Fourteenth International Conference on Grey Literature

National Research Council, Rome, Italy 29-30 November 2012

TRACKING INNOVATION THROUGH GREY LITERATURE

Conference Proceedings

ISSN 1386-2316

Host and Sponsors

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ SR

JST

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library

LIBRARY OF CONGRESS
FEDLINK

cnrs inist
INSTITUT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE

NEW YORK ACADEMY OF MEDICINE 1847

Springer

ELSEVIER

THOMSON REUTERS

KISTI
www.kisti.re.kr

GreyNet
Grey Literature Network Service

EBSCO
PUBLISHING

GL14 Program and Conference Bureau
TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com · conference@textrelease.com
Tel/Fax +31-20-331.2420

CIP

GL14 Conference Proceedings
Fourteenth International Conference on Grey Literature : Tracking Innovation through Grey Literature, 29-30 November 2012 at CNR Rome, Italy / compiled by D. Farace and J. Frantzen ; GreyNet, Grey Literature Network Service. Amsterdam : TextRelease, February 2013. – 192 p. – Author Index. – (GL conference series, ISSN 1386-2316 ; No. 14). - CVTISR, EBSCO, FEDLINK-Library of Congress, INIST-CNRS, ISTI-CNR, NIS-IAEA, NTK, and NYAM are Corporate Authors and Associate Members of GreyNet International.

These conference proceedings contain the full text of 25 papers presented during the two days of plenary and poster sessions. The papers appear in the same order as in the conference program book. Included is an author index with the names of over 70 contributing authors/coauthors along with their biographical notes. A list of more than 50 participating organizations as well as sponsored advertisements are likewise published in these conference proceedings.

ISBN: 978-90-77484-20-3

Foreword

TRACKING INNOVATION THROUGH GREY LITERATURE

Innovation is a process manifested in and through grey literature. Both have their origins in knowledge generation and both demonstrate value for government, academics, business and industry through their uses and applications. In a way, innovation and grey literature are two sides of the same coinage. Innovation is the catalyst for positive change and grey literature is the measure of benchmarks in the further process of research and development. Innovation and grey literature share parallel life cycles in which early growth is relatively slow until their use and application become recognized both within and later beyond their community of origin. Expected top-line growth and increased bottom-line results are achieved in part through new technologies, through redeployment and enhancement of existing products and services, which at times are unachieved. Nevertheless, the process shared by innovation and grey literature carries on.

The goal of the Fourteenth International Conference on Grey Literature set out to track the process of innovation by tracing the research life cycle and observing its convergence in the field of grey literature. Thirty-five presentations from authors and researchers from 17 countries worldwide are harvested in this Program Book.

Dr. Dominic J. Farace
Grey Literature Network Service

Amsterdam,
February 2013

GL14 Conference Sponsors

CNR, National Research Council, Italy
Central Library "Guglielmo Marconi"

IRPPS, Italy
Institute of Research on Population and Social Policies
National Research Council, CNR

ISTI, Italy
Institute of Information Science and Technologies
National Research Council, CNR

FEDLINK, USA
Federal Library Information Network
Library of Congress

EBSCO Publishing, USA

INIST-CNRS, France
Institut de l'Information Scientifique et Technique;
Centre National de Recherche Scientifique

JST, Japan
Japan Science and Technology Agency

GL14 Conference Sponsors

Reed Elsevier, Netherlands

Thomson Reuters, USA

NTK, Czech Republic
National Technical Library

NYAM, USA
New York Academy of Medicine

Springer Verlag, Germany

CVTISR, , Slovak Republic
Slovak Centre of Scientific and Technical Information

KISTI, Korea
Korea Institute of Science and Technology Information

Slovak Centre of Scientific and Technical Information SCSTI

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ SR

INFORMATION SUPPORT OF SLOVAK SCIENCE

SCIENTIFIC LIBRARY AND INFORMATION SERVICES

- technology and selected areas of natural and economic sciences
- electronic information sources and remote access
- depository library of OECD, EBRD and WIPO

SUPPORT IN MANAGEMENT AND EVALUATION OF SCIENCE

- Central Registry of Publication Activities
- Central Registry of Art Works and Performance
- Central Registry of Theses and Dissertations and Antiplagiarism system
- Central information portal for research, development and innovation - CIP RDI >>>
- Slovak Current Research Information System

SUPPORT OF TECHNOLOGY TRANSFER

- Technology Transfer Centre at SCSTI
- PATLIB centre

POPULARISATION OF SCIENCE AND TECHNOLOGY

- National Centre for Popularisation of Science and Technology in Society

IMPLEMENTATION OF PROJECTS

- National Information System Promoting Research and Development in Slovakia – Access to electronic information resources - NISPEZ
- Infrastructure for Research and Development - the Data Centre for Research and Development - DC VaV
- National Infrastructure for Supporting Technology Transfer in Slovakia - NITT SK
- Fostering Continuous Research and Technology Application - FORT
- Boosting innovation through capacity building and networking of science centres in the SEE region - SEE Science

www.cvtisr.sk
Lamačská cesta 8/A, Bratislava

Table of Contents

Program

Advertisements

Appendices

Session One – Tracing the Research Life Cycle

Chair, *Donatella Castelli, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", ISTI-CNR, Italy*

Session Two: Tracking Methods for Grey Literature

Chair, *June Crowe, Information International Associates, IIA, USA*

Session Three – Adapting New Technologies

Chair, *Daniela Luzi, Institute for Research on Population and Social Policies, IRPPS-CNR, Italy*

SESSION FOUR – REPURPOSING GREY LITERATURE

Chair, *Bertrum H. MacDonald, Dalhousie University, Canada*

POSTER SESSION

Chair, *Dominic Farace, Grey Literature Network Service, The Netherlands*

GL14 Program and Planning Committee

Planning Committee:

Rita Ciampichetti, Chiara D'Arpa, Raffaella Lalle,
Adelaide Ranchino, Flavia Cancedda, and Luisa De Biagi
National Research Council, Italy

Program Committee:

Stefania Biagioni **Chair**
Institute of Information Science and Technologies, ISTI
National Research Council, Italy

Rosa Di Cesare and Daniela Luzi, **Co-Chair**
Institute of Research on Population and Social Policies, IRPPS
National Research Council, Italy

Danielle Aloia
New York Academy of Medicine, USA

Christiane Stock
Institut de l'Information Scientifique et Technique, INIST
Centre National de Recherche Scientifique, France

Roberta Shaffer
Library of Congress, USA

Petra Pejšová
National Technical Library, Czech Republic

Joachim Schöpfel
University of Lille 3, France

Blane Dessy
Federal Library and Information Center Committee, USA

Dominic Farace
Grey Literature Network Service, Netherlands

Customized OAI-ORE and OAI-PMH Exports of Compound Objects for the Fedora Repository

Alessia Bardi, Sandro La Bruzzo, and Paolo Manghi
Istituto di Scienza e Tecnologie dell'Informazione,
Consiglio Nazionale delle Ricerche, Italy

Abstract

Modern Digital Library Systems (DLSs) are based on document models which surpass the traditional payload-metadata document model to incorporate further entities involved in the research life-cycle. Such DLSs manage graphs of interconnected objects, hence offer tools for the creation, visualization and exports of such graphs. In particular, objects in the graph are exported via standard OAI-ORE and OAI-PMH protocols, encoded as (XML) "packages of interlinked information objects", also known as compound objects. Fedora is a well-known repository platform, designed to support the realization of DLSs implementing modern document models. To date, Fedora does not provide tools to customize compound object exports from DLS object graphs. This paper presents Fedora-OAIzer, an extension of Fedora which allows DLS developers to customize the structure of compound objects to be exported from a given DLS document model – expressed in terms of Fedora Content Models – and to select the OAI protocol of preference. In order to prove the completeness of the approach, Fedora-OAIzer is compared to other solutions for exporting compound objects from Fedora repositories.

1 Introduction

In the past, Digital Library Systems (DLSs) adopted "traditional" document models representing collections of payloads of digitized or born-digital material (e.g., publications, multimedia files) and their digital descriptions (i.e., metadata records). "Modern" document models enhance the traditional document model to incorporate further entities involved in the research life-cycle and semantic relationships. Consequently, modern DLSs manage graphs of interconnected information objects, called *object graph*, rather than flat collections of objects. For example, a "traditional" publication-metadata document model can be enriched with further contextual information, such as the used and generated research data.

The complexity of modern document models introduces a number of challenges concerning the way graphs of information objects are displayed, encoded, and exported across different DLSs. In particular, sub-parts of the object graph are exported, rather than each single information object or the object graph in its whole, in order to disseminate a package containing semantically related information objects. Such packages, called *compound objects*, are usually encoded in XML or other machine-readable formats and exported via standard protocols. The most used standard protocols are promoted by the Open Archives Initiative [6]: the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) [4] and OAI-ORE (Open Archives Initiative Object Reuse and Exchange) [5].

The Fedora Repository [3] is a well-known platform supporting the realization of DLS. Its data model is designed to represent modern DLS document models thanks to its graph-oriented primitives. At the time of writing, Fedora features a non-customizable OAI-PMH publisher[1] which exports only Dublin Core metadata records. An additional module supports the OAI-PMH exports of datastreams with different XML formats. Two plugins for OAI-ORE export are also available online: oreprovider[2] and Fedora2ORE[3] . Both plugins do not allow to fully customize the structure of the exported compound object.

This paper presents Fedora-OAIzer, an extension of Fedora for the export of compound objects conforming to a given portion of the underlying DLS document model through the OAI-PMH or OAI-ORE protocols. Fedora-OAIzer implements a mechanism based on the concept of "OAI view" of a Fedora document model. An OAI view is the sub-structure of the document model that developers select to customize the shape of the compound objects to export. OAIzer interprets OAI views to automatically deploy web APIs capable of exporting compound objects compatible with the given structure and according to the preferred OAI protocol.

Outline

Section 2 introduces basic concepts and defines the addressed problem. Section 3 describes the Fedora repository platform and the available tools for exporting compound objects via OAI-PMH and OAI-ORE. Section 4 presents Fedora-OAIzer and its approach that enables developers to customize compound objects by selecting sub-parts of the document model at hand. We conclude and compare Fedora-OAIzer to other existing solutions in Section 5.

2 Document models, Digital Library Systems and compound objects

A Digital Library System (DLS) [2] is a DL-oriented software serving a particular DL community. A DLS offers functionalities for the management, access and dissemination of graphs of information objects whose structure is defined by a data model called document model. A document model is a formal definition of types of entities and relationships that a Digital Library (DL) wants to manage. An entity type typically describes properties of objects in terms of name, cardinality and value type. A relationship type usually include a semantic label expressing the nature of the association and the types of entities allowed as sources and targets of the relationship. Figure 1 shows a document model defining three types of entities (Article, PDF, and Data) and the available relationships (HAS PDF, USES, USED BY, GENERATES, GENERATED BY).

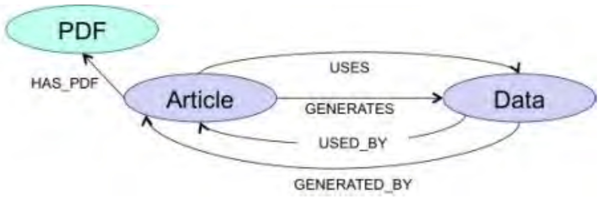


Fig. 1. A modern document model: articles linked with research data

For example, a DLS adopting the document model in fig.1, manages graphs of information objects as that in Fig. 2.

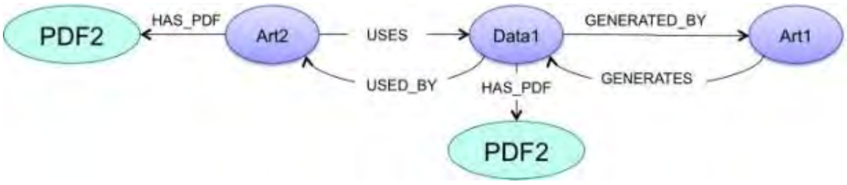


Fig. 2. An example of object graph

To clarify, it is possible to compare the above concepts with similar concepts in the relational database world. An entity-relationship model in the database world corresponds to a DLS document model. DLSs are comparable to applications realized on top of Relational Data Base Management Systems.

Since DLSs manage graphs of information objects, it is important to define the granularity of the data to export. Indeed, exporting each single information object separately (for example, exporting Art1 of Fig.2 without the generated data) leads to a loss of contextual information. Related information objects should be packaged and exported together as a single compound object, capable of maintaining all semantic relationships involving the packaged objects.

In order to exchange and re-use compound objects, interoperability issues must be tackled. The Open Archives Initiative [6] defines two standard protocols for data interoperability and information reuse: OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting [4] and OAI-ORE Open Archives Initiative Object Reuse and Exchange [5]. Both protocols allow to export compound objects of a DLS, but they adopt two radically different approaches. OAI-PMH is meant for the export of the descriptive metadata in the form of XML files; OAI-ORE is meant for the export of web-interpretable RDF representations of so-called aggregations, which are special web resources encoding compound objects.

OAI-PMH provides an application-independent interoperability framework based on metadata harvesting. Its data model has four main elements:

- Resource: an object described by one or more metadata records.
- Metadata record: XML data describing a resource. Each metadata record has its metadata format, often referred to as the XML Schema.
- Item: container of metadata records describing one resource. Each item must have at least one Dublin Core metadata record.
- Set: optional element used to group items.

In order to export via OAI-PMH, DLSs set up an OAI-PMH publisher capable of exporting a set of XML files encoding compound objects. Well-known metadata formats that can be adopted are METS[9] and XML-DIDL[8].

OAI-ORE defines standards for the description and exchange of Web resources called aggregations. An aggregation is a Web resource with its own identity and it represents a group of related Web resources.

The OAI-ORE protocol does not define an exchange protocol, but only a data model, while suggesting exchange formats such as XML/RDF and ATOM feeds. The OAI-ORE model captures four type of resources:

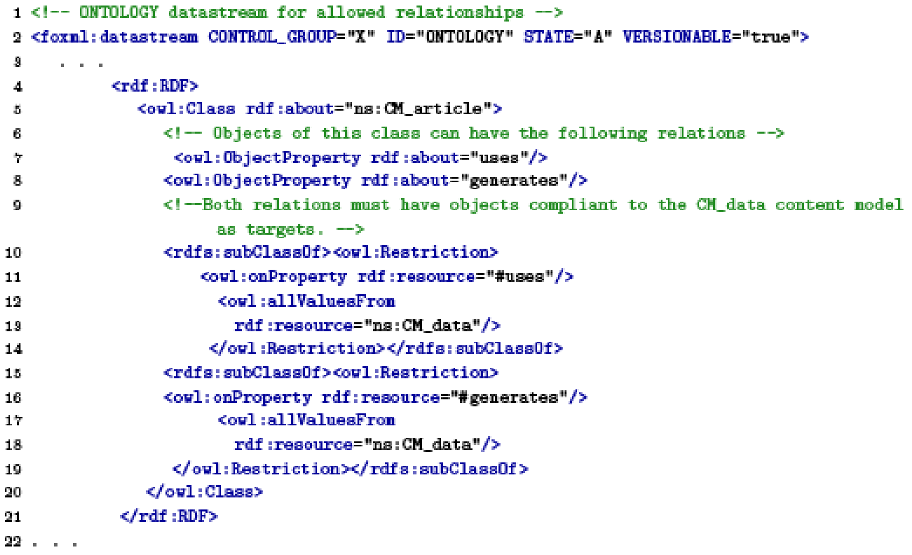
- Aggregation: resource that groups other resources, called aggregated resources.
- Aggregated resource: resource that belongs to an aggregation, that is the ORE representation of an information object in a compound object.
- Resource map: serializable description of an aggregation. A resource map:
    - lists the aggregated resources;
    - has properties about the aggregation and its aggregated resources, e.g., relationships among aggregated and other external resources.
- Proxy: resource that allows to assert relationships among aggregated resources in the context of one specific aggregation.

Among the functionalities offered by a DLS, we usually find support for data exchange according to the OAI standard protocols. However most DLSs do not provide means to customize the structure of the compound objects, but they rather fix statically the mapping between the document model and the OAI-PMH and OAI-ORE data models.

3 Fedora and exports of compound objects

Fedora (Flexible Extensible Digital Object Repository Architecture) is a well-known repository platform, designed to support the realization of DLSs. Its data model supports the representation of labelled graphs of information objects. Fedora manages objects of different kinds. “Fedora Data Objects” (FDOs) represent information objects. A FDO is composed by the following parts: an XML Dublin Core metadata record, a list of local or remote files called “datastreams”, and a list of relationships to other FDOs. The latter are serialized as RDF/XML[7] into a special datastream called RELS-EXT. “Fedora Content Models” (FCMs) are special objects devised for the definition of a document model in a Fedora instance. FCMs define constraints on the structure of FDOs, declaring which are the mandatory datastreams, relationships and operations (i.e., Web Service methods) of the FDOs compliant to that FCM. Figure 3 shows how the document model in Fig. 1 can be represented in terms of FCMs. CM article is the content model for the class article and defines one mandatory datastream called ART of mime type PDF. CM data is the content model for the class data and defines two mandatory datastreams: one for binary content called DATA, the other for the XML description of the data in DDI format[1]. By default, Fedora also includes one mandatory XML datastream called DC for metadata in Dublin Core format. The arrows between the two models represent the available semantic relationships as they are declared in the ONTOLOGY datastream of both content models. Listing 1.1 is an excerpt of the ONTOLOGY datastream for CM article.

