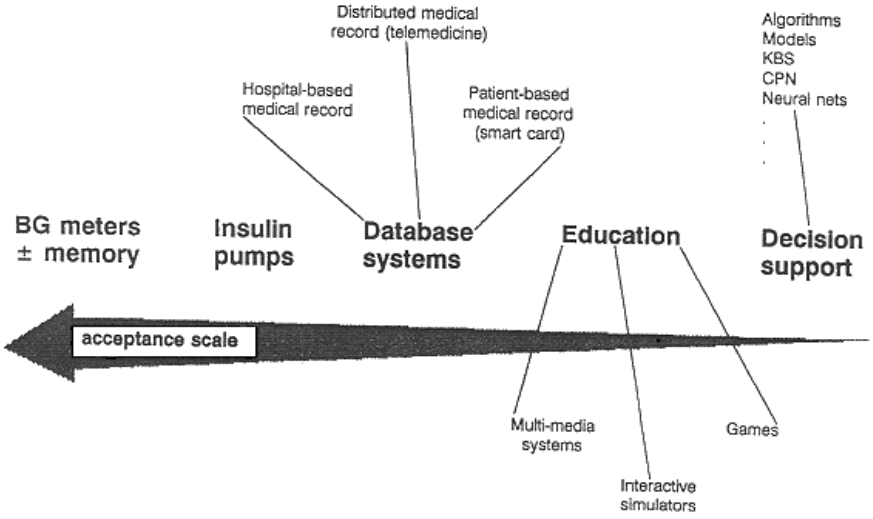
## Computer technology

While treating diabetes with insulin the key word is *information,* the same word used in many other technologies. Information is necessary because on the basis of the information important decisions are done many times, every day.

Question like following are asked … What is happening with me today? How high is the glucose level now? Am I going to eat a lot? Is my food going to

contain many carbohydrates? Am I going to perform an exercise and for how long? How much insulin should I apply? Which type of insulin should be applied and when? When is the proper time to measure blood glucose again? … and many others. It is not possible to have available physician (or some other health care provider) near the patient constantly in every moment of the day.
All diabetic patients with insulin therapy (or in some circumstance members of family, or people who are taking care about patient) should be educated to answer questions like those mentioned above. Physician is contacted in time of crisis (sudden illness, sudden elevation or drop of blood sugar) and during the regular check. Information technology gives possibility for rapid and easy exchange of information from patients to physician or from one physician to another, and allows interactive communication.
Computer technologies are used for collecting, viewing and interpreting home monitoring blood glucose and diagnostic glycation tests. They are in use for insulin delivery, artificial and bioartificial pancreas, educational software, databases, for use of new biomaterials, development of new sensors etc. Finally there are telemedicine techniques connecting patient and health care provider.
There are many scientific reports about diabetes and technology published in many journals; one of them is Journal of Diabetes Science and Technology, a bi-monthly, peer-reviewed scientific e-journal published by the Diabetes Technology Society (www.journalofdst.org).

Picture 1. Applications of information technology in clinical diabetes care




Lehmann E.D. Application of information technology in clinical diabetes care Part 2. Models and education. Medical Informatics, 1997; 22, (1), 1-120

Even in recent times there were many doubts concerning computer technologies in diabetes including papers with question like "Designing computer-assisted instruction programs for diabetic patients: how can we make them really useful?" (Juge CF, 1992) However, today there is more and more evidence confirming necessity of introduction new technologies into diabetics' daily routine. Picture 1 shows how widely some computer applications were accepted in routine clinical practice five years later. In further text computer technologies are described in same order.
The acceptance scale indicates how widely such applications were adopted in yr. 1997 into routine clinical practice. Education and decision support could be either for patients or healthcare students / professionals. (BG = blood glucose; KBS = knowledge based system CPN = causal probabilistic network).

## Algorithmic-based decision support systems

Medical algorithms are decision trees that help making choice on the basis of many information collected previously. Computerised algorithms can provide support of clinical decision while being adherent to evidence based guidelines. Evidence based decisions in those solutions are up to date Some of them cover variety of scenarios and could be helpful in everyday practice.
Though, algorithms are not capable to manage all situation that diabetic patient with insulin meets. The plan of algorithmic based support system is to provide set of schemes by which patient can adjust a therapeutic insulin routine and achieve the desired glicaemic control. Albisser (Albisser A.M, 1996) described algorithmic, telemedicine-based system called HumaLink. Patient who measures blood glucose regularly accesses HumaLink system from touch-tone telephone available 24 hours a day. Unique identification number opens a variety of verbal instructions from speaking system. Patients are entering data like for example: blood glucose level, current illness, physical activity, meals etc. The HumaLink system then relays instructions in accordance with an individualised treatment plan programmed by caller's physician.
The decision about insulin dosage is done in fully automated advisory mode or in manual recording/documenting mode. The first one, a fully automated mode, applies algorithms to modify insulin dosages within pre-defined limits set by the physician automatically. The other one, manual' recording mode logs the patient's readings and a physician reviews the data before leaving a verbal message for the patient on the system.

## Educational tutorials

As already mentioned, extremely important part of the insulin therapy is education: how to use insulin and control its effect. There are many types of education: a group education, a single education etc. (Zgibor JC, 2007). Mostly, patient gets first information in direct contact with health care provider-a medical educator. For further information or repetition Internet could be an option.

There are web sites designed for educational and teaching purposes. Some web pages describe basic instruction like "how to apply insulin", other more sophisticate ones provide examples of diabetes case scenarios with problems to be slowed (Reed K, 2006).
On the basis of educational level or current situation it is possible to choose a scenario (pregnancy, low glucose, flu etc.) and tray to solve it. Some tutorials combine written information with an interactive diabetes simulator. AIDA, available on the World Wide Web since 1997/1998 without charge is one of such sites (http://www.2aida.net/welcome). It′s purpose is strictly highlighted: *AIDA is only for general use, without intention to provide personal medical advice or substitute advice of doctor.* AIDA is constantly growing. Enhancements of the software are accompanying insulin development and innovations (Lehmann ED, 2007). Educational systems are also used for teaching purposes of health care professionals, especially students.

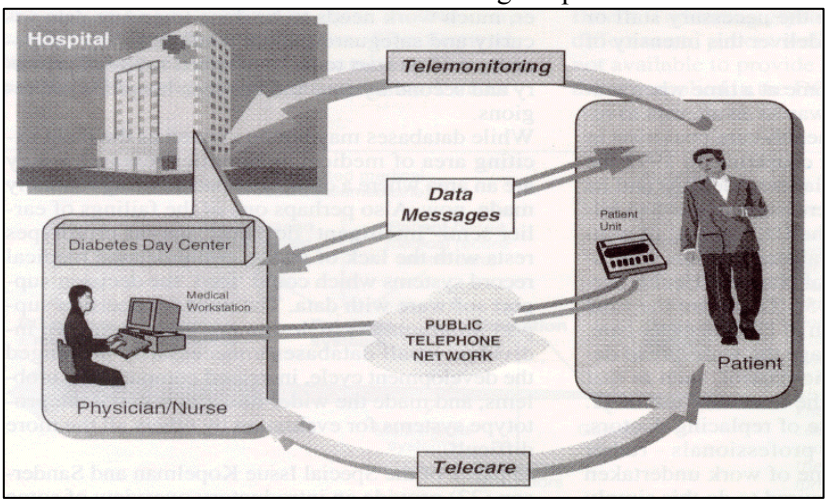
## Databases

Databases in a global view are the basis of modern health care service. Many different computer database programs that consider diabetes are being developed with more or less success-DIAMOND, DIABCARD, DIABTel etc. (Lehmann ED, 1995). The basic idea of such databases is that information about patient with diabetes is accessible locally, regionally and nationally for statistics, research and clinical practice. Some of the mentioned databases are accessible for primary, some for secondary care, or for both. The "must be" is data security and patient confidentiality while transferring the data.
In 1997, the same year when AIDA was introduced on www, a database system in Croatia called CroDiabNET was initiated for the first time (Metelko Ž, 2001). It is a first register in Croatia based on daily data entries from everyday clinical practice. The basic information sheet is the main part of the system. It contains all the data collected on regular check-ups (identification data of patient with diabetes, type of diabetes, year of diagnosis, treatment started date, oral drugs/insulin introduction, purpose of visit, risk factors, blood glucose self-control, education, body weight, height, blood pressure, laboratory data, complications of diabetes, other diseases etc.). Croatian National Diabetes Registry is firs public health registry where users can input data through www (http:// crodiab.continuum.hr).
Practical application started in 1999; at the end of 2003 data from 33,000 patients in 19 centres in Croatia have been collected. The program is currently being used in the referral centre, regional centres for diabetes, county centres and other diabetes centres. CroDiabWEB has been designed for the annual registration of diabetic patients at the general practitioners' offices.

## Telemedicine

Telemedicine is medicine at distance or use of computers and computer networks with the idea to facilitate communication among two or more medical professionals or among medical professional and patient. Telemedicine exchanges health care services without necessity to be present at the same time in the same place (McLaren P, 1995). The first experiences from 1960th in North America and Northern Europe intended to improve communications between patients and physicians, but now there is a trend to assist in everyday situation.
A good example is a small Italian study from 2003 that proved telemedicine provided benefit in treatment of insulin dependant diabetes. (D'Annunzio G, 2003) A system of two main components was used: a medical unit (web-based workstation) in the hospital close to the health care provider and a patient unit (patient logbook, therapy consultation, electronic messages, communication system; implemented as a PC based software). Both were connected trough a telecommunication system. Through telemedicine system patients were able to send easily their self-monitoring data of the glucose level, insulin dosages and urinalysis. The data were sent weekly. During mean follow-up period of 415 days, 901 blood glucose levels per patient were collected and analysed. An increase of links between patients and physicians were noticed.

Picture 2. Telemedicine for diabetes care using telephone communication




Gomez E., del Pozo F., Hernando E.: Telemedicine for diabetes care: the DIABTel approach towards diabetes telecare. Med. Inf. 1996; 21: 283-295

In particular, the medical unit sent to the patient unit an average of 56 messages; and an average of 35 messages were sent by the patient unit to the medical unit. The system seemed to be feasible and provided clinical benefits. The most in-

teresting thing was that the patents were children aged 9.9-15.8 years who were open towards new technologies.
Today, benefit of telemedicine is discovered in many countries, also in Croatia. 2007 began a pilot project called "Telemedicine on Croatian islands". The aim of the project is to connect and improve health care on the Croatian islands (http://www.mmpi.hr/userdocsimages). Picture 2 shows a simplified example of interaction between hospital and patient through telemedicine.

## Glucose monitoring and insulin pumps

Today's majority of diabetes computer technology is used in glucose monitoring and recently in insulin pumps. Usual method of checking glucose levels is pricking a fingertip and then using a glucose meter that measures the blood sample's glucose level.
Continuous glucose monitoring systems use a tiny sensor inserted under the skin that checks glucose levels in tissue fluid. The sensor in inserted in average for three days, during this time a transmitter sends information about glucose levels through radio waves from the sensor to a pager like wireless monitor. Special software is available to download data from some glucose meters, or continuous glucose monitoring systems. In that manner it is possible to display trend graphs on the monitor screen.
Data management systems can store hundreds of test results and other information (the time and date of analysis, types and doses of insulin, meals, and a log of exercise). Almost 10 years ago motivated patient could print the graphs and take them to the clinic (Lehmann ED, 1999), but today graphs can be easily send by mail.
Insulin pump is a device that delivers insulin continuously into the body. In that way it imitates better physiological insulin secretion then standard insulin application. Insulin pump consists of the pump itself (including controls, processing module and batteries), a reservoir for insulin inside the pump (patient is refilling the reservoir), a disposable infusion set (patient is changing the set every 2-3 days). A disposable infusion set has a short tube with a needle (cannula) placed under the skin and a tubing system that connects insulin reservoir to the cannula. Insulin pump delivers insulin for 24 hours*: a basal rate* continuously and short smaller doses before meals*: a bolus*. The physician predefines a basal rate, mostly after analysis of continuous blood glucose monitoring data sheet. A bolus doses are set before the mealy by the patient. An insulin pump eliminates individual insulin injections and improves glicaemic control. It requires maximal co-operation of the patient and some technology skills.
The latest, still experimental project is an artificial pancreas. Artificial pancreas integrates continuous glucose monitoring and insulin pump with a closed loop system that provides the right amount of insulin at the right time. The data from continuous glucose monitoring are providing blood glucose reading every few minutes; a sensor is connected via wire to the insulin pump. Blood glucose

variation is signalling automatically to the pump sending information how much insulin to deliver (Friedrich M J, 2009).

## Conclusion

There are many evidences that computer technologies are applicable in the insulin-based therapy of diabetes mellitus. A large meta-analysis of twenty-six studies (with over 4,811 participants) reported that interactive information technology in diabetes care improved health care utilisation, behaviours, attitudes and knowledge (Jackson CL, 2006).

Though, there is still much to learn and to improve. Technology sets new demands to the both patient and physician. Patient with diabetes needs to be more proactive and needs to learn both medical and computer skills.

At the end, I would like to quote Aaron Kowalski, research director of the Juvenile Diabetes Research Foundation's Artificial Pancreas Project: "We have data on hand today that suggests that you could get much better diabetes outcomes with the computer taking the lead instead of the person with diabetes doing it all themselves." In 1977, at the age of three, Dr. Aaron Kowalski's brother Stephen was diagnosed with type 1 diabetes. In 1984, at the age of thirteen, Aaron himself was diagnosed with type 1(http://www.jdrf.org).

## References

Albisser A.M ., Harris R.H., Sakkal S., Parson I.D., Chao S.C.E. Diabetes intervention in the information age. Med. Inf. 1996; 21: 297-316.

D'Annunzio G, Bellazzi R, Larizza C, Montani S, Pennati C, Castelnovi C, Stefanelli M, Rondini G, Lorini R. Telemedicine in the management of young patients with type 1 diabetes mellitus: a follow-up study Acta Biomed. 2003;74 Suppl 1:49-55 8

Friedrich M J. Artificial Pancreas May Soon Be a Reality JAMA. 2009;301:1525-1527

Jackson CL, Bolen S, Brancati FL, Batts-Turner ML, Gary TL A Systematic Review of Interactive Computer-assisted Technology in Diabetes Care J Gen Intern Med. 2006;21:105-110

Juge CF, Assal JP. Designing computer assisted instruction programs for diabetic patients: how can we make them really useful? In: Proceedings, 16th Annual Symposium on Computer Applications in Medical Care (Ed.), M.E. Frisse, IEEE Computer Society Press, New York, 1992;16: 215-219.

Lehmann ED. Database Programs for Use with Blood Glucose Meters Diabetes Technol Ther. 1999; 1: 391-393.

Lehmann ED, Chatu SS, Hashmy SSH. Retrospective pilot feedback survey of 200 users of the AIDA version 4 educational diabetes program. Diabetes Technol Ther. 2006;8:419-32, 602-8 and 2007; 9:122-32.

Lehmann ED, Deutsch T Application of computers in clinical diabetes care. Med Inform 1995; 20:303-29.

Metelko Ž, Pavlić-Renar I,. Poljičanin-Filipović T, Car N, Dumičić J, Hercigonja R. CroDiabNET -sustav za praćenje dijabetološke skrbi in Telemedicina u Hrvatskoj: dostignuća i daljnji razvitak Ed Kurjak A., Richter B., Akademija medicinskih znanosti Hrvatske Zagreb, 2001, 209-214

McLaren P, Ball CJ. Telemedicine: lessons remain unheeded. BMJ 1995; 310: 1390-1391

Reed K, Lehmann ED. Interactive educational diabetes/insulin tutorial at www.2aida.info. Diabetes Technol Ther. 2006 ;8:126-137.

The diabetes control and complications trial research group. The effect of intensive treatment of diabetes on development and progression of long term-complications in insulin dependent mellitus. N Eng Journal of Medical 1993; 329: 977-986.

UK Prospective Diabetes Study (UKPDS) Group. Lancet 1998;352:837–853

World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Report of a WHO Consultation: diagnosis and classification of diabetes mellitus. Geneva: World Health Organization, 1999

Zgibor JC, Peyrot M, Ruppert K, Noullet W, Siminerio LM, Peeples M, McWilliams J, Koshinsky J, DeJesus C, Emerson S, Charron-Prochownik D. AADE/UPMC Diabetes Education Outcomes Project. Using the American Association of Diabetes Educators Outcomes System to identify patient behavior change goals and diabetes educator responses. Diabetes Educ. 2007;33:839-842.

## Links

http://crodiab.continuum.hr (15.9.2009)
http://www.2aida.net/welcome/ (15.9.2009)
http://www.jdrf.org/ (15.9.2009)
http://www.journalofdst.org/ (15.9.2009)
http://www.mmpi.hr/userdocsimages/2007/vrh-252-09-otoci-zdravst.pdf. (15.9.2009)
http://www.who.int/diabetes/facts/world_figures/en/ entered (15.9.2009 )

# Information Architecture and e-Government

John Akeroyd
Information Management Consultant
London, UK
john.akeroyd@googlemail.com

**Summary**

*E-government has become pervasive in many parts of the world in various forms and can be defined in multiple ways. However it essentially looks at how ICT and the web in particular can be adopted to improve how government runs and especially how it interacts with citizens. It is possible to categorise e-government activities and what emerges is the central role information plays in enabling the delivery of effective services. This paper looks at existing and emerging e-government developments based on a specification of information architectures. It looks at how basic classifications can be used to ensure the flow of information across distributed services and enable accurate data management and goes on to present a case study of a government agency.*

**Key words:** information architecture, e-government

## 1. E-government – Scene Setting

E-government can be defined in many different ways but essentially looks at how ICT and the web can be used to improve how government operates especially through its interactions with citizens. It is possible to categorise e-government activities and what emerges is the central role information has in enabling effective services. This is more true for certain aspects of government business such as tax collection which is essentially an administrative task whilst in other cases it provides a more supporting back office role.

One of the fundamental precepts of e-government is the shift from services which essentially serve the internal needs of government administration to those which are more proactive in meeting citizen needs, whatever their status or financial position, As Wimmer as it "IT has played a major part in incrementally changing and shifting traditional and bureaucratic government models into the current e-government model where services are delivered according to customer needs" (Wimmer 2004.)

The drivers according to Weearakkody (2007) include improving internal costs and management efficiencies, encouraging citizen participation, promoting economic development and improving overall governance. We can add to these the idea of service integration or 'joined up' government as the Blair government

described it whereby citizens are not shunted from department to department, often undertaking the same basic processes before achieving their desired outcome or, more critically, where the needs of vulnerable people are not neglected through different government agencies not taking a holistic view of an individuals situation.

E- government is now widespread across the globe especially in developed countries but also in the developing world. There are case studies aplenty on the status of projects in the UK, Europe generally, the US, Asia etc. These show remarkable similarities and face essentially the same challenges. Denmark for example has proved particularly adept at initiating- government activities with Enterprise Architecture and the establishment of an interoperability framework with specific integration standards (Weerakkody, 2005). Murphy (2005) quotes information sharing, ensuring inclusivity, and managing inter-agency initiatives as common key challenges. But he also notes the differences between countries: for one, the degree of government centralisation, where more federal structures present greater challenges to implementing cross agency solutions than those with a highly centralised structure; secondly is the issue of differing political agendas – with some countries focusing on broadband in order to deal with heavily distributed populations; finally he notes the local legal context where existing laws can facilitate or block e-government projects: thus if there is already a heavy administrative burden for business it is unlikely to be reduced significantly through new processes whatever they are and whilst the degree to which the state keeps information on its citizens will enable easier web based interaction. His report identifies projects in Australia, the US, Canada, Japan Europe.

Some definitions of e-government see it as essentially about improved service to citizens and predominantly using the web. It is seen as a way of enabling improved interaction with citizens so that they can transact their business in a slick and efficient way to suit them and not the producer. It borrows from e-commerce developments over the past decade and has similarities in the way services are offered, particularly those which are charged for. It also parallels a greater emphasis on customer service throughout the public sector. Other commentators rightly see it as about increased business efficiency between government agencies and government agencies, within agencies and between agencies and business. In essence e- government is about the transformation of internal and external processes.

Lastly the term transformational governance is also widespread and implies a fundamental rethink of the way in which services are offered focussing on savings, efficiency and customers service as the primary drivers for e-government and indeed many successful projects have managed to achieve these goals. However what is noticeable is that whilst there has been significant progress in enabling web access to government IT there has been less success in joining that access up to the (often disparate) array of back end functional services.

## 2. Challenges to E-Government

E-government will not come about organically and requires investment in programmes and people to bring it about. Nor is it necessarily straightforward, in that there are many barriers and challenges to overcome - these are discussed below.

### 2.1. Information sharing

Information sharing is at the root of e-government – it is required to bring systems together, to ensure a holistic approach to customer service and to ensure efficiency through reducing duplication of effort in collecting and storing information. In most organisations there is a high, continuing overhead in duplicate files which in turn leads to inaccuracy and poor decisions. But information sharing is not easily achieved; there are real barriers to its deployment which are both technical and cultural.

### 2.2. People Issues

Managing the people aspects of e-government is an often underestimated challenge, one example is delivering a national electronic health record system, which in itself is a large scale challenge, but ensuring that health professionals are willing and able to use it is a people challenge on similar scale. This kind of change is often only brought about by fostering public confidence through the delivery of projects which are seen to work and secure user acceptance.

### 2.3. Power structures

> "Power conflicts over departmental boundaries and control of services will become more apparent as integration progresses"
> (Signore, O et al 2005)

All commentators suggest that one of the most common barriers to the effective update of e-government is the pre-existing power structures which tend to inhibit cross agency working, information sharing and a focus on citizen as customer.

Effective e-government requires a change in mindset of agency directors; many perceive their department as the most important and tend to disregard other agencies. And though this silo structure may well work well in business where the ultimate focus is always on profitability it could be deemed anathema to public service.

### 2.4. The Legal Context

A further issue in information sharing is the problem of sharing personal data between different agencies as well as within agencies. The idea is that from a citizen perspective there is only one point of access and information has only to be conveyed once for relevant distributed systems to be actioned, thus both

saving the user time in interaction with a range of target systems and ensuring maximum efficiency in data collection and updating across those systems. However the legal framework can often block such exchange.
Data protection and other privacy laws are an inhibition to the extensive re-use of information in a pan government context. All European countries have something equivalent to a Data Protection Act or at least legislation relevant to the re use of information in a context other than which it was arguably provided. Thus taking information which might have been gathered from, say, a parking violation and re-using it to investigate tax evasion could be seen by citizens to be invidious and counter to their interests. But in other circumstances where information could be re-used to a citizen’s advantage it would be pointless not to share it. Whichever way it is looked at, there should be a process in place which reflects citizens needs in an efficient and connected way and this will require an information sharing policy which must be supported by all, applied across all agencies and above all be legally sound.

### 2.5. Information Security

And whilst Data Protection laws are there to ensure that citizens rights over their personal information are protected, information security polices are a means of ensuring that personal data is not abused and is kept securely where it is deemed necessary. Information security has both technical and policy challenges, but is not the topic of this paper. Suffice to say that governments for whatever reason, do keep information confidential from their citizens and this requires systems which ensure maximum security where data can only be viewed by the relevant authority.

## 3. Information Architectures

So the argument is that e-government is underpinned by information and its effective management is a necessary prerequisite for service delivery. Information management is concerned with information quality, security, business processes and metadata and all of these need to be addressed to deliver good e-government. Good information management implies understanding what information assets are in place and what part they play in a particular business process and the first step in that regard is usually the compilation of an information audit which details the size and scope of the information available and its lifecycle. It comprises the total knowledge base of the organisation. This activity can be complex and is the subject of extensive literature.
Information architecture is a subset of information management and is not new but was previously more commonly used to describe Enterprise Architecture or Enterprise Information Architecture and was more concerned with infrastructure and applications than information per se. IA also leans towards describing what might be rather than dealing with the ‘as is’, so that it is more of a framework onto which future services and applications can be mapped.

So to understand government information, as well audit of the extent of it, we need to be able to model it and then to analyse, categorise and classify it and this is what we will turn to next.

### 3.1. Models for e-government

The seminal paper analysing models for e-government is that of Layne and Lee (Layne, K & Lee, J. Developing fully functional E-Government: a four stage model) Writing in 2001, they proposed a model which as a first stage, includes the cataloguing and presenting over the web of services and processes to inform citizens. This leads to a second phase which emphasises transactional government that is it supports forms processing, online transactions, e-payments etc.

It could be argued that many developed government agencies are now at, or very near this stage, though it requires users to be authenticated at some point so that the system knows who they are. These two models in turn lead to further phases of vertical and horizontal integration. Vertical integration envisages links to line of business systems whereby citizens can interact directly with back office systems so as to allow questions to be directly answered or a service secured. Business may be within an agency or straddle multiple agencies. Horizontal integration proposes that back office systems are themselves integrated to provide a one step approach to meeting users needs. Horizontal integration may not simply follow on from vertical integration but could well be run in parallel. Layne and Lee's model has stood the test of time albeit that other researchers have proposed enhancements to it such as Weerakody (2007)

Janssen and Veenstra (2005) propose a five-stage model for the development of architectures for local government agencies. Their model consists of 1) no integration 2) one to one messaging 3) warehouse base systems 4) brokering systems and 5) advanced broker architecture. The model moves from simple to an advanced process based architecture able to manage links and cross-organisational processes and supporting service oriented architectures.

Each of these models presumes organisational and technical challenges – and here we will look at those occurring at the level of horizontal integration as it is here where the most significant barriers are found and which need to be dealt with if we are to achieve a joined up approach to government.

### 3.2. Integration

What do we mean by integration? It could be argued that there are four different levels or types of integration which are current in information systems. Most integration at the moment is at Levels 3 and 4 but the ambition ought to be or is to move to 1 and 2.

- **Level 1 Functional Integration**
  Whereby a secondary application is accessed and used through a primary application to the extent that the secondary application is transparent to the user;

- **Level 2 Data Integration**
  Where data from one system is used to populate another either in near or real time usually using standard protocols /programmes such as SOAP/BizTalk
- **Level 3 Linked integration**
  Where a secondary application or dataset can be accessed/triggered via a primary application but which essentially appears as is to the user. Links might be hyperlinks or file paths
- **Level 4 Data exchange**
  Where data is moved from one application to another as the result of an operator initiated action. Data is usually structured as XML or CSV or XLS.

Integration at all levels is inhibited if underlying data structures are not using essentially the same data structures and the same descriptive metadata. (It is possible and indeed frequent to find that the same numbering system is used to mean entirely different things and only through standardisation).

## 4. Information Architectures

So there are a number of problems associated with system integration which need to be addressed for effective information sharing. These can be summarised as:

- In large organisations there are likely to be many application systems with high overhead on maintenance and complexity
- Data will probably be held many times in many places leading to confusion as data accuracy, currency and what to believe;
- There will be increasing complexity in understanding how data moves across and around the organisation;

How can these issues be overcome? Alternative architectures have been proposed which seek to answer these cases though none is perfect and each has problems. In this section I have documented four possible approaches.

### 4.1. SOAP/Web Services

In the web services approach, information assets essentially remain where they are within a functional business model but assets are then joined up using integration tools such as web services. Web Services is defined as ‘reusable components as services and which enable linking of these services between and across different systems using XML. It deploys three XML standards SOAP, UDDI (Universal Description, Discovery and Integration**)** and WSDL (the web services description language) to provide a platform for developing available web services (Weerakkody,V, 2007)

The benefits of the web services approach are modularity, accessibility and a well described implementation independent of any given system and are thus

highly interoperable (Fremantle 2002). Web services can cross not only internal boundaries but external components can be brought into play. But there are complexities in that web services can accommodate not just information flows but also business logic requiring a separate repository for ease of maintenance and re-use. (Zhao, JL 2008)

### 4.2. The Single Repository or Data Warehouse

A second solution is that of the single information repository or data warehouse where all crucial data relating to customer interaction is copied to a single place. It leads to the concept of 'one version of the truth' where all information is confined within a single unified place and master data sets which support all systems. There are multiple benefits to this approach, sometimes known as a data warehouse, whereby data is brought together to support customer front office negotiations whilst the back office retains operational control. Front office or certain related staff are able to see the whole picture around a certain transaction.

The main problems associated with a data warehouse are 1) the potential vulnerability from having a sole source to support both back end and front end systems with the potential for failure and 2) that data often has to be copied across regularly into the repository from back end systems with the potential for the data set to be out of date at any given time;

### 4.3. Information Flows

Thirdly but not mutually exclusive is an architecture which looks at the flow of information flows into an organisation and seeks to track and audit it at that point. Much information will be in one of write, email or electronic document formats and unified systems can ensure that all such systems are converted to the same e-format and then logged to an audit file before being steered to the relevant back end system for processing. This is more a matter of workflow than architecture in that in some ways it simply reinforces the warehouse model above but it does have a major benefit in improving efficiency in that documents get to where they need to be processed quickly and in a processable format. The downsize is that metadata must be added at the front-end and is likely to be limited unless there is a system in place which is capable of some intelligent semantic derivation, so that documents can be routed to where they belong and finding that data at a later date is rendered easier..

## 5. Information Categories

Any of these approaches to information management will not succeed unless there is a well described set of information assets so as to allow them to be used in different contexts in different parts of the business. To achieve this we need some way of defining those assets so that they can easily be re-used. As a starting point we can generally define government services as essentially the deliv-

ery of a service to a citizen (people) or organisation based on a certain place or property. This leads to the four key categories of:

- People,
- Organisations,
- Property & Place,
- Transactions.

### 5.1. Property & Place

Property & Place in some ways is easily defined in that we can define any location through its geographic coordinates and hence uniquely identify the building, property or even street article concerned. In government contact is more often through often property or buildings and these are thus more critical. The UK at least has defined a national system of property management known as the Loan and Property Gazetteer which documents every building and address in the UK and allocates to each a unique identity known as the UPRN or Uniform property Resource Name. The 12-digit Unique Property Reference Number (UPRN) covers every building and plot of land in the country and is defined by a standard BS 7666 which thus allows for the interchange of such information between agencies and between systems.