Listing 1.1. Excerpt from the FCM for the Article class: the ONTOLOGY datastream



```
<!-- ONTOLOGY datastream for allowed relationships -->
<foxml:datastream CONTROL_GROUP="X" ID="ONTOLOGY" STATE="A" VERSIONABLE="true">
  . . .
        <rdf:RDF>
          <owl:Class rdf:about="ns:CM_article">
            <!-- Objects of this class can have the following relations -->
             <owl:ObjectProperty rdf:about="uses"/>
            <owl:ObjectProperty rdf:about="generates"/>
            <!--Both relations must have objects compliant to the CM_data content model
                 as targets. -->
            <rdfs:subClassOf><owl:Restriction>
                <owl:onProperty rdf:resource="#uses"/>
                  <owl:allValuesFrom
                    rdf:resource="ns:CM_data"/>
                 </owl:Restriction></rdfs:subClassOf>
            <rdfs:subClassOf><owl:Restriction>
            <owl:onProperty rdf:resource="#generates"/>
                  <owl:allValuesFrom
                    rdf:resource="ns:CM_data"/>
             </owl:Restriction></rdfs:subClassOf>
           </owl:Class>
         </rdf:RDF>
. . .
```

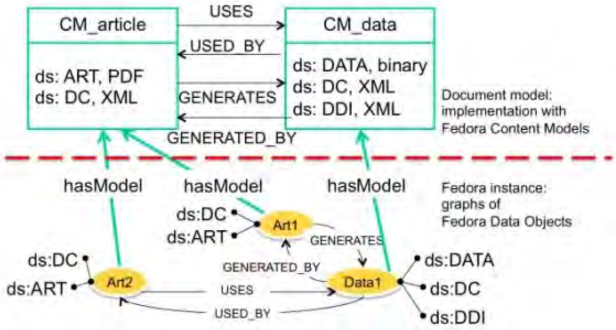


Fig. 3. Fedora content models and data objects example

3.1 Fedora and OAI

The Fedora data model is an expressive data model because its primitives allow to represent graph of information objects without imposing pre-defined structural or semantic constraints. Nevertheless, boundaries of compound objects are fixed, because Fedora's concept of compound object is that of an aggregation of datastreams. This means that in Fedora the notion of compound object matches the definition of a Fedora Object. Given the instance in Fig. 3, each of Art1, Art2 and Data1 are considered by the system as distinct compound objects. Considering a set of interconnected Fedora Objects as one compound object is not possible in Fedora without the realization of a new logic layer on top of it.

The rest of this section describes four existing solutions for the export of compound objects from a Fedora instance.

OAI-PMH Providers

Basic OAI-PMH Provider[4] is the built-in OAI-PMH provider for Fedora. It exports the mandatory DC datastream of each FDO. For the export of datastreams with other metadata formats, then the additional OAI Provider Service[5] module is required. OAI Provider Service supports any metadata format available through datastreams and interprets relationships with a given name to set up OAI Sets. Both tools provide a static mapping from the Fedora data model to the PMH data model. A FDO is mapped into an OAI-item, XML datastreams are mapped into metadata records.

OAI-ORE Providers

OREprovider[6] enables the export of FDOs as OAI-ORE aggregations with an object-oriented approach. It implements a static mapping from the Fedora data model and the OAI-ORE data model. Datastreams are mapped into ORE aggregated resource. For the generation of ORE aggregations there are two modes available. In "annotation mode" FDOs must be annotated with relationships that assert the identifier of the target ORE aggregation and the datastreams to be mapped into aggregated resources. In "auto-creation" mode each FDO is mapped into one ORE Aggregation, whilst each of its datastreams is mapped into one aggregated resource. In "autocreation mode" the tool is easy to use and no alteration has to be done to an existing repository. However, if a DLS developer wants to have control over the exported objects, FDOs must be appositely annotated, hence the document model must be aware of the special relationships exploited by OREprovider. Furthermore, the static mapping does not include the concept of ORE proxy and the tool does not provide any support for the modelling of relationships among aggregated resources.

The Fedora2ORE[7] tool adopts a navigation-oriented approach for the export of FDOs as OAI-ORE aggregations. ORE aggregations consist of one FDO together with the objects reachable by navigating its relationships up to a given depth. Fedora2ORE traverses the object graph starting from an object with a given identifier according to a variant of the breadth first search. A resource map is created to represent the visited sub-graph. The behaviour of the traversal can be customized by statically specifying in a configuration file which relationships, datastreams, FDOs are to be ignored. Each node of the resulting sub graph is an Aggregated Resource. Fedora2ORE is independent from DLS applications because there is no need to define special relationships as for OREprovider. It generates aggregations by following relationships between objects, but DLS developers can only define the boundaries of aggregations in terms of navigation depth rather than in terms of their preferred document model sub-structure. Furthermore, relationships among FDOs are not mapped into the OAI-ORE model, hence most of the semantics of the compound object is lost.

4 Fedora-OAIzer

Fedora-OAIzer is an extension for Fedora that allows to customize exports of compound objects via OAI-ORE and OAI-PMH protocols. As shown in Fig. 4, Fedora-OAIzer realizes a new layer on top of Fedora, capable of dynamically deploying OAI-PMH or ORE-ORE interfaces.

Fedora-OAIzer exploits the information about the Fedora Content Models to construct a representation of the document model of the DLS. We denote such a representation as *entity graph*. DLS developers can choose granularity, shape and properties of the compound objects to export by selecting interesting nodes and edges from the entity graph. Since the entity graph is a representation of a Fedora document model, the selected parts are a sub-structure of the document model. That sub-structure is called *OAI view* of a Fedora document model. The OAI view corresponds to the structure of the compound objects to export. OAIzer interprets OAI views to automatically setup OAI-ORE and OAI-PMH repositories. Repositories are here intended as ORE or PMH APIs available at a given URL, dynamically generated after an OAI view interpretation.

OAIzer exploits the built-in capabilities of Fedora, namely the REST APIs and the triple store, hence it is possible to plug Fedora-OAIzer in any standard Fedora instance.

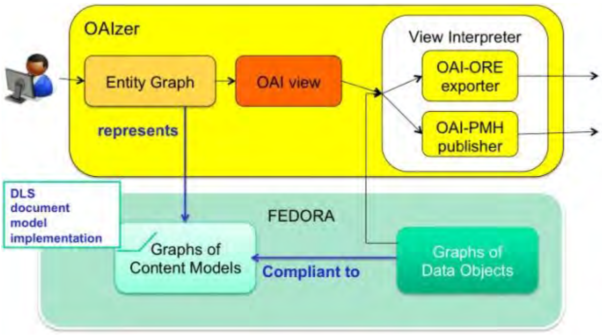


Fig. 4. View mechanism and OAI tools

4.1 Generation of the entity graph

When a DLS is based on Fedora, then developers define the document model in terms of Fedora Content Models (FCMs). Figure 5 shows an example of entity graph representing the document model of Fig.3: nodes represent classes of information objects, that is Fedora Content Models. Nodes are annotated with information about datastreams. Edges represent allowed relationships between objects of the connected classes.

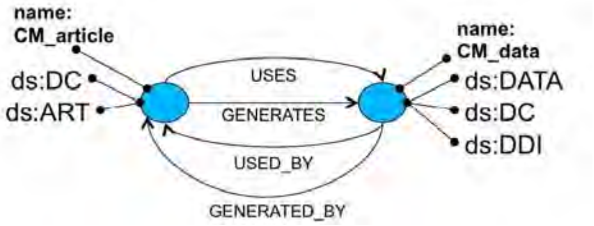


Fig. 5. An example of entity graph

Fedora-OAIzer generates the entity graph by performing the following macro steps:

1. Creation of one node of the entity graph for each FCM in the Fedora instance. The list of existing FCMs can be obtained by querying the triple store[8].
2. Obtain the full XML representation of each FCM using the REST API[9]. Information in that file is used in the next steps.
3. Enrichment of nodes with properties. Properties of a node reflect the declarations of datastreams in the corresponding content model. Such information is extracted from the standard XML datastream named DS-COMPOSITEMODEL.
4. Generation of edges. If the FCM declares a relationship to another content model, then an edge is created between the nodes corresponding to the content models involved in the relation. The edge is labelled with the name of the relationship as it is declared in the ONTOLOGY datastream (see Listing 1.1).

4.2 Definition of the view

OAIzer provides a graphical user interface where DLS developers can see the generated entity graph and define their OAI view, that is the shape of the compound objects to export by choosing the interesting nodes, properties, and edges.

The DLS developer first chooses the root node of the OAI view. The view interpreter navigates the Fedora object graph starting from every object compliant to the content model represented by the root node. After the selection of the root, the DLS developer performs iteratively the following steps until the OAI view is completed according to requirements:

- select one or more properties of the current node. The property names match the names of the corresponding Fedora datastreams: by selecting a property, the developer includes the corresponding datastream in the ORE resource relative to the current node.
- select one or more edges from the forward star of the current node. Each edge is labelled with the name of the corresponding Fedora relationship. If the DLS developer selects an edge, the target node is also included in the OAI view and an ORE relation is added between the ORE resources corresponding to the current and the target node.

The OAI view is eventually serialized into a formal language for the view interpreter.
Figure 6 shows a possible OAI view defined over the entity graph in Fig. 5. Compound objects are rooted in the class CM article and include the DC datastream of the article, the DC and DDI datastreams of the research data objects reachable through relationships labelled with USES, and the DC datastream of articles reachable from those research data objects via relationships labelled GENERATED BY.

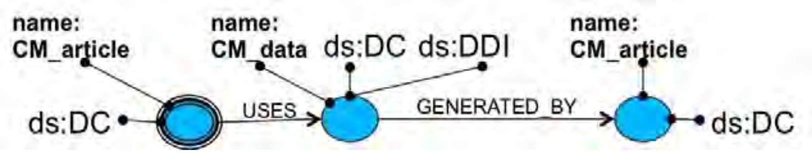


Fig. 6. An example of OAI view

4.3 Interpretation of the view

The View Interpreter is the component of OAIzer that deploys OAI-PMH and OAI-ORE APIs for the dissemination of compound objects whose structure is defined by an OAI view. The OAI view defines a "root" content model: each object compliant to the root content model is the entry point for an *instance of the OAI view.* An instance of the view is a sub-graph of the object graph. It includes all objects and relationships of one compound object.
The interpreter performs a visit on the FDO's graph for each entry point in order to get all FDOs that form an instance of the view. Moreover, the interpreter processes the FDOs of the view instance to keep track of the semantic relationships between them.
Figure 7 shows two instances of the view in Fig. 6 over the Fedora instance in Fig. 3.
For the customization of an OAI-PMH publisher, the interpreter maps the view into an OAI-PMH Set. Each instance of the view is mapped into an OAI-PMH Item with at least two metadata formats: OAIzer-XML and DC. OAIzer-XML is an idiosyncratic format for the representation of compound objects. OAIzer provides a default mapping from OAIzer-XML to DC in order to generate the DC records and be fully compliant to the OAI-PMH protocol. More metadata formats can be added by providing customized XSLT transformations from OAIzer-XML to the target metadata formats (see fig.8).

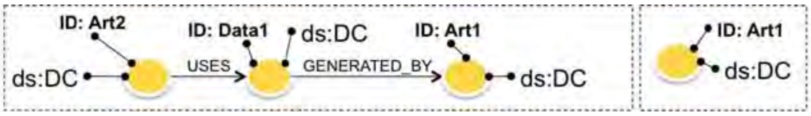


Fig. 7. Instances of the OAI view

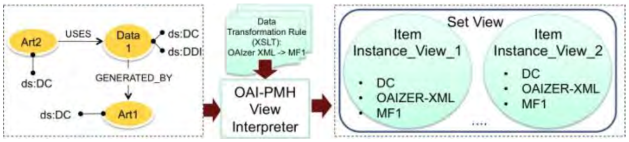


Fig. 8. Examples of OAI-PMH exports

For the customization of the OAI-ORE exporter, the interpreter creates for each instance of the view one ORE aggregation together with its corresponding resource map. For each FDO of the instance, an aggregated resource and an associated proxy are created. The aggregated resource is the OAIzer-XML representation of the FDO, including the datastreams selected in the view. Finally, the interpreter processes the FDOs of the view instance to add semantic relationships between aggregated resources. At this aim, the interpreter exploits the ORE proxy entities. The interpreter connects two proxies with a relationship r if the Fedora triple store contains the triple: fdo1 r fdo2 where fdo1 and fdo2 are the identifiers of the FDOs corresponding to the ORE proxies (see Fig. 9)

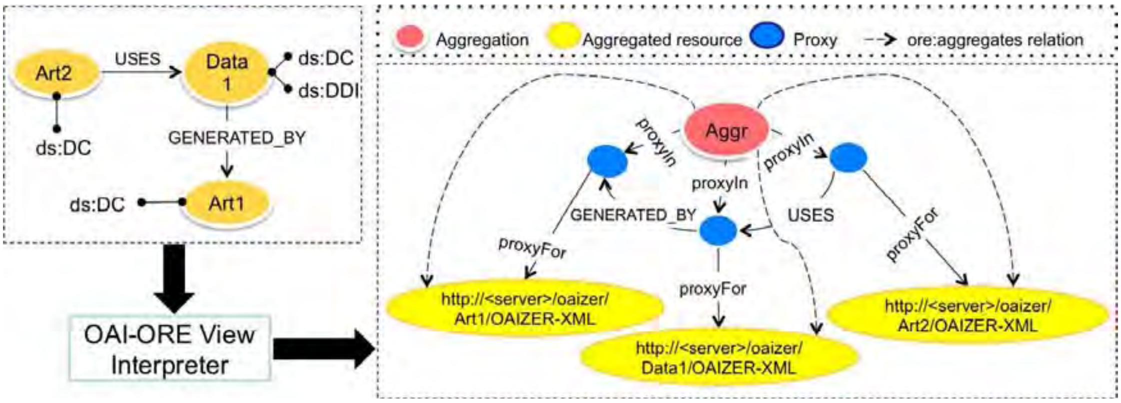


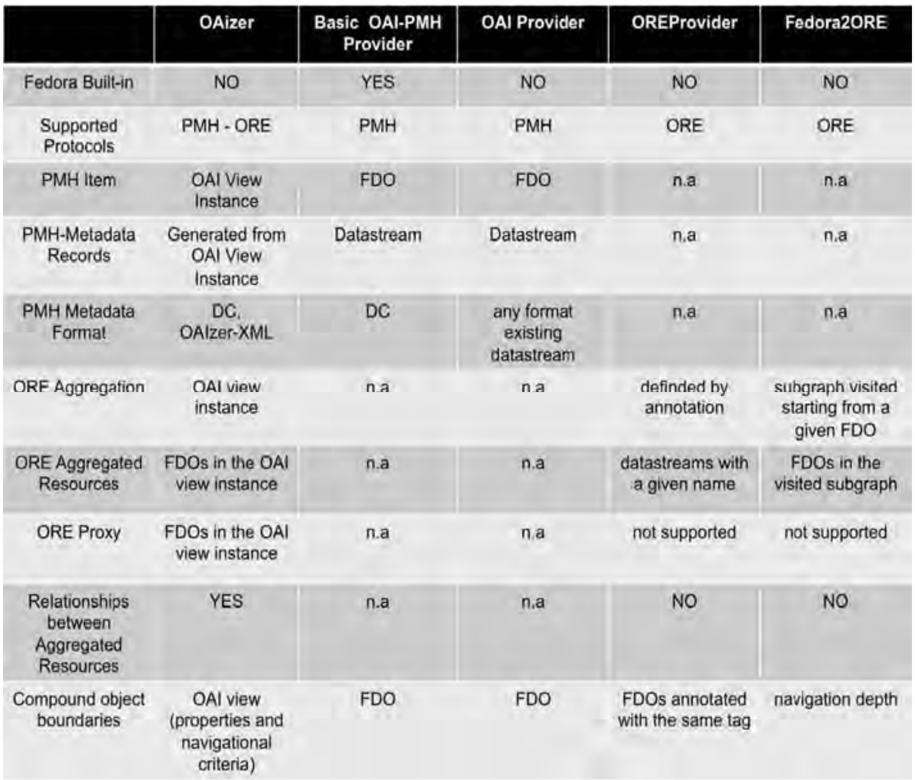
Fig. 9. Examples of OAI-ORE Aggregations

5 Conclusion

We discussed how the evolution of today's Digital Library Systems (DLSs) and the complexity of document models to represent led to data interoperability challenges. We addressed issues regarding the exchange of compound objects, i.e., packages of interlinked information objects with their own identity, among different DLSs via the standard protocols OAI-ORE and OAI-PMH.

We presented Fedora-OAIzer, a tool for the customization of OAI-PMH and OAI-ORE exports of compound object from Fedora repositories. Fedora-OAIzer creates a layer on top of Fedora in order to allow DLS developers to select the shape and boundaries of the compound objects to export. Fedora-OAIzer analyses a Fedora instance and infers the document model at hand from the Fedora content models. The document model is represented as an entity graph from which developers can select a subset of nodes and edges to define the OAI view. The OAI view defines the structure of the compound objects to export. Fedora-OAIzer then interprets the OAI view and, on demand, deploys OAI-ORE and OAI-PMH APIs delivering compound objects whose structure complies with the view definition.

Figure 10 summarizes the comparison among Fedora-OAIzer and the other existing solutions for OAI-PMH and OAI-ORE exports we described in Sect. 3.1.



	OAIzer	Basic OAI-PMH Provider	OAI Provider	OREProvider	Fedora2ORE
Fedora Built-in	NO	YES	NO	NO	NO
Supported Protocols	PMH - ORE	PMH	PMH	ORE	ORE
PMH Item	OAI View Instance	FDO	FDO	n.a	n.a
PMH-Metadata Records	Generated from OAI View Instance	Datastream	Datastream	n.a	n.a
PMH Metadata Format	DC, OAIzer-XML	DC	any format existing datastream	n.a	n.a
ORE Aggregation	OAI view instance	n.a	n.a	definded by annotation	subgraph visited starting from a given FDO
ORE Aggregated Resources	FDOs in the OAI view instance	n.a	n.a	datastreams with a given name	FDOs in the visited subgraph
ORE Proxy	FDOs in the OAI view instance	n.a	n.a	not supported	not supported
Relationships between Aggregated Resources	YES	n.a	n.a	NO	NO
Compound object boundaries	OAI view (properties and navigational criteria)	FDO	FDO	FDOs annotated with the same tag	navigation depth

Fig. 10. Comparing Fedora-OAIzer to other OAI solutions for Fedora

Acknowledgements

This work is partly funded by the European Commission as part of the projects OpenAIRE (FP7-INFRASTRUCTURES-2009-1, Grant Agreement no. 246686) and OpenAIREplus (FP7-INFRA-2011-2, Grant Agreement no. 283595).

References

1. D. D. I. Alliance. Data Documentation Initiative. http://www.ddialliance.org/.
2. L. Candela, D. Castelli, P. Pagano, C. Thanos, Y. E. Ioannidis, G. Koutrika, S. Ross, H.-J. Schek, and H. Schuldt. Setting the foundations of digital libraries: The delos manifesto. D-Lib Magazine, 13(3/4), 2007.
3. C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An Architecture for Complex Objects and their Relationships. Journal of Digital Libraries, Special Issue on Complex Objects, 2005.
4. C. Lagoze and H. Van de Sompel. The OAI Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html.
5. C. Lagoze and H. Van de Sompel. The OAI Protocol for Object Reuse and Exchange. http://www.openarchives.org/ore/.
6. C. Lagoze and H. Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries, pages 54–62. ACM Press, 2001.
7. F. Manola and E. Miller. RDF primer. Technical report, W3C Recommendation, 2004. http://www.w3.org/TR/rdf-primer/.
8. MPEG-21, Information Technology, Multimedia Framework. Part 2: Digital Item Declaration. Technical report, ISO/IEC 21000-2:2003, 2003.
9. The Library of Congress. Metadata Encoding and Transmission Standard. http: //www.loc.gov/standards/mets/, February 2002.

[1] Basic OAI-PMH Provider, https://wiki.duraspace.org/display/FEDORA36/Basic+OAI-PMH+Provider

[2] oreprovider Fedora module, http://oreprovider.sourceforge.net /

[3] Fedora2ORE, http://trac.eco4r.org/trac/eco4r/wiki/Fedora2ORE

[4] https://wiki.duraspace.org/display/FEDORA35/Basic+OAI-PMH+Provider

[5] https://wiki.duraspace.org/display/FCSVCS/OAI+Provider+Service+1.2.2

[6] http://oreprovider.sourceforge.net/index.html

[7] http://trac.eco4r.org/trac/eco4r/wiki/Fedora2ORE

[8] select ?o from <#ri> where
?o <fedora-model:hasModel>
<info:fedora/fedora-system:ContentModel-3.0>

[9] http://<fedoraServer>/objects/<id>/objectXML

Grey in the Innovation Process

Keith G Jeffery, STFC-Rutherford Appleton Laboratory, United Kingdom
Anne Asserson, University of Bergen, Norway

Abstract

The research lifecycle has multiple objectives materialised as outputs, outcomes and impacts. Typical outputs are research publications (including grey literature), patents and products such as research datasets and software, many kinds of art or prototype engineering artifacts. Outcomes include patent licence income, value of a company set up to exploit the output or trained research staff. Impacts include employment creation, a commercial product that saves lives or labour or development of a new field of knowledge and research such as genomics since the 1950s.

Commonly research in progress may be documented as grey literature – such as technical reports, laboratory notebooks or instructions for operating new equipment. There is a decision point when grey literature is produced.

One can innovate academically. The output is peer reviewed publications; the outcomes include developing trained researchers; the impact leading to a new field of research. This route provides academic recognition.

Alternatively one can innovate along the wealth-creation route. The output could be a patent; the outcome license income or a new company; the impact employment, dividends to shareholders or a new 'wonder product'. This route provides wealth and possibly improvement in the quality of life.

If research is published this usually precludes following the wealth-creation route since the novel idea is now in the public domain and not protected by patent(s).

Increasingly research funding and research performing organisations wish to demonstrate that the research they fund or do leads to impacts of relevance to society. Tracing of impacts back to the original research is not easy, partly because the eventual impact may not be known for many years. The key is an accurate recording of the research lifecycle including important dates so that the innovation cycle from idea to impact and back to further ideas can be demonstrated.

Recent work – especially in UK in the JISC-funded MICE project – has produced a taxonomy of outputs, outcomes and impacts. In parallel an extension to CERIF (Common European Research Information Format – an EU recommendation to member states) has been developed and approved by euroCRIS. This extension re-uses typical CERIF entities of persons, projects, organisations, publications, patents, products but relates them (with temporal validity and appropriate role) to the production or utilisation of outputs, outcomes and impacts. Naturally grey literature is a key component within this model.

1. Innovation

Innovation may be defined as the development of new customers' value through solutions that meet new needs, inarticulate needs, or old customer and market needs in new ways (Wikipedia). There are essentially three kinds of innovation:

1. Academic innovation: leading to new research techniques and results and providing also trained researchers;
2. Commercial innovation: using the results of research to create wealth via patents or products taken up by industry to provide a commercial product or service that generates the wealth;
3. Societal innovation: using the results of research for improvement in the quality of life in environmental, health, cultural or social aspects.

Innovation is a process with a time dimension and along that time dimension are produced outputs, outcomes and impact.

The grey literature output of research sits within a much broader context of other outputs (such as peer-reviewed publications, patents, products). Nonetheless, as argued in (Jeffery and Asserson 2007) grey literature objects are a very important output since they may lead to outcomes and impact. The outputs are not the innovation: this is achieved by utilising the outputs to produce outcomes and impact.

Outcomes are activities derived directly from the outputs: from research publications or grey literature one might have as outputs trained researchers and new research techniques. From patents (themselves considered grey literature) an outcome could be license income and possibly consultancy work (income) for the researchers to assist exploitation of the patent. From products an outcome could be the setting up of a spin-out company which employs people and produces products or services. Interestingly, many of the outcomes of research are documented by grey literature since patents, company technical reports, government internal (or external) reports are generated and these are not peer reviewed 'white literature'.

Impact is difficult to define but it is the effect the research has on society. Commonly impact is detected many years (commonly 10-15) after the research is completed. Some examples may illustrate:

1. An impact may be millions of lives saved due to a drug made available after extensive trialling, produced by a pharmaceutical company after further in house development and based on an output of research at a university.
2. Impact may be due to a policy change – possibly enforced by law – based on research. An example is the reduced deaths from lung cancer in Western society because of laws based on policies derived from research on the effect of tobacco on human health.
3. Similar examples exist in the environmental domain where research has led – in time – to policies concerning the provision of clean water supplies.

Of course in some of these cases there is associated wealth creation (e.g. for the pharmaceutical company in the first example) and associated provision of employment and hence further wealth creation.

2. BACKGROUND

2.1 Previous Work

For more than two decades, the authors have worked on research information in the widest sense comprising information not only about grey literature (grey objects) but also all the outputs of research (products, patents, publications) and the context within which the research was done including projects, organizations, funding, persons, facilities, equipment, events. Within the GL community we have highlighted the issues as we see them:

1. the need for formal metadata to allow machine understanding and therefore scalable operations (Jeffery 1999);
2. the enhancement of repositories of grey (and other) e-publications by linking with CRIS (Current Research Information Systems) (Jeffery and Asserson 2004);
3. the use of the research process to collect metadata incrementally reducing the threshold barrier for end-users and improving quality in an ambient GRIDs environment (Jeffery and Asserson 2005);
4. an architectural model for scalable, highly distributed, workflowed repositories of grey literature based on hyperactive 'intelligent' documents (Jeffery and Asserson 2006).
5. A 'from 10,000 metres altitude' view of the grey information landscape 'Greyscape' based on the hypothesis that grey literature is the foundation for the knowledge economy (Jeffery and Asserson 2007).
6. An analysis of interoperation architectures among research information systems 'INTEREST' (Jeffery and Asserson 2008).
7. A proposal that Grey Literature should be seen within the context of e-Science supported by a CERIF-CRIS (Jeffery and Asserson 2009).
8. A proposed architecture 'GLASS' using CERIF metadata to demonstrate transparency in the Grey process (Jeffery and Asserson 2010).

Although this corpus of work demonstrates how CERIF provides the required context for processing grey objects, in particular the 'Greyscape' paper (Jeffery and Asserson 2007) related Grey Literature described by CERIF metadata to the knowledge economy and is thus a relevant piece of previous work related to the current topic: Innovation.

2.2 The Requirement

The requirement is to record the innovation derived from research and in the context of this community research recorded as grey literature (or grey objects) meaning outputs not formally peer-reviewed. The innovation is recorded as outcomes and impact. However, the outputs, outcomes and impact need to be related back to the research project, the persons involved, the organisations involved (e.g. university/ faculty / department / research group or maybe a commercial company), the funding and funding organisation (including where appropriate research programme and topic), the facilities and equipment used in the research.

3. The Hypothesis

We assert that a solution – CERIF – exists already which covers these requirements. CERIF has already been in use widely in 42 countries for recording research activity and is an EU Recommendation to Member States. CERIF is maintained, developed and promoted by euroCRIS (www.eurocris.org) at the request of the European Commission. In particular, work done during the MICE (JISC-funded) project in UK extended the CERIF datamodel (which included already outputs) to include indicators and measurements which can record outcomes and impact (and aggregated outputs) (Gartner, Cox, Jeffery 2012). This proposed extension to CERIF was ratified through the euroCRIS process for inclusion in the CERIF datamodel, and provides a way to record unambiguously and in context outcomes and impact derived from research outputs – including grey literature.

4. Proposed Architecture

4.1 Introduction

The CERIF datamodel is already quite well-known in the Grey Literature Community but the overall model is reproduced here (Figure 1) to illustrate the entities that are recorded together with their relationships thus giving the context of the research.

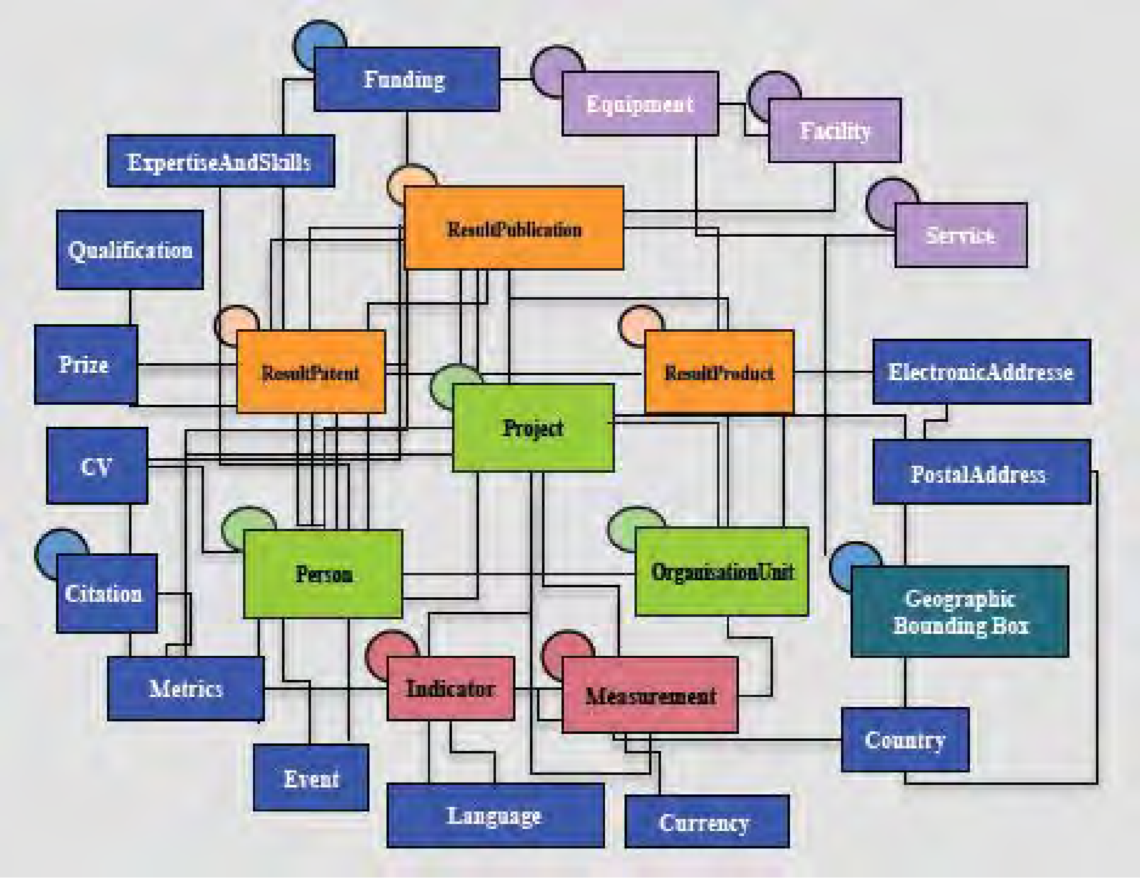


Figure 1: The CERIF Datamodel

The new entities for managing indicators and their measurements for the purposes of innovation (outputs, outcomes and impact) are indicted in the diagram above but dealt with in more detail below.

4.2 Indicators and Measurements for Outcomes and Impacts

The part of the datamodel concerning indicators and impacts is reproduced in detail here (Figure 2) and in particular it should be noted how the entities representing indicators and measurements relate to the base entities of CERIF (such as publication, patent, product and person, organisational unit, project etc) through the semantically rich temporally bound linking relations.

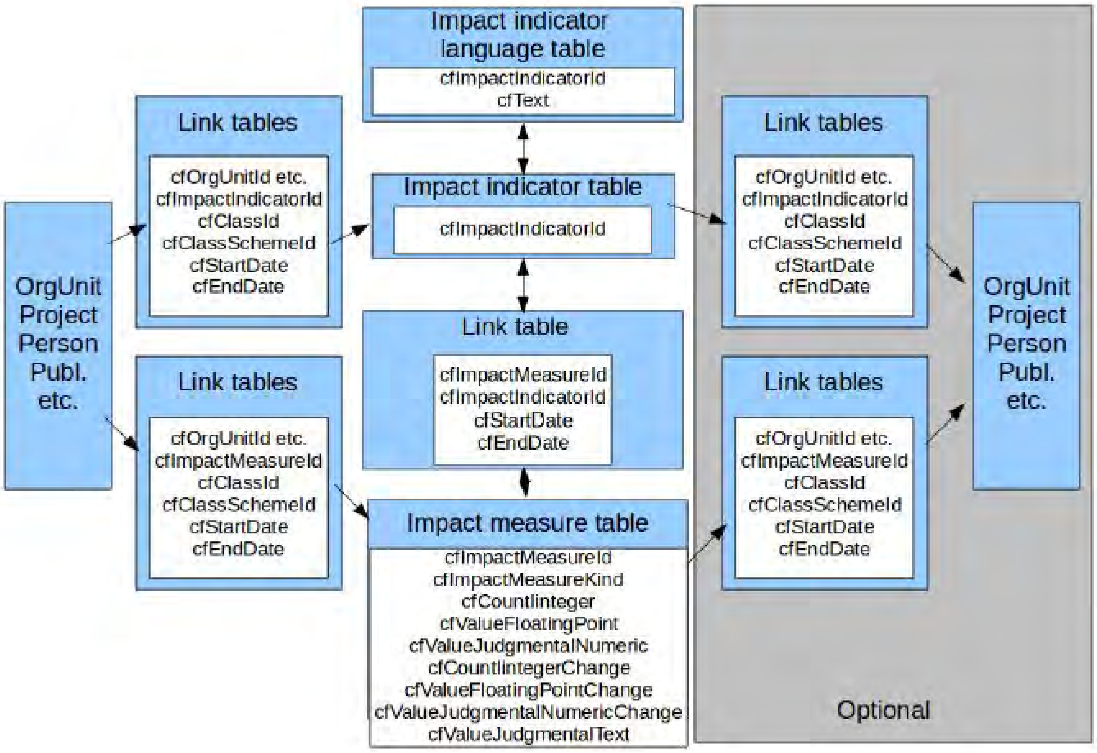


Figure 2: The MICE Datamodel within CERIF

It should be noted that the right side of the diagram (i.e. the instances of entities affected by or benefitting from the outcome or impact) is optional. Furthermore the instances of entities on the right hand side as e.g. beneficiaries of outcomes may appear on the left side as initiators of the transition from outcomes to impacts.