### 5.2. People or citizens

People or citizens are more difficult to define. The status or position of an individual is usually a good starting point; citizen or voter or tax payer are all roles that we might occupy. Each of these roles will have a functional system or process attached to it and possibly therefore some identifier associated with that person and role. The availability of a unique identity is enormously helpful in that it can be used as proxy for people, with consequent benefits in processing their data and joining up systems. This is sometimes referred to as ‘tell us once’ where information submitted or collated for one purpose is reused for another. If there is no unique identifier then a combination of name, date of birth or other identifier could be used as a substitute. The management of people data is a particular challenge where there are itinerant populations, constantly shifting and with citizens living outside of formal structures.

### 5.3. Organisations

Organisations are less easily defined than other entities, given that they are subject to hierarchies, subsets and also constant change. The value of an accurate record to government is possibly largely in tax and revenue collection where tax raising from business is often highly critical that financial standing becomes key piece of information. But organisations do have other roles vis a vis government such as the voluntary or charity sector who can complement and work with government. Which ever way an organisational directory is often a key component.

### 5.4. Transactions

By transactions is implied a record of say a transaction between a person and a service which might also relate to a property, such as land purchase, and which might be undertaken within the business unit or agency or even corporately. Transactions also encompass financial records which are basically transactional records of specific actions between a person and an organisation e.g. payment or receipt of money

### 5.5. Key metadata

Finally, key metadata e.g. property/people can be exemplified as:

- Property
  - UPRN available in all property records;
  - GIS coordinates
  - Fileplans standardised around street name number
- People
  - Citizens;
  - Tax payers;
  - Students;
  - Government workers;
  - Visitors
- Organisations
  - Businesses
  - Schools;
  - Community Centres
  - Churches
  - Support agencies;
  - Public agencies
- Transactions
  - Case file Numbers
  - Invoice reference
  - Purchase orders
  - Transaction reference

Good descriptive i.e. semantic metadata will come from adding controlled vocabularies to documents either at source or later. To that end in the UK there have been attempts to standardise vocabulary across systems via a pre determined set of taxonomies defining services, activities, transactions etc with some (though not overwhelming) success. The best example of this is the IPSV (illustrated in Figure 1), a combination of sets of vocabularies designed to describe all government operations. Such taxonomies do have the wherewithal to address one of the key areas of integration, that of semantic interoperability.

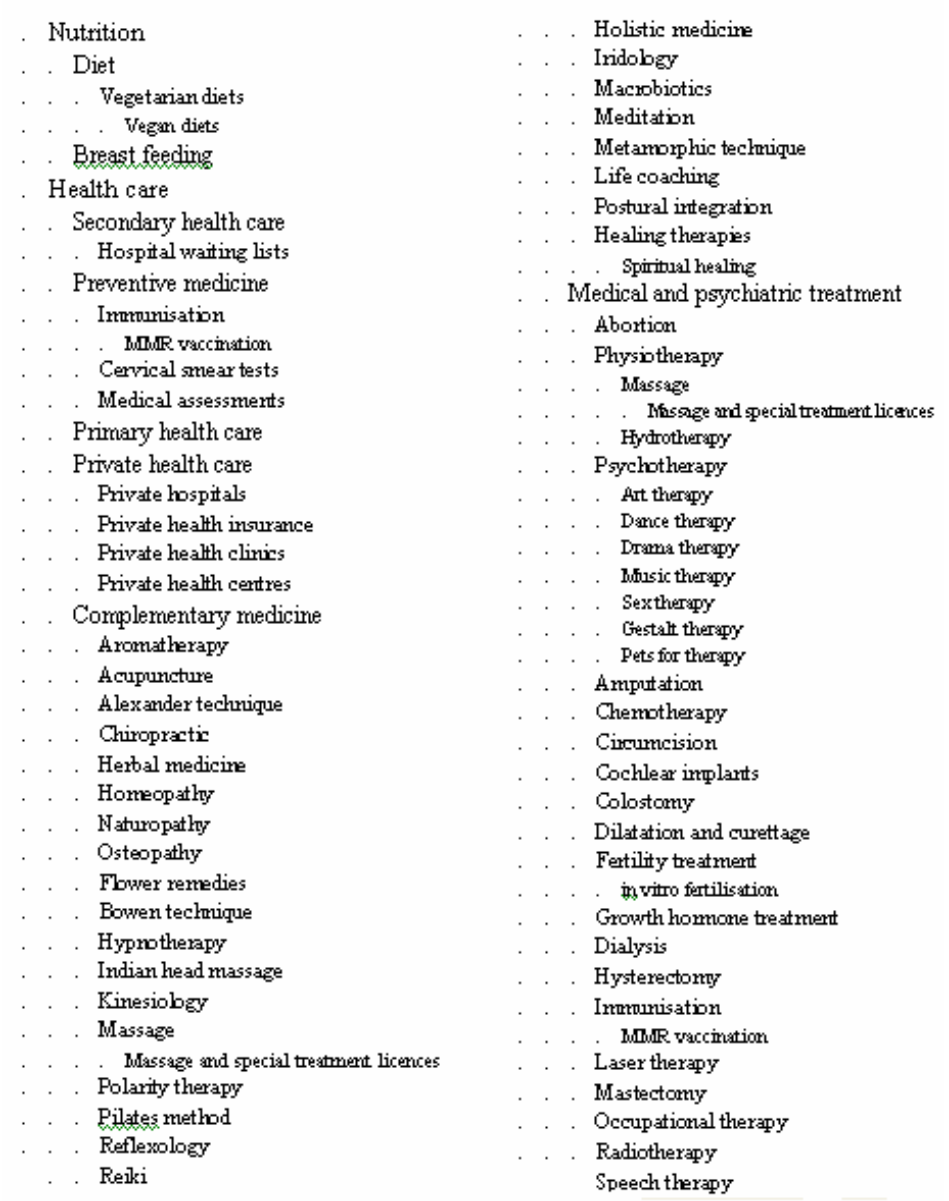



Figure 1. The Integrated Public Sector Vocabulary (IPSV)

## 6. Unstructured versus Structured Information

A further possible categorisation is that between structured and unstructured data whereby it can be argued that a percentage of organisation assets are held in an unstructured format such as documents or reports and wherein is a lot of key information needed for effective service delivery. Structured information by contrast is that contained within a database be it a spreadsheet or sql or other. Structured content is by its very definition well managed and can be easily manipulated in a digital sense. However it tends to be small scale in volume. By contrast unstructured information needs very good metadata to make it retrievable and whilst OCR technologies are enabling more information to be derived, good findability through web engines and other enterprise search engines will only come through enhanced indexing.

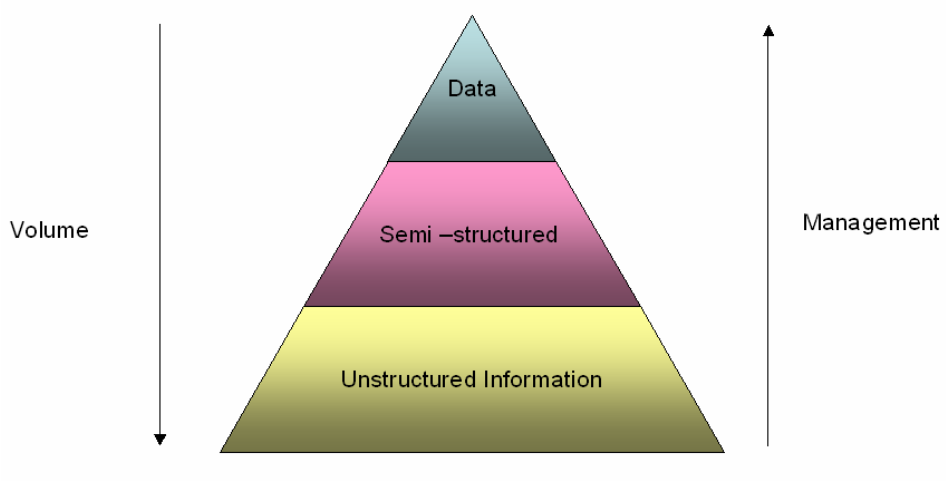



Figure 2. Structured and Unstructured Information

## 7. Case Study

In this section we look at a case study of a government agency in London providing a wide range of services including tax and revenue raising, environmental services, planning and building, road sand highways and social care. This mix is typical of such agencies. To provide context the agency supports over 100 system applications in over 50 key business units each of which supports a separate line of business. The challenge is to understand information architecture and set a strategy which will ensure that the agency is able to provide a coherent 'joined up' service to its community at the most economic cost. This strategy has to be underpinned by the guiding principle that information should

be held no more than once and should be re-used or repurposed for delivery to different channels rather than being recreated. That implies ensuring that all information assets are known and recorded and master data kept to ensure accuracy and integrity. To this end the strategy replicates the points made above – the first step is to ensure that all information assets are audited through an information audit and their availability and metadata recorded. This information asset base can then be categorised using the broad headings described in Figure 2 and built as below.

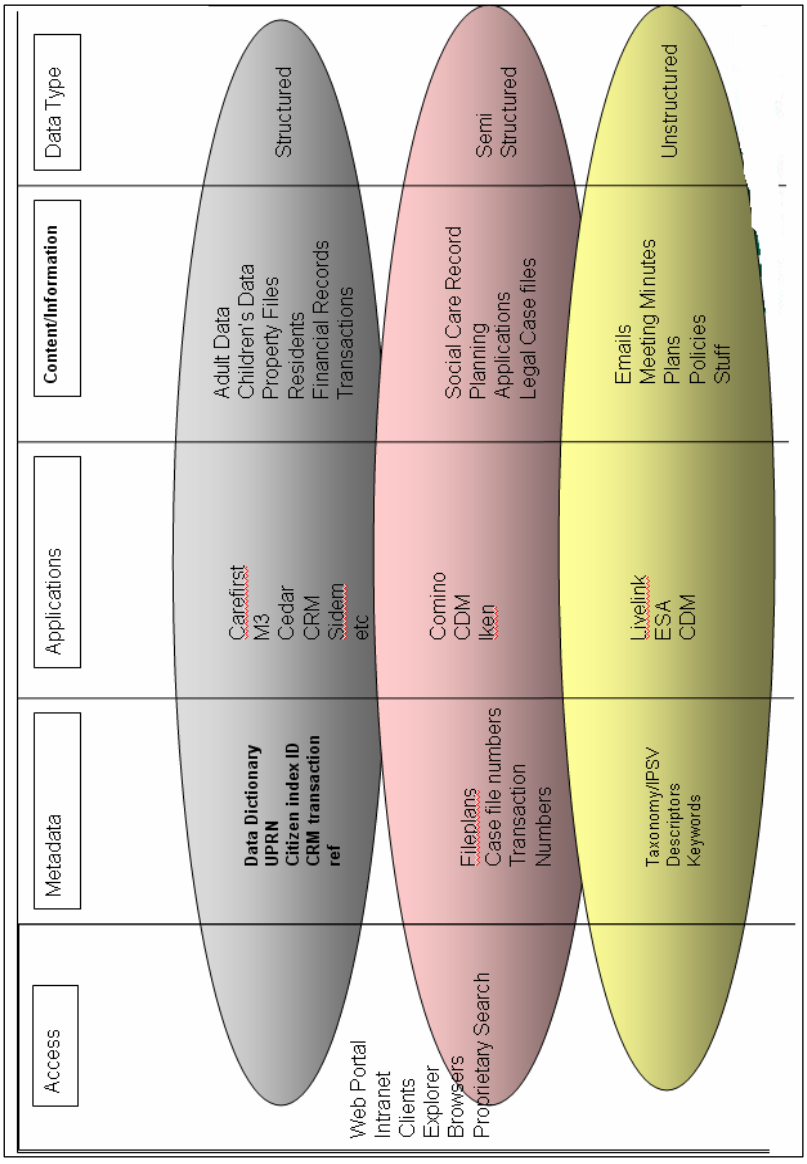



Figure 3. Categorising information and applications both horizontally and vertically

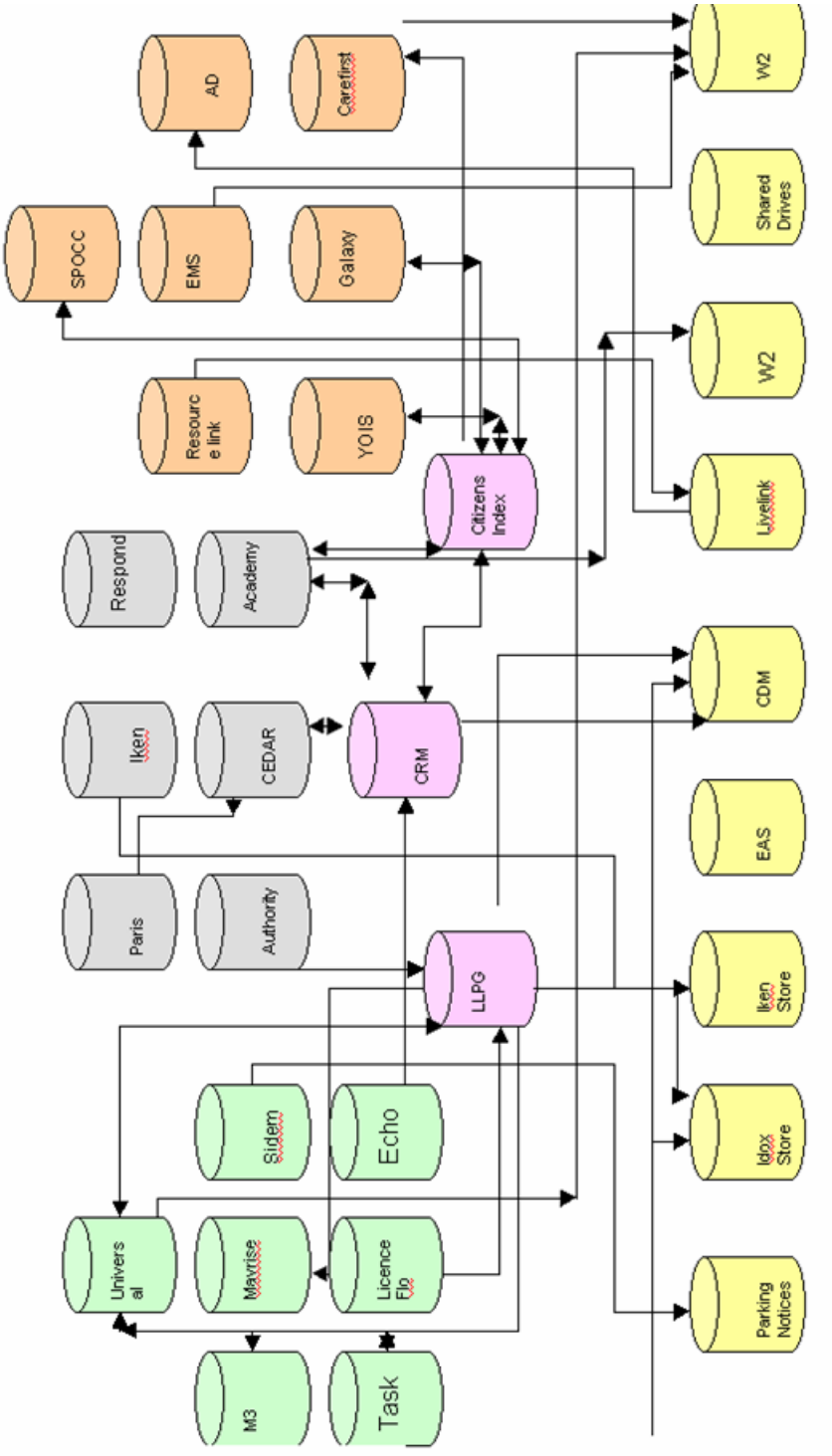



Figure 4. Applications categorised on the basis of people, property and transactions also showing integrations between each

This in turn leads to an asset analysis described above where key information and the applications, which drive that information, are placed in context depending on the type of information they support.
This can be made even more specific by looking at each independent application and categorising on the basis of the people, property, and transaction analysis described in Section. This is shown in figure 4. In the top left are property related data assets and systems and the link between show the possible integration between those systems so as to ensure that data is not replicated and is sued to best effect. On the right are people related systems whether they are citizens or employees or others. And in the centre transactions bring all these systems together through a customer related system.
This kind of architectural analysis can serve to direct the information strategy and ensure a programme of integrations (the integration plan) so as to maximise the re-use of information across all systems and services. Information should flow swiftly to where it is needed when it is needed with little effort on behalf of staff.
Finally Figure 5 illustrates possibilities for bring all systems together in a coherent structure where independent systems are replaced by overarching corporate systems and the whole can be addressed through an enterprise search engine capable of bring disparate information resources together.

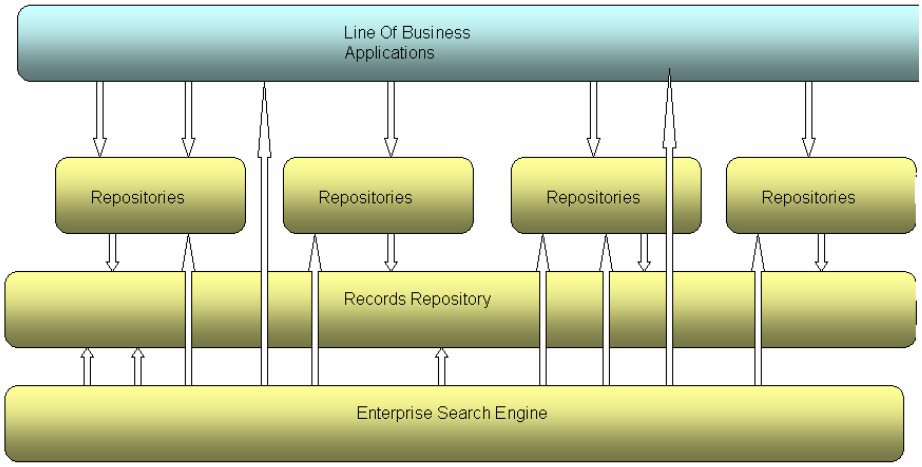



Figure 5. Possible future structure for a local government agency

## 8. References

BS7666 http://archive.cabinetoffice.gov.uk/govtalk/gdsc/html/noframes/BS7666Address-2.htm

Di Mariom, A Moving from e-government to government transformation. Business Issues 1-3 2006 http://www.gartner.com

Dmitriev, D Enterprise information architecture in London Borough of Hackney Sept 2008 MSc Dissertation University of Warwick.

Fremantle, P, Weerawarana, S and Khalaf, R 'Enterprise services. Examining the emerging field of web services and how it is integrated into existing enterprise infrastructures' Communications of the ACM, Vol 45 No 20 (2002) pp 77-82.

ISPV-Integrated Public Sector Vocabulary. Version 2.00 http://www.esd.org.uk/standards/ipsv/

Information Matters: building governments capability in managing knowledge and information HM Government Nov 2008

Jain, P The catch up state: government in Japan. Japanese Studies Vol 22 No 3 2002, p237 -255

Kamal, M et al The case of EAI in facilitating e-government in a Welsh authority. International Journal of Information Management Vol 298 209 p161-165

Janssen, M & Van Veenstar, A.F. 2005 Stages of growth in e-government; an architectural approach. Journal of E-government Vol 3 No 4 2005, p193-200.

Layne, K and Lee, J Developing fully functional models: a four stage model. Information Quarterly vol 18 2001 p122-136

Murphy, J. Beyond E-government: the world's most successful technology enabled transformations INSEAD 2005

NPLG National Land and Property Gazetteer http://www.nlpg.org.uk

Sarikas, O.D & Weerakkody, V Realising integrated e-government services: a UK local government perspective. Transforming Government: People, Process and Policy Vol 1 No 2 2007 p153 – 173

Signore, O et al. Egovernment: Challenges and Opportunities CMG Italy X1X Annual Conference 7-9 June 2005 Florence Italy. http://www.w3c.it/papers/cmg2005Italy.pdf

Weerakkody, V et al Integration and enterprise architecture challenges in E-government: a European perspective. International Journal of Cases in electronic commerce Vol 3 No 2 2007 p13-35

Weerakkody, V & Dhillon, G Moving from E-Government to T-Government A study of Process Reengineeing Challenges in a UK Local Authority Context. International Journal of Electronic Government Research Vol 4 No 4 2008 p1-16

Weerakkody, V & Dhillon, G. Moving from E-government to T –government: a study of process re-engineering challenges in a UK local authority context. http://www.igi-global.com/downloads/exceptrts/33216.pdf

Wimmer,M A European perspective towards online one stop government. The e-gov project. Electronic Commerce Research and Applications p92 -103

Zhao, J.L. & Cheng, H.K. Web services and process management: a union of convenience or a new area of research? Decision Support Systems 2008 Vol 20 no1 p1-8

# KNOWLEDGE MANAGEMENT

# Knowledge Sharing and the Process of Comprising Post-modernism and its Indeterminacy

Theodora Stathoulia
Independent expert on information technologies
23 Avenue de Hinnisdael
1150 Brussels, Belgium
tstath@gmail.com

**Summary**

*While it is generally accepted that knowledge management and information sciences are interdisciplinary fields, relying upon the foundations of the twentieth century post Kantianism deducting reasoning and the hermeneutic research hypotheses, the digital content and the web-based information services have imposed new claims in the domain of information science.*
*Structural changes in knowledge representation, multilingualism in digital content representation, semantic and syntax problems overcome prevailing knowledge organization morphologies.*
*This paper supports postmodernism and its indeterminacy as appropriate and integral in the new agenda of digital resources management and examines some theories about knowledge organization and representation.*
*Paper's objectives fall within the historical context of knowledge organization, not as ontology of objects not either as metadata schemata but as the basis of understanding the post -modernist social epistemology.*

**Key words:** knowledge management, modernism, postmodernism, social epistemology

## Introduction

Culture of modernism fulfilled human expectations for certainty about theories, disciplines, and models in information studies. Standing points deriving from the optimism of quantitative deductive positivistic models have supported diachronically solid research frameworks, introducing the term interdisciplinary in 'informational' theory and culture. A creative framework from dominated disciplines in social and humanistic sciences has stimulated curricula redesign, educational objectives and new directives on vocational skills for the information professional. The 'inevitable' technology and the basic assumption of technological determinism *'that a new technology [...] changes the society or the*

*sector into which it has 'emerged' […] 'we' adapt to it because it is the new modern way'*[1] seems to drive modern history.

Modern theorists found a comprehensive context to develop the outline of the 'information era' within which all research questions should meet the underline societal movements. Technicalities, such as indexing and classification in clear structures for both modernism fluidity and post modernism indeterminacy exercise new approaches to meta -industrial world[2]

In between, *'new concepts are needed to comprehend the world and its transformations. New words are needed to designate or to go along with these concepts […] the concept moves gradually to the center of terminology'*[3]

However, information and knowledge models based on conceptualisation of information should be supplemented by considerations on information and knowledge theory.

Within this framework this paper argues that knowledge representation and knowledge management cited as core issues in the meta-information era need a new epistemic standing point. Social epistemology here, as the *'social path'* dimension[4] indents to contribute to the theoretical construction of this work. To sign the importance of this need we argue that the system of knowledge that is, for Leibniz a system of truths should be deductively based on the division and the analysis of concepts and symbolisms within the meta-modern world where

[1] Williams, Raymond. The politics of modernism. London: Verso, 1989

[2] Hanson, Alan. From Classification to Indexing: How Automation transforms the way we think //*Social Epistemology,* Vol. 18 (4), October-December, 2004 pp.333-356

[3] Gresser, J-Yves. Terminology and Information Science(s)// *3rd International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA* , 2008

[4] Social epistemology refers here to the concept as it has been defined by Goldman, Alan in his work *'Knowledge in a Social World'*. Goldman, Alvin I. Regents Professor of Philosophy, University of Arizona. Print publication date: 1999. Published to Oxford Scholarship Online: November 2003. *"Traditional epistemology, especially in the Cartesian tradition, was highly individualistic, focusing on mental operations of cognitive agents in isolation or abstraction from other persons. […] But given the deeply collaborative and interactive nature of knowledge seeking, especially in the modern world, individual epistemology needs a social counterpart: social epistemology. In what respects is social epistemology social?* First, *it focuses on social paths or routes to knowledge. That is, considering believers taken one at a time, it looks at the many routes to belief that feature interactions with other agents, as contrasted with private or asocial routes to belief acquisition. This "social path" dimension is the principal dimension of sociality that concerns me here. Second, social epistemology does not restrict itself to believers taken singly. It often focuses on some sort of group entity – a team of co-workers, a set of voters in a political jurisdiction, or an entire society – and examines the spread of information or misinformation across that group's membership. Rather than concentrate on a single knower, as did Cartesian epistemology, it addresses the distribution of knowledge or error within the larger social cluster. Even in this second perspective, however, the knowing agents are still individuals. Third, instead of restricting knowers to individuals, social epistemology may consider collective or corporate entities, such as juries or legislatures […]"*

the individual should be taken in front of our thought. Current bibliography is occupied with an intense empiricism. There is a need for reconceptualisation of the basic components of our driven thinking.

In this line, important theoretical contributions should be recognized at the basics of information science. *'Information and library education has, as a result, become preoccupied with the 'management' of information and knowledge and its associated technologies of performance maximization. […] such a focus on information management needs to be balanced by new reconceptualised information science curricula. Such curricula it is claimed need to be responsive to the flux and creative potential of the post modern networked age, but also underpinned by principles of humanism, empowerment and critical distance'*.[5]

In this context of fundamental rethinking the re-approaching of structural rules in reforming information science, namely, the recapitulation of old perceptions should be deployed with the new approaches of the meta information era . Simply, as in the composition of a sonata, the musical themes that were introduced earlier should be repeated.

## Knowledge capture and knowledge representation models

*'An increasing amount of work is now focused on a knowledge-based paradigm in which knowledge is captured as past experiences in the form of case-specific knowledge. This type of knowledge forms the basis for case-based reasoning (CBR) methods, in which past problem solving episodes - cases - are recalled and used to solve new Problems.'*[6]

A typical approach to knowledge capture could be defined as the methodological approach presented above. Knowledge capture, as *case-specific knowledge* within the collective existing experience, the so-called *past experience* constitutes the core of a generic paradigm where past experience develops itself the reasoning framework to fulfilling new knowledge representation issues. It seems that a historic human knowledge acquisition can be multiplied in recent times. The highlighted paradigm taking for granted that human knowledge is transmitted throughout the same cultural and social mechanisms and understandings underestimates the invention of the new radical social, scientific and technological breakthroughs.

Moreover, according to this approach all we really needed to implement is the paradigm of the semantic values of the new social backgrounds through the new ontological schemes. That is to say, philosophy of language and semantics would help serve as the bridge between our past knowledge and the new lan-

[5] Muddiman, Dave. Towards a postmodern context for information and library education// *Education for information*; 1999, Vol. 17 Issue 1, p1-19

[6] Aamodt, A. A knowledge representation system for integration of general and case-specific knowledge //*Tools with Artificial Intelligence*, 1994. Proceedings., Sixth International Conference, 6-9 Nov. 1994 Page(s):836 - 839

guage representation. In other words, construction in the way humans from different knowledge and cultural backgrounds understand words and phrases could be used to create an interesting information tool for incorporating the distinction between word's 'meaning' and 'form'. For instance, the following work referring to a multilingual information system demonstrates a proposition of the discussed paradigm.

*'The words and the phrases in a natural language are symbolic representations of real world concepts. Information systems have traditionally associated semantics with keywords to index and retrieve information. However, ambiguity of word meanings and variation of user vocabulary result in unsatisfactory performance of these information systems. An online lexical database, such as the WordNet, distinguishes the "word meanings" (the intended concepts) from the "word forms" (the utterances) in English and establishes several lexical and semantic relations between the word meanings. The database has been used in several knowledge-based applications that attempts to "interpret" a message containing some user request or other forms of information. Similar lexical databases have been developed in other languages also. The major drawback of such lexical databases is that they are confined to a single language' The proposed model 'A domain ontology needs a medium for expression, which usually consists of terminology borrowed from a natural language. Thus, a Knowledge based application becomes susceptible to linguistic and cultural context. In this paper, we present a new knowledge representation technique that distinguishes between the abstract concepts in a domain and their expressions. It can associate expressions from different languages with the concepts in an ontology network. Non-textual symbols and media property specifications can also be used to express the concepts using this technique."*[7]

Abstract concepts and their expression from different languages create a new technique in accomplishing knowledge capture and representation. The reference presented extensively above preserve the same methodological paradigm in knowledge representation and capture as the majority of works. It should be underlined that there have been significant ways of understanding the problem of incorporating knew knowledge into dynamic mark up language schemes though unanswered questions of meta-information knowledge sharing era and its conveying has to be answered. In other words, the new world of division and ambiguity reshapes research questions though answers are investigated within the technicalities of knowledge representation. However, the modernism of coherence is gradually and dramatically replaced by the insignificance of the meanings; It is the strangle of an interactivity which paves to respond to the new world and its varieties by incorporating the suspend to the order of the

[7] Ghosh, H.; Rajarathnam, N.; Chaudhury, S. Knowledge representation for Web based services in a multi-cultural environment //*Web Site Evolution, 2001. Proceedings. 3rd International Workshop.* 10 Nov. 2001 Page(s):7 - 13

socio-economic development forms. The belief that *"[…] a knowledge based application becomes susceptible to linguistic and cultural context"*[8] is an insubstantial interpretation of the new concerns of the postmodernism era. Again, the cultural context is partly connected with the complexity of the new knowledge mechanisms and the representation queries. The prevailing model of a technological determinism and its administration seems to obstacle the manufacturing of new approaches; the argument that new issues in knowledge capture should reconceptualize the whole of cultural processing leaves aside the ontologies that should be built upon the core of a new paradigm.

## Discussion

Our inquiry focuses on knowledge theory, knowledge capture and representation in relation to post modernism new imperatives.

This paper proposes the necessity of a new paradigm shift to re-engineering the way the scientific legacy has introduced *'paradigm change'*. What about the incapacities of languages *'especially the ordinary languages of common life due to its preoccupation with the sense world and its consequent vagueness on ultimate matters"*.[9] Again, are *'concept maps'* the *'ultimate matters"*? Here, then, *'Knowledge models'* should be examined within the doctrine of philosophy to express the foundations of language and ontology.