For each indicator there are one or more measurements. Not all measurements are relevant for every indicator; in fact usually only one measurement is appropriate for an indicator. The measurements include an integer count (e.g. how many publications were produced by a person or group in a given time period or how many lives were saved by a new drug); a floating point measurement (e.g. amount of licence income for a patent) and a judgement expressed numerically (e.g. quality on a scale of 1-10). In addition there are the 'delta' measures which record change and compare the value for one period of time with another. Examples would include the increased number of publications, the increased licence income or the improvement in quality. Finally an attribute is made available for a textual statement on judgement to justify the measure for an indicator or to express less precisely the estimated quality.

The MICE project produced a detailed taxonomy of indicators (Gartner, Cox, Jeffery 2012) that could be used, but there are others in the scientometrics and bibliometrics fields. CERIF can, of course, allow the use of any scheme of indicators due to its flexible semantic layer feature.

Another UK project, Snowball, (SnowballProject) has produced a set of indicators for university benchmarking. However, the indicators are dominantly to record performance and less to record outcomes and impact. The 'Snowball Recipe Book' (produced by Elsevier which was a project partner) (SnowballRecipes) was launched at a recent euroCRIS Members' Meeting.

The recently initiated Indicators Task Group of euroCRIS is exploring and researching the available techniques and intends to produce a canonical set of indicators and associated measurements that can be used for benchmarking and comparison across outputs, outcomes and impact. This will provide the basis for measuring innovation and especially innovation in the Grey Process.

5. Conclusion

From the above we may conclude:

1. CERIF provides an appropriate data structure for recording innovation, including within the GREY process;
2. It is being used in significant systems tracking outputs, outcomes, impact related to contextual, temporal, geospatial metadata;
3. euroCRIS has an Indicators Task Group dealing with new scientometrics (including bibliometrics) and new methods for detecting impact (backward chaining).

References

(CERIF) www.eurocris.org/cerif

(Gartner, Cox, Jeffery 2012) Richard Gartner, Mark Cox, Keith Jeffery 'A CERIF-based schema for encoding research impact' The Electronic Library xxx.yyy 2012 Emerald Publishing (in press)

(Jeffery 1999) Jeffery, K G: 'An Architecture for Grey Literature in a R and D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999 http://www.konbib.nl/greynet/frame4.htm

(Jeffery 2004b) Jeffery, K.G.; 'The New Technologies: can CRISs Benefit' in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 77-88 (available under www.eurocris.org)

(Jeffery and Asserson 2004) K G Jeffery, A G S Asserson; Relating Intellectual Property Products to the Corporate Context; Proceedings Grey Literature 6 Conference, New York, December 2004; TextRelease; ISBN 90-77484-03-5

(Jeffery and Asserson 2005) K G Jeffery, A G S Asserson 'Grey in the R&D Process'; Proceedings Grey Literature 7 Conference, Nancy, December 2005; TextRelease; ISBN 90-77484-06-X ISSN 1386-2316

(Jeffery and Asserson 2006c) Keith G Jeffery, Anne Asserson: 'Hyperactive Grey Objects' Proceedings Grey Literature 8 Conference (GL8), New Orleans, December 2006; TextRelease; ISBN 90-77484-08-6. ISSN 1386-2316 ; No. 8-06-X

(Jeffery and Asserson 2007) Keith G Jeffery, Anne Asserson: 'Greyscape' Opening Paper in Proceedings Grey Literature 9 Conference Antwerp (GL9) 10-11 December 2007 pp9-14; Textrelease, Amsterdam; ISSN 1386-2316

(Jeffery and Asserson 2008) Keith G Jeffery, Anne Asserson: 'INTEREST' Proceedings Grey Literature Conference Amsterdam 8-9 December 2008 in Tenth International Conference on Grey Literature : Designing the Grey Grid for Information Society, 8-9 December 2008, Science Park Amsterdam, The Netherlands ed. by Dominic J. Farace and Jerry Frantzen ; GreyNet, Grey Literature Network Service. - Amsterdam : TextRelease, February 2009. GL-Conference series, ISSN 1386-2316; No. 10. - ISBN 978-90-77484-11-1.

(Jeffery and Asserson 2010) Keith G Jeffery, Anne Asserson 'GL Transparency: Through a Glass, Clearly' Proceedings Twelfth International Conference on Grey Literature, Prague December 2010 Ed Dominic Farace and Jerry Frantzen 95-100, TextRelease Amsterdam Series ISSN 1386-2316; no 12; ISBN 978-90-77484-16-6

(SnowballProject) http://www.snowballmetrics.com/

(SnowballRecipes) http://www.snowballmetrics.com/wp-content/uploads/Snowball-Metrics-Recipe-Book.pdf

(Wikepedia) www.wikepedia.org

Characteristics and use of grey literature in scientific journals articles of Algerian researchers: Case study of University of Science and Technology Houari Boumediene (Physics, Chemistry and Computer Sciences)

Lydia Chalabi,
DRDSI, CERIST research center, Information Science Department, University of Algeirs 2, Algeria

Abstract

This paper examines grey literature in research articles of Algerian teachers-researchers across the STM(Physics, Chemistry & Computer science) sciences at Algerian University of Science and Technology "USTHB". Each of these disciplines is respectively high ranked according to the report of Algerian general directorate of scientific research and technological development about Top emerged disciplines in Algeria. The purpose of the study is to reports the important and use of electronic resources particularly grey literature resources in their scientific journal papers and to identify the characteristics of these kinds of resources. The methodology is based on quantitative and qualitative components. The quantitative part of our sample consists of five articles per year for each discipline: physic, chemistry, computer science. 1028 references were examined; derived from 60 Algerian teachers- researchers articles journals published in international scientific journals and reflected in Scopus. The study attempts to provide insights into the characteristics of grey literature in a range of disciplines.

Keywords: Grey Literature, Open Access, Algerian University of Science and Technology Houari Boumediene (USTHB), Algerians researchers, Algerians teachers, STM, Physics, Chemistry, Computer science.

Introduction

The advent of the World Wide Web and the Internet in particular has a huge impact on the publishing of science where new forms of circulation and access to knowledge have emerged namely open access, as well as , a strong need to assess the effectiveness of scientific research has emerged. The latter must be consistent with current communication modes knowledge and the changing of users needs.

In reality, issues of significant amounts allocated to electronic resources integrated into the academic sphere and for which a return on investment analysis and is indispensable, have allowed the development of studies uses in order to guide the offer and services to offer [1].

However, when studies of electronic resources uses and open access platforms proliferate, it should be noted, that few studies, in this regard, on the ground in developing countries in general and Algeria in particular.

In fact, the Algerian scientific research makers are aware that the evaluation of scientific research is an absolute necessity [2], where a favorable policies with infrastructure and digital services to promote a national policy research and development was created. We also notice several incentive and financing scientific research programs and investment in some large commercial scientific publishers (Boukacem, 2010).

Therefore, among of documents used by researchers which belong to the so-called grey literature consisted of a set of scientists variable contents where the most are academic. This type of literature is not easily accessible, especially for developing countries that many studies and researches are made on issues of vital importance to the growth and development, but which are not widely distributed and difficult to find [3].

In this regard, our study focuses on the analysis of teachers-researchers patterns in physics, chemistry and computer science, at University of Science and Technology Houari Boumediene « USTHB » - first Algerian university in terms of publications (Thomson Reuters in January 2012) - in terms of information retrieval, characteristics and place of the grey literature in the research activity and communication results, in the era of electronic resources pay and open access.

What are the Algerian researchers habits regarding information retrieval?

How important and characteristics is grey literature publications in scientific research?

Interest and objective of study

This study is the first of its meaning, it can be seen as a contribution to stimulate interest in an issue that we consider important to the Algerian researchers, it also concerns the assessments of research institutions to decide how to orient and improve their acquisitions, information policy, and information dissemination. It also contributes to study the impact of open access on publication of Algerian researchers.

Hypothesis

To achieve these objectives, we made two assumptions that will form the basis for the collection of information:

- Grey literature is used in the research of Algerian researchers, particularly in scientific articles.
- The Algerian researchers use, much more, the open access grey literature retrieved on the web.

Methodology

Our study sample is Algerian University of Science and Technology Houari Boumediene "USTHB" first active university in terms of publications according to the report of Algerian leadership of research and development about 10 Top Algerian universities [4]. Therefore the discipline choice is focused on the physics, chemistry and computer science, ranked among the top 10 emerged discipline in Algeria [5].

Our methodology consists of two parts: quantitative and qualitative.

Quantitative part

Our sample is consisted of five articles per year for 2009 to 2012.

These journals articles are referenced in one of the biggest databases of citations and referencing known worldwide « Scopus » [6].

The selection of items is based on the following points:

- Period: 2009-2012
- Most cited journals articles
- First author affiliated to USTHB

1028 references were examined and derived from 60 Algerian teachers- researchers articles journals published in international scientific journals and referenced in Scopus for 2009 to 2012.

Quantitative part

After the quantitative part, it was necessary to find "on the ground" qualitative confirmation of diversity and convergence of practices observed among the researchers interviewed.

In this section it was discussed to achieve semi-structured interviews with a sample of 12 teachers, researchers, men and women, PhD, lecturers and teachers at Physic, chemistry and computer science faculties of Algerian University of Science and Technology Houari Boumediene "USTHB"

The interview grid was based on the following:

- research and teaching disciplines
- time devoted to research and teaching
- habits and practices regarding information retrieval
- use and characteristics of grey literature

The interviews lasted between 40 minutes and 1 hour 30 minutes on October 2012 to November 2012 in offices of faculty, which has allowed reading the material environment and organizational documentation and resources used.

During these meetings, teachers-researchers explained their habits and practices related the information research: site consulted characteristics and the materials used...etc.[... they show reasoning, arbitrations, strategies and objectives that underlay](Boukacem, 2010).

This part has also been able to shed light on the context in which teachers-researchers work at university and understand if it affect or not the integration of resources in their research.

However the biggest obstacle that we faced was to meet teachers to achieve the interviews. Obtaining the contact details of researchers has been very difficult. The latter are not always visible on directory of researchers and if they exist, not always up dated for both appointments.

Results and discussion

The goal of our study is to define the use and understanding of the researchers habits looking for information, how electronic resources are received, how important grey literature is in their scientific publications.

1. Patterns in information research

The majority of interviewed teachers-researchers admit having used the web to find the scientific literature from their homes or internet cafes in particular. This is due of connection problems at the university and if it exists, it does not permit necessarily a good information retrieval and the download of full text because of its slow speed. The interviews confirmed that the almost exclusive use of Google as a search engine and allowed to understand that consultation electronic resources platforms is limited and insufficient. Contingency access provided by institutions (difficulty, slow connections) is cited as an obstacle to the use of electronic documentation. Also for the Algerian online national system of documentation"SNDL"[7], the interviewers report that over its shortcomings in terms of supply and services, adding connection problems that drive most of the time to give up.

Another aspect has to be emerged from this analysis, that teachers-researchers manage their information research, by using Google for the supervision of doctoral thesis, through contacts with colleagues or an internship. We noticed a random and inefficient information search which indicates a web training needs as well as platforms of electronic resources.

2. Place and characteristics of gray literature

According to the quantitative information as shown below the use of grey literature differs from discipline to another. The higher level citation of this kind of literature comes from computer science and the last one form chemistry. (tab.1)

This difference is due to the characteristics of information research for each discipline. The Interviews related with what we said earlier, in some discipline, researchers prefer simple research, which guarantee wider answers, even if they are not well organized.

In other discipline, as chemistry, the interviews allowed us to understand that the majority of teachers-researchers prefer to include items most cited and not difficult to identify.

The teachers-researchers have a limited time to read, they have a teaching activity parallel to them research, they have administrative and educational responsibilities and it have yet to publish regularly in places recognized by the community.

The information we have gathered during the interviews shows clearly that this practice is more prevalent among the more experienced users, with scientific and administrative responsibilities, and combining the roles of "reader-author-reviewer".

Therefore the characteristics of grey literature used, is consisted to five documents type: thesis, conference, working paper, reports and data collection, where the conference take the majority of uses. (tab.2). For the accessibility and grey literature full text retrieval, qualitative and quantitative results revealed that most of full text can be found on the web freely. (tab.3)

Table 1. Volume of grey literature references

Year	Physic			Chemistry			Computer science		
	N° of all references	N°of grey literature references	Percentage	N° of all references	N°of grey literature references	Percentage	N° of all references	N°of grey literature references	Percentage
2009	94(100%)	9	9,57 %	62(100%)	1	1,62 %	57(100%)	24	42,06 %
2010	90(100%)	6	6,66 %	68(100%)	4	5,88 %	120(100%)	45	37,5 %
2011	79(100%)	4	5,06 %	99(100%)	1	1,01 %	102(100%)	23	22,54 %
2012	75(100%)	8	10,66 %	68(100%)	0	0 %	114(100%)	50	43,85 %

Table 2. Characteristics of grey literature

Discipline	Year	N° of GL references	N° of conference	N° of thesis	N° of rapports	N°of working papers	N° of data collection
Physic	2009	9(100%)	88,88%	11,11%	0 %	0 %	0 %
	2010	6(100%)	66,66 %	16,66 %	0 %	16,66 %	0 %
	2011	4(100%)	75 %	0 %	25 %	0 %	0 %
	2012	8(100%)	87,5 %	0 %	12,5 %	0 %	0 %
Chemistry	2009	1(100%)	1 %	0 %	0 %	0 %	0 %
	2010	4(100%)	50 %	0 %	25 %	0 %	25 %
	2011	1(100%)	100 %	0 %	0 %	0 %	0 %
	2012	0(100%)	0 %	0 %	0 %	0 %	0 %
Computer science	2009	24(100%)	79,16 %	20,83 %	0 %	0 %	0 %
	2010	45(100%)	77,77 %	11,11 %	6,66 %	4,44 %	0 %
	2011	23(100%)	95,65 %	0 %	4,34 %	0 %	0 %
	2012	50(100%)	82 %	14 %	2 %	2 %	0 %

Table 3. Accessibility of documents

Discipline	Year	N° of GL references	Accessibility by a pay platform	Open access	unpublished
Physic	2009	9(100%)	3(33,33%)	3(33,33%)	3(33,33%)
	2010	6(100%)	1(12,5 %)	1(12,5 %)	4(66,66%)
	2011	4(100%)	1(25%)	3(75 %)	0 %
	2012	8(100%)	4(50 %)	4(50 %)	0 %
Chemistry	2009	1(100%)	0 %	0 %	1 %
	2010	4(100%)	0 %	3(75 %)	1(25%)
	2011	1(100%)	0 %	0 %	1(100 %)
	2012	0(100%)	0 %	0 %	0 %
Computer science	2009	24(100%)	9(37,5 %)	10(41,66 %)	6(25 %)
	2010	45(100%)	10(22,22%)	20(44,44%)	15(33,33%)
	2011	23(100%)	5(21,74 %)	15(65,21 %)	3(13,04 %)
	2012	50(100%)	13(26 %)	30(60 %)	7(14 %)

Finding

This is an exploratory study that can be considered difficult in this type of work.Even the limits; this study reveals characteristics of Algerian teachers-researchers practices regarding the information research. The reality is that:

Even quantitative data reveals the use of electronic resources, difficulties that the teachers- researches could face, show a marked lack of uniformity due to the environment in which they work and consult electronic resources. Like autonomy and isolation that leads to develop unraveling practices that is not homogeneous and inactive. Despite the offer content which it’s not enough.

As was clearly indicated by Schöpfel (2010) and Chalabi (2012), the grey literature has found a means of communication on the web especially in institutional open repository, but its uses differ from discipline to another. More grey literature studies covering other developing African nations are necessary.

References

[1] K. Salima, H. Hakim, D. Samia, *Revue RIST* **18**, 7 (2010).

[2] C. Boukacem-Zeghmuri, Abd-Allah Abdi, Mohamed Ben Romdhane, *usages des ressources électroniques dans pays du Maghreb*, C . BOUKACEM-ZEGHMOURI, ed. (ADBS éditions, 2010), pp.281-300.

[3] P. Muswazi, *journal of special libraries* **35**, 217 (2001).

[4] D. G. de la Recherche Scientifique et du Développement Technologique Algérienne, Disciplines émergentes en Algérie : TOP 10, *Tech. rep.* (2012).

[5] D. G. de la Recherche Scientifique et du Développement Technologique Algérienne, Top 10 universités algériennes.

[6] Scopus: http://www.scopus.com/home.url

[7] SNDL: https://www.sndl.cerist.dz/

[8] J. Schöpfel, Hélène Prost, *Les statistiques d'utilisation d'archives ouvertes : Etat de l'art ,* C . BOUKACEM-ZEGHMOURI, ed. (ADBS éditions, 2010), pp.147-164.