Furthermore, are Kaminsky's *'ontological commitments'*[10] enough to represent the *'ultimate matters"*? The ontological commitment to subject by maintaining sentence structure and fundamentals of language is part of knowledge capture and representation, or knowledge syntax and semantics do not need such commitments? The use of the term 'knowledge capture' suggests that we have already placed in orbit the solar bodies round the observer in opposition to the evolutionary theory of Copernicus who wanted the observer to move round the solar bodies. By reducing our view to the tradition of rationalism, for that matter to Kantian *a priori* Knowledge, and furthermore to Leibnizian objective idealism, that time and space are 'orders' and 'relations', not entities or existences[11] we secure our scientific future. On the other hand, how are we going to capture and represent 'relations' in an extremely ambiguous knowledge environment? Can we find the answers in the construction of S*emantic schemata?*

However, so far, the belief that all propositions are steadily introduced when knowledge is identified within its conceptual context deriving from the histori-

[8] Ibid.

[9] Morrow, G. 'The theory of Plato' seventh epistle'// *The Philosophical Review*, vol. 38 (4), 1929, pp. 326-349.

[10] Frye, M. (Book review). Language and Ontology by J. Kaminsky//*The Philosophical Review*, vol. 80(3), July 1071, pp.394-396.

[11] Cassirer, E. Newton and Leibniz//*The Philosophical Review*, vol. 52(4), July 1943, pp.366-391.

cal background o rationalism, its linguistic parameters, and why not, its controversies as well as its own *a priori* existence seems to disregard our unfamiliarity with the new world of uncertainties. Thus,the three aspects of knowledge engineering, that is knowledge capture, knowledge storage and knowledge deployment have to be implemented within the new evidently unsecured scientific human environment.

*"[…] postmodernism is indefinable is a truism. However, it can be described as a set of critical, strategic and rhetorical practices employing concepts such as difference, repetition, the trace, the simulacrum, and hyper reality to destabilize other concepts such as presence, identity, historical …"*[12]

To examine the paper's argument within this definition is equal to an attempt for someone to go into the deep ocean to discover the lost ring. However, our argument proposes a standing point for a new attempt highlighted here as "*Destabilizing other concepts"* in a post modern perspective. This is the key concept that moves our argument to the field we feel confident to discuss. We argue that knowledge engineering should bring into attention the need for such a *"destabilization"*. Concepts taken as part of an existing historical framework, identical to the human achievements in the context of the *Enlightenment* forward current technology and culture assessment and propose that rationalism and science development have liberating human progress. In this context, new developments are placed firmly on this historical development. Any inherent irrationalism to human progress is considered as an instant in the advance of the human advanced movement.

Undoubtedly, knowledge engineering is grounded in social epistemology, where *'social paths'* guide to the certainties of modernism. However, the belief that all concepts are well-grounded in the socio-historical context and all they need is the conditions for reshaping the forms of expression overcomes current needs in knowledge sharing. In this way, modernism and the persistence on its theoretical premises leave behind conceptual misconstructions of the rationalist era. Therefore, postmodernism as a movement to 'destabilize' concept expression and representation within the social and epistemological constructions should be associated with current knowledge theory demands.

Finally, by attempting to remain in the certainties of the modernism we might fail to understand current requirements on knowledge theory.

## Conclusion

In order to line around the boundaries of our argument, a rough outline of the two tendencies on knowledge theory was attempted. The foundation of our proposal is the reconciliation of the empiric supervisory knowledge with the purely conceptual. The system of knowledge that is for Leibniz a system of truths

[12] Gary Aylesworth. http://plato.stanford.edu/entries/postmodernism/

should be deductively based on the division and the analysis of concepts within the *'disturb'* of modernism. The ideal of knowledge representation supported by the Kantian deductive method should incorporate the uncertainty of the new world. Knowledge schemata deriving from a system of truths universally accepted in the classic hierarchical representation of the scientific knowledge it is proposed to be uploaded under the postmodernism and its indeterminacy.

We have a long way ahead to connect, however not merely ontologically, the various knowledge engineering epistemological approaches. To connect empiricism and the rationalism, and moreover to find the structures and functions of languages we need to serve our knowledge representation. The reconciliation of the *'observer'* with the *'object'*, the *'harmony'* between the capacities of the knower and the nature of the known, the Kantian *'transcendental idealism'*, after all the belief that the existence following *a priori* knowledge serves fundamental aspects should be incorporated into future research work altogether with efforts in understanding the new meta information world.

However, the new perspective in knowledge sharing and knowledge management should be examined by the replacement of facts by more *"symbolism in thoughts"*. Now, we need more concepts and syllogism and less pragmatism. This prospect is supposed to bring to the surface the topics all along with the forms that should be addressed in the new knowledge landscape that is shaped by the emerging needs for Knowledge sharing.

In particular, to serve the emerging needs for the unification of concepts and case-based experiences (external knowledge) and for the corporation of fact-based knowledge (implicit knowledge). The new world might recollect the historical outcome as a questionable fact. In this process knowledge engineering should be present.

Original scientific paper

# Predecessors, Scholars and Researchers in Information Sciences. Contribution to Methodology for Bibliometrics Analysis of Scientific Paradigms

Đilda Pečarić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
dpecaric@ffzg.hr

Miroslav Tuđman
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mtudman@ffzg.hr

**Summary**

*Development of Information Science paradigm is researched on the corpus of most cited references retrieved from doctoral dissertations in Information Science (from 1978 to 2007). New approach for analysis of scientific paradigm by empirical display of dominant zones within scientific paradigm is proposed: empirical knowledge zone, conceptual knowledge zone and research front zone. Alterations of scientific paradigm are followed across three time periods by display of most cited authors in librarianship, information systems, communicology, archivistics and documentation, museology and information science. Besides the data about most cited authors, the data about most cited references according to periods and disciplines are shown. Analysis of most cited references resulted with discovery of the dominant research topics in particular periods. Based on changes in research topics it can be concluded: a) which research topics were interesting for Information Science researchers in Croatia, and b) changes within Information Science paradigm, by tracking changes of key authors which are cited during period of thirty years. Suggested methodology can serve as a model for tracking the development of scientific paradigm in other research disciplines as well.*

**Key words:** Information Science, Scientific Paradigm, Communication Models, Knowledge Zones, Methodology, Bibliometrics Analysis

## Introduction

It is possible to analyze the development of Information Science and the role of key authors and key publications in Information Science community by bibliometrics methods.

We start from the assumption that doctoral dissertations in Information Science are a good sample for the analysis of Information Science development in Croatia, because doctoral dissertations are original scientific publications which are using up to date world key literature.

## Methods

We analyzed 134 doctoral dissertations in Information Science done on Universities in Croatia from 1987 to 2007. The doctoral dissertations were done on Croatian Universities that have postgraduate studies in Information Science, i.e. The Senate of the University of Zagreb / Znanstveno-nastavno vijeće Sveučilišta u Zagrebu (from 1978 to 1981), Zajednički studij informacijskih znanosti (from 1985 to 1987)/, Faculty of Organization and Informatics (from 1987), Faculty of Humanities and Social Sciences (from 1990)/ and The University of Zadar (from 2001).

The classification of doctoral dissertations according to disciplines is based on the classification of scientific disciplines and fields used by the Ministry of Science, Education and Sports of The Republic of Croatia. According to that classification Information Science is divided into following disciplines: Archivistics and Documentation, Librarianship, Communicology, Lexicography and Encyclopedics, Museology, Information Science and Information Systems (According to classification of Ministry of Science ‘Information Systems and Information Science’ are the same discipline, but for the purpose of our analysis we divided them into two disciplines, Information Systems and Information Science, in order to separate doctoral dissertations done on the Faculty of Organization and Informatics and Faculty of Humanities and Social Sciences). For the analysis of citation corpus of 22,210 bibliographic units in 134 doctoral dissertations we used cluster analysis. Clusters are formed according to the frequency of cited authors and titles. The obsolescence of literature was important for our analysis. Therefore we used usual criterion of citation “half-life” which is determined as period of time in which 50% of references are cited.

In previous papers we presented the criteria that can more precisely describe the development of the Information Science. We advocate that is possible to identify dominant fields of scientific influence inside scientific paradigm, i.e. empirical knowledge zone, conceptual knowledge zone and research knowledge zone (M. Tuđman, Đ. Pečarić, 2009.). Further analysis of relationships between authors’ in research and in conceptual knowledge zones (Đ. Pečarić, 2009.) indicates that in spite of constantly changing position and role of authors, it is possible, with citation obsolescence criteria, to identify three different groups of authors: group of predecessors, group of scholars and group of researchers.

The development of Information Science in Croatia, ie. Information Science disciplines in the last thirty years will be analyzed by prepared methodology.

## The Most Cited Authors in Information Science Disciplines

Tables 1 to 3 show the most cited authors in museology, information science and information systems[1]. Authors of papers written in different languages are not grouped in the same cluster. Why? We wanted to stress the fact that there exists a difference between citation and reference. Although both are formed from the same bibliographic data and both can be and are the same, the important difference between citation and reference lies in the manner of their usage: reference is "acknowledgment which one author gives to another", whereas citation is "acknowledgment which one document receives from another" (J. Petrak, 2003.). Because of language barrier it is possible to assume and advocate inequality that exists between citation and reference. It is evident that authors of doctoral dissertation acknowledged the authors who published their papers in foreign languages. At the same time, the authors who are not familiar with the "small" languages can not respond in the same way. Because of that asymmetry of citation usage publications published in foreign languages are shown in the right top corner of the table, and publications published in Croatian language are shown in the left bottom corner of the table.

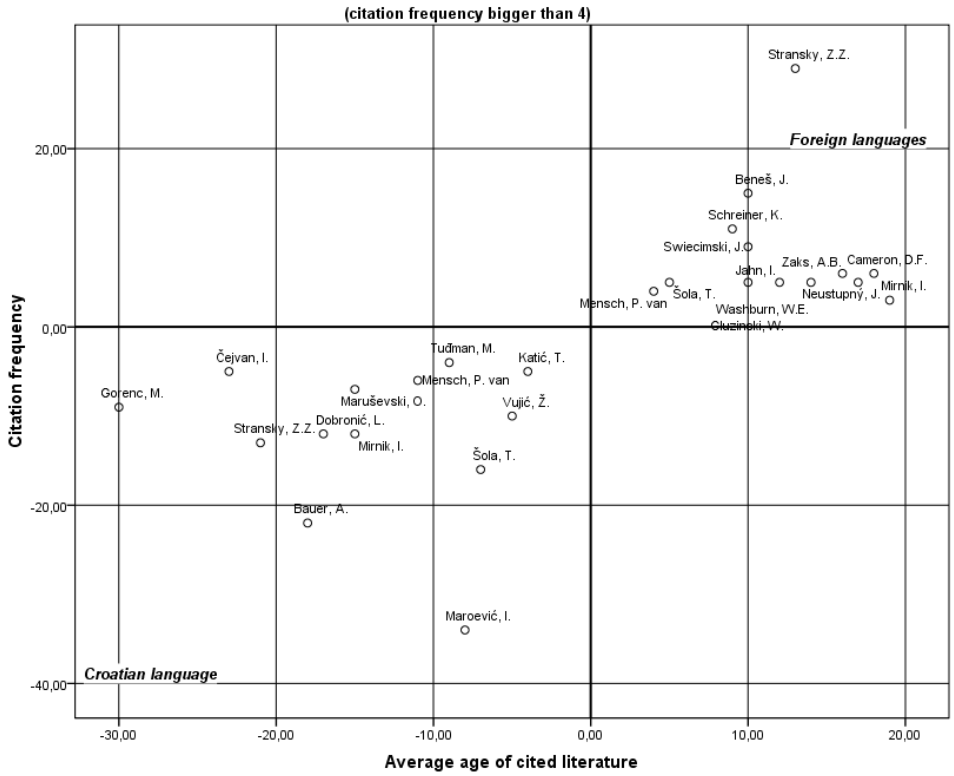
In order to be able to make conclusions about development of Information Science paradigm, it is important to evaluate the sample of the most cited authors in certain disciplines that are shown in tables 1 to 3.

Of the overall number (1279) of all cited authors in museology, 22 of the most cited authors make only 1.7%. However, these 22 authors hold 10.2% of citations from overall number of cited documents in museology. There are 972, or 76%, of authors that are cited only once in museology. But, in order to be precise, these percentages should be corrected, because the number of documents (both anonymous and those having an author) that are cited only once is 51.9%. Therefore, it is more precise to say that almost 1/5 of all multiple citations hold 1.7%, that is, 22 most cited authors.

In other two disciplines frequency of citations behave in a similar manner. In information science, first 32 authors or 1.8% authors (from 1770 most cited authors) hold 7.7% of citations. In information science there are even 80.8% of authors that are cited only once. However, since in this discipline a large number of documents without authors (16%) are cited, the overall number of all documents (with or without authors) cited only once is 62.9%. So, the conclusion is similar to previous one, i.e. a small number of authors (1.8%) holds 1/6 of all multiple citations.

[1] Because of the lack of space, in this paper, we are not able to show the most cited authors in all disciplines. In previous paper (M. Tuđman, Đ. Pečarić, 2009.) the most cited authors from librarianship and communicology are shown.

Table 1: 27 most cited authors in Museology from 1988 to 2007

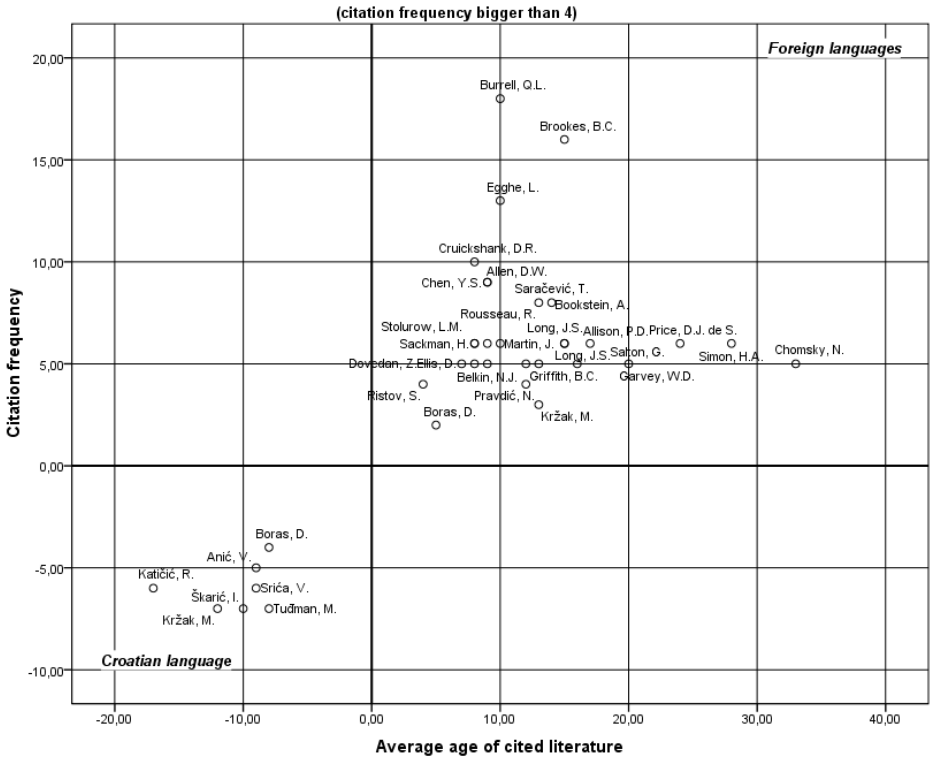



Although the number of most cited authors in information systems (table 3) is similar to the previous discipline, the differences are following. These 31 most cited authors make only 0.8% of 3662 authors cited in this discipline. In information systems 2981 authors or 81.4% are cited only once. Also in this discipline 11.6% of cited documents are without authors, so it is more realistic to accept that 61.1% documents are cited only once. But, in comparison with this information, 0.8% authors hold almost 1/6 of multiple citations.

From this data it can be concluded that a small number of authors (in our example between 1.6% and 1.8%) receives between 8% and 12% of all authorial citation. However, it is realistic to start from the fact that in these disciplines about 60% of cited documents are cited only once (regardless of the authorship status), so it can be concluded that 1.6% to 1.8% authors hold 1/5 or 1/6 of all multiple citation.

In three analyzed disciplines 90 authors hold 1/6 of all citation. However, it should be taken into account that out of 90 most cited authors in all three disciplines, 50% of authors is “mutual”; namely, 44 authors are cited in two or three disciplines.

Table 2: 32 most cited authors in Information Science from 1978 to 2007

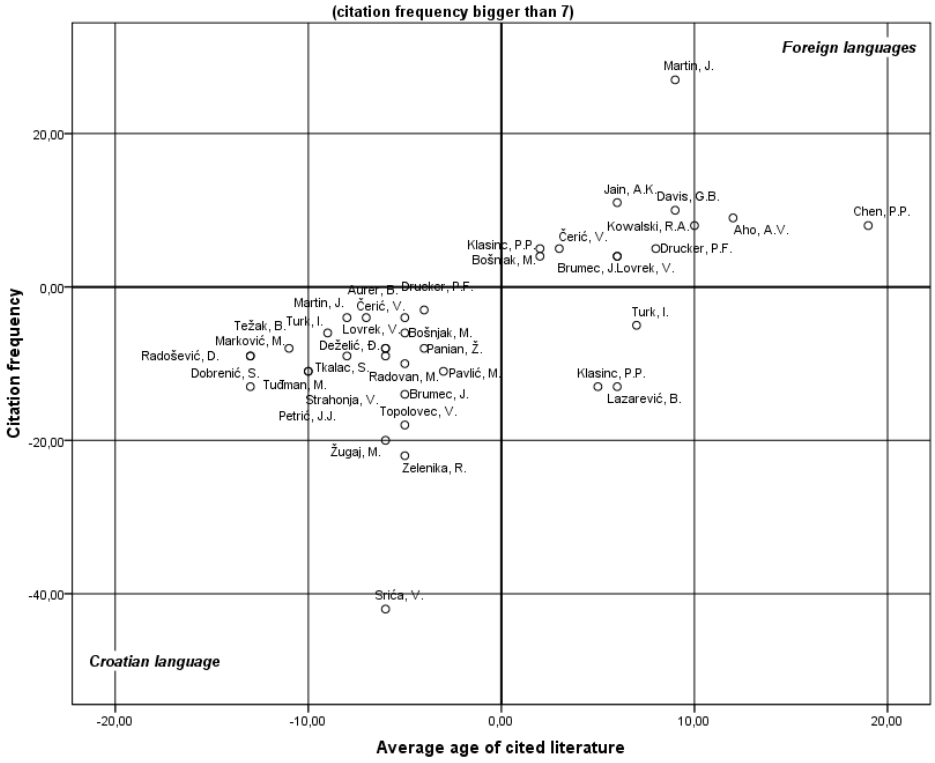



It is also important to know how many cited authors in three disciplines (table 1 to 3) are cited in other Information Science disciplines. In museology only 4 out of 27 authors are cited in other disciplines. However, 26 out of 32 most cited authors in information science are cited in other disciplines, whereas 19 out of 31 most cited authors in information systems are cited in other disciplines.

We can raise the question in how many disciplines are present the most cited authors from museology, information science and information systems[2]? Only one author (M. Tuđman) is cited in all seven disciplines. Four authors (N. J. Belkin, G. Salton, T. Saračević, A. I. Mihajlov) are cited in five different disciplines. Seven authors are cited in four different disciplines (V. Anić, M. Kržak, D. de S. Price, V. Srića, B. Težak, S. Tkalac, M. Žugaj). Nine authors are cited in three, and 23 authors are cited in two different Information Science disciplines.

[2] In this analysis we use partition of Information science into following seven disciplines: archivistics and documentation, librarianship, communicology, lexicography, museology, information science and information systems.

Table 3: 31 most cited authors in Information Systems from 1980 to 2007




It is evident that joint core of authors who are often cited in several Information Science disciplines exist. Based on these data we can conclude that joint theoretical baseline in Information Science also exist.

## Predecessors, Scholars and Researchers in Librarianship, Communicology and Information Systems

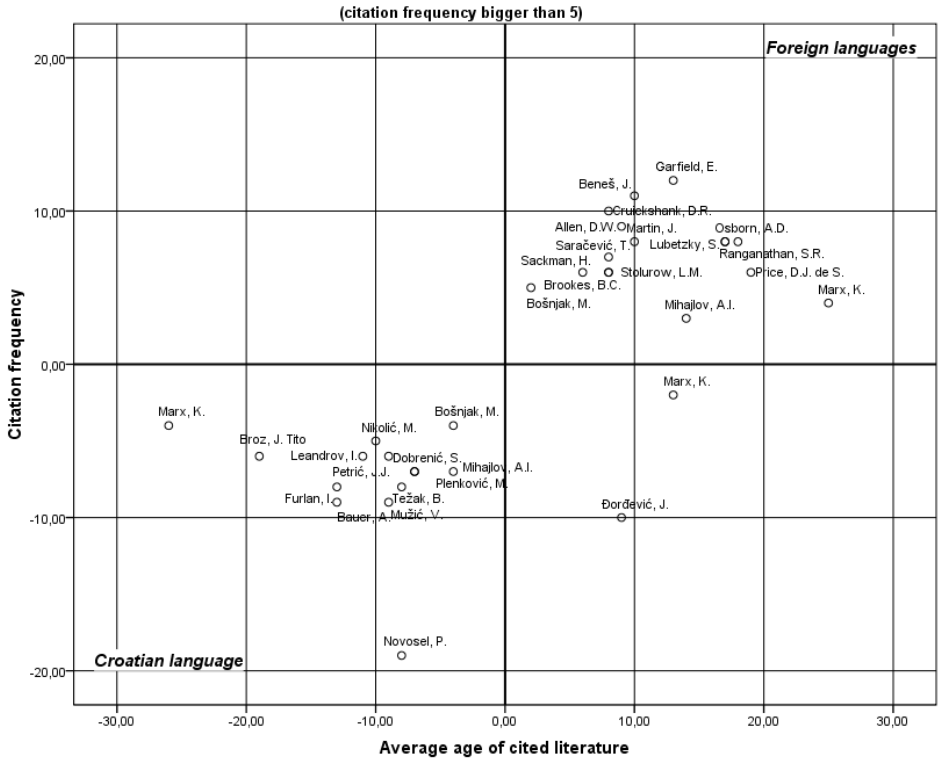
According to the criteria of cited literature obsolescence, and according to location in the cluster of cited authors, we recognize several groups of authors: predecessors, scholars and researchers (Đ. Pečarić, 2009.).

In the group of predecessors we can include authors that are continuously cited after double citation half-life, i.e. those that belong to the last 20% of citations cycle.

The group of scholars form the authors that are cited after the period of citation half-life and until the end of double period of citation half-life. Those two groups of authors (scholars and predecessors) are defining conceptual knowledge zone. According to T. Kuhn these two groups of authors are key representatives of dominant scientific theories.

Researchers are the most cited authors in the first half of citation half-life. They belong to the research front. Their publications are mostly cited immediately after publishing – and if they remain permanently present in scientific community, then during the time they become part of the dominant scientific paradigm.

Table 4: 28 most cited authors from 1978 to 1989



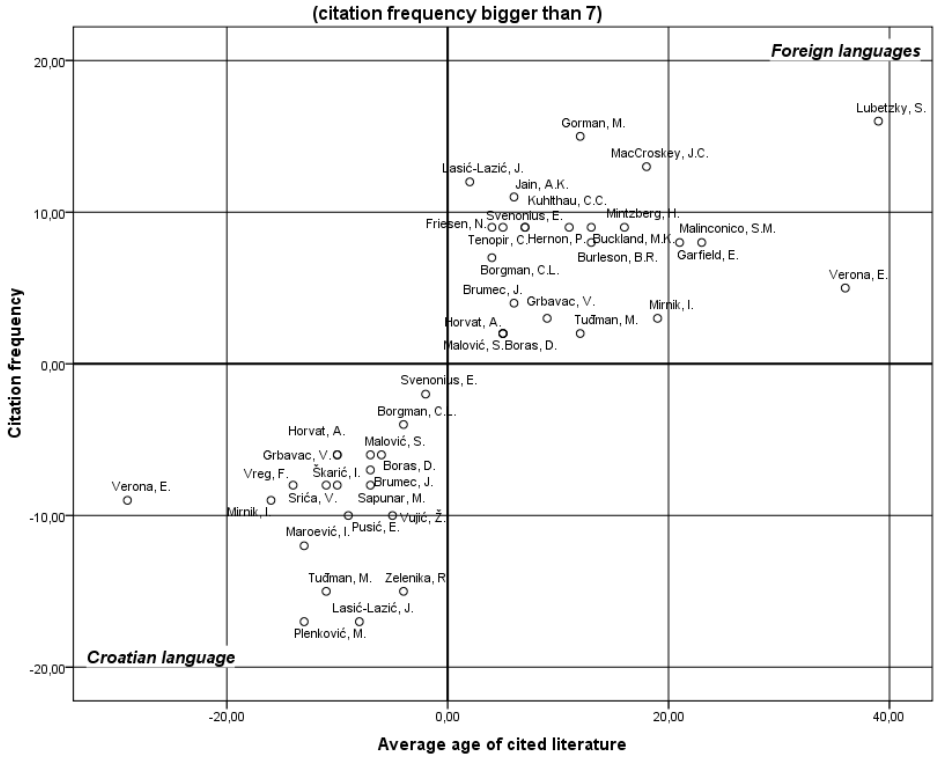

The group of authors that form predecessors in museology are both founders and key authors. According to obsolescence of cited literature the group of predecessors in museology is: R. Horvat, M. Gorenc, I. Čejvan, Z. Z. Stràsnký, I. Mirnik, A. Bauer[3].

The group of predecessors in information science form: N. Chomsky, H. A. Simon, D. J. de S. Price, W. D. Garvey, K. Katičić, P. D. Allison, G. Salton, J. S. Long, B. C. Brookes.

[3] Average obsolescence time in museology is unrealistically high (12.6 years) because it was not possible to discern the citation of documentation's source material from the citation of relevant literature. That is why the authors whose cited literature is around 20 years old are included in this group, and not only those whose cited literature is more than 24 years old.

The group of predecessors in information systems forms: W. D. Garvey, G. Salton, S. Dobrenić, D. Radošević, A. I. Mihajlov, A. V. Aho.
According to formal criteria of most cited authors and literature obsolescence, the group of scholars in museology form: D. F. Cameron, L. Dobronić, J. Neustupný, O. Maruševski, W.E. Washburn.
The group of scholars, according to formal criteria of most cited authors and literature obsolescence, in information science form: A. Bookstein, T. Saračević, N. Pravdić, M. Tuđman, M. Kržak, Q. L. Burrell, L. Egghe, J. Martin, D. W. Allen, Y. S. Chen, V. Anić, N. J. Belkin, V. Srića, D. R. Cruickshank, R. Rousseau, H. Sackman, L. M. Stolurow, D. Boras, Z. Dovedan.
According to same criteria, the group of scholars in information systems form: R. A. Kowalski, M. Tuđman, I. Turk, G. B. Davis, P. F. Drucker, I. Martin, J. J. Petrić, V. Strahonja, V. Srića, V. Čerić, Đ. Deželić, B. Aurer, M. Žugaj, S. Tkalac, J. Brumec, A. K. Jain, V. Lovrek.
The group of researchers in all three disciplines is formed from the remaining authors that we did not list as scholars or predecessors.

Table 5: 33 most cited authors from 2000 to 2007



In this paper we were unable to indicate change of place and authors' role within the paradigm in all Information Science disciplines over time. In tables 4 and 5 we show different positions of certain authors – at the beginning and at the end of analyzed development cycle of Information Science. Data fragmented according to document citation and periods[4] indicates that certain authors are being cited for a long period of time. But usually citation period is not longer than two periods of time[5]. In fact only one author occurs in all three periods - M. Plenković. Authors that are cited in the first and second periods are A. Bauer (13, 22), B. C. Brookes (6, 19), J. Martin (10, 8), P. Novosel (8, 11), D. de S. Price (19, 23), T. Saračević (8, 13)[6].

Authors cited in the second and third periods are: I. Maroević (6, 13), V. Srića (7, 11), M. Tuđman (9, 12). It is interesting that E. Garfield (13, 21) and S. Lubetzky (17, 39) are cited in the first and third, but not in the second period.

With these examples it has to be taken into account that there is approximately the same small number of most cited authors in all three periods[7].

Some of these most cited authors are cited in other periods as well, but with not so high frequency. Therefore, the absence of cited frequency indicates the oscillations of the authors' influence and alterations of authors' position in scientific knowledge zones.

## Predecessors, Scholars and Researchers' Key Publications According to Disciplines

We can provide empirical data for qualitative analysis of Information Science development, specifically data about who key authors in specific time periods were, as well as the publications crucial for the education and scientific development of information science. But we have to establish the criteria for the selection of those authors and publications. Only after that we can make conclusions about main topics that were dominant in certain Information Science disciplines during thirty years.

Citation criterion, i.e. insight in most cited authors, is not sufficient alone and can lead to wrong conclusions. For example: among most cited authors there are publications of: T. Mušnjak, P. Strčić in arhivistics; P. Selem, E. Laszowki, G. Novak, I. Uranić, etc. in museology; P. Rudan, A. Sujoldžić, D. Horga, etc. in

[4] Analyzed cited literature corpus is divided into three periods: 1st period is from 1978 to 1989; 2nd period is from 1990 to 1999; 3rd period is from 2000 to 2007

[5] What we have in mind here is the "durability" of the most cited author, i.e. on their presence among the most cited authors in empirical and conceptual knowledge zones.

[6] The numbers in brackets symbolize the average age of cited literature in the 1st and 2nd periods.

[7] The first period embraced 28 authors whose citation frequency was bigger than 5; the second period embraced 28 authors, too, but their citation frequency was bigger than 11; the third period embraced 33 authors whose citation frequency was bigger than 7.

information science. However, each of this authors’ are cited in only one dissertation and therefore it is realistic to assume that these publications or authors are not crucial for Information Science paradigm.
An overview of key authors and their publications can be presented according to several criteria, or combination of criteria, so far described as:

a) overview of most cited authors and their publications according to disciplines;
b) overview of most cited authors and their publications according to periods;
c) overview of most cited authors according to location and authors’ role in scientific community: predecessors, scholars, researchers;
d) overview of most cited authors and their publications according to the number of disciplines in which they were cited.

Since in this paper is not possible to elaborate the presentation of all these overviews, i.e. implementation of all analysis’ criteria, this approach will be illustrated only with a few fragmentary examples.

**Overview of the most cited authors and their publications according to disciplines**

First five most cited authors and publications in museology:

- Stràsný, Z.Z.: *Pojam muzeologije; Temelji opće muzeologije; Prezentacija najnovije historije u čehoslovačkim muzejima*.
- Maroević, I.: *Uvod u muzeologiju; Predmet muzeologije u okviru teorijske jezgre informacijskih znanosti; Sadašnjost baštine.*
- Bauer, A.: *Muzeologija; Mreža muzeja i međumuzejska suradnja.*
- *Šola, T.: Prilog mogućoj definiciji muzeologije; Marketing u muzejima : ili o vrlini i kako je obznaniti; Od obrazovanja do komunikacije.*
- Mirnik, I.: *Numizmatička zbirka; Skupni nalaz novca iz Krupe.*

First five most cited authors and publications in information science:

- Burrell, Q.L.: *The analysis of library data; A note on ageing in a library circulation model.*
- Brookes, B.C.: *The foundations of information science; A New Paradigm for Information Science.*
- Egghe, L.: *Introduction to informetrics: quantitive methods in library, documentation and information science; Consequences of Lotka's law for the law of Bradford.*
- Tuđman, M.: *Teorija informacijske znanosti; Struktura kulturne informacije; Obavijest i znanje.*
- Kržak, M.: *Serbo-Croatian Morpho-spelling; Rječnička baza hrvatskoga književnoga jezika; Opisna, stohastička i relacijska gramatika na primjeru morfologije hrvatskog književnog jezika.*

First five most cited authors and publications in information systems:

- Srića, V.: *Uvod u sistemski inženjering*
- Strahonja, V. M. Varga, M. Pavlić: *Projektiranje informacijskih sustava*
- Lazarević, B., V. Jovanović, M. Vučković: *Projektovanje informacijskih sistema*
- Radovan, M.: *Projektiranje informacijskih sistema*
- Tkalac, S.: *Relacijski model podataka*

It is not hard to conclude that overview based only on citation frequency of authors and publications is not sufficient for conclusions that would make us better to understand key authors in Information Science. This list should be corrected and presented in such way that authors can be grouped, not just according to citation frequency, but according to place and role in scientific community, in order to recognize whether they are researchers, scholars or predecessors.

**Overview of most cited authors according to periods**

From overall number of cited authors in all disciplines, in the first period (from 1978 to 1989) first five most cited publications are:

- Mihajlov, A.I.: *Uvod u informatiku i dokumentaciju.*
- Vreg, F.: *Društveno komuniciranje.*
- Dworatschek, S.: *Uvod u obradu podataka.*
- Eco, U.: *Kultura, informacija, komunikacija.*
- Novosel, P.: *Delegatsko informiranje.*

In the second period (from 1990 to 1999) first five most cited publications are:

- Tuđman, M.: *Teorija informacijske znanosti.*
- Srića, V.: *Uvod u sistemski inženjering.*
- Plevnik, D.: *lnformacija je komunikacija.*
- Žugaj, M.: *Osnove znanstvenog i stručnog rada.*
- Grad, J., G. Resinović, V. Rupnik: *Ekonomika informacijskih sistema.*

In the third period (from 2000 to 2007) first five most cited publications are:

- Lasić-Lazić, J.: *Znanje o znanju.*
- Tuđman, M.: *Obavijest i znanje.*
- Žugaj, M.: *Temelji znanstvenoistraživačkog rada.*
- Boras, D.: *Teorija i pravila segmentacije teksta na hrvatskom jeziku.*
- Eco, U.: *Kultura, informacija, komunikacija.*

An overview of the most cited publication is also not sufficient for the complete understanding of Information Science development. The reason for that is that the overviews of the most cited authors and the most cited publications often differ. In fact, often the most cited authors are the authors that have large number of publications. That amount of publications is what, in the end, puts them in the leading position on the citation scale. In other words, authors that publish larger number of publications cover larger number of topics, and that is the rea-

son why they get cited more often. Overview based only on citation frequency of single publication does not take into account continuous presence of authors that publish large number of publications and their relevance for the broader field of Information Science.

**Overview of most cited authors and their publications according to the number of disciplines in which they are cited**

Earlier we stated that a small number of authors are cited in more than three Information Science disciplines. That is why we can also display those authors and their papers which are cited in several disciplines.

Authors cited in five or more Information Science disciplines[8]:

- M. Tuđman (21): *Teorija informacijske znanosti; Struktura kulturne informacije; Obavijest i znanje.*
- N. J. Belkin (12): *Information concepts for information science; The cognitive viewpoint in information science; Information science and the phenomenon of information.*
- G. Salton (10): *On the Development of Information Science.*
- T. Saračević (24): *Relevance. A Review of and a Framework for the Thinking on the Notion in Information Science; An Essay on the Past and Future (?) of In-formation Science Education; The impact of information science on library practice.*
- A. I. Mihajlov (9): *Uvod u informatiku i dokumentaciju; Uvodni tečaj o informatiki i dokumentaciji.*

Authors cited in four different Information Science disciplines:

- V. Anić (5): *Pravopisni priručnik hrvatskoga ili srpskoga jezika*
- M. Kržak (12): *Serbo-Croatian Morpho-spelling; Opisna, stohastička i relacijska gramatika na primjeru morfologije hrvatskog književnog jezika; Rječnička baza hrvatskoga književnoga jezika.*
- D. de S. Price (10): *Little Science, Big Science; Networks of Scientific Papers.*
- V. Srića (21): *Informacijski sistemi; Informatički inženjering i menadžment; Od krize do vizije skice - za jugoslavensku tehnološku utopiju.*
- B. Težak (13): *Informaciono-dokumentaciono-komunikacioni (INDOK) sistem.*
- S. Tkalac (7): *Relacijski model podataka.*
- M. Žugaj (10):*Osnove znanstvenog i stručnog rada.*

It is obvious that citation of a larger number of key authors and their publications in several Information Science disciplines, would make a better foundation for joint theoretical basis, because of the fact that scientific community quotes

[8] The number of cited publications is given in brackets behind the authors' name. Further, we give the titles of first or next several titles of most cited publications for each author.

and shares same sources. Nevertheless, even in that case one could perceive a lack of insight into the inner dynamics of Information Science development: according to time periods and according to roles that specific group of authors has in specific time period. The lack of insight into the inner dynamics of Information Science development can be perceived even if we expand roles that specific groups of authors have in specific time period: Are these authors a part of research front? Are these authors scholars that dominate in scientific community? or Are these authors predecessors whose knowledge is the authority, but also a part of historical knowledge?

## Instead of conclusion

The task of this paper was not to give precise answer on who were the key authors and what were the key publications in Information Science that the Croatian scientific community from 1978 to 2007. Our intention was to prepare possible methodology for the research of Information science development.

Usage of quantitative bibliometrics methods, to make qualitative conclusions could be rather risky. However, with the combination of a variety of quantitative criteria it is possible to process data in such a way that a large number of data (in our research 22,210 cited documents) can be reduced. Using empirical method to find set of key data (several dozens of key authors and publications) we can provide reliable data for qualitatively analyzed.

In our analysis of Information Science development we advocate several starting points. First of all, we demonstrate how it is possible to identify dominant field of scientific influence inside the scientific paradigm (i.e. we recognized empirical knowledge zone, conceptual knowledge zone and research knowledge zone).

Second, we propose criteria for the recognition of several groups of authors, with different influence and roles in described zones: predecessors, scholars and researchers.

Third, based on the examples given in this paper we uphold the use of several criteria that can serve as a filter for data selection: a) citation of authors according to disciplines; b) citation of authors and their publications according to periods; c) classification of authors according to location and role in scientific community (i.e. on predecessors, scholars, researchers); d) overview of authors and their publications according to the number of disciplines in which they are cited.

We are convinced that with this kind of approach it can be possible to obtain empirical data relevant for research and qualitative analysis not only for Information Science development but also for some other disciplines in social sciences.

## References

Capurro, Rafael; Chaim Zins. Knowledge Map of Information Science. Rafael Capurro's responses to Chaim Zins. (2006). http://www.capurro.de (2009.)

Garfield, Eugene. Citation indexes to science: a new dimension in documentation through association of ideas. // Science. (1955.): 122; 108-111. http://garfield.library.upenn.edu/essays/v6p468y1983.pdf. (2005.)

Garfield, Eugene. The History and Meaning of the Journal Impact Factor. //JAMA. 295 (2006.) 1, 90-93

Garfield, Eugene; Sher, Irvin H. Genetics Citation Index. Philadelphia, Pa: Institute for Scientific Information; 1963. http://www.garfield.library.upenn.edu/essays/v7p515y1984.pdf. (2005.)

Jokić, Maja. Bibliometrijski aspekti vrednovanja znanstvenog rada. Zagreb: Sveučilišna knjižara, 2005.

Kuhn, Tomas S. Struktura znanstvenih revolucija. Naklada Jasenski i Turk. Hrvatsko sociološko društvo. Zagreb. 1999.

Pečarić, Đilda. Razvoj informacijske znanosti u Hrvatskoj. Bibliometrijska analiza doktorskih disertacija iz informacijskih znanosti 1978.-2007. Zagreb: Filozofski fakultet, doctoral dissertation, manuscript, 2009.

Petrak, Jelka. Citati i njihova analiza. Vidi: Hrvatsko informacijsko i dokumentacijsko društvo, 2003. http://www.hidd.hr/articles/citati.php (URL 10.4.2009).

Tuđman, Miroslav; Pečarić, Đilda. Prilozi dubinskoj analizi komunikacijskih obrazaca. // Informatologija 42, 2009., 2, 87-92.

Original scientific paper

# Frequency of Scientific Production in Information Sciences

Emil Bačić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
emil1811@gmail.com

Tihomir Pleše, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
tihomirplese@gmail.com

Ivana Tomić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
ivanatomic76@gmail.com

**Summary**

*The aim of this research is to investigate the development of the Information Sciences in Croatia. The specific research goal was to follow scientific production of the doctoral candidates after completion of their doctoral studies according to discipline of research (primarily in the area of Information Sciences) and frequency of scientific production. 2,402 relevant scientific papers written by 107 doctoral candidates were found or 22.45 scientific papers per author. The majority of papers were written in the field of the Information Sciences and after the doctoral graduation. The results of this research may be considered to be a solid indicator of the scientific production in the area of Information Sciences.*

**Key words**: Information Sciences, scientific production, doctoral candidates, CROSBI

## Introduction

Considering the fact that scientific production is one of the major indicators for development of a scientific discipline, the starting point of this research was the list of all candidates for Ph.D. in Information Sciences from 1978 to 2007 at

Croatian universities[1]. Their scientific production had been analyzed according to Croatian Scientific Bibliography database. This is a database sponsored by the Ministry of Science, Education and Sports of the Republic of Croatia, which contains the list and description of those scientific papers published within projects financed by the Ministry of Science, Education and Sports of the Republic of Croatia.

Croatian Scientific Bibliography (CROSBI) database is an electronic database that allows for data input and search through a web interface, offers fast access to scientific publications from a particular science project or scientist and enables the user to find scientists working in narrow, highly specialized scientific fields in Croatia. CROSBI was initiated in 1997 with the fundamental goal of collecting in one place all publications resulting from scientific projects financed by the Ministry of Science, Education and Sport of the Republic of Croatia. Authors themselves create the bibliography, while librarians, computer and information specialists supply the forms, standards and monitoring of the entire process.

CROSBI is a database open to all types of science papers, journal articles, books, book chapters, proceedings of scientific symposia as well as all types of papers, expositions, technical reports, manuscripts et al. CROSBI currently stores data on 190,000 scientific and professional papers along with 4,000 complete works by Croatian authors.

The specific research goal was to follow scientific production of Ph.D. candidates after doctoral graduation according to discipline of research (primarily in the area of Information Sciences) and frequency of scientific production.The analysis covered 2,402 relevant scientific papers written by 107 Ph.D. candidates or 22.45 scientific papers per author. The data from the papers that had been analyzed was introduced into Microsoft Excel worksheets, where most important data included: publication year, scientific field and paper category (scientific papers, professional papers…)

## Data analysis

Having collected the data, it was necessary to perform basic data analysis so the collected data could be used and studied. The first step was establishing a relationship between the dates papers had been published and their authors' graduation. This was done in Microsoft Access format, which produced new data – a numerical indicator of the age of published literature. For example, papers written 10 years after completing the doctoral study were marked as 10, while those written 4 years prior to completion of doctoral study were marked as -4.

The next step of the analysis focused on scientific areas covered in the papers. All the works from the domain of Information Sciences (including those which

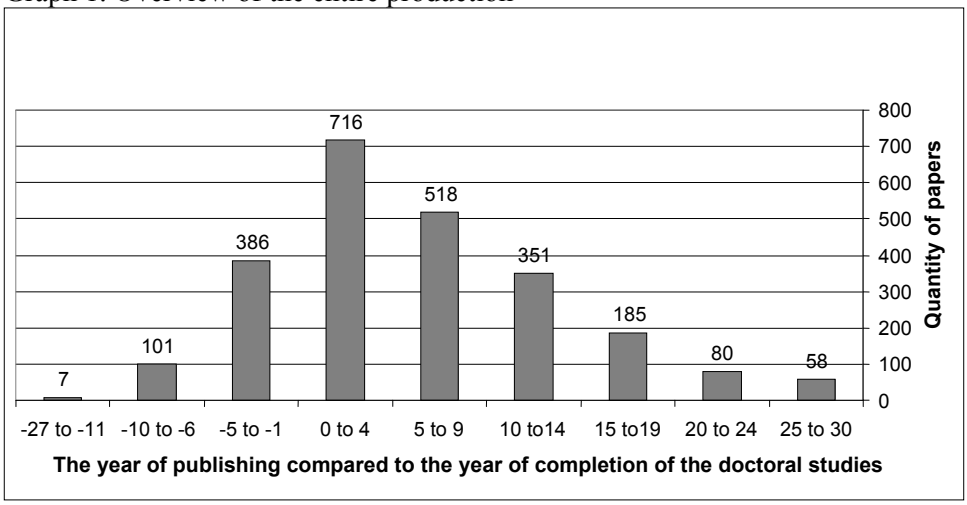
[1] All data on Ph.D. graduates in Information Science taken from Đ. Pečarić (2009).

covered other areas such as for example: Information Sciences and Economics) were labeled as Information Sciences while papers addressing different areas were labeled as Other. A number of papers dealt with non-scientifically definable areas and were therefore labeled as Unknown/Unclassified.

## Presentation of the frequency of the scientific production

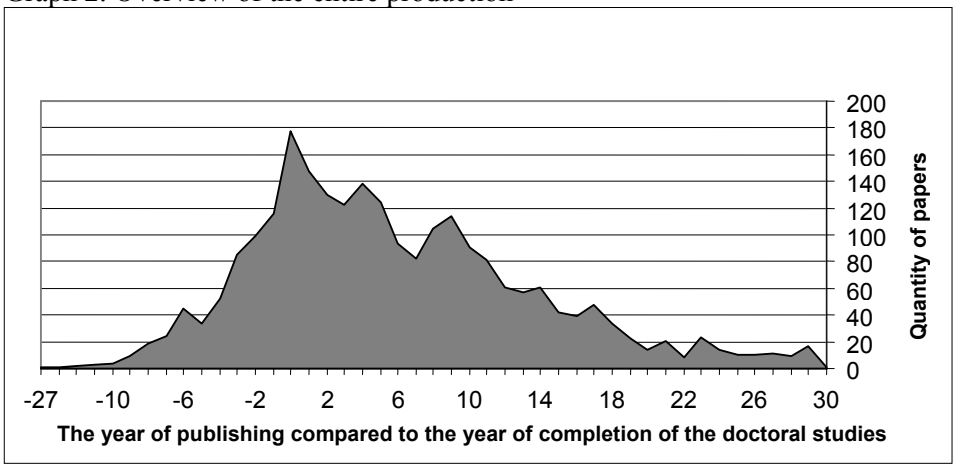
Having analyzed the data, it was possible to use graphs to present the findings. Graphs 1 and 2 show the entire scientific production in relation to completion of doctoral studies.

Graph 1. Overview of the entire production




The graphs indicate that the "first" paper was written 27 years prior to completing the doctoral study, the "last" one dating 30 years after the completion of the doctoral study. This should not come as a surprise bearing in mind the age of the oldest doctoral candidates 75 and 71 as it is fair to assume that the majority of their work had been done prior to the completion of their doctoral study. On the other hand, the youngest doctoral student completed the study at the age of 28 which suggests that their works are yet to be submitted during the course of the forthcoming career. The ideal option for producing a graph that illustrates the scientific production before and after completion of doctoral studies would be based on data from a reference database to contain a number of doctoral students of the same age and who have completed their studies in the same year. Apart from the above mentioned information, the graphs clearly suggest that the majority of the scientific production was submitted during the year of completing doctoral studies, marked as 0, as many as 178. Another strong indication is of the increase in production closer to the time of completion of studies and the decrease coinciding with the lapse of time.

Graph 2. Overview of the entire production

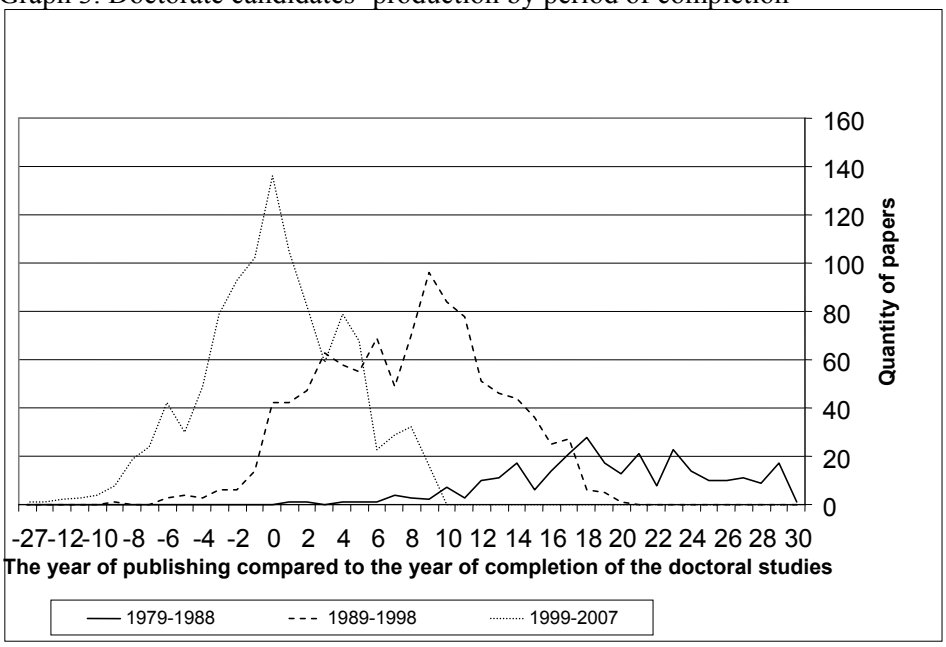



This finding can be explained partly by the limited database. As previously mentioned, the doctoral students completed their studies during the period 1978-2007 (this period is 1979-2007, if only published work is taken into consideration). The papers contained in the database were published between 1980 and 2009. In order to achieve more precise indicators of the scientific production, the database should include all the works produced within the period of 30 years from the time of the completion of studies of the oldest doctoral student and 30 years after the youngest candidate completed the studies.

Since these conditions were impossible to establish the doctorate graduates have been divided into three groups based on the year they completed their doctoral studies (Graph 3), into ten-and nine-year periods. The goal of this division was to demonstrate there was no actual drop in scientific production after the completion of doctorate studies. This graph shows that the reason for the seeming drop in production shown in Graphs 1 and 2 lies in the relatively small sample of published papers. Graph 3 demonstrates an increase in production after the completion of doctorate studies in doctors from the period of 1979 to 1988 (solid line), an increase in production after the completion of doctorate studies in doctors from the period of 1989 to 1998 (dashed line), as well as increased production before completion of doctorate studies in doctorate candidates from the period of 1999 to 2007 (dotted line).

Graph 3 demonstrates that Information Sciences are marked by steady development and that scientific production grows after the completion of doctorate studies. Even better results would have been obtained with a demonstration of production for each individual year; however that was not possible with this research paper for practical reasons.

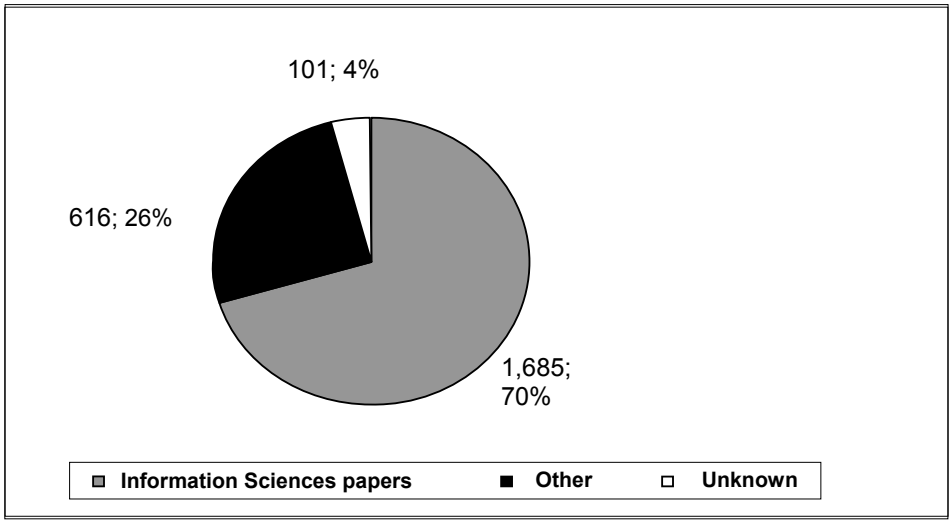
Graph 3. Doctorate candidates' production by period of completion



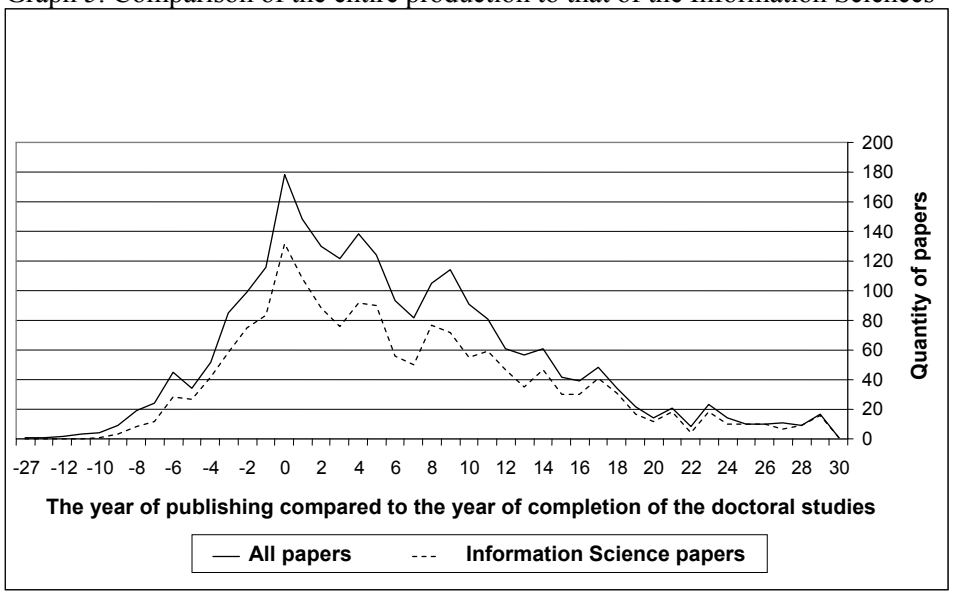

Graphs 4 and 5 illustrate the production of scientific papers in relation to the respective scientific areas. As expected, the Information Sciences doctoral students produced the majority of their works within the specific domain of their research. In total, 1,685 papers were submitted in the area of Information Sciences and they account for 70% of all the works. Graph 5 tells us that ratio between production in information sciences and other science is constant. In other words, scientists and explorers of information sciences are always present, with almost one third of their scientific productivity in at least one scientific area.

Graph 5 compares the entire production to that of the Information Sciences. It suggests the absence of any deviation within periods of publication of the works. Consequently, the most productive year was the year zero (the year of completion of the doctoral study) in which as many as 132 works were written. Those years that saw fewer works being published equally reflected lower difference in production. It is important to draw attention to the fact that the „oldest" work from the field of Information Sciences was published 10 years prior to its author completing the doctoral studies.

Graph 4. Production of scientific papers in relation to the respective scientific areas



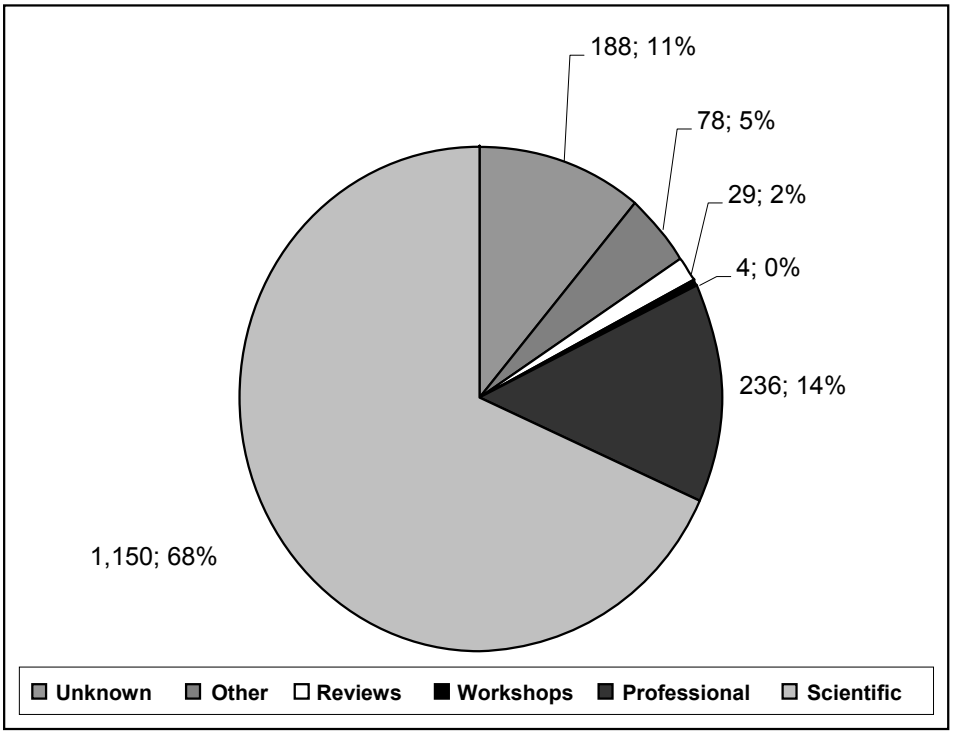

Graph 5. Comparison of the entire production to that of the Information Sciences




The following graphs clearly illustrate the relation of the works contrasted against the categories of the Information Sciences. Graph 6 shows the overview of the distribution of all categories where scientific papers are represented with 68% or 1,150 papers, well above the others. The second most represented works

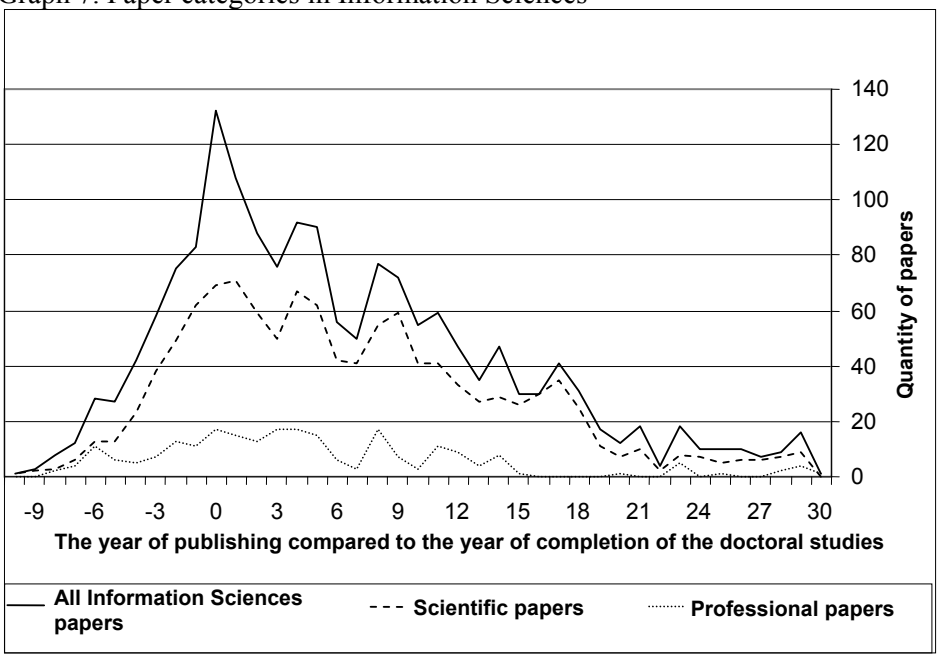
are expert papers with 14% or 236 papers. There were only 29 reviews or 2%, whereas there were an insufficient number of workshops to merit the comparison, only 4 papers. It is interesting to observe that as many as 11% of all papers have no qualification of category, which is a sizeable omission for a database.

Graph 6. Paper categories in Information Sciences




Graph 7 represents the comparison of the Information Sciences works with scientific and professional works that were available in sufficient numbers. It clearly documents that the works follow the trend of the whole production with the Information Sciences and reiterates that the scientific works account for the majority of the scientific production as well as the year zero being the most productive.

Graph 7. Paper categories in Information Sciences




## Discussion

Based on the data analysis of the scientific production of doctoral students in Information Sciences in the period between 1978 and 2007 several conclusions have been made despite the misgiving that the data for the initial ten-year period was incomplete as a result of the Croatian Scientific Bibliography having been established as late as 1997.