[9] C. Lisée, V. Larivière, E. Archambault, *J. Am. Soc. Inf. Sci. Technol.* **59**, 1776 (2008).

[10] C. Boukacem-Zeghmouri, *Documentaliste-Sciences de l'Information* **47**, 4+ (2010).

[11] L. Zhang, *College & Research Libraries* **72**, 167 (2011).

[12] I. U. Rajgoli (NISCAIR-CSIR, India, http://nopr.niscair.res.in/handle/123456789/13484, 2011), vol. ALIS Vol.58.

[13] S. Halima, *Revues Sciences Humains* pp. 77–83 (2006).

[14] D. J. Brown, *Aslib Proceedings* **62**, 112 (2010).

[15] I. Derfoufi, *The Candian Journal of Information and Library Science* **36**, 122 (2012).

[16] J. Schopfel, ed., *La publication scientifique : analyses et perspectives*, Traité des sciences et techniques de l'information (Lavoisier, 2008).

[17] J. Schöpfel, C. Boukacem-Zeghmouri, *Grey Literature in Library and Information Studies*, D. Farace, J. Schöpfel, eds. (De Gruyter Saur, 2010), pp. 227–238.

[18] L. Zhang, *College & Research Libraries* **72**, 167.

Appendix

Links offering free full text

http://www.pnas.org
http://scripts.iucr.org/
http://rspa.royalsocietypublishing.org
http://arxiv.org
http://scitation.aip.org
http://royalsocietypublishing.org
http://www.aipuniphy.org
http://scitation.aip.org
http://eresearch.qmu.ac.uk/806/
http://www.assta.org
http://www.afcp-parole.org/doc/Archives_JEP
http://www.svms.org/learnability
https://online.tugraz.at
http://wam.inrialpes.fr
http://reference.kfupm.edu.sa
http://www.cnbc.cmu.edu
http://www-clips.imag.fr
http://www.speech.kth.se
http://www.aipuniphy.org/Portal/Portal.aspx
http://www.clean-auto.com/spip.phparticle1334
http://www.soe.uoguelph.ca
http://www.irisa.fr/
http://www.dtic.mil
http://www.asel.udel.edu
http://www.speech.kth.se
http://acustica.ing.unibo.it/
http://ijrte.academypublisher.com
http://research.microsoft
http://eresearch.qmu.ac.uk/806/

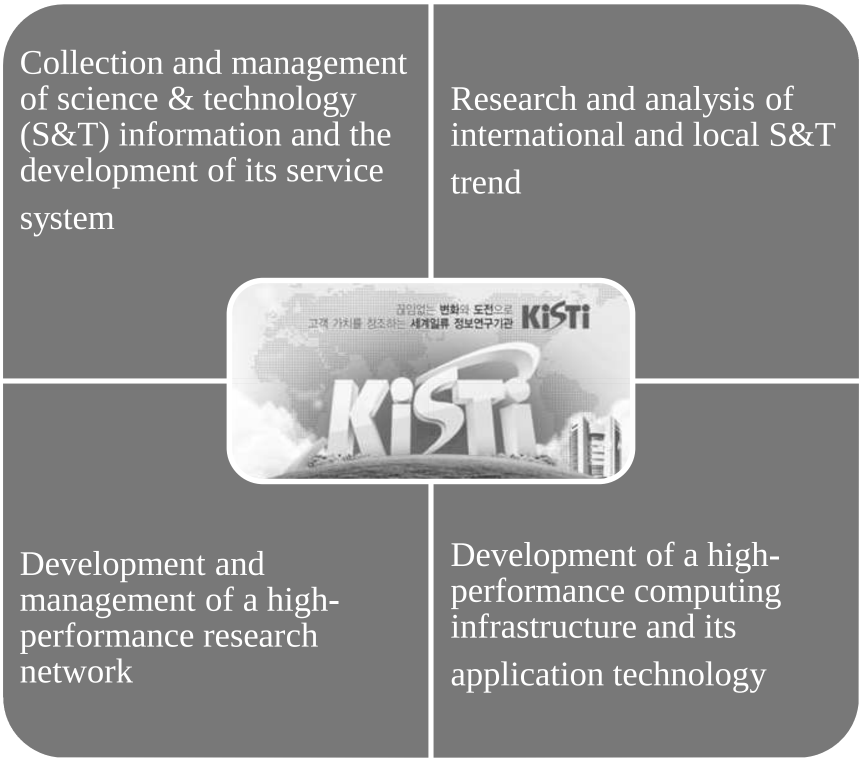
Korea Institute of Science and Technology Information (KISTI)

English version - http://en.kisti.re.kr/

*** Vision**

World-class information research institute creating values for customers

*** Main functions**



Collection and management of science & technology (S&T) information and the development of its service system

Research and analysis of international and local S&T trend

Development and management of a high-performance research network

Development of a high-performance computing infrastructure and its application technology

*** Management and service of Korean R&D reports**

KISTI exclusively manages, preserves, and serves Korean R&D reports for citizens and government officials. It provides Korean R&D reports and their information with National science & Technology Information Service (NTIS) and National Discovery for Science Leaders (NDSL).

***Contact information**

KISTI email address: hcpark@kisti.re.kr

Headquarters: Tel : +82-42-869-1004, 1234 Fax: +82-42-869-0969

The research life cycle and innovation through grey literature in nanotechnology in Korea

Seon-Hee Lee and Hye-Sun Kim
Korea Institute of Science and Technology Information, KISTI, Korea

Abstract
This paper studied the research life cycle of nanotechnology (NT) and its innovation through grey literature in the form of technical reports in Korea. The changes in the numbers of publications of grey literature and white literature in NT show the process of innovation, as technical reports and journal articles contain research results. Numbers of publications of Korean technical reports on National Discovery for Science Leaders (NDSL) and journal articles on Web of Science Science Citation Index Expended (WoS (SCIE)) are compared year by year. In general, the technical reports were produced at an earlier date than journal articles on the Web of Science. Grey literature contains creative ideas and research output, and reflects the early stage of nanotechnology and thus provides a means of tracing innovation in a specific field of science and technology in Korea.

1. Introduction

Grey literature such as technical reports, conference proceedings, etc. contains creative ideas, suggestions and research results. Researchers in science and technology fields use them for literature reviews and produce them to share their research output in the research life cycle. Scientists and researchers conduct R&D and write technical reports and then publish the research output in domestic and overseas journals. In this process, innovation in a certain field can be traced through an analysis of grey literature. The process of nanotechnology development can be revealed through comparison of literature publications such as grey literature and white literature.

Nanotechnology emerged in the 1980's and has become one of the six technologies funded by the Korean government. Nanotechnology in Korea reflects the nation's innovation in science and technology. A similar situation can be seen in other developed countries. MIT selected 'nanopore sequencing' as one of 10 emerging technologies for 2012 and the World Economic Forum chose 'nanoscale design of materials' as one of the top 10 emerging technologies for the same year. Nanotechnology became one of leading technologies to change world.

This paper verified that the grey literature has contributed innovation of nanotechnology in Korea through case study. The goal of this study is to analyze the research life cycle of nanotechnology (NT) and trace innovation in Korea through grey literature in Korea. First, the Korea Institute of Science and Technology Information (KISTI) conducted in-depth interviews and close observations of 24 researchers working in NT to analyze the NT research life cycle. Second, to trace innovation through grey literature, the numbers of publications of Korean technical reports in NT appearing on National NDSL and journal articles on Web of Science Science Citation Index Expended (WoS (SCIE)) were countered and compared. The numbers of technical reports funded by Korean government on NDSL and journal articles written by Korean researchers on WoS (SCIE) from 1980 to 2011 were counted and compared. I assumed that the changes of numbers of publications of grey literature and white literature in nanotechnology show the process of innovation, because technical reports and journal articles contain research results. Research output is often evaluated by numbers of publications in renowned journals listed on Web of Science, Scopus, etc. The research output in the literature can be read, highly cited, and developed by other researchers.

2. The Research Life Cycle in Nanotechnology in Korea

2.1. Needs of analysis of the research life cycle in Korea

As a national information center for science and technology, KISTI should be aware of the information environment of domestic researchers and provide a stable system. To prepare for the changeable information environment, recognition of the R&D research life cycle is important. The R&D research life cycle in nanotechnology is to reveal the Korean situation.

According to Encyclopedia Britannica, 'nanotechnology' is the manipulation and manufacture of materials and devices on the scale of atoms or small groups of atoms. The 'nanoscale' is typically measured in nanometers, or billionths of a meter (*nanos*, the Greek word for "dwarf," being the source of the prefix), and materials built at this scale often exhibit distinctive physical and chemical properties due to quantum mechanical effects. Although usable devices this small may be decades away (microelectromechanical system), techniques for working at the nanoscale have become essential to electronic engineering, and nanoengineered materials have begun to appear in consumer products such as nano-silver toothbrush, nano laundry detergent, etc. The field of nanotechnology and nanoscience covers a broad area of expertise. Classical fields of physics, chemistry, material science, electrical/mechanical/chemical engineering, and medicine, are all involved in the new field of nanoscience. Furthermore research and development in this area is naturally multi-disciplinary. Nanotechnology includes nanoelectronics, nanomechanics, nanomaterials, nanomedicine, bionanotechnology, etc.

2.2 Conducting in-depth interviews and close observations

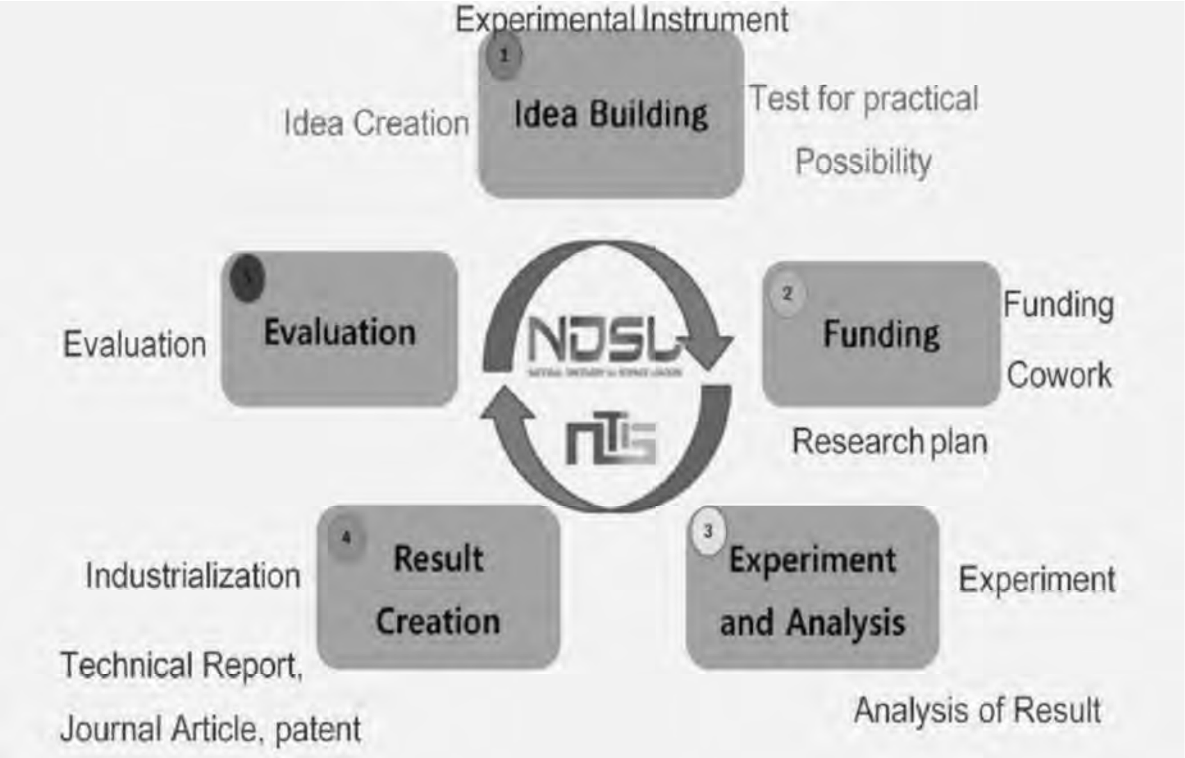
In-depth interviews and close observations were conducted in 2011 to shed light on the R&D research life cycle in nanotechnology. Twenty-four researchers working in the field of nanotechnology at the universities and research institutes were interviewed. The interviews were conducted from March to April 2011 for a month.

2.3. **Analysis of research life cycle**

According to the study, the research life cycle in nanotechnology can be divided into 5 stages (Fig. 1): idea building, funding, experiment and analysis, result creation, and evaluation. In idea building stage, researchers create idea, design experiment with instruments, and then test for practical possibility. In funding stage, researchers look for funding, write research plan and proposal, and seek out coworkers. In experiment and analysis stage, researchers conduct experiment and analyze research results. The results of creation appeared in technical reports, journal articles, and patents and became industrialized. The last stage of the research life cycle is the evaluation. In evaluation stage, research results are evaluated by funding agencies or other researchers. The stage doesn’t end in the evaluation but it is influencing the new idea building stage for new projects. Therefore new research life cycle will begin. Research drives innovation in science and technology as well as human life. Research is undergoing revolution.

Needs are different in every stage of the research life cycle but needs for literature reviews through technical reports, patents, trends, and journal articles are evident in every stage of the research life cycle. The researchers use grey literature and white literature in every stage of the research life cycle and also produce both forms of literature.

<Fig. 1> the Research Life Cycle in NT



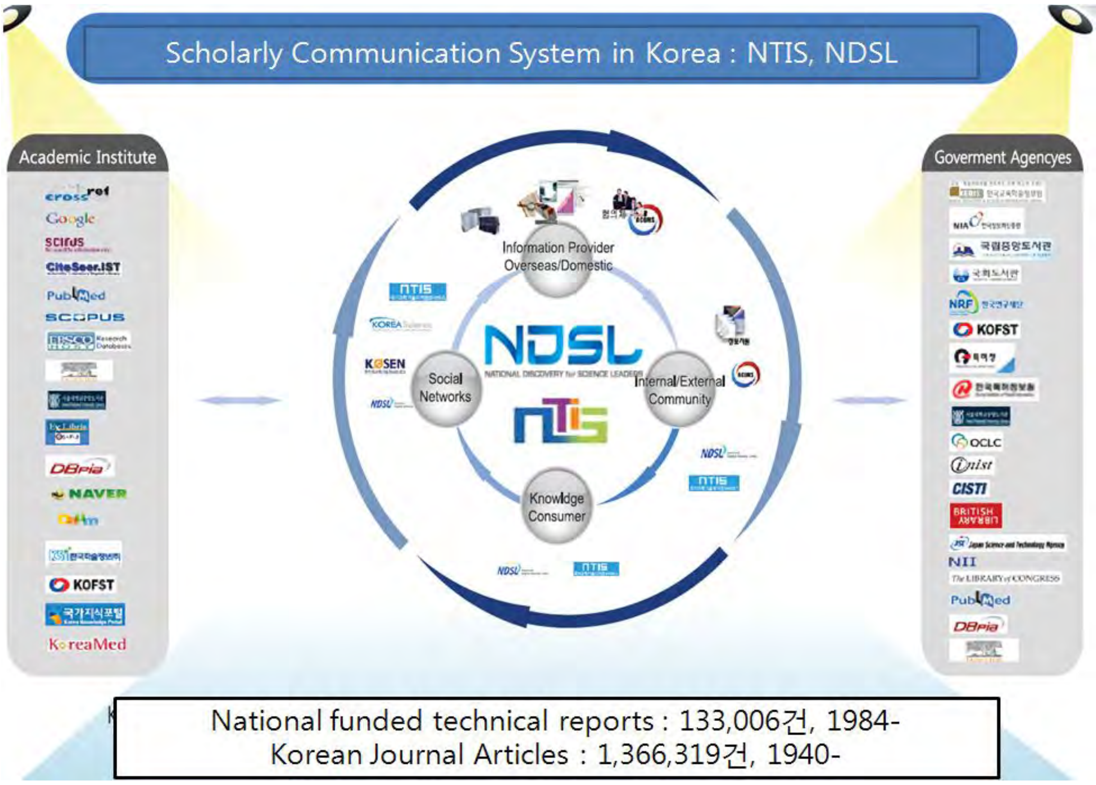
3. Tracing innovation through grey literature

3.1 Scholarly Communication System based on NDSL and NTIS in Korea

The scholarly communication system in Korea is based on National Science and Technology Information System (NTIS) and National Discovery for Science Leaders (NDSL). The research results by scholars and researchers are collected in the forms of technical reports of national research and development projects through NTIS and journal articles are archived through the Article Contribution Management System (ACOMS). Researchers are should upload their technical reports and journal articles on the systems via the internet directly. The grey literature and white literature uploaded then provided to the public through the NDSL portal system. NDSL provides 133,006 titles of technical reports funded by the Korean government since the 1980'- and 1,366,319 Korean journal articles since 1940'-. NDSL also provides other domestic and overseas science and technology information such as patents, standards, fact data, etc. Researchers are both knowledge creators and consumers for NDSL, NTIS, and other domestic and overseas networks. They also communicate through internal and external communities and social networks in the research life cycle to share ideas and new discovery.

.

<Fig. 2> Scholarly Communication System in Korea through NTIS and NDSL (As of 2012)



Scholarly Communication System in Korea : NTIS, NDSL

National funded technical reports : 133,006건, 1984-
Korean Journal Articles : 1,366,319건, 1940-

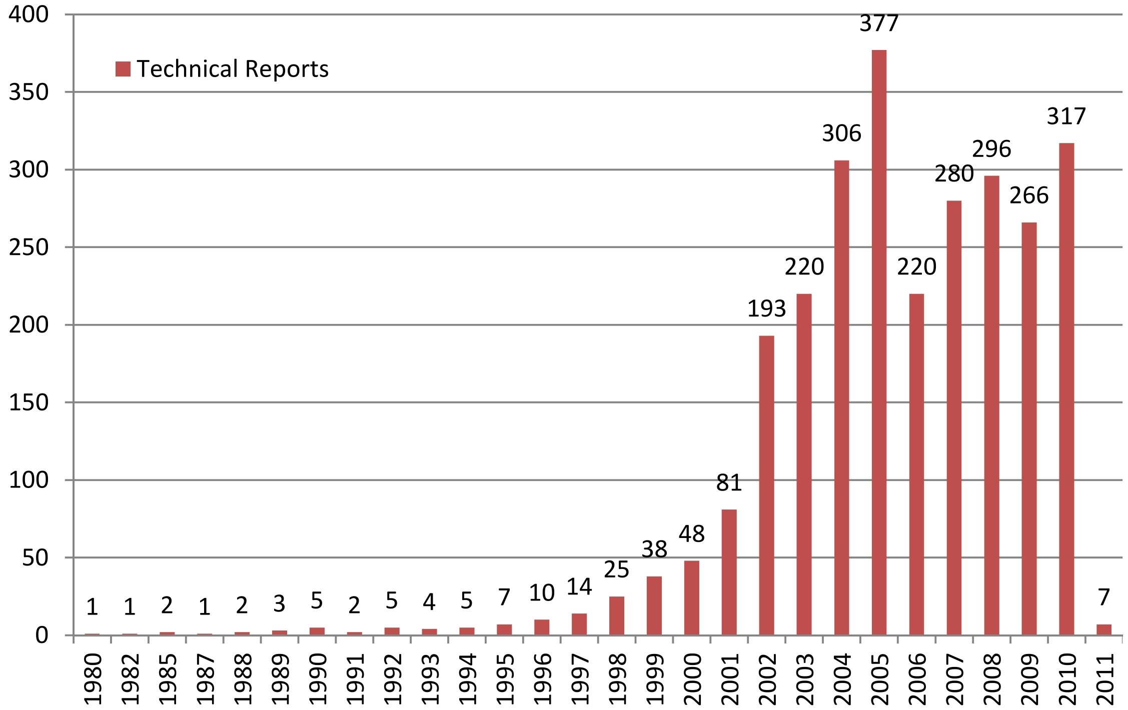
Academic Institute
Goverment Agencyes
Information Provider Overseas/Domestic
Social Networks
Internal/External Community
Knowledge Consumer
NDSL
NATIONAL DISCOVERY for SCIENCE LEADERS
NTIS

3.2 Comparison of Technical Reports on NDSL and Journal Articles written by Koreans on WoS (SCIE)

3.2.1 Data of Technical reports on NDSL and Journal Articles on WoS (SCIE)

- Data: technical reports funded by Korean government in nanotechnology on NDSL Technical reports on nanotechnology funded by the Korean government have been published since 1980'-. Technical reports published by private enterprises were not included in this study. To extract records of technical reports on nanotechnology on NDSL, key word searching was used. The keywords were 'nano*' or '나노*'. The total numbers of technical reports was 2,736**.** The peak year of publication of technical reports was 2005. 377 titles were published in 2005. The number fell to 7 in 2011 (Fig. 3). The early stage of research development of NT appears in the grey literature in the form of technical reports.

<Fig. 3> Korean Technical Reports in NT on NDSL (1980-2011)



As a multidisciplinary field, nanotechnology is related to numerous subjects such as applied physics, engineering and allied operations, chemical engineering and related technologies, metalworking processes and primary metal products, physics, ceramics and allied technologies, organic chemistry, precision instruments and other devices, etc. (Table 1).

<Table 1> Subjects of technical reports related to NT on NDSL

No	Subjects	Records
1	Applied physics	1,021
2	Engineering and allied operations	423
3	Chemical engineering and related technologies	299
4	Metalworking processes and primary metal products	272
5	Physics	255
6	Ceramic and allied technologies	251
7	Organic chemistry	234
8	Precision instruments and other devices	156
9	Technology applied sciences	111
10	Chemistry and allied sciences	103
11	Sanitary and municipal engineering Environmental protection engineering	85
12	Textiles	76
13	Other branches of engineering	60
14	Miscellaneous branches of medicine Surgery	47
15	Crystallography	44
16	Technology of other organic products	41
17	Physical chemistry	37
18	Elastomers and elastomer products	32
19	Food technology	30
20	Technology of industrial chemicals	29
21	Pharmacology and therapeutics	28
22	Systems	27
23	Heating, ventilating, air-conditioning engineering	22
24	Manufacturing	20
25	Civil engineering	19
26	Cleaning, color, coating, related technologies	19
27	Life sciences, biology	17
28	Inventions and patents	16
29	Light and infrared and ultraviolet phenomena	15
30	The others	

- Data: Journal Articles written by Korean in NT on WoS (SCIE)

In WoS (SCIE), journal articles that are written by Koreans, funded by Korean government and sorted by the subject 'Nanoscience nanotechnology' are selected. The numbers of journal articles on Web of Science (SCIE) are used as evaluation criteria of R&D productivity at institutional, national, and global levels. There were no articles in the 1980's but journal articles written by Koreans have been increasing dramatically since 1990'-.

The subjects of journal articles on WoS (SCIE) related to nanoscience and nanotechnology include materials science multidisciplinary, physics applied, chemistry multidisciplinary, physics condensed matter, chemistry physical, engineering electrical electronic, metallurgy metallurgical engineering, optics, biotechnology applied microbiology, etc (Table 2).

<Table 2> 10 Subjects related to NT Journal Articles on Web of Science

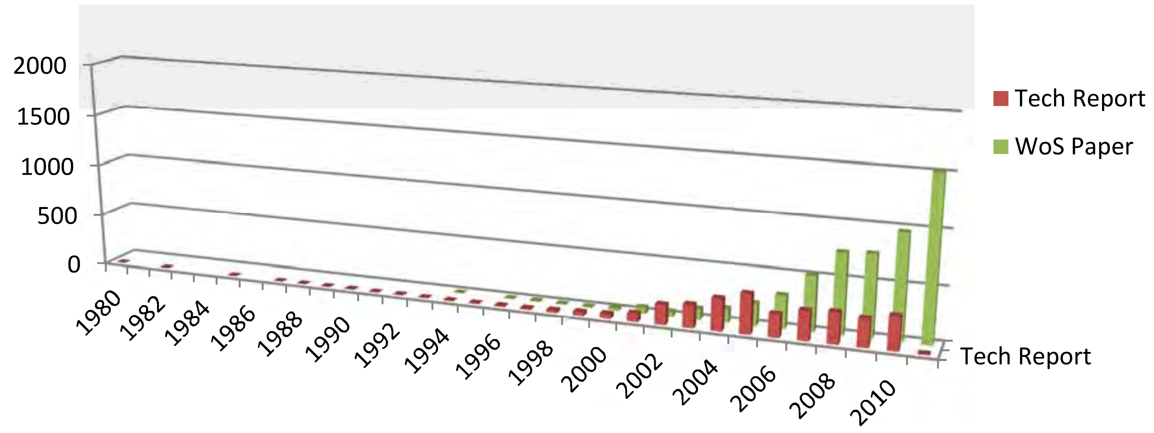
No	Web of Science Categories	records	% of 5428
1	Nanoscience nanotechnology	5428	100
2	Materials science multidisciplinary	4660	85.851
3	Physics applied	3644	67.133
4	Chemistry multidisciplinary	2659	48.987
5	Physics condensed matter	2365	43.57
6	Chemistry physical	1682	30.987
7	Engineering electrical electronic	535	9.856
8	Metallurgy metallurgical engineering	354	6.522
9	Optics	161	2.966
10	Biotechnology applied microbiology	138	2.542

3.2.2 Comparison of Technical Reports on NDSL and Journal Articles on WoS (SCIE)

2006 was the turning point for dominant literature in nanotechnology. The trend moved from grey literature to global level white literature (Fig. 4). The numbers of WoS (SCIE) publications has exceeded those of technical reports since 2006. WoS (SCIE) journal articles written by Korean in NT are still increasing. Nanotechnology became one of six technologies supported by the Korean government and the second most funded technology by the Korean government. In 2011, nanotechnology became the 10th most published subject in WoS journal articles written by Koreans.

However, technical reports fell dramatically in 2011. Technical reports or grey literature provided creative ideas and research results in the early stage of development of NT in Korea. When global level R&D was conducted, the research results appeared in WoS (SCIE). Grey literature was the driving force behind increasing of white literature in NT. These results indicate that innovation in NT has been taking place through grey literature. Grey literature contains the birth and early development of innovation in NT in Korea.

<Fig. 4> Comparison of Publication for White Literature and Grey literature (1980-2011)



Conclusion

The R&D research life cycle in nanotechnology in Korea can be divided into 5 stages. R&D researchers in nanotechnology use grey literature and white literature throughout the research life cycle. In general, researchers produced technical reports when they finished their R&D projects in the research creation stage. Grey literature contains important information, but it is difficult for other researchers to access it due to its method of distribution. The NDSL provides research output in technical reports, paper articles, etc through internet. published in Korea. Grey literature played an important role in the early stage of development of nanotechnology in Korea. Researchers later try to submit and publish their research output in well known journals such as those on the WoS, so that their results can be read and highly cited by other researchers.

Innovation in nanotechnology can be traced through grey literature, especially technical reports on NDSL. Technical reports in nanotechnology emerged the 1980s, peaked in 2005, and then reduced rapidly. 2006 is the turning point of the changing dominant literature in NT from grey literature and white literature. Grey literature contributed development of NT in Korea in early stage of research. Journal articles written by Korean in nanotechnology on the WoS (SCIE) appeared in the 1990's and has increased dramatically until now. NT became the 10th most published subject in WOS (SCIE) papers written by Koreans in 2011. NT became one of six major technologies supported by the Korean government in 2011. NT was the second most funded technology by the Korean government in 2011. NT has influenced to other subjects, created knowledge, and changed human life. Innovation through grey literature is in progress in other area of science and technology.

References

Encyclopedia Britannica <http://www.britannica.com/EBchecked/topic/962484/nanotechnology>

Keith G Jeffery, Anne Asserson. 2005. Grey in the R&D Process. Proceeding of the international conference on Grey Literature. Nancy, December 2005. <http://epubs.cclrc.ac.uk/work-details?w=35461>

Keith G. Jeffery. Architecture for grey literature in a R&D context.' International Journal on Grey Literature, Vol.1, no.2, pp.64 -72.

Heeyoon Choi, Seon-Hee Lee, Hyekyong Hwang. 2006. The collection development and usage strategy for grey literature in digital information environment. Seoul: KISTI.

Hye-Sun Kim etc. 2011. A study on R&D life cycle in science and technology. Knowledge Report no. 27. Seoul: KISTI.

Nanotechnology Wikipeia < http://en.wikipedia.org/wiki/Nanotechnology>

NDSL < http://www.ndsl.kr>

MIT Technology Review's top 10 emerging technologies 2012 <http://ceramics.org/ceramictechtoday/2012/05/08/mit-tech-reviews-top-10-emerging-technologies/>

Pardelli, Sara Gpggi, Manuela Sassi. 'Grey Literature between tradition and innovation: Is there a continuum?' <http://connection.ebscohost.com/c/articles/74313201/grey-literature-between-tradition-innovation-there-continuum>

Wan-Jong Kim, etc. 2011. Biometric analysis on SCI journal articles written by Korean scientists in 2010. Knowledge Report no. 16. Seoul: KISTI.

Web of Science <http://www.isiknowledge.com/?DestApp=WOS >.

Witte, Joachim de (Senter). 1998. Grey Literature and innovation: a strategic approach third international conference on Grey Literature: perspectives on the design and transfer of scientific and technical information, proceeding of the international conference on Grey Literature, Amsterdam, November 1997.

What goes up must come down: Publications from developing countries in the Aquatic Commons

Maria Kalentsits and Armand Gribling
Fisheries & Aquaculture Branch Library, FAO
Food and Agriculture Organization of the United Nations, Italy

Abstract

During 2010-2012, the Fisheries and Aquaculture Branch Library (FBL) of the Food and Agriculture Organization of the United Nations (FAO) was involved in a project that included the selection, digitizing, web-optimization, creation of metadata, and uploading into the Aquatic Commons (AC) digital repository of grey literature published by, amongst others, issuing agencies in several African countries, and a regional project in Asia - the STREAM Initiative. Furthermore, links to these full text online versions have been added to the Aquatic Sciences and Fisheries Abstracts (ASFA) bibliographic database.

The AC is a thematic digital repository covering the marine, estuarine, brackish and freshwater environments. This repository is directed by the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC) and hosted by the UNESCO/IOC International Oceanographic Data and Information Exchange (IODE) project office in Belgium.

The AC repository is built on Eprints software. In addition, it uses the related Interoperable Repository Statistics (IRS) software for usage statistics. The paper presents the results of the analysis of the access to the grey literature from selected issuing agencies in developing countries, whereby it focuses specifically on the Technical Documents, published by the Lake Victoria Fisheries Research Project and those published by the Support to Regional Aquatic Resources Management (STREAM), which was based in Bangkok, Thailand.

Examples of digital preservation and repatriation to countries of origin, two of the main objectives for FAO's participation in the repository, are discussed.

By using the IRS software the paper seeks to evaluate FAO's contribution to the content development of the repository and it finds a confirmation of the increasing visibility of and access to some baseline information in the field of fisheries and aquaculture published by institutions and/or projects in developing countries. Finally, the paper describes some of the utilities and limitations of the software.

1. Introduction

As is mentioned in one of the publications by Jean Collins, former FAO Fisheries Librarian,

> *One of the characteristics of the literature of fisheries and aquaculture – in particular the practical and management rather than the scientific aspects – is that it does not easily find its way into commercial journals. The results of research and the development lessons learned are often lost because of inadequate opportunities to publish, especially but not only in developing countries.*[1]

So, often scientific and practical knowledge generated in fisheries and aquaculture is presented in multiple formats of grey literature. The amount of information available as grey literature, which includes project reports and technical documents, training manuals and practical guides, workshop and conference materials, thesis, etc., is rather high and widely spread through marine and fisheries institutions in both developed and developing world. The value of grey literature is recognized and it is widely used in fisheries and aquaculture management.

Aquatic grey literature originating from developing countries can be often considered documents at high-risk of getting damaged due to frequent natural disasters and improper storage conditions or lost due to various human reasons. Access to these documents that are of limited distribution, often remains fragmented and inconsistent not only for the international community but also for internal users.

In the years 2010-2012, FAO's Fisheries and Aquaculture Branch Library (FBL) has been involved in a project that included the selection, digitizing, web-optimization, creation of metadata and uploading into Aquatic Commons (AC) repository of grey literature published by, amongst others, issuing agencies in several African countries and a regional project in Asia, the STREAM Initiative. Links to these full text online versions have been added to the Aquatic Sciences and Fisheries Abstracts (ASFA) bibliographic database.

In speaking about digital information sharing with countries of origin, we use the term "repatriation". By retrospective scanning and uploading of grey literature from developing countries we are carrying out a digital repatriation of information.

Besides thematic repositories like Aquatic Commons and OceanDocs, there are several other open access digital archives related to fisheries and aquaculture which have been developed. Relevant examples from developing countries are the institutional repositories from the Central Marine Fisheries Research Institute in India, the Kenya Marine and Fisheries Research Institute, the Aquaculture Department from Southeast Asian Fisheries Development Center, based in the Philippines, and the Digital Library from the Fisheries, Aquaculture and Marine Ecosystems Division (Secretariat of the Pacific Community), based in New Caledonia. Furthermore, there are multidisciplinary full text resources which include aquatic sciences, like Global Agricultural Research Archive (GARA), related to CAB Abstracts.

The paper presents the results of usage analysis performed with the Interoperable Repository Statistics (IRS) software. The focus is on the access to the grey literature from selected issuing agencies, specifically the Technical Documents published by the Lake Victoria Fisheries Research Project and the documents published by the Support to Regional Aquatic Resources Management (STREAM), which was based in Bangkok, Thailand.

2. Resource Sharing

Information resource sharing is a way to increase availability of documents through optimizing collections' usage while minimizing expenses. With regards to grey literature, the cost of the document is often a secondary issue to limited availability which becomes a determinative and negative factor in the dissemination of fisheries grey literature. Resource sharing programmes, such as traditional interlibrary loans or innovative repositories of digital collections such as the AC, are required to overcome this challenge.

Shared use of available resources is a concept that has been implemented into marine information management several decades ago by establishing national, regional and international resource-sharing networks as well as through facilitating sharing of marine and aquatic information by several UN organizations and programs.