The database registers the scientific production of 107 (79.8%) of 134 doctors of Information Sciences. It is necessary to point out that these are doctors of Information Sciences who have graduated in Croatia and not all doctors of Information Sciences active in Croatia. The question is where are the “missing” 20% of the doctors of Information Sciences. Are they lost to scientific work? In other words have they relinquished scientific activity altogether or only their activity in Croatia?

Another important realization is that scientific production is relatively large (22 scientific papers per doctor). The data for individual years shows that every doctoral candidate on average published 1.66 papers in the year they completed their doctoral studies, 1.38 in the first year thereafter, 1.21 in the second year, 1.14 in the third, 1.29 in the fourth and 1.16 scientific papers in the fifth year after having completed doctoral studies.

There is an evident difference in scientific production before the completion of doctoral studies between particular scientific groups. Those who completed their doctoral studies in the first ten-year period (1978-1988) show a longer sci-

entific productivity before attaining their doctorate. The reason for this is found in the fact that Information Sciences have been present in Croatia since the 1960ies, but it was not possible to achieve a doctorate in that discipline until a lot later. It logically follows that scientific production before the completion of doctorate studies would be greater for that particular group in that period than it is for the group in the latest period observed. Doctors from the latest period enter the field of Information Sciences as junior researchers, normally immediately upon completing their studies.

The third significant factor of the scientific production is a relatively large number of original scientific papers (68%) and a relatively small number of professional papers (only 14%) and merely 2% of reviews.

This data leads to the question of character and goals of scientific research activity in Croatia. The small number of professional papers points to the fact that scientists are rarely involved in applied science and developmental research. This leads to conclude that Information Sciences in Croatia are not focused on applied and developmental research, but on fundamental research. The question is whether this kind of research is at all useful. Another way to look at the data is to wonder whether the number of original scientific papers, which takes up two thirds of all scientific production, is, perhaps, overrated. In other words, perhaps the classification of original papers is not realistic.

## Conclusion

The production of scientific works is one of the most significant indicators of the development of any scientific area. The same applies to the Information Sciences. The aim of this paper was to establish whether Information Sciences follow the positive trend and if it is possible to present and predict the frequency of the scientific production. The given results indicate that the aim has been achieved. Although the works used for this research were published over a relatively short period (as mentioned before), the collected data suggest the advent of the Information Sciences. The same data can be used for the future research which would be most valuable and welcome as it could echo the reality and contrast it against the optimistic forecast.

## References

Croatian Scientific Bibliography. http://bib.irb.hr/ (Access from April 1, 2009 to May 4, 2009)

Glänzel, W. Bibliometrics as a research field. A course of Theory and application of bibliometric indicators. Course handouts. 2003.

Pečarić, Đilda. Razvoj informacijske znanosti u Hrvatskoj. Bibliometrijska analiza doktorskih disertacija iz informacijskih znanosti 1978.-2007. Zagreb: Filozofski fakultet, doctoral dissertation, manuscript, 2009.

Original scientific paper

# Telling the Future of Information Sciences: Co-Word Analysis of Keywords in Scientific Literature Produced at the Department of Information Sciences in Zagreb

Siniša Bosanac, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
sbosanac@ffzg.hr

Marija Matešić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
mmatesi1@ffzg.hr

Nino Tolić, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
nitolic@ffzg.hr

**Summary**

*Research fields dealing with information from various perspectives have been rapidly developing throughout the last few decades. Information science is one of the most prominent among them. The aim of this article was to investigate how concepts related to information sciences in Croatia change over time, and by doing this to show the development of the field. For this purpose, co-word analysis has been used. Using this method, the most important concepts of information sciences that appeared in the 14-year period 1995-2009 were indentified, and the results have been visualized. The concepts are visualized in the form of a network along with their respective clusters for the whole time frame, and also separately for two seven-year periods, 1995-2002 and 2002-2009. The analysis has shown concepts centred on education and community to be the most prominent and stabile. New concepts in the network appear independently, as a replacement for similar concepts, or as a result of braking down of general concepts into more specific ones. Results presented in this paper are of purely quantitative nature and, if combined with observations of relevant external*

*factors, can serve as a basis for a study of trends in scientific production, and ultimately, their prediction.*

**Key words:** co-word analysis, clustering, data visualization, information science, key words, scientific literature

## Introduction

Predictions are a key factor in every decision-making process. In every aspect of life we operate using predictions and in most cases we are not even aware of it. Short-term predictions are more accurate, but long-term ones enable us to make decisions that lead to much greater gains. All professional long-term predictions are based on previous trends and regularities. In order to recognize trends and regularities, we need detailed and precise data on previous development in the area on which we focus. The time-span and the quality of data determine the quality of predictions.

Human action is arguably the most difficult phenomenon to predict. In science, things are made somewhat easier by the fact that science is a very structured activity which records even the smallest steps it makes in the form of scientific literature. The analysis of these records is done using bibliometric methods. It is only natural that those methods are used to describe the field in which they originated – information sciences.

One of the key characteristics of the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb is its interdisciplinary area of activities and increased dynamic of its development, which is also typical for other academic organizations in this field. Being a scientific and educational organization, the Department has two kinds of output. The first kind, the *scientific output*, is principally measured by the number of completed research projects and produced scientific literature. The second kind*, educational output*, can be measured by the type and number of available courses and the number of its graduated students and their specializations.

In this paper, we have established a framework for representing the development of its *scientific output*. For this purpose we have used co-word analysis, an established technique for mapping the structure and dynamics of science[1], to analyze keywords in scientific papers produced by members of the Department in the 1995 – May 2009 period catalogued in the Croatian Scientific Bibliography.

The aim of this article was to describe the methods used and to provide quantitative results. In order to interpret them and to draw qualitative conclusions about the development of information sciences, it would be necessary to take into account various other factors. One of those factors is the question whether keywords provided by authors accurately represent the actual topic of a scien-

[1] Qin He. Knowledge Discovery Through Co-Word Analysis. // *Library Trends*. Vol. 48 (1999), No. 1; pp. 135-159.

tific paper's content.[2] We have not dealt extensively with this issue in our article, and for its purposes, description using author-provided keywords is taken to be accurate.

## Method

Co-word analysis is a bibliometric technique that examines co-occurrence of keywords. (Glänzel, 2004) Words are the most important in this analysis, and can be extracted from various types of scientific publications. They can be mined from titles, abstracts, full texts or keyword lists of various types of scientific publications. The main purpose of this technique is to show the dynamics of scientific field's development by visually representing the co-occurrence matrix of words chosen according to their frequency in the corpus. Higher co-occurrence frequency of two keywords indicates closer and stronger links between them. The closer links between two keywords represent closer relationships between the concepts they refer to.

### Data harvesting

A total of 376 articles authored by 35 members of the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb were harvested[3] from the Croatian Scientific Bibliography (CROSBI)[4] from 1995 to the present. This period was selected because information on publications dating earlier than 1995 were unavailable at the time the data harvesting was conducted. Publications were selected according to type of publication, field of science, and author. Types of publications that were included in the harvesting process were primarily scientific articles. Books, book abstracts, book chapters, conference reports, unpublished papers, or graduation thesis were not included.
The publications' title, keywords and year of publishing were harvested according to the chosen 35 authors that are members of the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Zagreb. That was chosen harvesting filter to include information science related publications. Publications that did not have English keywords listed were excluded from the analysis. Keywords were not standardized because a thesaurus was not available, so there is a possibility of inconsistencies with standard terminology of information sciences.

[2] For a discussion on this problem see Whittaker et al. Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis. 1989.

[3] Harvesting process was automated by using CURL module in manually written PHP script.

[4] The website of Croatian Scientific Bibliography (CROSBI) can be found at http://bib.irb.hr/

**Data processing**

A total of 689 unique keywords from 1136 tokens (keyword forms) were harvested out of 367 articles covering the 1995-2009 period. Keywords unmistakably denoting the same concept, or occurring in different forms were standardized through the process of normalization and lemmatization. As a part of lemmatization process all inflected forms were reverted to their base form, except when changing the form would change the meaning of the whole keyword. During normalization, abbreviations were converted into their full form, e.g., CAL + Computer assisted learning = Computer assisted learning; CALL + Computer Assisted Language Learning = CALL; EU + European Union = European Union; HMM + Hidden Markov Model = Hidden Markov Model; IT + Information Technology = Information Technology; LIS + Library and information science = Library and information science; LMS + Learning Management System = Learning Management System.

After lemmatization and normalization data were converted to a data format supported by Bibexcel[5] – freeware software for bibliometric analysis, and co-word analysis in particular. Within Bibexcel as a tool, word frequency is calculated. Finally, words with frequency more than two were selected for the next step - co-word analysis. A co-occurrence matrix was formed that shows relationships between phrases or words.

To provide a very clear view what is happening in co-occurrence matrix a visualization tool called Pajek[6] was used to map co-occurrence data. Pajek is an open source program for large network decomposition, visualization and clustering. Co-occurrence data were visualized using Kamada & Kawai algorithm as it available in Pajek. This draws general graphs with minimal energy[7]. In order to keep visualization readable, the analysis was limited to words that co-occurred with frequency more than two.

## Results

In order to show the development of the observed scientific field it is necessary to show how the results changed over time. To facilitate this, the results were divided into three parts. The first part shows the results for the whole period 1995-2009. The other two show results for two seven-year periods, 1995-2002 and 2002-2009. The analysis was done independently for each period.

[5] Bibexcel is publicly available at: http://www.umu.se/inforsk

[6] The homepage of Pajek can be found at http://pajek.imfm.si/doku.php

[7] Wikipedia Contributors. Force-based algorithms. Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Force-based_algorithms (Jul 24, 2009), Aug 16, 2009.

**1995 - 2009**

A total number of 376 scientific articles by 35 authors were collected for the whole period. These articles contain 689 unique keywords. Finally, a data matrix of 64 co-occurring words was created and visualized using Pajek.

Visualization of co-occurring words for whole period shows the existence of five main clusters gathered around strongest connected nodes which are *museology, Croatian language, scientific communication, information literacy and education* as it is presented in Figure1. Cluster *museology* includes topics relating to users, museums and repositories. Cluster *Croatian language* is related to natural language processing concepts and topics as the field of artificial intelligence. *Information literacy* as a core cluster contains knowledge and library, which are public administration oriented concepts. Cluster *scientific communication* includes topics on scientific activities. Cluster *education* includes concepts related to networked society, which is a result of the impact of the Internet and information technology. An isolated cluster which contains keywords *computer-assisted language learning*, *web application* and *Croatian old dictionary portal* can also be observed. This cluster is not related to any of the given clusters.

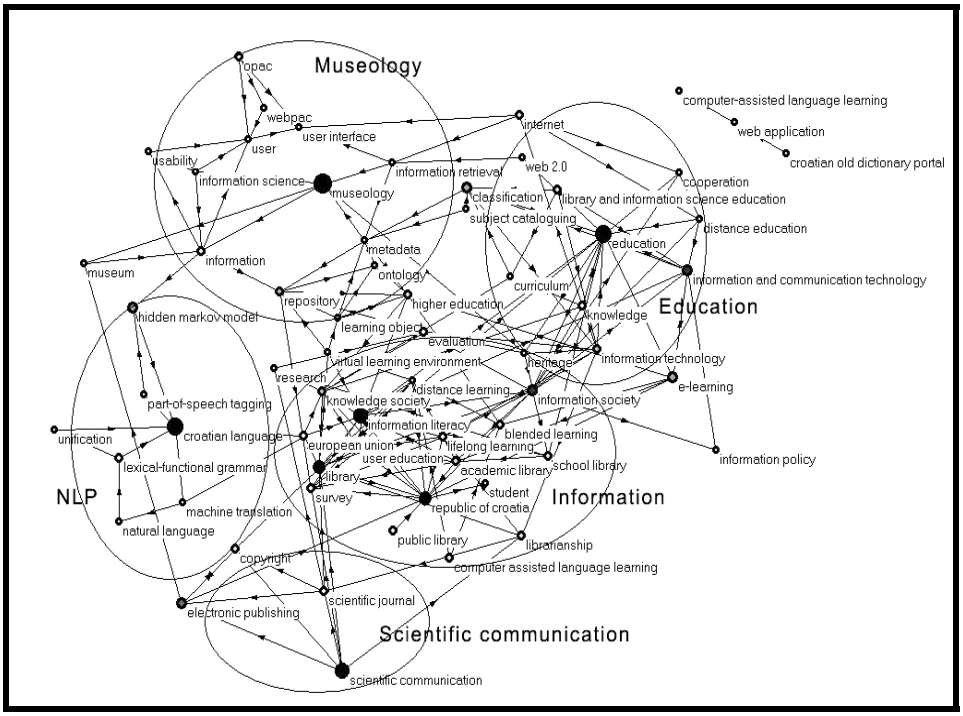



Figure 1: Visualization of 64 words with frequency greater than 2 for 376 information science related articles for the whole period (1995-2009)

**1995 – 2002**

In the first time period from 1995 to 2002 there is a total of 35 items shown in the diagram (Figure 2). The three largest nodes are as follows: *education, information communication technology* and *library*. As shown on Figure 2, five larger clusters are visible: *information technology, education, NLP, community* and *Interfaces.* Cluster *information technology* contains nodes which are linked to technology itself, as well as possible applications of technology, such as *classification* and *computer assisted language learning.*

Cluster *community* contains nodes related to general public such as *library, knowledge society, library users* and the like. This cluster is worth examining in greater detail, due to the possible relevancy to the time frame in question.

Cluster *education* contains nodes related to education itself and its evaluation such as *comparison* and *characteristics.* Close relationship this cluster shares with cluster *information technology* could be worth examining further.

Cluster *interfaces* contain nodes related to information search and retrieval such as *OPAC* and *WEBpac*. Isolated cluster *NLP* contains nodes related to natural language processing of the Croatian language.

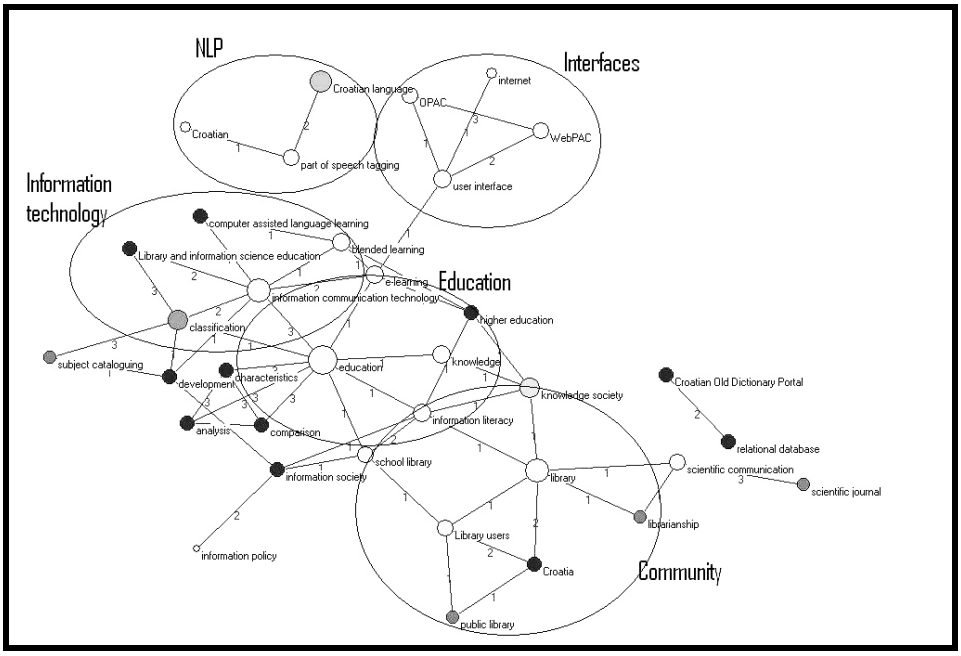



Figure 2: Visualization of 35 words with frequency greater than 2 for 187 information science related articles for the time period between 1995 and 2002

**2002 – 2009**

In the second time period from 2002 to May 2009 there is a total of 37 items shown in the diagram (See Figure 2). The five most interconnected nodes are the following: *education*, *repository*, *information literacy, higher education*, and *information and communication technology*.

Visualization of co-occurring words for time period between 2002 and 2009 contains four clusters: *Education*, *Repository*, *Web 2.0*, and *Community*. The largest is the one centered on *education*. Its strongest connections are with nodes that could be roughly described as analytical in meaning: *analysis*, *comparison,* and *characteristic.* The rest of nodes in the cluster could be described as web-related: *content management system*, *web 2.0*, *semantic web*. Links with strength 1 are those with *information and communication technology*, *e-learning*, *information literacy*, *school library*, and *knowledge*.

Second largest is the cluster with *repository* as its central node. Its members could be described as mostly maintenance-related; *technical support*, *valorization*, *evaluation*, *standardization*, and there is also *digital educational material*, which is slightly different by its meaning. The weaker links with its central member are with *information*, *metadata*, *ontology*, *higher education*, and *survey*.

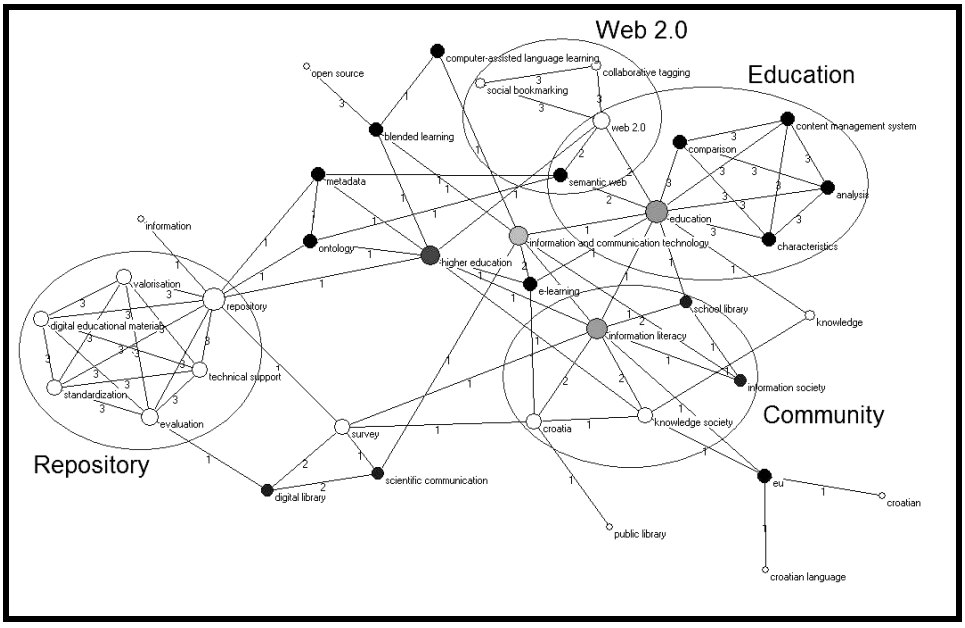



Figure 3: Visualization of 37 words with frequency greater than 2 for 189 information science related articles for time period between 2002 and 2009

The third cluster is that centered on *web 2.0*. Its member-nodes are typical subordinate concepts of web 2.0: *social bookmarking*, *collaborative tagging* and *semantic web*. There are also links with *education* and *higher education*.
The fourth primary cluster, *Community* is the one centered on *information literacy*. Its members are *school library*, *knowledge society*, and *Croatia*, with links to *EU*, *information society*, *knowledge society*, *survey*, *higher education*, *education* and *information and communication technology*.
There are also several secondary clusters: *digital library - scientific communication - survey, blended learning – open source - computer-assisted language learning.* Links between *education - knowledge* and *knowledge society* could also be very interesting to examine, same as those between *EU – Croatian – Croatian language – knowledge society,* and *information literacy*.

**Time period comparison**
It is important to note that these represent only the most frequent concepts in a relatively broad time period.
The most obvious difference between the two time periods is the number of clusters (first period 5, second 4), and the fact that the later time period does not have isolated clusters and nodes.
The *Education* cluster is the most prominent in both time periods. Its main members and links have remained mostly unchanged – particularly members belonging to group of analytical concepts, and links with *information and communication technology*, *knowledge, knowledge society, school library,* and *e-learning*. The difference in this cluster is that it lost its links with *classification*-related nodes, and formed several strong ones with *Web 2.0* cluster.
Although the nodes from the larger part of *Information technology* cluster from the first period are also present in the second, it was not designated as an independent cluster in the second period because the links between its members were weaker. A significant difference in this group is the disappearance of three prominent nodes with strong connections - *classification*, *subject cataloguing*, and *library and information science education*.
Cluster *Community* has kept a significant number of its nodes, namely, *knowledge society*, *school library*, *Croatia*, and *information literacy*. The cluster in the second period is centered on *information literacy*, which is much more pronounced than in the first period. Other changes were the reduction of library-related concepts, and the emergence of *European Union* as a concept interconnected with cluster *Community*.
*NLP* cluster, which was isolated in the first period, dissolved in the second period with the disappearance of its central member – *part of speech tagging*, while its remaining members formed links with *European Union* node.
*Interfaces* cluster also dissolved with all of its members disappearing. *Internet* node, which had strong connections, also disappeared, but it was replaced by an entire cluster named *Web 2.0*, whose members can actually be considered as

subordinate concepts of the *Internet*. This would mean that the node *Internet*, actually multiplied and became more specific.
Cluster *Repository*, which appeared in the second time period, consists of newly formed nodes, but also has connections with some of the "old" ones, such as *scientific communication* and *higher education*.
Some independent nodes denoting more specific concepts, such as *information policy*, *scientific journal,* and *Croatian Old Dictionary Portal*, are not present in the second time period, while new ones, such as *open source*, *metadata,* and *digital library,* have appeared. More general concepts, e.g. *knowledge*, and those denoting scientific methodology, such as *survey*, *analysis*, *comparison*, and *characteristics*, also did not change. From this we can conclude that specific concepts are more dynamic than the general ones, and those describing methodology.

## Conclusion

The goal of this article was to investigate the main topics and trends within the field of information science during the time period from 1995 to 2009. Using co-word analysis and by comparing the two time periods, first from 1995 to 2002, second from 2002 to 2009, along with the overarching period, we have endeavored to present our findings through quantitative analysis. Several interesting topic shifts were uncovered, all meriting further qualitative and quantitative research.
In order to improve upon our research we propose several modifications. For a more exact analysis, it would be necessary to include more publications, to display the results in more segmented and narrower time spans as to increase the resolution of graphical representations. A guideline for author-added keywords that would prescribe the classification of keywords according to a hierarchy of concepts, and whether they describe the method, or the actual topic of the article, would greatly improve not only the quality of their retrieval, but also the precision of similar analyses. Also, it should be noted that improvements in the storage, classification and general usability policies on CROSBI servers are in order.
With these modifications the research would be more precise and perhaps, would uncover more. Hopefully, the topic in question will be revisited on a broader scale than was possible in this article.

## References

Ding Ying, Matthew; Gobinda G. Chowdchury, Foo, Schubert. Bibliometric cartography of information retrieval research by using co-word analysis. 2001. http://www3.ntu.edu.sg/home/assfoo/publications/2000/00ipm_fmt.pdf (Jun 24, 2009)

Glänzel, W. Bibliometrics as research field. 2004. http://www.norslis.net/2004/Bib_Module_KUL.pdf (May 10, 2009)

Jokić, Maja. Bibliometrijski aspekti vrednovanja znanstvenoga rada. Zagreb: Sveučilišna knjižara, 2005

Leydesdorff, Loet. The university-industry knowledge relationship: Analyzing patents and the (May 10, 2009)

Qin He. Knowledge Discovery Through Co-Word Analysis. // *Library Trends*. Vol. 48 (1999), No. 1; pp. 135-159

Whittaker, John; Courtial, Jean-Pierre; Law, John. Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis. 1989. http://www.jstor.org/stable/285083 (May 13, 2009)

Wikipedia Contributors. Force-based algorithms. Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Force-based_algorithms (Jul 24, 2009), Aug 16, 2009.

Professional paper

# Scientific Publication Productivity of the Information Sciences' Doctors in the Republic of Croatia

Lucija Šćirek, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
lscirek@ffzg.hr

Ines Novosel, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
inovosel@ffzg.hr

Matija Latin, student
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
malatin@ffzg.hr

**Summary**

*In this paper the authors conducted an analysis on the Information Sciences' doctors' productivity. As a source, the Croatian Scientific Bibliography (CROSBI) database was used, in which 106 of 134 doctors, who got their Doctoral Degrees from Croatian Universities in the period between 1978 and 2007, were listed. The research included published papers mostly, articles, lectures, posters and chapters in books, which amounted to 2,406 papers. The analysis of scientific productivity showed that 21% of Information Sciences' Doctors are not scientifically active. On the other hand, Information Sciences' Doctors are active in several scientific fields, with a 70% share in the field of Information Sciences and 30% share in other fields. Scientific papers were categorized by type, form and scientific field, put into tables and presented both numerically and graphically, giving us a complete review of Information Sciences' Doctors' publication productivity. This is the first paper that analyzes research activities of one part of the scientific community, that is, published researches of Information Sciences' Doctors that are financed by the Ministry of Science of the Republic of Croatia.*

**Key words:** Information Sciences, doctors, publication, productivity, analysis

## Introduction

This paper is an analysis of scientific productivity of authors with PhD degrees in Information Sciences. As a source, the Croatian Scientific Bibliography (CROSBI) database was used, in which 106 of 134 doctors, who got their Doctoral Degrees from Croatian Universities in the period between 1978 and 2007, were listed.[1] Data analysis began in April 2009 and was planned to end by the end of May. Upon the completion of analysis, a database with scientific papers written by Information Sciences' Doctors was done. The data was categorized in tables and presented both numerically and graphically using Microsoft Office Excel and Access tools for a better overview. With this paper the authors want to analyse the scientific productivity of Information Sciences' Doctors by paper type and category, and show the paper ratio between the Information Sciences field and other fields. The analysis is referred to paper type: mostly books, journals and conference proceedings, and categories: scientific or professional papers, or review articles. The CROSBI database contains 241,720 papers from various fields, and there are 38,834 papers within the Social Sciences field. This is the first paper that analyzes research activities of one part of the scientific community, that is, published researches of authors with PhD degrees in Information Sciences that are financed by the Ministry of Science of the Republic of Croatia.

## Starting point for analysis

The doctors' papers were extracted from the Croatian Scientific Bibliography (CROSBI) database, which currently contains data for 190,000 scientific and professional papers and 4,000 *in extenso* papers from Croatian authors, and 243,802 papers altogether. The database, which has paper input possibility and web-interfaced searching, was made as an electronic bibliography by establishing adequate technological and communication conditions with joint initiative of computer and information specialists of the Ministry of Science of the Republic of Croatia and the Ruđer Bošković Institute library. The database is updated by the authors themselves, while the librarians, computer and information specialists secure the forms, standards and monitoring of the whole process. To ensure the information credibility, from 01.03.2007 new papers can be added only by using the AAI electronic identity. This AAI identity is owned by every scientist, teacher or a student, along with every employee of the Ministry of Science of the Republic of Croatia. The database provides guidelines for paper categories classification, determines that the scientific paper contains unpublished results of an original scientific research, and the scientific information is presented in a way that enables verification of the analysis and the data on

[1] The data about Information Sciences' Doctors, who defended their dissertations on Croatian universities from 1978 to 2007, were taken from Đilda Pečarić's dissertation (2009).

which it is based. Professional papers contain already known, published results of scientific researches and focus on the practical use of data or their spreading in educational purposes. Professional papers also contain useful information in fields which are not related with the author's original research, and presented observations are not necessarily news in the given field. They must be written in a systematic and comprehensive manner, in accordance with reader's profile. Review article is a scientific paper which contains an original, concise and critical view of a field or a part of a field, in which the author is actively participating. The author's direct contribution to the field, regarding the already published papers, must be accentuated. It can be written by one or a group of authors, and it is usually written at the request of an editor. The database discerns the difference and the definition of grouping into books or conference proceedings. In specific situations, when the editor (or more editors) undertakes a complex task of collecting and processing materials from a conference, he lists the conference proceeding as a book, and himself as the editor. Individual conference presentations are added as "Conference Papers", and never as "Book chapters". The difference between a paper in the proceedings and in a magazine is the quality and extent of the paper[2]. Data about the researchers, projects and institutions are updated quarterly, respectively when the Ministry of Science of the Republic of Croatia publishes that data.

## Methodology and sample

The analysis of Information Sciences' Doctors' productivity was conducted according to the list of 134 Doctors, who got their Doctoral Degrees in the period from 1978 to 2007. The sample was comprised of 53 women and 81 men of all ages. After receiving the list of Doctors, the database was being built from mid-April to mid-May and it would have ultimately contained all papers of all Information Sciences' Doctors, who were found in the used database (CROSBI). The database was searched using a basic author search, and each given result, except papers in the publishing process and paper abstracts, was imported into a database which was built in Microsoft Office Excel, and later transferred to Microsoft Office Access for easier processing of the collected data. The papers were processed in several categories, out of which the most important ones are: paper title, bibliographical unit, paper type, category (scientific, professional or other) and the scientific field to which the paper belongs. After paper categorization, the mentioned Microsoft Office tools were used to extract the information which was considered relevant for the research. Those were primarily the percentages of papers per categories, paper types, scientific fields and such. Results are contained in tables and presented with graphs for a better overview.

[2] Source of categorization were taken from CROSBI database site http://bib.irb.hr/faq

While processing the CROSBI database, the authors often encountered the problem of incomplete data, i.e. for some bibliographical units some data was not listed. It was mostly data on paper category, or the scientific field of the paper. In such cases the authors marked these categories as "unknown".

## Results

From 134 authors with a PhD in Information Sciences, 106 of them are scientifically active. During the CROSBI research, 2,406 papers were registered. They included various articles, comments, conference reports, popular articles, professional papers, studies, book chapters. The research showed that 21% of authors are not scientifically active. On the other hand, Information Sciences' Doctors are active in several scientific fields, with 70% share in the Information Sciences, and 30% share in other fields. The research was divided by paper type, and showed that the most of papers are scientific papers, respectively 69% or 1,677 papers, followed by professional papers which have a total of 14% or 334 papers, and 46 review articles, which make 2% of all paper types. 9% or 214 papers were listed as "unknown", and 5% or 135 papers were listed as "other" (Chart 1).