2.1 IAMSLIC

The International Association of Aquatic and Marine Science Libraries and Information Centres (IAMSLIC), which consists of more than 300 members, plays an important role in aquatic and marine information resource sharing.[2] More than 90 participating libraries from more than 25 countries offer their local holdings to other member libraries through *IAMSLIC Z39.50 Distributed Library*. The libraries that do not have online catalogs that can be searched via Z39.50 can add selected serials holding information to *The Union List of Marine and Aquatic Serials.* The Union List provides access to several thousand titles including a large number of report series originating from developing countries. The program was implemented in 2002 and since its inception more than 38,000 requests have been submitted through the system. In 2011/2012, 124 IAMSLIC libraries in 45 countries used the service; whereas the volume of activity was continuously high for Latin America and increased significantly for African countries while remaining moderate in the Pacific Region. The IAMSLIC Resource Sharing Program includes *duplicate exchange*. Members of the Association use the IAMSLIC Discussion List to share information about duplicates available to be sent free of charge or for the cost of mailing to other libraries on their request.

The Resource Sharing Program took a new approach with the development of the Aquatic Commons digital repository, which is complementary to the OceanDocs repository project in which many IAMSLIC members also participate.

2.2 IODE

The program "International Oceanographic Data and Information Exchange" (IODE) of the Intergovernmental Oceanographic Commission (IOC) of UNESCO was established in 1961 with the objective to stimulate the management and exchange of data and information on a regional and international scale in the area of marine science and oceanography.[3]

In addressing its mission, IODE has developed a number of marine information projects, two of which can be mentioned as examples of aquatic resource sharing efforts:

OceanDocs is an electronic repository developed to collect, preserve and facilitate access to research outputs from members of Ocean Data and Information Networks (ODINs). In addition to IODE working documents (525 items), the repository currently includes eight Latin America country collections (1855 documents) and 16 African country and program collections (1618 documents).[4]

The Open Science Directory IOC/IODE has been developed by EBSCO and the Hasselt University Library; it provides users in developing countries with a comprehensive search tool for all open access and research program serial titles. The Directory offers open access to about 13,000 scientific and scholarly journals. Among the main collections available through the Directory are DOAJ, BioMed Central, HighWire Press and PubMed Central as well as the special programs HINARI, AGORA and OARE.[5]

In 2011, IODE and IAMSLIC signed a Memorandum of Agreement on cooperation in the field of marine information management which, among other objectives, aims at “promoting the capacity of libraries and information centers to disseminate and provide access to marine scientific literature for the benefit for marine scientists and other relevant users”.[6] As part of the Agreement, the Aquatic Commons is hosted by the IOC Project Office for IODE in Oostende, Belgium.

2.3 FAO Fisheries and Aquaculture Branch Library (FBL)

The Library has a mission of providing specialized, high quality library and information services to FAO staff, particularly in support of the activities of the Fisheries and Aquaculture Department, and to stakeholders involved in fisheries and aquaculture, especially those in developing countries, and to disseminate globally FAO information to fisheries and aquaculture organizations.[7]

Resource sharing and library networking is one of the core activities of FBL, which includes an interlibrary loan service (ILL) and retrospective digitization of documents for inclusion into the FAO Corporate Document Repository.

ILL services. The FBL fisheries and aquaculture serials collection (including all FAO fisheries and aquaculture report series as well as over 600 serials from developing countries) is available to aquatic science libraries worldwide through the IAMSLIC Resource Sharing Program.

The FAO Corporate Document Repository contains FAO documents and publications in electronic format. Currently, there are more than 6,000 digital documents on fisheries and aquaculture in the repository; this collection represents slightly more than 25% of all documents published by FAO Fisheries and Aquaculture Department over the years.

Selective Retrospective Digitization of regular FAO fisheries and aquaculture series, project reports as well as papers presented at many FAO technical meetings, which collect and analyse unique information and data, is an important library activity. Many of these documents are published in limited numbers, have narrow distribution and are not readily available in electronic format, although frequently requested by specialists in fisheries and aquaculture around the world. The Library makes every effort to identify, acquire and digitize FAO project documents. Selections of these documents have also been published on CD-ROM.

FAO and IAMSLIC signed a Memorandum of Understanding (MoU) on fisheries information systems and services in 2005. Within the framework of the MoU, FAO and IAMSLIC collaborate to enhance aquatic information resource sharing through improving coverage and access to fisheries and aquaculture publications from developing countries. The IAMSLIC Z39.50 Distributed Library and the Aquatic Sciences and Fisheries Abstracts (ASFA) database, as well as promoting the development and use of the Aquatic Commons e-repository and improving linkages from ASFA to full text open access resources, e.g. FAO fisheries documents and records in Aquatic Commons, are the tools and techniques used to achieve this goal.

2.4 Aquatic Sciences and Fisheries Abstracts (ASFA)

The Aquatic Sciences and Fisheries Information System (ASFIS) is engaged in the collection and dissemination, of information covering the science, technology and management of marine, brackish water, and freshwater environments.[8] The ASFA bibliographic database is the main product of ASFIS. Input to ASFA is provided by international network of co-operating institutions and organizations. ASFA partnership is composed of 66 co-sponsoring, national and international partners responsible for monitoring, selecting, abstracting and indexing publications for inclusion in the ASFA bibliographic database. The database includes more than 1.6 million bibliographic records (Oct. 2012), covering a large number of grey literature, which is one of its important comparative advantages with respect to other information sources. The Secretariat, provided by FAO, develops and maintains the ASFIS system. ASFA, through its Trust Fund, which is the collective property of ASFA partners, supports the projects initiated by partners and aimed at digitization of grey literature published in their countries. The ASFA partners deposit these documents in open access repositories such as Aquatic Commons and link them to the relevant records in the ASFA database, which increases the utility and value of the database.[9]

The ASFA Thesaurus is a freely available indexing and searching tool. It contains the subject descriptors used to index the records which are contained in the Aquatic Sciences and Fisheries Abstracts (ASFA) bibliographic database.

3. Aquatic Commons

The Aquatic Commons repository, established by IAMSLIC in 2007, is a subject based, thematic digital repository covering the natural marine, estuarine, brackish, and freshwater environments. It includes the science, technology, management, and conservation of these environments, their organisms and resources, and the economic, sociological and legal aspects.[10] It contains a growing collection of published and unpublished research, organizational publications, and other materials, including an increasing number of (grey) literature from developing countries.

The repository is funded and directed by IAMSLIC, through a Board of IAMSLIC members and representatives of the UNESCO/IOC project Office for IODE, which is hosting the repository. It is powered by EPrints software. Its multi-language interfaces are available in English, French and Spanish. Aquatic Commons supports self-submittal of digital resources by authors or issuing agencies, or deposits are done by third parties, with permission from the copyright owners. The repository offers archiving of digital copies where local information and communication technologies (ICT) are lacking or inadequate.

Three-quarters of the uploaded documents are published before the year 2000; therefore, the major part of the repository is the result of retrospective scanning. The more than 90 issuing agencies give proof of the international focus of the repository. Examples are the following national and international organizations:

From the USA:

- California Department of Fish and Game (594 uploaded documents)
- United States National Marine Fisheries Service (539)
- United States National Ocean Service (127)
- Florida Cooperative Fish and Wildlife Research Unit (84)

From Latin America:

- Facultad de Ciencias Naturales y Museo, Univ. Nacional de La Plata, Argentina (112)
- *From Africa:*
- Fisheries Society of Nigeria (431)
- Centre de Recherches Océanographiques, Côte d'Ivoire (128)

From Europe:

- Freshwater Biological Association, United Kingdom (552)
- Environment Agency, UK (Freshwater Biological Association) (102)
- German Federal Research Centre for Fisheries (2158)

From international projects and organizations:

- Charles Darwin Foundation (353)
- Inter-American Tropical Tuna Commission (233)
- Support to Regional Aquatic Resources Management (126) (Asia)
- Aquatic Plant Management Society (101)
- North Pacific Marine Science Organization (PICES) (93)
- Lake Victoria Fisheries Organization (77) (Africa)

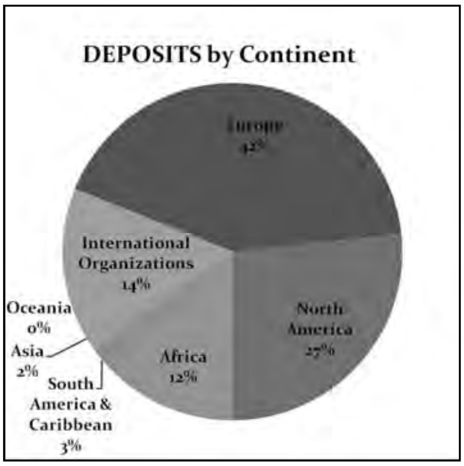
The German Federal Research Centre for Fisheries has uploaded 2158 documents: this is one third of the total repository. This Research Centre has no plans to create its own repository with long-term storage, but has preferred to use the AC as its home repository. One of its main contributions is the deposit of articles of their magazine, *Informationen aus der Fischereiforschung* (Information on Fishery Research). The AC is the primary source of their online journal and the URLs will be linked in the DOI of each publication.

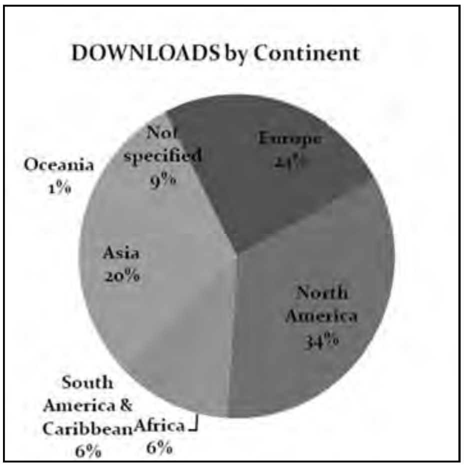
The number of uploads in the AC is increasing from a monthly average of 48 additions in the first year (2007) to 215 monthly deposits in 2012. The usage of the repository has also been growing over the years, with an average number of 13.5 thousands of downloads per month in 2012. By the end of October there were 575,306 downloads from a total of 7973 deposits. The IRS allows for presentation and interpretation of these statistics.

The IRS (Interoperable Repository Statistics) software was implemented in 2008 and permits ongoing monitoring of the repository. The advantage of the software is in its presentation possibilities, offering various kinds of graphs and tables. IRS creates raw data and other usage statistics, for the whole repository as well as for individual papers and authors. Unfortunately, it is not yet possible to get detailed statistics for a particular issuing agency and these calculations have to be done “manually” which can be rather time-consuming.

The majority of AC users are directed to the repository from a Google search (Google 47% and Google Scholar 8%); a number of searches was performed using simple or advanced option provided by the AC search engine, which shows name recognition among users.

Avano has been the sole marine and aquatic sciences OAI harvester. Developed by IFREMER, the French Research Institute for Exploration of the Sea, it gives access to more than 200 open archives. With the increasing importance of Google and Google Scholar, it seems now that the efforts of Avano have been superseded and the harvester will cease in the near future after five years of activity.[11]





A large majority of deposits are from Europe, Canada and the USA, while the usage statistics show a broader, more varied picture with higher representation of developing countries. For example, Asian and South American deposits constitute subsequently 2% and 3% of a total number, while downloads by users in these regions are 20% in Asia and 6 % in South America. Downloads by African countries still remain moderate with only 6 percent of a total number for the period 2007-2012.

4. ASFA Trust Fund Project on Grey Literature for AC

ASFA Trust Fund provides funding for small projects that aim to increase visibility of grey literature from partner institutions through filling gaps in the ASFA bibliographic database as well as by digitization of documents and adding full text links to ASFA records.

For FAO, a long-term goal of its participation in Aquatic Commons is to assist in providing access to legacy collections from institutions and projects in developing countries that have never been easily accessible. FAO's support for the Aquatic Commons is primarily aimed at content development and mainly intended to assist institutions in developing countries to improve visibility and access, sharing and preservation of fisheries and aquaculture management publications.

The ASFA Trust Fund Project "Published and Grey Literature of African Aquatic/Fisheries Institutions for Aquatic Commons" aimed to retrospectively digitize grey literature from developing countries, and to provide links to the full text for those documents already cited on the ASFA Database or to create new ASFA records where they are not yet available. The project was executed by FAO's Fisheries and Aquaculture Branch Library between 2010 and 2012. In total, almost 15,000 pages have been scanned, compressed and web optimized, metadata of the documents have been created and 754 PDF files were uploaded and links to the ASFA Database have been added. Several participating issuing agencies have also received CD-ROM's of digitized documents for offline consultation, both in TIFF and PDF format.

Seven issuing agencies participated in the project: Fisheries Society of Nigeria (FISON); Institute of Marine Biology & Oceanography (IMBO), Sierra Leone; Nigerian-German Kainji Lake Fisheries Promotion Project (NGKLFPP), Nigeria; Centre de Recherches Océanologiques (CRO), Côte d'Ivoire; Instituto de Investigação Pesqueira (IIP), Mozambique; Lake Kariba Fisheries Research Institute (LKFRI), Zimbabwe; and Lake Victoria Fisheries Research Project (LVFRP- now Lake Victoria Fisheries Organization, LVFO), Kenya-Uganda-Tanzania.

5. Statistics and Usage Analysis: Examples of Repatriation

A lifetime of uploaded documents is still too short to conduct valid citation analysis for evaluation of the documents impact. Also, knowing that many of the project documents (for example, those from STREAM) are of technical rather than scientific nature, it can be assumed that those documents have their main impact in the area of practical application of knowledge, for example, in implementing new technologies in aquaculture. Other documents are related to fisheries management and will be of importance to managers and other stakeholders.

By applying IRS software for the usage analysis of the grey literature from selected issuing agencies in developing countries, the paper seeks to address a further evaluation of FAO's contribution to the content development of the repository, confirming the increasing visibility of and access to some baseline information in the field of fisheries and aquaculture published by institutions and projects in developing countries.

Firstly, we focus on the Technical Documents published between 1998 and 2001 by the Lake Victoria Fisheries Research Project, a regional project on the African lake.

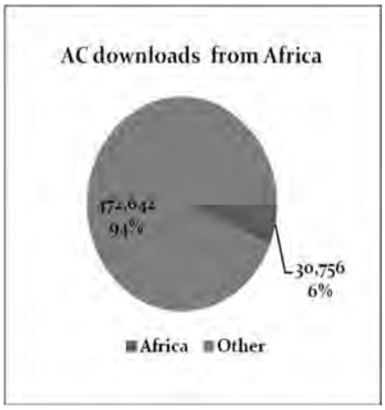
Secondly, we analyze the usage of the documents published by the Support to Regional Aquatic Resources Management (STREAM) during 2000-2009, which was based in Bangkok, Thailand.

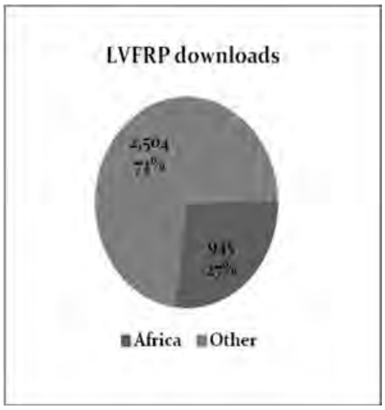
These project documents are examples of digital preservation and repatriation to the countries of origin, two of the main objectives for FAO's participation in the AC repository.

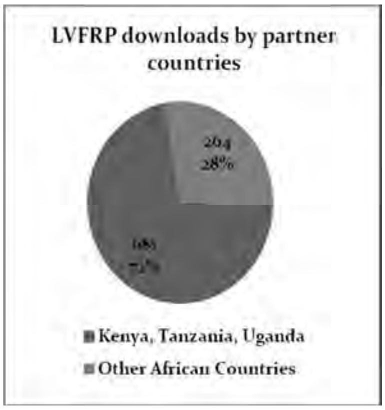
Downloads of all project documents were analyzed for geographic origin. Whereas downloads that originated from the project region were paid closer attention than downloads that originated from other areas, in order to evaluate impact of the project documents for the whole region and for the country of their origin in particular. Raw data analysis allowed conclusions to be drawn regarding user search strategies and dynamics of uploading and downloading activities, as well underpinning some assumptions on the repository name recognition and web crawlers downloads. The analysis shows that significant number of downloads for the period 2010-2012 comes from a range of IP addresses in France: these are most probably web crawlers. Starting with 3,2 % in 2009, at present every 10th download is done by a robot. In order to maintain the reliability of the analysis of LVFRP and STREAM documents, and taking into account that 87% of all French downloads are from this IP range, downloads originated from French IP addresses were excluded from our analysis.

5.1. *Lake Victoria Fisheries Research Project*

The Lake Victoria Fisheries Research Project (LVFRP) aimed at creating a framework for the management of the Lake's fisheries, with the objective of improving the management of its resources. The Lake Victoria Fisheries Management Plan, published in 2001, is based on the concept of co-management, which includes stakeholders at local, regional, national and international level; the Plan has been adopted by the Council of Ministers of the Lake Victoria Fisheries Organization (LVFO). The LVFRP Technical Documents have been digitized by the Trust Fund project and in total 75 records have been added to the repository.