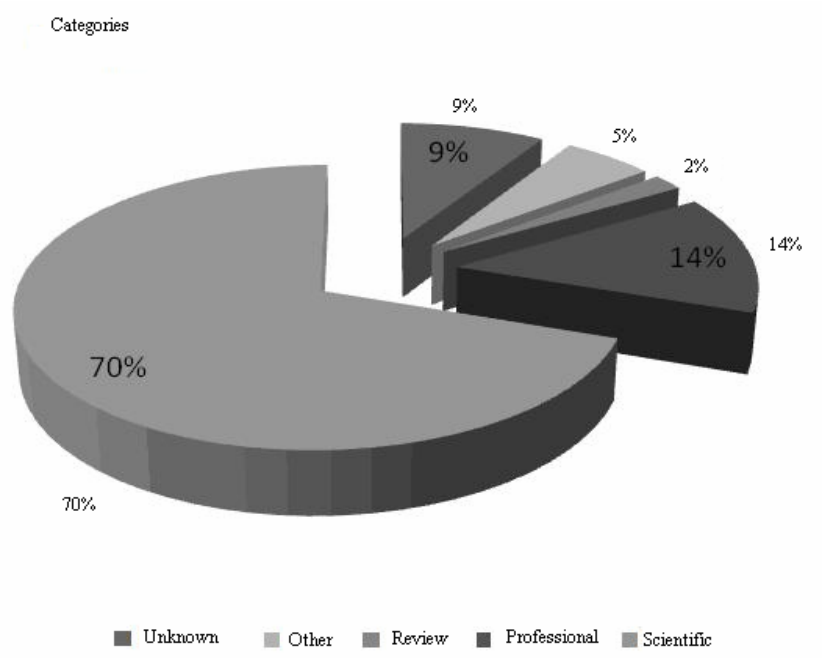



Chart 1.

From all 2,406 papers, 58% are listed as conference proceedings, 36% as magazines and 6% as books (Chart 2).

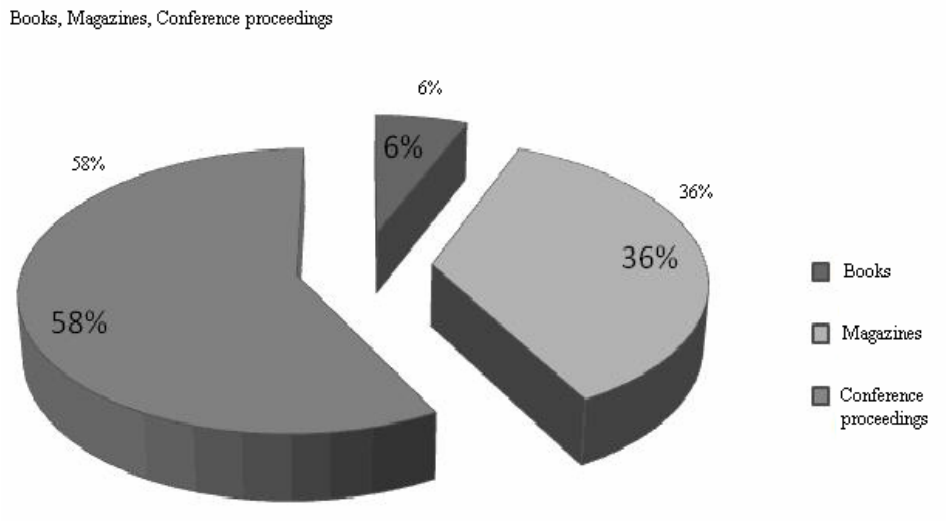



Chart 2.

Although most of the papers in our database are from the Information Sciences field, there are also 179 papers from the Economics field, 88 papers from the Graphic Technologies field and 82 papers from the Computer Science field. In the total are 101 papers listed as "unknown field" (Chart 3).

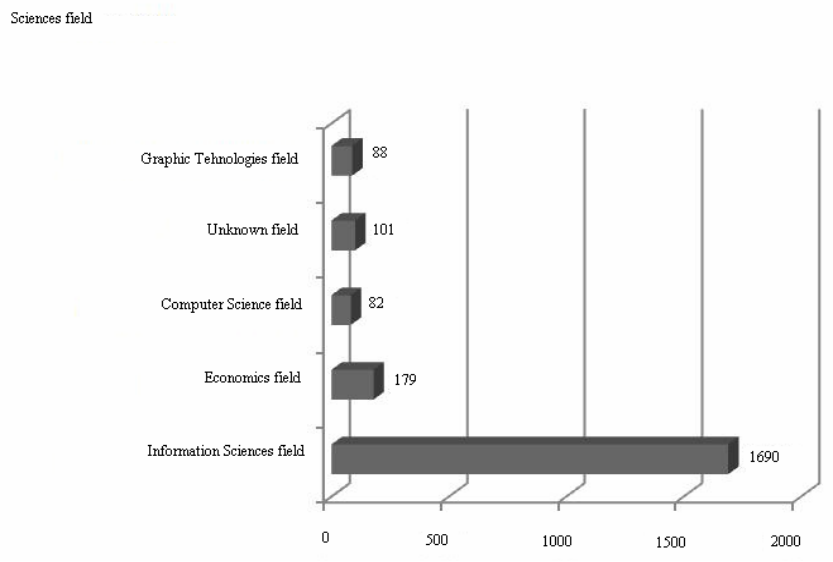



Chart 3.

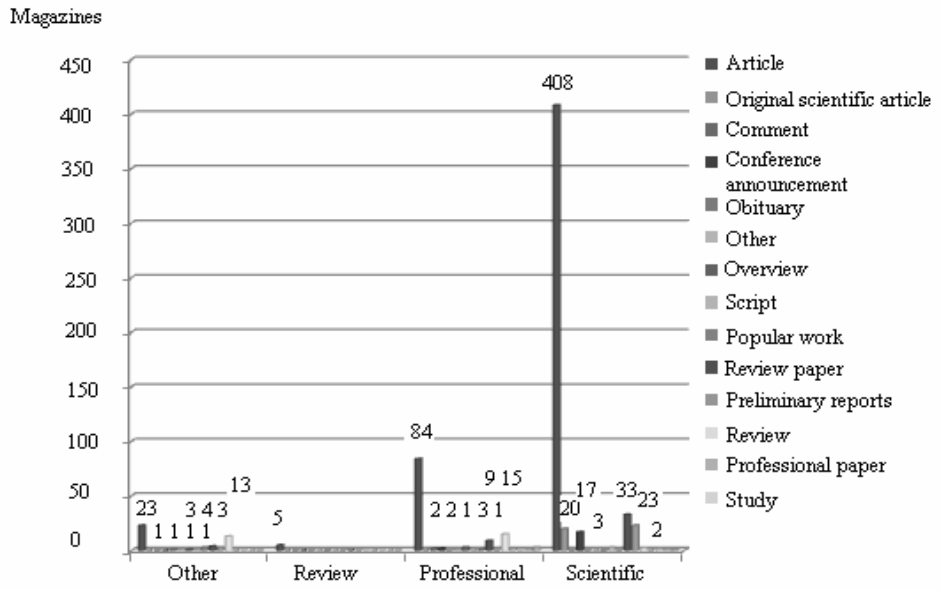



Chart 4.

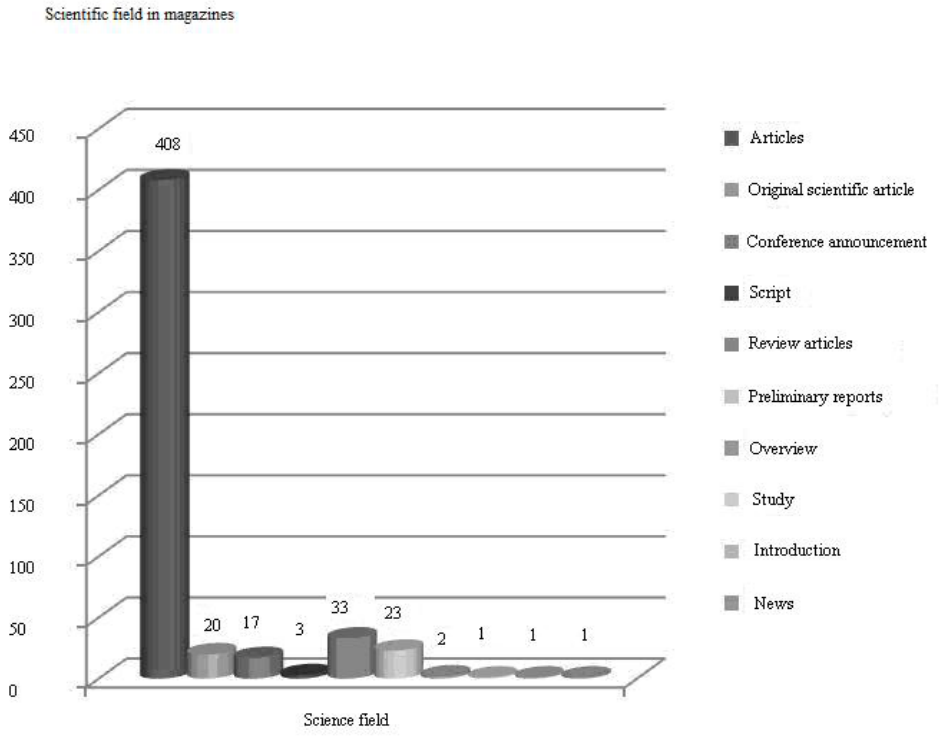



Chart 5.

The papers within magazines are divided in categories: scientific, professional, review and other. There are 509 papers categorized as scientific, 123 categorized as professional, 5 categorized as review, and 50 categorized as "other", which add up to 687 papers in magazines altogether. This shows that the majority of papers are in the category of scientific papers (Chart 4).
Out of 509 scientific papers in magazines, most of them, i.e. 408, are articles, 33 papers are review articles and 23 papers are preliminary reports (Chart 5).
The professional field contains 123 papers, with 84 articles as the majority, 15 overviews and 9 review articles. (Chart 6.)

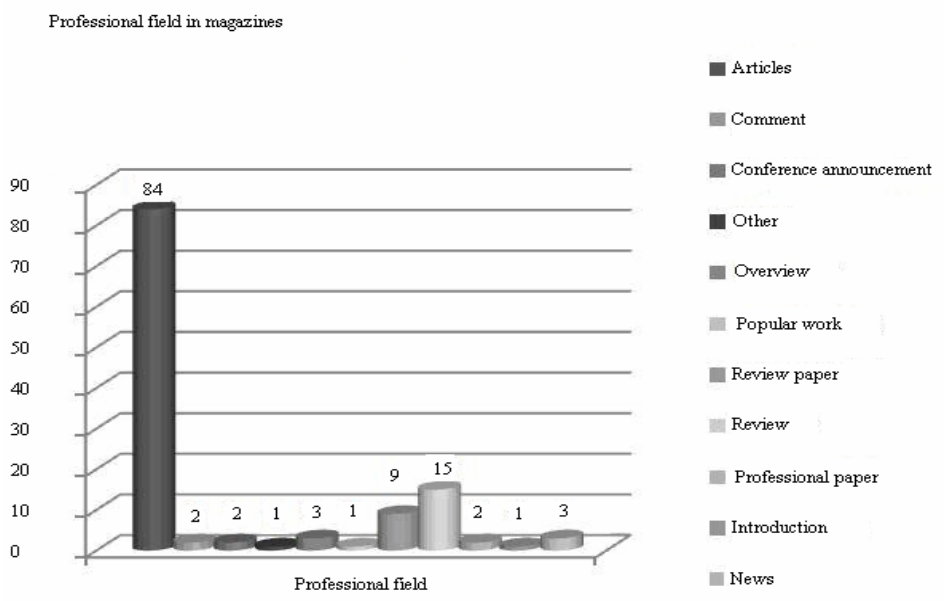



Chart 6.

The conference proceedings are also divided into categories. There are 1,105 papers altogether, divided into 867 scientific, 164 professional and 20 review papers. Again, the scientific category makes the majority of all categories (Chart 7).
Within the conference proceedings there are 867 scientific papers, and 164 professional papers. In both categories there are mostly papers in the form of lectures. There are 780 lectures in scientific category, and 149 lectures in professional category, followed by posters and invited lectures in both categories. The scientific category contains 41 posters, and the professional category contains 5 posters. Finally, there are 30 invited lectures in the scientific, and 4 invited lectures in the professional category (Chart 8).

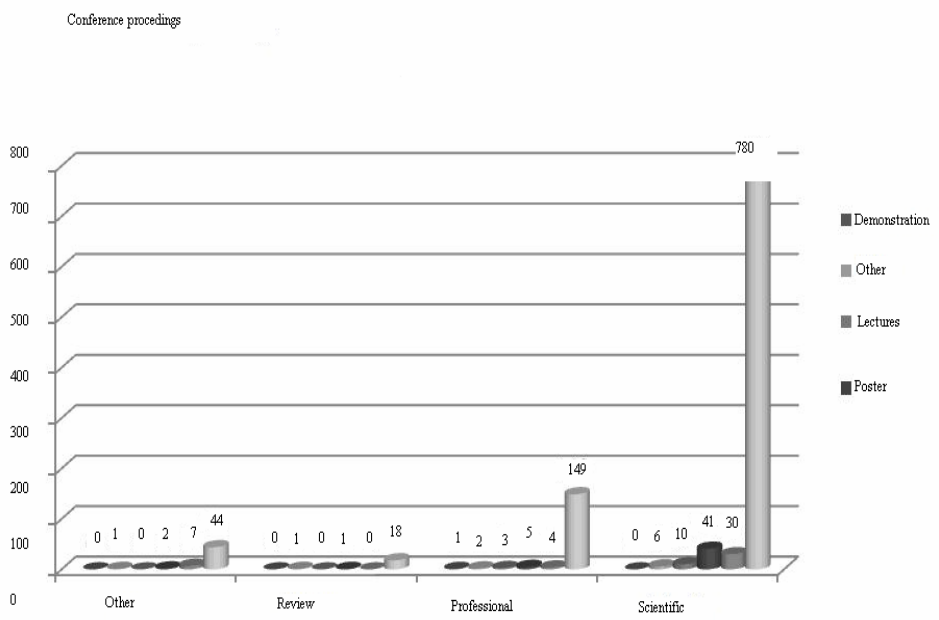



Chart 7.

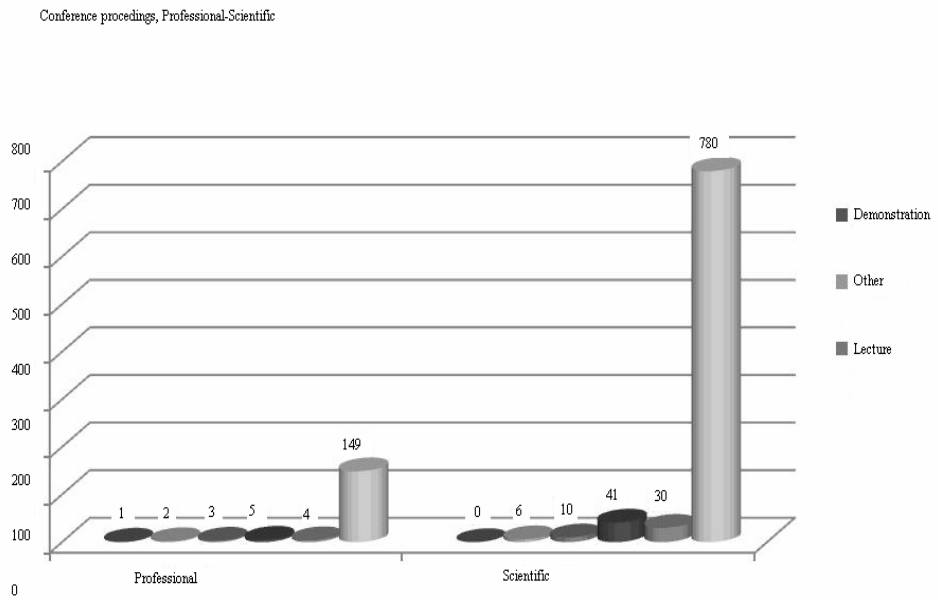



Chart 8.

## Conclusion

The analysis of scientific productivity showed that 21% of Information Sciences' Doctors are not scientifically active and are not listed in Croatian Scientific Bibliography (CROSBI) database. Papers are categorised by type, science field and form. The analysis was conducted by paper type and showed that the most of papers are scientific papers, respectively 69% or 1,677 papers, followed by professional papers which have a total of 14% or 334 papers, and 46 review articles, which make 2% of all paper types. 5% or 135 papers were listed as "other" and 9% or 214 papers were listed as "unknown". Analysis by science field shows that Information Sciences' Doctors are active in several scientific fields, with a 70% share in the field of Information Sciences and 30% share in other fields, mainly Economics, right after Information Sciences. Authors analysed database by form and found that the majority were conference proceedings (58% or 1,105) then magazines divided into all types of articles (36% or 509) and a minor number were books (6%). This paper shows that the most common field is Information Sciences, by the type that is scientific papers and in form it is conference proceedings with lectures, posters and demonstrations. With this paper authors gained unique database with papers of Information Sciences' Doctors with detailed information's. This is the first paper that analyzes research activities of Information Sciences' Doctors that are registered in a CROSBI. The problem with the database which was used is accuracy and missing data. Doctors have to be registered in the database and they have to submit their papers. This database only contains scientific papers which are made as part of project financed by Ministry of Science of the Republic of Croatia. Maybe this scientific productivity and research results are less accurate because only this database was used. On the other hand, it is the biggest database made by Ministry of Science of the Republic of Croatia and the Ruđer Bošković Institute library. Even though this research has some problems, it shows important side of the scientific community and its productivity.

## References

Pečarić, Đ. (2009) Razvoj informacijske znanosti u Hrvatskoj: Bibliometrijska analiza doktorskih disertacija iz informacijskih znanosti 1978-2007, Dissertation. Zagreb: Faculty of Humanities and Social Sciences

Stojanovski, Jadranka; Batistić, Ivo. 2002. http://bib.irb.hr/faq (05.08.2009)

Original scientific paper

# Ethical Questions in the Work of Hans Jonas in Informatics and Information Science

Đorđe Nadrljanski
University of Philosophy, Split
Viška 8, Split, Croatia
djordje44@yahoo.com

Mila Nadrljanski
Faculty of Maritime Studies
Zrinsko Frankopanska 38, Split, Croatia
milamika60@yahoo.com

Mira Zokić, Ph.D. student
Zvornička 9, Zagreb, Croatia
mira.zokic@gmail.com

**Summary**

*Society in which we live today is on a lesser scale determined by the manufacture of material goods, and increasingly determined by the exchange of information and knowledge, as well as by communication over-networking. Introduction of ethics into informatics proves that modern science should also comply with ethical standards, adjusted to technology, as the grounds of basic human values. The ethics of informatics determines an ethical framework in the procedure of collecting, processing and the use of data, and it is based upon unquestionable ethical premises, as well as on those that have imposed themselves during informatics development: information privacy, openness, safety, availability and justification of their violation. The main goal of this paper is to bring those theses into connection with informatics and handling computers, and then to discuss about possible consequences connected with jobs of informatics experts, that primarily refer to their area of expertise. Ethics is, in general communication, rather loaded or even overloaded by morality, so its practical meaning – or practical meaning of ethical efforts – is often no longer recognizable.*

**Key words:** responsibility, ethics, informatics, information science application of computers.

## Introduction

The modern society, organized according to western cultural frameworks, progressively considers itself to be an informatical society. It is also understood that such a society communicates primarily within electronically organized relations, and that it is, when it comes to relevant issues, communicationally organized via media systems (which always includes the public and public control). Furthermore, it is understood that the social changes are observed via the media. In the same way, those changes are formed and changed by the society. In his book "The principle of responsibility", philosopher Hans Jonas poses a question of "the new technological civilization ethics" (Jonas, 1984). Its necessity is explained by "the new dimensions of responsibility, imposed to people by the existence and the danger of application of modern technics". To irresponsibility in handling the technique, Jonas opposes "the command of caution" which, formulated as "the principle of responsibility", says: "act so that your actions are in line with the requests of pure human life on Earth." (Jonas[1], 1984). Although Jonas does not explicitly mention informatics and computer application, a great deal of his theses refers to that domain, so one could think that informatics is seen as a "technological discipline par excellence". It is proved in some of Jonas's theses. This paper deals with informaticians' ethical problems, which serve as a basis to answer the questions of responsibility. In this short survey of ethical problems related to engineers generally, the following topics will be elaborated: ethical problems which are encountered on a daily basis, problems related to computer system, problems of availability/ shortage of information, privacy and confidentiality of ethical data at a workplace – ethical principles which should be pursued by programming engineers and a modern view on computer morality called the Hacker ethics. This ethics offers an expanded view on a philosophy which might make ethics and its rules irrelevant, since it would be of concern to everybody. The Hacker ethics is new work ethics, which questions Protestant work ethics – an attitude towards work that has been present for so long. Protestant ethics was elaborated in a book by Max Weber, "The Protestant ethics and the essence of capitalism" (1904 – 1905). In short, Protestant work ethics highlights the importance of the feeling of responsibility towards work duty, and accentuates an attitude towards work that has to be liberated from constant calculation on how to earn a salary with maximal comfort and minimal effort. The work has to be done for its own cause, as a life profession. Since computer ethics is a rather new discipline, there are no company rules for every possible situation, and it more or less boils down to the individual and his moral principles. Therefore, the goal of this document is not to impose strict behavior rules, but to indicate a wide range of problems appearing

[1] The Imperative of Responsibility: In Search of Ethics for the Technological Age (trans. of Das Prinzip Verantwortung) trans. Hans Jonas and David Herr (1979). (University of Chicago Press, 1984)

in everyday life. For that reason, there are a lot more real-life examples than concrete instructions on how to behave in a given situation. Since there are often more asked than answered questions, it is advisable to apply an advice from ethical principles: "Ethical disproportion is best solved by careful consideration of basic principles, and not by obeying strict details of rules."

## What does responsibility imply?

The principal basis consists of several theses posed by a German-American philosopher, Hans Jonas, in his book "The principle of responsibility – an attempt of an ethics for technological civilization" (Jonas, 1984). The theses must be brought into connection with informatics and handling computers, and then it must be discussed about possible consequences related to jobs of informatics experts, that primarily refer to their area of expertise.

Before explaining Jonas's theses, his understanding of "responsibility" and "responsible work" must be explained more closely. It can be best explained by quoting the answer of a president of a big concern to the question: what is indispensable to prevent catastrophic acts of technics?: "International security standards, (…) international control (…) and then a necessary faith in God, that the technics will become yet more secure in its further development". By means of this quote we want to explain the fact that we do not have such an idea when it comes to responsible relation towards technics. For us, the question whether technique functions and whether it will be useful or harmful to individuals or humanity as a whole, is not a thing of faith in God, but exclusively an issue of people who, directly or indirectly, participate in its development. The objection here is that responsibility, as something which is a human purely a work, is transferred to an intangible, and therefore a higher, instance that cannot be attacked. In this way, any responsibility could be easily neglected, like the one for "the remaining risk", which is often gladly reduced. Our understanding of responsible action is emphasized in the following quotation: "Only when there are more engineers listening to the voice of their conscience, who take into consideration whether their actions lead to the ordinary or to the divine, to the ugly or to the beautiful, to the good or to the bad – only then can the shadows of destruction be sent away from us" (Jung[2], 1963).Let us turn to different forms of responsibility which we encounter in our work in informatics. We would like to illustrate them in several examples:

1. An associate would like to test his product (compiler) better, but the work must end due to a deadline. Who takes the responsibility?
2. Can we justify cooperation on a data input system which is created to measure the work efficiency of a worker who is in charge of data col-

[2] Carl Gustav Jung was a Swiss psychiatrist, an influential thinker and the founder of analytical psychology known as Jungian psychology.

lection? Or a cooperation on a project which can be proved to leave many people unemployed?

3. Associates, whose beliefs do not allow them to work on military projects, worked on an experimental product of general usage. The client with most interest in the product is a military person. Is he given moral support?
4. An instrument of an expert system can be equally used for development of medical diagnostic system and for the development of a battle system. Can its usage be limited?
5. Can cooperation on a project like SDI (strategic defense initiative) be justified?

These examples show us that it is necessary to distinguish:

- responsibility for a result, meaning that it concretely and according to a schedule, fulfills previously determined tasks (example 1); from
- responsibility for a set goal and the effect of a project, as well as the results gained thereby (examples from 2 to 5)

We would like to briefly take stance on the first kind of responsibility, "the result of responsibility". We will completely skip the legal aspect. As for the moral aspect, i.e. "the feeling of responsibility for a result", we believe there is no principal, but a gradual difference between developing computer programs and other wanted results, since perception and estimation of mistakes and shortcomings in computer programs is extremely difficult. Anyhow, the acceptance of one of them also includes the moral responsibility to finish it as better and as more careful as possible. I believe it is irresponsible to accept a task superficially, just to fool others. The focus of my theses should be based on ethical and (in a wide range) political aspects, as can be seen in examples from 2 to 5. It is important to note a difference between the projects (i.e. the results) with direct, obvious action, or applicable possibilities (example 2); and projects with indirect, hardly obvious consequences. In the latter we can include all kinds of "metaresults", e.g. software developing instruments (compare examples 3 and 4). We must note a close connection existing between responsibility for a result and the one related to setting a goal. It is obvious from
example 5 and the attitude of David Parnas[3] about cooperation with SDI (compare with Parnas, 1995). He explains his decision not to cooperate with SDI by the fact that nobody can take responsibility for the result, and the set goal is not only questionable, but also dangerous. This standpoint, based primarily on technical arguments, must be distinguished from any other standpoint which refuses SDI because of its possible political goal, such as combat ability and request for hegemony that lies under it. In short, this would mean that, if we want to inves-

[3] David Lorge Parnas is a Canadian early pioneer of software engineering, who developed the concept of information hiding in modular programming, which is an important element of object-oriented programming today. He is also noted for his advocacy of precise documentation.

tigate whether it is justifiable to work on a completion of a task, we must first pose a question of its goal. If the answer is positive, another question arises: shall the expected result achieve that goal? If that is not the case, or if we must count on various side effects, the goal must be modified accordingly. The modification must assure that the new goal can be achieved by eliminating the side effects, or that, when formulating the goals, inevitable goals and side effects must be counted on. The justifiability of such modified goals must then be re-examined. Technical civilization ethics. Our projects are a part of constant attempt to solve the problems of the entire human kind or a group of people by technical means. We are all aware of wanted and unwanted, as well as local and global, effects of technical solutions. Does this require a special "technical civilization ethics"? Hans Jonas studies this question in his book and answers it affirmatively (see Jonas, 1984). In continuance I would like to consider his theses that I found extremely important more closely.

## The new dimensions of responsibility

In his historical discussion Jonas compares the current situation with the one from the previous "untechnical age", approximately the medieval times. The area of their responsibility included an exclusively clearly defined and a very limited life space – "the city". The limits of responsibility were clearly marked by the city walls. Outside this "human state area" there was nature, intact and left to exist on its own. Today, almost the entire planet became a "global city". On Earth, there are almost no more large areas uninfluenced by people, no more areas existing on their own. Thus, the area of human responsibility is drastically increased. "Technics" (Greek: "techne") originally meant "skill", and primarily referred to objects, e.g. agricultural, domestic and hunting tools. Such technics had rare and slight effects on people (except the tools that have always been present). Today, technics almost always affects people (even the producers themselves), whether directly – as with genetic modulation or behavior control techniques – or indirectly, e.g. environmental changes. The increase in the area of human responsibility is equal to the increase in the area of inevitable consideration: if that area had earlier been spatially and temporally limited and clear (thus limited to an area of one's "city" and the length of a human life and possibly another human life), today it is spatially and temporally unlimited. In includes the entire Earth and even a part of the universe surrounding it, and encompasses many future generations, until their life space is completely destroyed. That changed the subject domain of ethics. If earlier ethics was limited to direct human relations, today it must comprise the indirect consequences of human actions, including the unknown and the unborn. Jonas includes these ideas, thereby expanding Kant's imperative: "do so that you can wish your maxim becomes the general law" by his own imperative: "do so that the consequences of your actions are in accordance with the requests of human life on Earth" (Jonas, 1984). While Kant's ethics refers to human interaction in direct

contact, Jonas is trying to explain "the new dimensions of responsibility" by means of his expanded imperative. Of course, many unanswered questions still remain, such as:

- What is "real human life? Is life in underground bunkers still considered "real"?
- How wide is the concept of "permanence"? It is certain that today we can still imagine the conditions of life and the possibilities of humans in the year 10,000? Maybe this Jonas's imperative could be determined like this: do so that the capability of deciding on life conditions of the next generation is unlimited (do not thus create the confusion about the real condition).

"The advantage of bad prognosis over good prognosis". Jonas introduces this seemingly odd thesis on grounds of the following probability: in great technical projects, there is a large number of "failures" as opposed to one great "success". Thus, the risk of failure is major. Even when the risk is decreased by verification measures, a great problem remains. Jonas explains it by comparing it to a competition ("the element of competition in human actions"): great technical projects in a great deal simulate competitions where, with a (probable) "chance of final victory", exists (significantly less probable, but not excluded) "a danger of infinite loss". As an example of such competition we can give the allegedly "secure" roulette system, consisting of duplicating the stakes after every lost game (only betting on a pair – *rouge-noir* chances). It is certain that the probability of a small prize is increased, but in case of a miss (which is relatively impossible – when minimal stake amount determined by the casino is exceeded), the loss is rather great. The logical conclusion for Jonas, after these reflections, is the "command of caution". No "ultimate goal" justifies "infinite total investment". The mere thought of the possibility of "the infinite loss", although impossible, should be enough to discourage us from such intent.

This means the following: where damage affecting larger portion of humanity cannot be completely excluded for the next several generations, the limit of responsible technics is reached, if not exceeded. This statement, in the meantime supported by the influential Church, is in opposition with the statement of the leading politicians, who modified Jonas's "danger of infinitive loss" to "the remaining risk". "The utopia of technical improvement dynamics and excessiveness of responsibility". In this chapter Jonas compares great technical projects to the activity of nature during the evolution. The nature takes a lot of time – it makes a great deal of slight mistakes, progresses slowly and does not affect the whole. As opposed to that, people are trying to reach the goal within their reach by means of great technical projects. There is no time for the mistaken ones (not even with great projects). Mistakes are not allowed – if too many risks are taken, the natural advantages are resigned. So, technical projects do not develop "communicative dynamics". The positive effects of reverse action ensure progress only if the first step had already been made. For instance, if construction

of a channel, a bridge or a tunnel is considered, a partial solution, which may only be obvious in the initial step of the construction, makes no sense. This means that the rest must be built (maybe even contrary to the knowledge gained in the meantime) in order to justify the initial initiatives. That leads to the famous "potato syndrome": "the potatoes are on the table, so now they must be eaten". Also, in relation to this, negative effects of reverse action must be mentioned: what must be included is the technics that would limit or diminish the harmful effects of past technical developments, e.g. effects of the removal of dangerous remains. From such dangerous effects Jonas derives the commitment of "watchfulness from the beginning". His thoughts, among others, are modified in the following demands:

"The demand of political philosophy". Due to complex and dangerous relations, the new ethics must turn to public politics more than to private behavior (compare Kant).

"The changed essence of human action changes the essence of politics". Hereby, politics should not be understood in sense of party of daily politics, but in the original Greek meaning of "the community of citizens".

"Representatives for the future" are indispensable. Daily politics only cares for present interests. "Nonexistence does not have a porch and the unborn are powerless." While discussing such demands and their practical consequences, we must not give in. Among many possible prognoses, the most favorable is often encountered. Whoever raises a voice is being denied by arguments such as: "We still know so little" and "There is more time". But, that is exactly what is not true because of the already mentioned dynamics. If "we still know too little", that means that we should by all means study all the possible consequences before we indulge into a technical adventure. We have an "obligation towards the future", primarily obvious in the "obligation towards our descendants. For Jonas, this is the "original form of responsible action". In the end of his every consideration, Jonas takes a stance towards the possible "pessimistic reproach". The previously introduced theses should not be misunderstood as pessimism. The greatest pessimist is the one who feels that the situation nowadays is so bad that, in order to change it, some very risky technical projects should be done. The role of informatics. Jonas does not in any occasion mention informatics and the computer appliance. Still, many of his theses relate to our domain, so it could be concluded that it is being seen as a "technical discipline par excellence". Let us look into some of his theses and in that sense, at the same time, try to find touching points for "new thoughts in informatics".

## Computer systems and the thoughts on evolution

Programming and evolution. The goal set by a programmer is from the very beginning hard to relate to the evolution of thought. The evidence is in the very word "programming", which means planning of long term predetermined procedures and streams. A program always requires the "exact solution", and often

the solution is found on the basis of early thoughts, i.e. their clients. Mistakes are undesirable – they can be tolerated, but can hurt further development. This is possibly where the change of thought in informatics started: lately, the incorporation of thoughts on evolution into software-engineering has been tried, using the so called "prototype". It is yet to be seen whether this approach is successful. Expert systems and evolution. Related to this topic, arises a question of a class of expert systems that needs to be the basis for human decision making in unformulated and completely formulated areas. By this we mean all the systems in which (most often on believable grounds) the decision-maker is allowed a "space for mind games", e.g. doctors' diagnosis; psychological advice (compare Weizenbaum/ELIZA!); marginal cases with the issues of presentation (lecturing); even legal areas. What shall we say when in a bank X we get the same data as in bank Y, because both of them operate with the same expert system and have the same scientifically-creative operating method?
In possible human form standardization I see a danger for further evolutionary development of human kind. I believe it is based on a large number of possible decisions – including many "wrong decisions". The next problem lies in the decision to shorten the time for decision making. Evolution allows itself a great deal of time. I do not think we should embrace the shortening when it comes to very significant decisions. Why is the saying that "all the decisions must be slept over "?

## Conclusion

On "communicative dynamics" of technical projects. At first sight, it may seem that communicative dynamics in ICT systems (data analysis systems) is not the same as in other technical projects. Physically, it is easier to remove a computer or a magnetic track than a nuclear power station or a big airport. But, the first sight is often deceptive. Most frequently, really important ICT system are greatly "embedded", i.e. one of the integral parts of its surrounding, without which it is amputated and incapable to operate. Everyone is familiar with examples of computer "breakdowns" in airports, banking money transfers and the military systems of preliminary alerts. ICT is made of facts that cannot be changed by people any more, such as trivial examples of a four-digit zip code, or a flight code consisting of two letters only.
"A change of opinion" would here, for example, mean that in initial phase of our projects we ask the clients whether they want themselves or others to be made dependent by the ICT system. A "technical evaluation of consequences" is necessary here – understood generally. Professional activity and the question of responsibility. How can a man, who in his professional activity, confronts the questions of responsibility, react? Some of possible behavior patterns are:
to withdraw; to change work, i.e. to decrease business engagement; to engage politically outside his company (public jobs, parties, FIFF, business initiatives, peace groups…); to take responsibility in his own company; to engage politi-

cally in his company (political companies), e.g. as an advisor; to convince people by talking to them within (or outside) the company; to ignore the facts; to give over to resignation.

We must immediately state: we can not and will not give recipes, or "recommend a choice" when it comes to the patterns above. Instead, we would like to give several examples and, based on them, discuss different options (and their limits).The first example – the activity of advisor in a software company. In this example we shall talk about the work of an advisor, reflecting our own experience. For me, to candidate for the advisor in my company, the critical criterion was the idea of my "political engagement inside the company". In the election period there was incertitude in further development of the company, especially concerning military projects related to security. The previous advisor in the company did not succeed with his idea of making a poll for every issue. The task of the new advisor was thus to find another method for talking about the subject, provided that the idea does not fail again and that nobody is provoked by it. Hesse continues by saying: "We have tried to openly talk to the company management, in order to signal the wishes and the ideas of the team. The idea was that a poll must be made among team members – with the question of which themes, and in which order, the advisor should deal with". On the theme list – beside classic themes like working hours, the cafeteria and traffic connections – the theme complex "Company development" and "Social aspects" was highlighted. The poll was organized anent one of the company meetings. There, the project manager reported that he could not find associates for a certain military project and asked the management if they considered it reasonable to keep acquiring projects from that area under such circumstances. A long, mutually open and vivacious discussion followed, and it did not end in a conclusion, but both sides gave word to the advisor.

Hereby I would like to express my appeal for both sides, i.e. the union members and the workers who are not in the union, to exchange opinions and end the hostility. The second example – from weapon production machinery. Some employees, mostly the ones engaged in production for military needs, took part in an SDI program. Peace group supporters from the company and union representatives tried to talk about the subject. The reaction served as an internal "guide" to discard the appeals in forms of paroles for workers' personal co-responsibility, as well as the attempt to "create an attitude of repulsion" among the union members. Such "political actions (…) which might significantly disrupt the peace in the company should not be tolerated".

One group paid a lot of attention to "responsibility" (legal), which was given before it was even clearly familiar with "the new dimensions of responsibility" (Jonas, 1986). In a heated, confronted climate, political approach to the topic recommended by Jonas and which, naturally, has to start in the company (it has nothing in common with forbidden party-political actions in the company), can not work.

However, that is not all. Along with time optimization which is directed to work, protestant ethics understands an organization of time directed to work. With protestant ethics, the idea of regular working hours came into the center of life. Self-organization is lost and transferred into the area left after work: evening as remain of the day, the weekend as remain of the week and retirement as what remains after a working life. In the centre of life there is work, which is regularly done and which organizes all the other ways of spending time. Weber describes how, in protestant ethics, “inconstant work, to which ordinary worker is obliged, is often an inevitable, but always an unwanted inter-state. What a man ‘without profession’ is missing is exactly a systematic-methodical character, which, as we have seen, is required by the world”. So far, that time organization in informatical economy has not changed a lot. However, rare people can still deviate from strict, regular working hours, despite the fact that new information technologies do not only compress time, but also make it more flexible (Castells calls it “time desequencing”). With technologies like the Internet and mobile phones, we can work wherever we want and whenever we want.

## References

Himanen, Pekka; Torvalds, Linus; Castells, Manuel. *The Hacker Ethic and the Spirit of the Information Age*

Jonas, Hans. Ethical aspects of experimentation with human subjects. Boston: American Academy of Arts and Sciences, 1969. OCLC 19884675.

Jonas, Hans. Heidegger and theology, 1964. OCLC 14975064 (included in The Phenomenon of Life)

Jonas, Hans. Immortality and the modern temper: the Ingersoll lecture, 1961. Cambridge: Harvard Divinity School, 1962. OCLC 26072209 (included in The Phenomenon of Life)

Jonas, Hans. The Gnostic Religion: The Message of the Alien God & the Beginnings of Christianity. Boston: Beacon Press, 1958. ISBN 0-8070-5801-7

Jonas, Hans. The Imperative of Responsibility: In Search of Ethics for the Technological Age (trans. of Das Prinzip Verantwortung) trans. Hans Jonas and David Herr, 1979. ISBN 0-226-40597-4 (University of Chicago Press, 1984.) ISBN 0226405966

Jonas, Hans. The Phenomenon of Life: Toward a Philosophical Biology. New York: Harper & Row, 1966. OCLC 373876, Evanston Ill.: Northwestern University Press, 2001. ISBN 0810117495

Nadrljanski, Djordje. Obrazovni softver – hipermedijalni sistemi, Univerzitet u Novom Sadu, 2000.

# Relationship between Scientific Paradigm and Research Front. On Example of Information Science Research Production

Đilda Pečarić
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10 000 Zagreb, Croatia
dpecaric@ffzg.hr

**Summary**

*Empirical indicators about time distribution and "obsolescence" of cited references are retrieved from doctoral dissertations in Information Science from 1978 to 2007 at Croatian universities. Cluster of most cited authors who are recognized as key authors for Information Science paradigm is gained from citation and co-citation analysis. Domination of these authors constitutes "conceptual knowledge zone" which is placed according to time distribution at the end of axes of obsolescence of cited literature. As opposed to that, research front, which is the period of most intensive research activity, is placed in the first time quarter (from zero to four years old cited literature), and there are 10% of most cited authors. Our research follows transit of researcher from research front to zone of paradigmatic knowledge. Hypothesis is that new authors who enter conceptual knowledge zone suppress "old" authors; so it can be concluded that incomers in conceptual knowledge zone are holders of new theoretical approaches and solutions for new problems. Duration and importance of "old" authors ensure paradigm and methods for solution of "old" problems, respectively for the production of professional papers and traditions of profession.*

**Key words:** Information Science, Research Front, Scientific Paradigm, Obsolescence of Cited Literature, Conceptual Knowledge Zone

## Introduction

At the beginning of 1960's E. Garfield and Irving H. Sher defined "journal impact factor" as a criterion for journals selection for Science Citation Index, which Institute for Scientific Information (ISI) in Philadelphia began publishing in 1961. Since then, the impact factor has become one of the basic, most prevailing and most used criterion in the Information Science for the evaluation of scientific journals and scientific papers. After that strong pressure on research-

ers, scientific journals and scientific institutes begins, because their "effectiveness" and relevance in scientific community is determined by measuring impact factor and journal citation frequency. Entire scientific community lives under pressure that they have to be cited in the shortest possible time, because impact factor became key criterion for promotion of scientists, evaluation of scientific institutes, financing of scientific journals[1].

On the other hand, the research of scientific development ie. "maps of sciences" and "cognitive structures of science" display and recognise the most cited authors in sciences and scientific disciplines as key authors in prevailing scientific paradigms[2]. Paradox is that the Information Science has not examine relationships between first and second group of authors: research front determined by impact factor, i.e. speed of citation on one side, and continuity of dominant authors in certain scientific paradigm, on the other side. Our interest is to explore this relationships and alterations inside first and second group of authors respectively within "research front" and "zone of conceptual knowledge".

Research is done on 134 doctoral dissertations in Information Science at Croatian universities from 1978 to 2007 (Đ. Pečarić, 2009).[3]

## About constants in scientific communication and about differences in communications models

In another research, we explored, by citation analysis, features of communication models that are dominant in scientific communication (M. Tuđman, Đ. Pečarić, 2009.). The corpus of 22,210 cited bibliographic units is analyzed[4]. On the basis of citation frequency according to the age of cited literature we determined existence of constant in scientific communication models.

[1] See E. Garfield, 2006.; Respectively, M. Jokić (2005) "Fundamentally, impact factor is ratio between citations and recent citable items published in the same period."

[2] Overview of those research are in H.D. White i K.W. McCain (1998)

[3] In the period from 1978 to 2007 at Croatian universities 134 doctoral dissertations were done in seven different Information Science disciplines: 20 in librarianship, 21 in information science, 53 in information systems, 22 in communicology, 9 in museology, 8 in archivistics and documentation, 1 in lexicography. The majority of doctoral dissertations were made at the Faculty of Organization and Informatics in Varaždin (FOI) – 69, followed by the Faculty of Humanities and Social Sciences in Zagreb – 49 doctoral dissertations. According to the periods of production: 21 doctoral dissertations were made until 1989; 62 doctoral dissertations from 1990 to 1999; 51 doctoral dissertations from 2000 to 2007.

[4] From total number of citation (22,210), there are 17,178 cited units with authors, that is, the total number of cited authors is 10,683. There are 8,296 (77.65%) authors that are cited just once, and 2,387 or 22.34% of authors are cited more then once. Those 2,387 authors, that are cited more than once, hold 51.71% of citation. The rule that a small number of authors are often cited repeats again: 451 authors that are cited 5 or more times hold 23.61% of citations; 118 authors that are cited 10 or more times hold 11.71% of citations. First 29 most cited authors hold 5% of citations, that is, first 49 authors hold 7% of citations (M. Tuđman, Đ. Pečarić, 2009).

Communication model[5] has several unchangeable characteristics. Regardless of the variable used (citation frequency, self-citations, citations according to languages or distribution of citation that are cited only once) citation distribution curve is always the same or similar. This is also confirmed when the data are fragmented according to scientific discipline, as well as time periods or faculties on which doctoral dissertations are made.

Second, when we know cited half-life (**t/2**), period in which 50% of documents are cited, then first 25% of documents are cited until half of cited half-life (**t/4**). In time period **t/4** maximum frequency from overall number of cited documents is reached. Therefore, citation curve has log-normal distribution, with maximum in time period **t/4**.

Third, perceived regularity is that in time period between **t/2** and **t**, i.e. second time period of cited half-life, following 30% of documents are cited. After double cited half-life – in which 80% of documents are cited, the last 20% of documents are cited. For those documents, ages cannot be statistically predicted.[6]

Based on previously described regularity we could identify three communication zones based on nature of citation usage. These zones are shown in Graph 1.

We named the first zone **empirical knowledge zone**[7], which is sequential and extends through entire communication process. This zone consists of citations of authors and documents that are cited only once. This group holds 60% of citations. Their distribution is equally distributed and presented during entire communication process.

Second zone is named **research front zone** and it is placed in communication space and time that we marked as **t/4**. In time **t/4** first 25% authors and documents are cited. In this period maximum frequency of overall document's citation is reached. Attendance and citation of authors in research front zone implicate their understanding of problem and communication with everyone in their surroundings relevant for the problem. This is the space of authors bidirectional communications in which empirical and conceptual knowledge are being overlapped, compressed and reinterpreted. In the nature of research activities, it is typical that research time is shorter from document citation half-time, although scientific, formal and informal communication can last much longer.

We named third zone **conceptual knowledge zone.** The most cited authors are in this zone, obviously because of their influence. It is logical that influence is

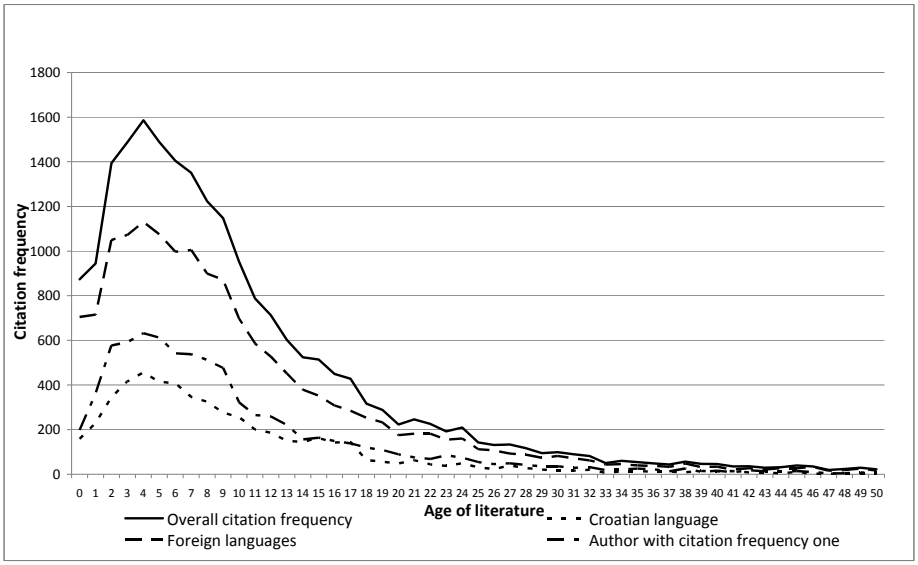
[5] See R. Vrana (2003)

[6] See M. Tuđman, Đ. Pečarić, 2009.

[7] Basic concept comes from Capurro's (2006) definition of empirical knowledge: it is information that is the result of the process of selection in communication process. The documents and authors that are cited only once have value as empirical or theoretical information. Since these citations are used only once, we assumed that their value is more empirical than conceptual.

bigger as it is more permanent. And that is why it is not strange that the age of cited literature of the most cited authors is older then citation half-live. Kuhn thesis[8] implicates that the most cited authors are cited primarily for referencing on dominant theories, for solutions of scientific problems. Referencing on mutual scientific paradigm, which is defined by influential scientist, binds members of certain communication community.

Graph 1. Citation frequency and knowledge zones.




## Authors alterations in conceptual knowledge zone

In Graph 2, 45 most cited authors in Information Science from 1978 to 2007[9] are displayed. If we know that half-time of cited documents is 7.5 years[10] it can be concluded that the most frequently cited authors are those whose cited publications are from 7.5 to 30 years old.

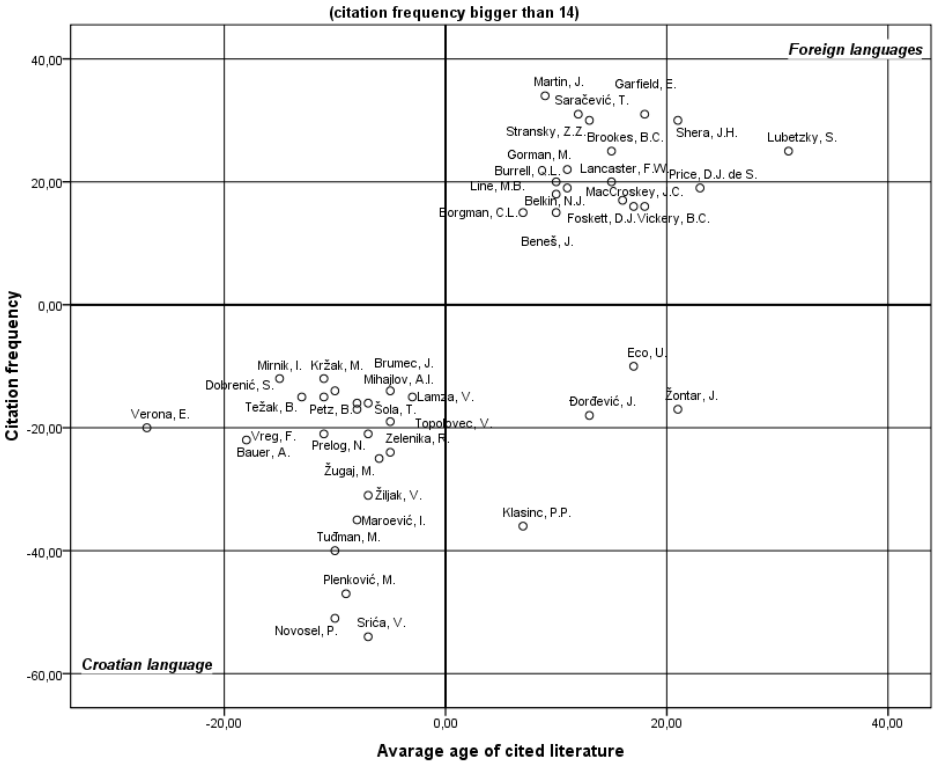
[8] See Kuhn, T. 1999.

[9] Table 2 shows authors that are cited 14 or more times. In the right top corner of the table authors cited in foreign languages are shown. Authors cited in Croatian language are shown in the left bottom corner of the table. In the bottom right corner of the table there are publications that are published in SFRJ, but outside of Croatia.

[10] Certain differences exist according to the type of cited documents: for monographs' cited half-life is 9.1 years, for journals it is 7.2 years, and for semi-publications it is 9.3 years. The difference exists also among disciplines. The shortest cited half-life is in information systems 5.9 years, and the longest in museology 12.6 years. For information science it is 7.1 years, for librarianship it is 7.8 years, for communicology it is 8.5 years, and for archivistics and documentation it is 8.6 years.

Since we are interested in the relationship between authors in research front zone and conceptual knowledge zone, primarily we have to establish how authors' alteration in conceptual knowledge zone occurs. That is why we analyzed the most cited authors according to periods of development of Information Science in Croatia from 1978 to 2007. The data are compared with several criteria: according to Information Science disciplines and according to faculties on which doctoral dissertations are made. In this paper we give only basic determinants, in order to indicate trends and to make conclusion about authors' alteration, if it exists, in the conceptual knowledge zone.

Graph 2: The 45 most cited authors in Information Science from 1978 to 2007




Graph 2 displays 45 most cited authors in Information Science from 1978 to 2007 in 134 doctoral dissertations. Clearly, the display of the most cited authors according to disciplines, or according to time periods, will be different than it is shown in Graph 2. Yet our interest is not in cited authors, but in regularity upon which authors' alteration in certain knowledge zone happens.

This paper will discuss primarily the differences that occur during three different time periods[11]. Time periods are arbitrarily divided into 10-year periods. In the first period (from 1978 to 1989), 31 authors out of the 45 most cited authors are cited. It is important to notice that these 31 most cited authors' hold 6.28% citations from overall number of cited documents in that period. Since most citations are older than 7.5 years, it means that those authors hold more than 10% of the citation in second part of cited half-life[12]. In this period first 7 the most cited authors are: P. Novosel, E. Garfield, J. Beneš, A. I. Mihailov, J. Đorđević, A. Bauer, B. Težak.

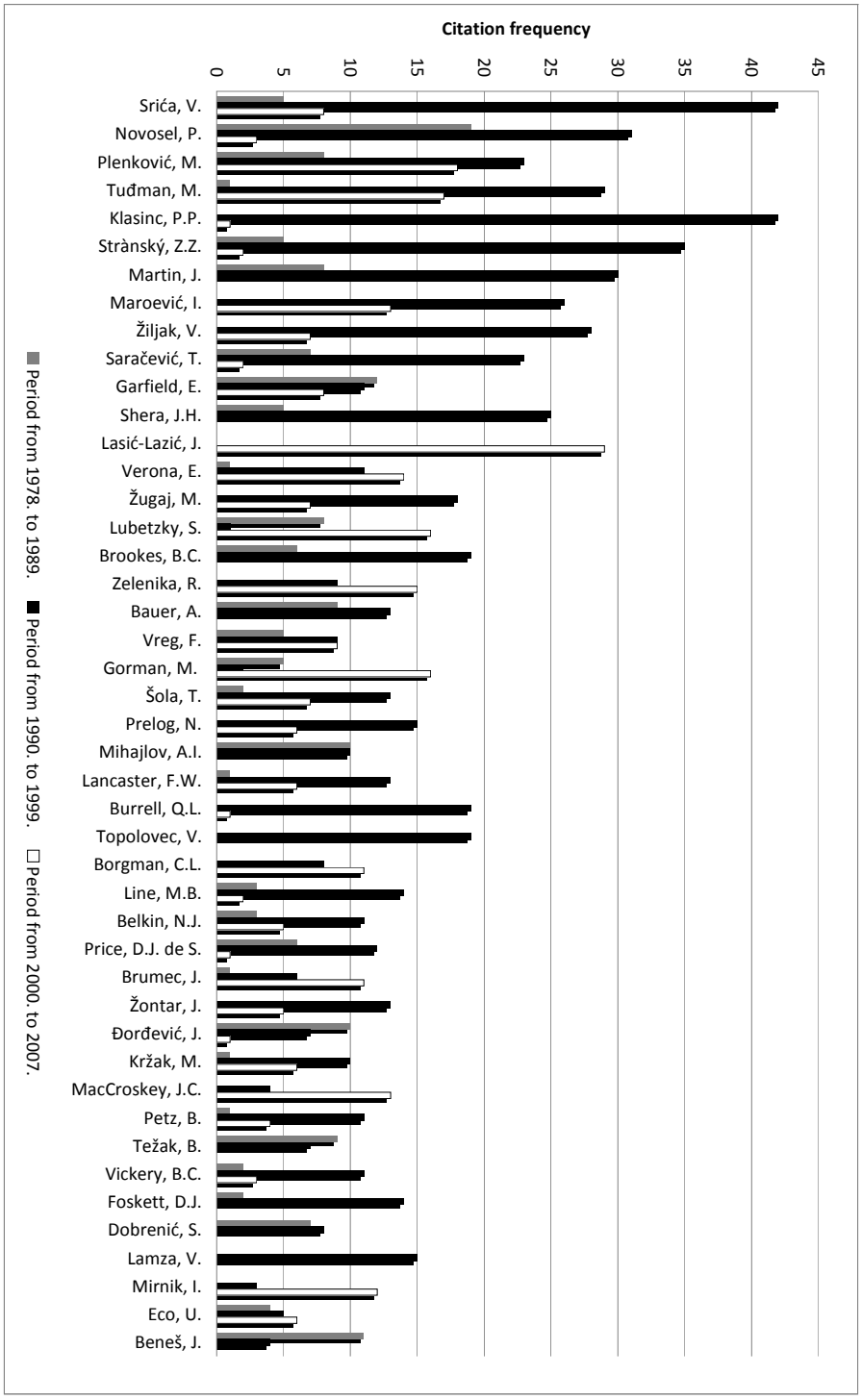
In the second period (from 1990 to 1999), 44 out of the 45 most cited authors are cited. Even 8.19% citations from all citation in this period are held by these 44 authors, i.e. 0.9% authors out of 5094 authors cited in second period. Since most citations are cited in second part of citation obsolescence half-life it means that 44 authors hold almost 15% of citation. First 7 of the most cited authors in this period are: V. Srića, P. P. Klasinc, Z. Z. Stràsnký, P. Novosel, J. Martin, M. Tuđman, V. Žiljak.

In the third period (from 2000 to 2007), 39 out of 45 the most cited authors are cited. Those 39 authors from overall of 4611 authors cited in third period hold 4.28% citations from overall number of citation. As the majority of these authors are cited in the second part of citation obsolescence half-life, we can see again that a small number of authors, 0.8% of the most cited authors, hold 8% of citation in the second part of citation obsolescence half-life. First 7 of the most cited authors in this period are: J. Lasić-Lazić, M. Plenković, M. Tuđman, S. Lubetzky, M. Gorman, R. Zelenika, E. Verona.

Authors' space and placement in conceptual knowledge zone are neither constant nor lasting. In analyzed range of 30 years, only 22 authors out of 45 of the most cited authors are cited in all three periods. This means that 50% of authors are not cited in all periods. In the first period, 31 out of 45 authors are cited, i.e. 14 authors are not cited, some of which are among the most cited authors in following periods (e.g. P. Klasinc, I. Maroević, J. Lasić-Lazić, V. Žiljak, N. Prelog). In the second period, 44 out of 45 most cited authors are cited. And in the third period, among 34 cited authors there are not present 11 authors, some of which are founders of Information Science (such as J. Martin, J. Shera, B. C. Brookes, A. Bauer, A. I. Mihailov, B. Težak, D. J. Foskett, S. Dobrenić).

[11] All of the following data are from Đ. Pečarić (2009, manuscript)

[12] In this period (1978.-1989.) overall number of cited authors is 1952, out of which 1582 authors are cited only once, which means that from overall number of cited authors 1.6% (from 31 most cited authors) of authors hold 6.28% of citation.




Graph 3: 45 most cited authors according to periods

These quantitative determinants cannot be the foundation for taking qualitative conclusions, because data can be re-arranged according to different criteria. For us it can be interesting to know which of 45 of the most cited authors are cited in doctoral dissertations from different disciplines in Information Sciences. Only one author (M. Tuđman) is cited in all seven disciplines; six authors (V. Žiljak, T. Saračević, J. Lasić-Lazić, N. Prelog, N. J. Belkin, and B. Petz) are cited in five disciplines. Ten authors are cited in four disciplines, while five authors are cited in three different disciplines. Remaining 23 out of 45 the most cited authors are cited in one or two disciplines.
These data alone indicate that neither authors' "lasting", i.e. their presence in all three periods, nor citation frequency are sufficient argument for evaluation of certain authors influence. The same rule applies for authors' citation in different Information Science disciplines. In fact, the most cited authors are not cited in all disciplines; authors that are cited in most disciplines are not cited in all periods. Exactly this authors' "vicissitude" is what we want to detect and describe.

## Researchers, Scholars and Predecessors

Table 1 shows 22 out of 45 most cited authors that are cited through all three periods. When the data are analyzed according to average age of cited literature within certain period, it can be concluded that every analyzed period consists of three time zones. This can be illustrated by following examples.

### Predecessors' time

In all three periods after the obsolescence of cited literature in time (**t**) i.e. double half-life of citation obsolescence, which is 15 years for our corpus of cited literature, the authors highly relevant for the development of Information Science are cited. The fact that they are present and cited after the time of obsolescence of cited literature gives us right to call this group 'predecessors of Information Science'.
It is clear that group of authors that belong to predecessors are not the same from one period to the next, in spite of the fact that there is often overlapping. Therefore, in the first period the following authors belong to this group: E. Garfield, A. Bauer, F. W. Lancaster, Z. Z. Stranski, E. Verona, S. Lubetzky, B. C. Vickery, D. de S. Price.
In the second period 13 authors belong to the group of predecessors, of which the "oldest" according to citations are: E. Verona, D. de S. Price, J. H. Shera, S. Lubetzky, A. Bauer, B. C. Vickery, E. Garfield, etc.
In the third period 11 authors (from 34 most cited authors) belong to the group of predecessors, among which the "oldest" according to citations are: S. Lubetzky, D. de S. Price, Z. Z. Stranski, E. Verona, etc.

Table 1: Authors cited through three periods according to average age of cited literature in certain period.

| | Author | Average age of cited literature | | |
|---|---|---|---|---|
| | | Period from 1978 to 1989 | Period from 1990 to 1999 | Period from 2000 to 2007 |
| 1 | Price, D.J. de S. | 19 | 23 | 38 |
| 2 | Vickery, B.C. | 17 | 21 | 6 |
| 3 | Lubetzky, S. | 17 | 22 | 39 |
| 4 | Verona, E. | 15 | 27 | 31 |
| 5 | Stràsnský, Z.Z. | 14 | 14 | 37 |
| 6 | Lancaster, F.W. | 13 | 13 | 21 |
| 7 | Garfield, E. | 13 | 21 | 21 |
| 8 | Gorman, M. | 10 | 8 | 12 |
| 9 | Eco, U. | 10 | 14 | 24 |
| 10 | Đorđević, J. | 9 | 16 | 23 |
| 11 | Novosel, P. | 8 | 11 | 8 |
| 12 | Vreg, F. | 8 | 12 | 13 |
| 13 | Belkin, N.J. | 8 | 11 | 14 |
| 14 | Saračević, T. | 8 | 13 | 17 |
| 15 | Srića, V. | 7 | 6 | 11 |
| 16 | Plenković, M. | 7 | 8 | 13 |
| 17 | Line, M.B. | 7 | 9 | 14 |
| 18 | Petz, B. | 5 | 10 | 4 |
| 19 | Šola, T. | 4 | 5 | 9 |
| 20 | Brumec, J. | 2 | 3 | 7 |
| 21 | Tuđman, M. | 2 | 9 | 11 |
| 22 | Kržak, M. | 0 | 10 | 14 |

**Time of researchers**

In every period authors whose publications are not older than cited half-life (in our corpus a half-life of citation obsolescence is 7.5 years) can be found. We recognize this time as time of research, and the authors in this period as the group of most cited researchers. This period contains 11 authors that we recognized as the group of researchers, from overall of 31 most cited authors in this period. The youngest citations of authors are in this group: M. Kržak, M. Tuđman, J. Brumec, T. Šola, etc.

In the second period, the group of researchers, i.e. the most cited authors according to half-life of citation obsolescence, consists of 11 researchers also, and the youngest are (in sequential order): J. Brumec, V. Lamza, T. Šola, V. Topolovec, V. Srića, etc.

In the third period, the group of researchers consists of 8 out of 34 most cited authors: Q. L. Burrell, R. Zelenika, B. Petz, C. L. Borgman, J. Lasić-Lazić, etc.

It is clear from these indicators alone that the group of authors in "time of researchers" can only be formally determined. For more profound content analyses it is necessary to determine if the authors are cited really new publications or just new translations or reprints of old publications.

### Time of scholars

On time scale between time of researchers and time of predecessors, which is between citation half-life and life of literature's obsolescence, third group of authors, which we named scholars[13], is positioned. Authors that belong to the group of scholars in three analyzed periods can be recognized in Table 1. It is visible that alterations of authors do not happen only from one period to the next, but from one group of authors to the next. Usually, path of scholars goes from researchers to the group of scholars in order to end up in the group of predecessors that future researchers and scholars will refer to.

## Conclusion

With analysis of cited bibliographic references from 134 doctoral dissertation made in Croatian universities from 1978 to 2007, several zones in scientific communication are recognized. Three zones are permanently present: empirical knowledge zone, conceptual knowledge zone and research knowledge zone. In this paper particularly is discussed authors' relationship between research knowledge zone and conceptual knowledge zone. We identify, by citation frequency and percentage in overall number of citation, which authors are dominant in conceptual knowledge zone. In addition, we identify alterations of dominant authors in conceptual knowledge zone. On time axis determined by the age of cited literature, we identify three groups of authors in conceptual knowledge zone. According to chronological order first group of authors consists of predecessors, i.e. those authors that precede scholars and researchers. Publications of those authors are mostly older than double obsolescence half-life of cited literature. The second group consists of scholars, i.e. authors whose publications are most cited in period between cited half-life and time of knowledge obsolescence. The third group of authors consists of researchers, i.e. authors that are most cited in knowledge obsolescence half-life[14]. Based on empirical data it can be concluded that the influence of certain authors from researchers via scholars to predecessors does not depend on the publication obsolescence time, but on the sequence of factors that were not the topic of our analysis. We identify three different groups of authors in conceptual knowledge zone, as well as regularities i.e. why some authors can occur in a certain group, but not necessarily in all groups that we identified in this analysis. Just by looking at the titles of most cited authors in conceptual knowledge zone we can confirm Kuhn's hypothesis about scientific paradigms that "incomers" suppress "old" authors, regardless of whether they work on old scientific problems in a new way or they deal with new problems. In this paper this hypothesis is shown

[13] Scholar – in the authentic meaning of the word, it is a person who has improved knowledge, a learned person, scientist. (Anić, 2003)

[14] See Kuhn, T. 1999.

only by quantitative indicators on alteration of authors in described zones. Only qualitative analysis of the publications of most cited authors would prove our hypothesis completely.

## References

Anić, Vladimir. Rječnik hrvatskoga jezika. Zagreb: Novi Liber, 2003

Capurro, Rafael; Chaim Zins. Knowledge Map of Information Science. Rafael Capurro's responses to Chaim Zins. (2006). http://www.capurro.de (2009)

Garfield, Eugene. The History and Meaning of the Journal Impact Factor . //JAMA. 295 (2006) 1, 90-93

Jokić, Maja. Bibliometrijski aspekti vrednovanja znanstvenog rada. Zagreb: Sveučilišna knjižara, 2005

Kuhn, Tomas S. Struktura znanstvenih revolucija. Naklada Jasenski i Turk. Hrvatsko sociološko društvo. Zagreb. 1999

Tuđman, M., Tudor-Šilović, N. Boras, D., Milas-Bracović, M. A literature measure of scientific communication: Co-citation analysis of masters theses in informatin sciences in Yugoslavia. 1961-1984. In: N. Tudor-Šilović, & I. Miheal (eds.), Information research methods in library and information science (pp. 225-247). London: Taylor Graham, 1988

Pečarić, Đilda. Razvoj informacijske znanosti u Hrvatskoj. Bibliometrijska analiza doktorskih disertacija iz informacijskih znanosti 1978-2007. Zagreb: Filozofski fakultet, doctoral dissertation, manuscript, 2009

Tuđman, Miroslav; Pečarić, Đilda. Prilozi dubinskoj analizi komunikacijskih obrazaca. // Society and Technology 2009 Zadar (Informatologija, in press)

Vrana, Radovan. Utjecaj mrežnih izvora informacija na razvoj znanstvene komunikacije u društvenim znanostima u Hrvatskoj. Zagreb: Filozofski fakultet, doctoral dissertation, manuscript, 2003

White, Howard D. McCain, Katherine W. Visualizing a discipline: An Author Co-Citation Analysis of Information Science, 1972-1995. Journal of the American Society for Information Science 49 (4)(1998) 327-355

Professional paper

# Development of Spatial-visual Intelligence

Mila Nadrljanski
Faculty of Maritime Studies
Zrinsko Frankopanska 38, Split, Croatia
milamika60@yahoo.com

Marija Buzaši
Faculty of Education
Podgorička 4, Sombor
buzasi@sbb.rs

Mira Zokić, Ph.D. student
Zvornička 9, Zagreb, Croatia
mira.zokic@gmail.com

**Summary**

*Due to development of mass media, greater attention has been shown to an image as a visual object. Development of visual competence has also been promoted so as to avoid the danger of visual analphabetism, and in order not to succumb to the power of the image. With the development of digital media and discussion about iconic turn, the awareness of the significance of images – not only in art – but also in natural and social sciences has been getting stronger. For active, critical and conscious perception of images, as well as for selection in yet bigger number of images, special competence, which is developed as basic competence in the sense of aesthetic education is essential. The subject matter of this paper's research is a possibility of developing spatial-visual intelligence. During learning and thinking process, visual intelligence plays a great role in creating visual image, memorizing by means of images and, accordingly, it should be developed in as greater measure. Spatial-visual intelligence can be developed via reading geographical maps, navigation, park planning, model construction, constructive games, plotting, making itineraries, or visual maps of any information (Mindmap). The degree of visual intelligence helps students to acquire knowledge not only through static images, but also to engage their imagination, to imagine objects from different angles. Through the dynamicity of video-clip they visualize gained information. Child's spatial-visual intelligence can be deformed by non-critical acceptance of mass culture's products, i.e. by predimensional visual impulses via media, films, the Internet, videogames.*

**Key words:** visual IQ, intelligence, perception, spatial-visual intelligence.

## Introduction

In the age of complete predominance of visual media and visual communication – that refers to computer and the Internet, mobile telephony, television, newspapers, advertising posters and shop windows etc. – we can notice that the awareness of visual education importance is very low. The reason can be found in insufficient art culture and in long-lasting decrement of importance of art subjects in school in general. In human perception, eyesight has the crucial role. We are able to aggregate and bank up an incredible amount of visual information. The question of identification of shapes and forms is subject-matter of psychologists, artists, but also of engineers and informaticians. Every information system contains reception and coding of information. In modern world, a need for fast decision making, choice skill, active initiation, creativity and constructive thinking is made. Fast, successful and opinionated thinking is precious in everyday work of engineers, managers and teachers. Each child is born with certain abilities. Though, a great number of children are not able to use their abilities, regardless the level of inborn talent. According to Bloom's taxionomy[1], different levels of knowledge can be distinguished – from simple knowledge of facts, through apprehension, implementation, analysis, synthesis, to most complex evaluation, which all require a more complex ability of thinking. Symbolically, for presentation and quality gaining knowledge of graduated students, we have modified a model of Multiple intelligence[2]. Students' desirable features can be presented in a form of Multiple intelligence:

1. Linguistic
2. Logical-mathematical
3. Spatial
4. Physical-kinetic
5. Musical
6. Interpersonal
7. Intrapersonal

## Perception and visual communication

Human perception is a process of gaining, interpretation, selection and organization of stimulative information and, as such, presents a source of all knowledge. It can be viewed as an observation of relationships between reality (enchantment) and mind (intellect) reflection (thinking, judgement etc.) Depending on which sensory organ is dominant, we can distinguish visual, auditive, tactile,

[1] In 1956 Benjamin Bloom developed the classification of levels within cognitive domain, together with a group of educational psychologists. Also, he discovered that the pupils in 95 per cent of cases come across questions which refer to the bottom level – recollection of information – during testing. Those six levels are: knowledge, comprehension, implementation, analysis, synthesis and evaluation.

[2] Theoretically elaborated by Howard Gardner in book *Frames of Mind*.

gustative and olifactive perception. In human perception, a crucial role belongs to eyesight. We are able to aggregate and bank up an incredible amount of visual information. The question of identification of shapes and forms is subject-matter of psychologists, artists, but also of engineers and informaticians. Perception is not only noticing of enchantment from surrounding, but also a complex process which takes place through organization and interpretation of meanings and provoking organism's reaction. Perception and comprehension of scene will depend on man, i.e. on his previous experience, motive and interest. Only things that we know, that we "see" with our mind, have influence on perception. In this way, for example, a simple sentence written on a piece of paper can be interpreted individually and in different ways. For those who are illiterate, paper contains only different lines and patterns, and someone who is not familiar with, for example, Croatian language, is able to distinguish only letters, without their real content. Every information system contains reception and coding of information. The written word "blind" is a carrier of meaning, and it is called 'signifier', and the meaning of the word is a term that signifies a man who is not able to see. Many of signs used in communication process incurred arbitrarily. Their decoding indicates how to learn the structure and language manners of their use. Communication is inevitable without people, sign and objects of universe. Contemporary science, which deals with signs and sign systems, has developed from two sources. In 1916 Swiss linguist Ferdinand de Saussure introduced presumptions about general science semiology. Opposing De Saussure's linguistic approach to issues, Anglo Saxon scientist Charles Sanders Peirce researched interrelationship between physical sign, object and human from a logical point of view, and is regarded as founder of semiology. The smallest element in De Saussure's semiology is a sign, which is composed of form and concept ('signified' and 'signifier'). According to him, majority of signs are symbolic and accidental, and their meaning can be understood as a process that connects the 'signified' with the 'signifier'.Communication is transmission of thoughts and messages. Elementary forms of communication are based on signs and sounds. Phrase 'communication' can be understood as comprehension of all procedures with which one mind can influence another. Such an extended definition contains not only speech, but music, art, theatre, ballet, etc. Communication with environment in which we live, and throughout it with broader human society, represents communication of ideas, attitudes or mental/psychological reactions on given conditions in society, problems and methods of their solving.

Expression 'communication' is derived from Latin word 'communis', which means 'common'. Therefore, communication can be specified as a process of communion or union of addresser and addressee. *Visual communication* manifests in linguistic and non-linguistic form. Linguistic form is, for example, language of the deaf and the dumb. Pitch of visual linguistic communication development is the development of alphabet. Communication systems as Morse

alphabet and traffic lights belong to this group. Typical non-linguistic forms of visual communication are facial expression (for example: smiling, scowl) and gestures (for example: shrugging). Visual-spatial intelligence refers to orientation in space, ability of figurative and abstract visualization, thinking via imagery (scenic) conception, capability of thinking in the third dimension, redefining and recomposition of existing art compositions into new ones. Art creativity affects imagination, ability of forming in different two-dimensional and three-dimensional materials, creating different practical works (drawing, pictures, reprints, sculptures, reliefs, installations). Visual language is made of art elements (paradigms) and compositional principles (syntagms). Visual presentations are based on the system of conventions which has developed in process of attempts and mistakes throughout millenniums, to unite sign and signed. In that context, in relation between sign and signed, our reaction on image should respond to reality that the image represents. Creator of visual message demonstrates its presentation in scenic form, which is made of visual elements. Communication can take place only in some social context and linguistic environment. Visually coded message is subject to influences of social and cultural environment, as well as recipient’s message interpretation is subject to same influences. Just as the recipient, communicator has to posses certain knowledges that have been acquired through learning and are predisposed by collective memory. Throughout the process of socialization, we acquire different kinds of knowledges. Similarly, we become sensitive to problems, so we develop an ability to perceive shortcomings or needs for changes, or improvements in the existing things. Also using redefinition, i.e. the capability of abandoning old ways of perception, a new and broadened meaning is given to familiar objects for some new purposes. The contemporary society demonstrates a need for acquisition of languages and techniques of visual media, as well as exploring the mechanism of their influence on forming a personality.

## Intelligence

Many people identify thinking and intelligence, even though the relationship between intelligence and thinking is like relationship between car and the driver. Intelligence is the car, and car’s driver will, via thinking, decide which route he will take. Therefore, intelligence is ability that we can use via experiential thinking, or on contrary, it will stay uncultivated (Mozart in rainforest). Intelligence, as a badly used tool, can be a barrier for thinking.

Gardner’s theory (Howard Gardner) of multiple intelligences based on understanding that different parts of brain are connected with those different types of intelligences, is well known. With that theory, Gardner abandoned contemporary theory according to which intelligence is conditioned and unchangeable. According to Howard Gardner’s theory of Multiple intelligence, intelligence is not a homogenous mind skill. Gardner’s model of Multiple intelligence indicates that there are numerous learning styles and cognitions of the world.

1. Linguistic intelligence refers to verbal expression (understanding of terms, reading, writing texts). Stories, puzzles, and debates are a motivation for learning.
2. Logical-mathematical intelligence, which is most frequently connected to scientific thinking (abstract thinking, mathematic problems, logical conclusion. Problem tasks, experiments, and problem solving encourage learning.
3. Visual-spatial (spatial) intelligence (visual thinking and memory, interpretation of maps, orientation in space, navigation). Tasks of visual expression, making mind maps, and activities of visualization motivate learning. Images, graphs, films, and demonstrations should be used in teaching.
4. Musical intelligence (singing, playing instruments, sense of rhythm). Listening to music for relaxation which inspires for visualization and rhythmic games should be used in teaching.
5. Bodily-kinesthetic intelligence (dancing, sport, body language in expressing emotions). The best way to acquire knowledge is via movements – dancing, movement, dramatization, learning in nature.
6. Interpersonal intelligence (empathy, pupils are communicative, understanding of knowledge, emotions, motives, they enjoy other people's company). Learning in pairs and group work motivate them for learning.
7. Intrapersonal intelligence (self cognition, understanding of oneself – who we are, how we can change). Independent activities and learning motivate learning. Writing diaries should be promoted in teaching.
The first modern test of intelligence was made by Binet in 1905, for measuring child intelligence. Joy Paul Guilford believed that intelligence tests can not measure the full extent of human intelligence. They are based on convergent thinking and are not able to measure creativity, which is so important for social development. He created a model of intellectual abilities. That three-dimensional model contains 4x5x6 mutually independent factors of cognitive abilities (1959). Each factor is marked by a letter which refers to operation, content and production.Guilford interpreted and distinguished convergent and divergent thinking in his three-dimensional model.

## Visual intelligence

Visual intelligence signifies ability of seeing things in the mind, i.e. it is an ability of visual perception of world that surrounds us and of creating an artistic view of the world. Visual intelligence signifies colour, line, form and space sensibility in itself. According to Gardner (1983), people who posses visual intelligence are great collectors, who satisfy their need for visual impulses. They encircle themselves with images of their own imagination, as well as with objects that enchant them. The influence of surrounding is indisputable.
In persons with expressed visual intelligence, a verbal deficit is also noticed. Since in traditional IQ tests logical-mathematical and verbal intelligence are primarily measured, persons with great visual intelligence often have worse re-

sults. Jola Sigmond first created tests for measuring visual intelligence. According to him, each individual possesses a great potential of intelligence, which is never used to its maximum. Yet in childhood we learn that we "learn" by imitating and that we don't use all our hidden potentials in that way. By training of visual thinking, "certain muscles of the body of the mind" can be built up, creating a condition for logical thinking, three-dimensional seeing and four-dimensional solving of problems (Jola, 2004). He was the first one who used colours in semantic sense, not only for illustrations. Through line of tests, trainings, and games, it is possible to stimulate visual, mathematical and logical intelligence, practice and extend mental skills, such as logic and attention, improve mental effectiveness, develop logical understanding, improve concentration. Influence of exercise on results in intelligence tests:

- effects of exercise can be produced in three ways:
  1. by giving a particular test or its parallel forms a number of times
  2. by analysis of mistakes in doing tests
  3. by discussion about principles of solving tests or coaching
- ability of reasoning and g-factor can not be significantly improved by exercise, and positive effects are obtained only on limited specific factors
- effects of exercise are different from individual to individual
- Milwaukee project – a goal was to prevent a development of mental retardation in a group of potentially endangered children in upgrowth through a program of enrichment of family environment
- no differences are obtained in accomplishment in mathematic and reading, in relation to comparative group
- there was no real increase of g-factor
- beside the influence of different forms of exercises on the result expressed in numbers, changes in ability structure that is enhanced with those tests occurred
- empiric data about the fact that no great progress was ever realized by exercises, show that testing of intelligence is justified

During learning and thinking process, visual intelligence plays a great role in creating visual image, memorizing by means of images and, accordingly, it should be developed in as greater measure. Spatial-visual intelligence can be developed via reading geographical maps, navigation, park planning, model construction, constructive games, plotting, making itineraries, or visual maps of any information (Mindmap). The degree of visual intelligence helps students to acquire knowledge not only through static images, but also to engage their imagination, to imagine objects from different angles. Through the dynamicity of video-clip they visualize gained information. Child's spatial-visual intelligence can be deformed by non-critical acceptance of mass culture's products, i.e. by predimensional visual impulses via media, films, the Internet, videogames. Many illusions of perception are no more than simple mistakes. Ac-

cording to Helmoholz, a simple rule of each illusion is to believe that we see those objects which we could see in normal circumstances.

## Experimental research

The subject matter of this paper's research is a possibility of developing spatial-visual intelligence. We involved students of Faculty of Philosophy in Sombor and Faculty of Maritime Studies in Split in our research. Sample was made of students from The Faculty of Philosophy, from different courses:

a) design of media, third year
b) design of media, fourth year
c) elementary school teachers, third year
d) elementary school teachers, fourth year
e) students of second year, pre-school teacher
f) students of Faculty of Maritime Studies who attend course Communicology

250 students of Faculty of Maritime Studies in Split and Faculty of Philosophy in Sombor were tested. The examinees were split into five groups. The test consisted of 16 problems that identify perception of form and background in our test. The results of each group are presented in the table down (Table1).

Table 1. Results of successfully solved problems on the test

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 21,43% | 100% | 92,86% | 85,71% | 92,86% | 42,86% | 78,57% | 92,86% |
| 2 | 16,66% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| 3 | 0% | 71,43% | 66,07% | 39,29% | 64,29% | 10,71% | 17,86% | 85,71% |
| 4 | 14,29% | 100% | 85,71% | 71,43% | 100% | 14,29% | 28,57% | 100% |
| 5 | 2,94% | 91,17% | 52,94% | 14,71% | 73,53% | 5,88% | 23,53% | 50% |
| | 11,06% | 92,52% | 79,52% | 62,23% | 86,14% | 34,75% | 49,71% | 85,71% |

| 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | Total |
|---|---|---|---|---|---|---|---|---|
| 71,43% | 28,57% | 50% | 57,14% | 64,29% | 92,86% | 85,71% | 21,43% | 57,86% |
| 100% | 66,66% | 33,33% | 66,66% | 100% | 100% | 100% | 50% | 81,25% |
| 50% | 21,43% | 26,79% | 37,50% | 16,07% | 58,93% | 35,71% | 12,50% | 30% |
| 57,14% | 71,43% | 28,57% | 100% | 42,86% | 100% | 0% | 100% | 63,93% |
| 47,06% | 14,71% | 44,12% | 14,71% | 20,59% | 52,94% | 35,29% | 2,94% | 22,65% |
| 65,13% | 40,56% | 36,56% | 55,20% | 48,76% | 80,95% | 51,34% | 37,37% | 51,14% |

One of the problems that was used in the test is so called Rubin's mirage. Rubin thought that any part of view field can be seen as an image, while other parts will represent the background. That is particularly true where the image is black or white. On borders of white or black field, especially in vertical direction,

they can be read as a white shape on a black background and vice versa (two human side faces or vase). If results for the group of problems which are based on the same principle are compared – almost everybody identified the form and the background of the first problem (92.52%), the percentage was smaller in the third problem (79.52%), and the percentage is almost twice lesser in the fourth problem (62.23%) in which the identification of letters, i.e. text, complicated the problem. The fifth problem is so called Hering illusion[3], and many examinees solved it successfully (86.14%). Problem 9 and 10 identifies 3D seeing – to solve a problem, examinee has to move given objects in his imagination. Successful solving of problems 12 and 13 assumes visual imagination. Some of the problems are well known mirages (problems 5, 8 and 15). Judgment of one image greatly depends on background. Constant form indicates that form is judged by a constant, apart from size, colour or other characteristics. That means that we are not able to recognize a letter and geometric form if it is not in its usual colour, size or form (problem 8).

## Conclusion

Many social psychologists think that visual communication is one of the most important canals, if not the most important, of nonverbal communication. Visual communication refers not only to watching and eye contact, but also to seeing available and useful social signs. Visual message took over the role of verbal message in today dominant western civilization. Nonverbal communication has a lot of advantages over verbal communication, because an agent for interpretation of the meaning of the verbal message is not necessary. Human perceptive system is sensitive on images and graphic presentations of data, and in that way is able to process shown information easily. We come across terms such as infographics and data visualization/information lately. During last years, visualization is present in a field of information searching. We can also say that it is a recognizable characteristic of semantic web. Visual style of learning is dominant for students who learn the fastest when information is visually presented in a form of text. During visual learning, they mostly use information from course books and notes. Students with emphasized visual style of learning prefer to learn on their own.

Psychologist dr. Howard Gardner studied the ways in which adults and children learn for a long time. He discovered that there are different forms of intelligence. You may have musical intelligence, or social intelligence, or you are naturally intelligent. Gardner's theory of multiple intelligence changed methods

[3] The Hering illusion is an optical illusion discovered by the German physiologist Ewald Hering in 1861. It looks like bike spokes around a central point, with vertical lines on either side of this central, so-called vanishing point. The two vertical lines are both straight, but they look as if they were bowing outwards. The distortion is produced by the lined pattern on the background that simulates a perspective design, and creates a false impression of depth.

of teaching throughout the world. Encouragement and development of visual-spatial intelligence in our experiment manifested in visual perception, coding and decoding visual messages in so called semiotic transfers (for example, identification and visualization of different natural shapes).

Visual communication has two tasks. The first one is expressive, that is, it refers to transmission of attitudes and emotions. The second one is informational – it operates with social meetings and supervises them. Since exactly these functions are said to be basic, when we speak about nonverbal communication in general, we can conclude that seeing of available social signs – visual communication – is really the most important channel which enables nonverbal interactions and makes verbal easier.

Our researches have showed that visual intelligence is still not identified as an important issue that needs to be researched within study programs on faculties, and indicate that learning of visual intelligence is essential, because it is very significant for future graduates on our faculties.

## References

Bloom, B. – Krathwohl, D. R. Taxonomyof Educational Objectives, Handbook 1: Cognitive Domain. David McKay, New York, 1956

De Bono, E. CoRT Thinking Programme. Science Research Associates, Henley, 1987

Gardner, H. Disciplinirani um. // Educa. (2005)

Gardner, H., Kornhaber, M.L. & Wake W.K. Inteligencija – različita gledišta. Jastrebarsko. // Naklada Slap. (1999)

Gardner, H.: Frames of Mind: A theory of Multiple Intelligence. 1983; The Mind's New Science: A History of the Cognitive Revolution. 1985; The Unschooled Mind. 1988; Art, Mind and Brain: A Cognitive Approach to Creativity. 1982; Creating Minds. 1993. Basic Books, New York.

Jensen, E. Poučavanje s mozgom. // Educa. (2005)

Jensen, E. Super-nastava. // Educa. ( 2003)

Jola, Sigmond. Visual IQ Tests, Sterling Publishing Co.Inc, New York, 2004.

Lipman, M. Harry Stottlemeier's Discovery. Institute for the Advancement of Philosophy for Children, 1974

Mayer, R.E. Multimedia Learning. New York: Cambridge University Press. (2007)

Richter, Sigrun. Grundlinien des Unterrichts in der Grundschule der Zukunft. //In: Grundschulmagazin. 11 (1999), pp. 37-40

# List of reviewers

Vjekoslav Afrić, Faculty of Humanities and Social Sciences, University of Zagreb
John Akeroyd, South Bank University, London
Mihaela Banek Zorica, Faculty of Humanities and Social Sciences, University of Zagreb
Ana Barbarić, Faculty of Humanities and Social Sciences, University of Zagreb
Bob Bater, London
David Bawden, City University, London
Božo Bekavac, Faculty of Humanities and Social Sciences, University of Zagreb
Helen Boelens, The Netherlands
Andrew Cox, University of Sheffield
Blaženka Divjak, Faculty of Organization and Informatics, University of Zagreb
Senada Dizdar, Faculty of Philosophy, Sarajevo
Zdravko Dovedan, Faculty of Humanities and Social Sciences, University of Zagreb
Zvjezdana Dukić, Hong Kong
Alexander Fraser, Institute for Natural Language Processing, Stuttgart
Aleš Gačnik, Znanstveno-raziskovalno središče Bistra - Ptuj
Aleksandra Horvat, Faculty of Humanities and Social Sciences, University of Zagreb
Neven Kranjčec, IBM Croatia
Steven Krauwer, Utrecht institute of Linguistics UiL-OTS, The Netherlands
Jadranka Lasić-Lazić, Faculty of Humanities and Social Sciences, University of Zagreb
Nikolaj Lazić, Faculty of Humanities and Social Sciences, University of Zagreb
Maurizio Lunghi, Fondazione Rinascimento Digitale, Florence
Luisa Marquardt, Roma Tre University
Vladimir Mateljan, Faculty of Humanities and Social Sciences, University of Zagreb
Nives Mikelić Preradović, Faculty of Humanities and Social Sciences, University of Zagreb
Anja Nikolić Hoyt, Croatian Academy of Sciences and Arts
Đorđe Obradović, University of Dubrovnik
Krešimir Pavlina, Faculty of Humanities and Social Sciences, University of Zagreb
Jelka Petrak, School of Medicine, University of University of Zagreb
Andreas Rauber, Vienna University of Technology (TU-Wien)
Sanja Seljan, Faculty of Humanities and Social Sciences, University of Zagreb
Karolj Skala, Ruđer Bošković Institute, University of Zagreb

Aida Slavić, UDC Consortium
Sonja Špiranec, Faculty of Humanities and Social Sciences, University of Zagreb
Hrvoje Stančić, Faculty of Humanities and Social Sciences, University of Zagreb
Marko Tadić, Faculty of Humanities and Social Sciences, University of Zagreb
Božidar Tepeš, Faculty of Humanities and Social Sciences, University of Zagreb
Jože Urbania, Faculty of Arts, Ljubljana
Radovan Vrana, Faculty of Humanities and Social Sciences, University of Zagreb
Aleksandra Vraneš, Faculty of Philology, Belgrade University
Kristina Vučković, Faculty of Humanities and Social Sciences, University of Zagreb
Žarka Vujić, Faculty of Humanities and Social Sciences, University of Zagreb
Mirna Willer, Department of Library and Information Science, University of Zadar
Goran Zlodi, Faculty of Humanities and Social Sciences, University of Zagreb
Danijela Živković, Faculty of Humanities and Social Sciences, University of Zagreb

INFuture2009 Conference was co-organised with


INFuture2009 Conference was sponsored by