Statistics show a correlation between download activity and national involvement into specific regional projects by the riparian countries - Kenya, Tanzania and Uganda. While indicating a comparatively low download activity for the entire repository by the African continent (6 % of the total number of downloads), the statistics indicates higher usage (27 %) of the repository by African users for the Lake Victoria regional project based in Africa. A close look into this project statistics reveals a high interest towards those documents by countries involved in the project. 72 percent of the total number of project document downloads comes from the countries bordering Lake Victoria - Kenya, Tanzania and Uganda.

Some of the project documents were quite intensively downloaded by users in developed countries. For example, tracking one of the referring IP addresses for a document on ownership and co-management of Lake Victoria, we discovered that this document was included in the list of compulsory reading for university students in the Netherlands; 47 % of all downloads were presumably done by students between September-December 2011.

5.2 Support to Regional Aquatic Resources Management Project

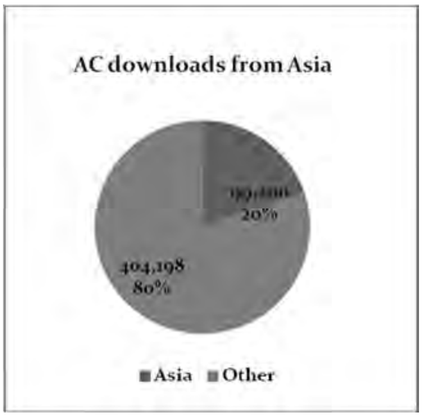
STREAM, the project "Support to Regional Aquatic Resources Management" was an initiative executed within the framework of the Network of Aquaculture Centres in Asia-Pacific (NACA). Besides NACA, partners included FAO, the Department for International Development of the UK (DFID), the Voluntary Service Organization (VSO) from the United Kingdom, and the Australian Government Overseas Aid Program (AusAID). The project aimed to support agencies and institutions to: 1) utilize existing and emerging information more effectively, 2) better understand poor people's livelihoods, 3) enable poor people to exert greater influence over policies and processes that impact on their lives, and 4) develop policies and processes of mediating institutions and capacity building. It adopted an approach where stakeholders engaged in aquatic resources management participated actively in the development of the Initiative.

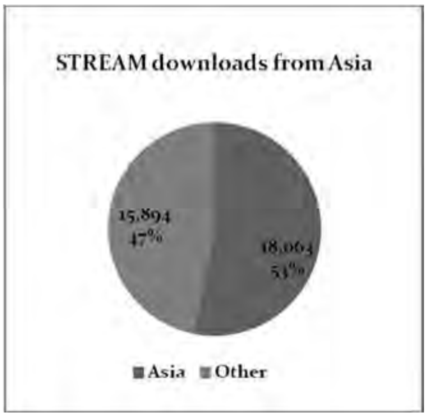
The STREAM Initiative was based at the NACA Secretariat in Bangkok, but operated in several Asian-Pacific countries, including Cambodia, India, Indonesia, Lao PDR, Myanmar, Nepal, Philippines, Vietnam and China.

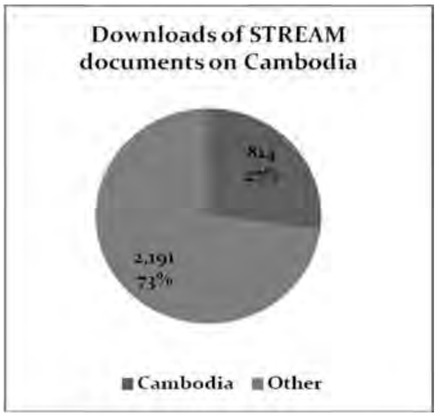
A large part of the published outputs of the project have been added by FAO to the AC between 2008 and 2010. All 126 uploaded documents were so-called "born digital" and were originally available on the STREAM Web site.

The website of the STREAM Initiative had a short lifespan and was only up from July 2002 to June 2008. Since then electronic copies of the documents produced by STREAM have been hard or even impossible to find, if it were not for the AC. Documents from the STREAM Initiative are good examples of preservation; without the AC, hardly any of the documents would still be available on the Internet.

The usage statistics for the STREAM documents show a similar picture as that of the documents of the Lake Victoria project. Although one-fifth of all downloads from the repository are from the Asian continent, more than half of all downloads of the STREAM project documents come from users in Asia. Once again a high percentage of downloads comes from the "country of origin" or the country covered by the documents.







6. Conclusions

Based on the results of the usage analysis of the repository it can be concluded that the digitization of grey literature published by fisheries and aquaculture institutions in developing countries increases overall **access and sharing** of information, contributes to **preservation** and enables **repatriation** to the countries of origin.

By participating in the AC, institutions in developing countries are challenging some of the information constraints, outlined in FAO Technical Guidelines for Responsible Fisheries 12, such as lack of awareness of and access to historical and baseline information, poor opportunities to publish and disseminate the results of research as well as dispersion of information between various government agencies, scientific and academic institutions and industry.[12] Through the Aquatic Commons, fisheries and aquaculture grey literature has found a new and much wider audience, and is integrated into the international information exchange.

Providing access to, and sharing of this historical and baseline information through the AC further enables issuing agencies to solve a problem of inadequate ICT support as well as to avoid costly and wasteful duplication of effort.

Destructive natural disasters which heavily affect tropic and sub-tropic coastal areas in developing countries may cause serious damage to grey literature print collections. Also, damage caused by human factor, such as lack of or poor knowledge transfer, short-sighted managerial decisions, language barriers, can be a reason of a permanent loss of the only available hard copy of the document. Digital preservation of grey literature on fisheries and aquaculture originating from developing countries ensure availability of these unique documents for future generations.

Repatriation of digital documents was carried out in two ways. Firstly, after signing the permission to upload the documents by third party, the issuing agency, retaining its copyright, has access to their documents in the AC repository. Additionally, they receive a CD containing electronic copies of the documents. The agency can create links on their website to the full text of these documents in the AC repository. Secondly, in a broader context we can talk about the repatriation to Internet users in the country of origin.

FAO's Fisheries and Aquaculture Branch Library will continue to promote the participation of institutions in developing countries in the Aquatic Commons and will seek to undertake digitization projects in the future.

7. References

1. Collins, J. 2007. Information sharing via Aquatic Commons. *FAO Aquaculture Newsletter.* 37:12-13 [online], Available through: FAO Fisheries and Aquaculture Department website <www.fao.org/fishery/publications/fan/en> [Accessed 02.11.2012].
2. Butler, B.A., Webster, J., Watkins, S., & Markham, J.W. 2006. Resource sharing within International Library Network: using technology and professional cooperation to bridge the waters. IFLA Journal 32(3): 189-199[online], Available through IFLA website
< http://archive.ifla.org/V/iflaj/IFLA-Journal-3-2006.pdf> [Accessed 02.11.2012].
3. Nieuwenhuysen, P. & Pissierssens, P. 2009. A UNESCO Agency offers professional development across geographic and generational boundaries. *In:* J. Varlejs & G. Walton, eds. *Strategies for Regenerating the Library and Information Professions: the Eight World Conference on Continuing Professional Development and Workplace Learning for the Library and Information Professions*, 18-20 August 2009, University of Bologna, Italy. IFLA Publications 139. pp. 317-327. [e-book] Available through: De Gruyter website
< http://www.degruyter.com> [Accessed 02.11.2012].
4. IODE. 2008 *OceanDocs: E-repository of ocean publications.* [online] Available at: <http://www.oceandocs.org/> [Accessed 02.11.2012].
5. IOC/IODE. 2008. *Open Science Directory.* [online] Available at:
< http://www.opensciencedirectory.net/> [Accessed 02.11.2012].
6. IAMSLIC. 2012. *IAMSLIC – the International Association of Aquatic and Marine Science Libraries and Information Centers.* [online] Available at: <http://www.iamslic.org> [Accessed 02.11.2012].
7. FAO. 2012. *Fisheries and Aquaculture Branch Library.* [online] Available at
< http://www.fao.org/fishery/library//en> [Accessed 02.11.2012].
8. FAO. 2012 *Aquatic Sciences and Fisheries Abstracts (ASFA).* [online] Available at: <http://www.fao.org/fishery/asfa/en> [Accessed 02.11.2012].
9. Garnica Carreño, J.L., Gribling, A. & Wibley, H. 2010. Visibility and access through the Aquatic Commons. *In: Proceedings of the 36th Conference of the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC), 17-21 October 2010, Mar del Plata, Argentina,* pp. 119-128. [online] Available through: WHOAS: Woods Hole Open Access Server
<https://darchive.mblwhoilibrary.org/handle/1912/4581> [Accessed 02.11.2012].
10. IAMSLIC. 2007. *Aquatic Commons.* [online] Available at: <http://www.aquaticcommons.org> [Accessed 02.11.2012].
11. Haas, S., Merrikin, P., Gribling, A., & Merceur, F. 2009. The Aquatic Commons Model: the roles of IAMSLIC, FAO and IFREMER in supporting open access to fisheries and aquaculture research and management. [online] Available through IFLA website
< http://conference.ifla.org/past/ifla75/101-haas-en.pdf> [Accessed 02.11.2012].
12. FAO. 2009. Information and knowledge sharing. *FAO Technical Guidelines for Responsible Fisheries.* 12. 97 p. [online] Available through: FAO Corporate Document Repository
< http://www.fao.org/docrep/013/i0587e/i0587e00.htm> [Accessed 02.11.2012].

Data sharing in environmental sciences:
A survey of CNR researchers

Daniela Luzi
Istituto di Ricerche sulla Popolazione e le Politiche Sociali, IRPPS-CNR, Italy

Roberta Ruggieri
Senato della Repubblica, Italy

Stefania Biagioni
Institute of Information Science and Technologies, ISTI-CNR, Italy

Elisabetta Schiano
Institute of Marine Sciences, ISMAR-CNR, Italy

Abstract
The paper presents the results of a survey on researchers' attitudes and practices of data sharing in the area of Environmental sciences. It is based on an online questionnaire submitted to CNR researchers active in this disciplinary field, that has proved to be data intensive, collaborative and multidisciplinary. The study lies within the framework of other international analyses that consider this complex process exploring different aspects that may influence the propensity of a consistent and effective data release. Therefore, motivations, perceived barriers and enablers to data sharing are analysed together with the outline of research context and practices in this field.

1. Introduction

Today the free availability of research data is considered an important driver of innovation and of new scientific insights. Due to the increasing amount of data collected as well as to the variety of purposes, process of acquisition and formats this is not an easy task. It implies the development of policies that promote data curation and preservation, the recognition of the value of research data as "first-class publication", the enforcement of clear rules for open access, copyright and ownership. It is also necessary that the scientific community agree on the development and use of common interoperability standards related to data models, format and exchange protocols. Last but not least, it requires that suitable infrastructures be developed at national and international level considering discipline specificity.

There is now a vast literature devoted to the definition and importance of research data [Borgman, 2012, Kowalczyk & Shankar 2011]. Many studies consider the technical aspects of preservation and management [Tjalsma & Rombouts, 2010, Graaf et al. 2011], while official documents and whitepapers outline current changes in the research process and propose policies and infrastructure that can promote data sharing [Hey et al. 2009, NSF, 2005]. Moreover, various surveys have been carried out to explore researchers' practices and perceptions towards data acquisition, curation and preservation, focusing in particular on perceived barriers and enablers of data sharing. Some surveys conducted within European projects, have analysed attitudes and opinions of different stakeholders: researchers, data managers, publishers, funding organisations [PARSE.Insight, 2009, Dallmeier-Tiessen et al, 2012], as well as libraries, national and local governments [EU Directorate, 2012]. They rely on different methods in the collection of results: questionnaires, interviews, desk research. They gained insight into differences between disciplinary fields across various countries with the aim of developing roadmaps or setting up a participatory process for the construction of international e-Science infrastructures [Tenopir et al. 2011]. Among the surveys that were particularly focused on specific research areas (Pinowar, 2011, Milia et al. 2012), it is also worth mentioning the studies related to biodiversity that combine the analysis of researchers' attitude (Enke et. al., 2012) with the evaluation of technical and information resources available in this multidisciplinary field [Bach et al., 2012, Bendix et al. 2012].

Most of these studies have a common vision on data lifecycle as closely connected with the research process, where data sharing "begins with good data practices carried out in all phases of the data lifecycle" (Tenopir et al., 2011). Moreover, researchers' propensity to data sharing largely depends on the research context, synthesized by Kim as the combination of technological infrastructure, institutional support and interpersonal interactions (Kim & Stanton 2012).

Based on this view our survey intends to analyse researchers' attitude in data sharing posing a particular emphasis on the exploration of research practices and context within the broad multidisciplinary field of Environmental sciences. In our vision the understanding of the complexity of data sharing embedded in a specific research environment can bring to the fore opinions, beliefs, concerns and practices that may contribute to the development of suitable information systems tailored on researchers' needs as well as to the introduction of policies that may promote their consistent and long- term diffusion.

2. Methods

Among the different CNR departments devoted to different disciplinary fields, we choose to analyse attitudes of researchers belonging to the Institutes of the Department of Earth and Environment because this research field has proved to be data intensive and multidisciplinary in nature. Moreover, in this area there are several initiatives both at international and CNR level that are promoting and setting up infrastructures for data sharing.

The survey makes use of a semi-structured questionnaire of 40 questions that consists of two main parts. Reflecting the survey hypotheses, the first one aims to gain insight into research practices that may influence data sharing. Based on the chosen target group we identified *ad hoc* questions to explore in particular:

- The general research context (research lines, types of funds, types of collaboration);
- Data acquisition (type or research carried out, data used, modes of data acquisitions and instrumentation);
- Data management (availability of standards, use of descriptive metadata, adoption of preservation procedures, presence of dedicated personnel for data management);
- Data re-use and availability (propensity of using data produced by others and related evaluation of its reliability; available resource to store own data, practices in data sharing).

The second part is specifically focused on capturing perceived barriers to as well as conditions that may motivate data sharing. This part contains a selection of questions submitted in large-scale international surveys (PARSE.Insight, 2009, Tenopir et al., 2011, Enke et. al., 2012) in order to explore commonality and differences in attitudes.

Additionally, respondents were asked for information on gender, age, length of CNR service and occupational position. Most of the questions are multiple choice, while two plain text answers were also included in the questionnaire to collect researchers' free opinion on this topic. Both questionnaire and survey data are available at: https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:53292.

The survey was Internet-based and the link was sent out via e-mail using LimeSurvey open source software that also supports invitations, reminders, and makes answers anonymous. The survey period was June-September 2012.

3. The sample

1087 questionnaires were sent to all researchers affiliated to the 13 CNR Institutes belonging to the Department of Earth and Environment. We included researchers with both permanent and temporary contracts and also external collaborators, who are generally researchers coming from universities that closely collaborate with CNR Institutes. The response rate was 48% (i.e. 523 responses) that can be considered satisfactory given the voluntary basis of the survey.

There are 13 CNR Institutes that belong to the Earth and Environment Department. These institutes are different in size, ranging from 40 researchers to more than 100 and are organised in various research units located all over Italy. Their area of interest ranges from land and water ecosystems to climate change, from the use of resources to the monitoring of anthropogenic risks, from biodiversity to the development of methods and technologies for environment protection.

The distribution of responses by Institute (tab 1) shows that five Institutes out of 13 reach a response rate higher than 50%. Most of them are in the area of marine sciences and water resources.

Table 1 - Distribution of responses by Institute

Institutes	***Questionnaires sent (No.)***	***Questionnaires received (No.)***	***%***
IAMC - Institute for coastal marine environment	93	53	***57.0***
IBAF - Institute of agro-environmental and forest biology	53	27	***50.9***
IDPA - Institute for the dynamics of environmental processes	58	25	***43.1***
IGAG - Institute of environmental geology and geo-engineering	103	36	***35.0***
IGG - Institute of geosciences and earth resources	132	63	***47.7***
IIA - Institute for atmospheric pollution research	67	29	***43.3***
IMAA - Institute of methodologies for environmental analysis	83	31	***37.3***
IRPI - Research institute for geo-hydrological protection	88	40	***45.5***
IRSA - Water research institute	69	40	***58.0***
ISAC - Institute of atmospheric sciences and climate	94	46	***48.9***
ISE - Institute of ecosystem study	55	35	***63.6***
ISMAR - Institute of Marine sciences	149	80	***53.7***
IVALSA Tree and timber institute	43	18	***41.9***
Total	**1,087**	**523**	

3.1. Respondents' profiles

An overview of the respondents' profile is given in table 2. The majority of respondents to the survey are male. They fall mainly into two age groups (from 41 to 50 and over 50 years old). The length of service at CNR is concentrated in two groups: from 11 to 20 years and over 20 years of CNR service. The majority of respondents have a permanent contract.

Table 2. - Respondents' profile

Respondents' profile		
	No.	***%***
Gender		
F	204	***39.4***
M	314	***60.6***
	518	
Age		
< 30	31	***5.9***
30 - 40	139	***26.6***
41 - 50	170	***32.5***
> 50	177	***33.8***
	517	
Length of service		
> 5 years	117	***22.4***
6-10 years	107	***20.5***
11-20 years	144	***27.5***
> 20 years	142	***27.2***
	510	
Position		
Permanent	327	***62.5***
Temporary	129	***24.7***
Training	9	***1.7***
External collaboration	56	***10.7***
Other	2	***0.4***
	523	

4. Research context

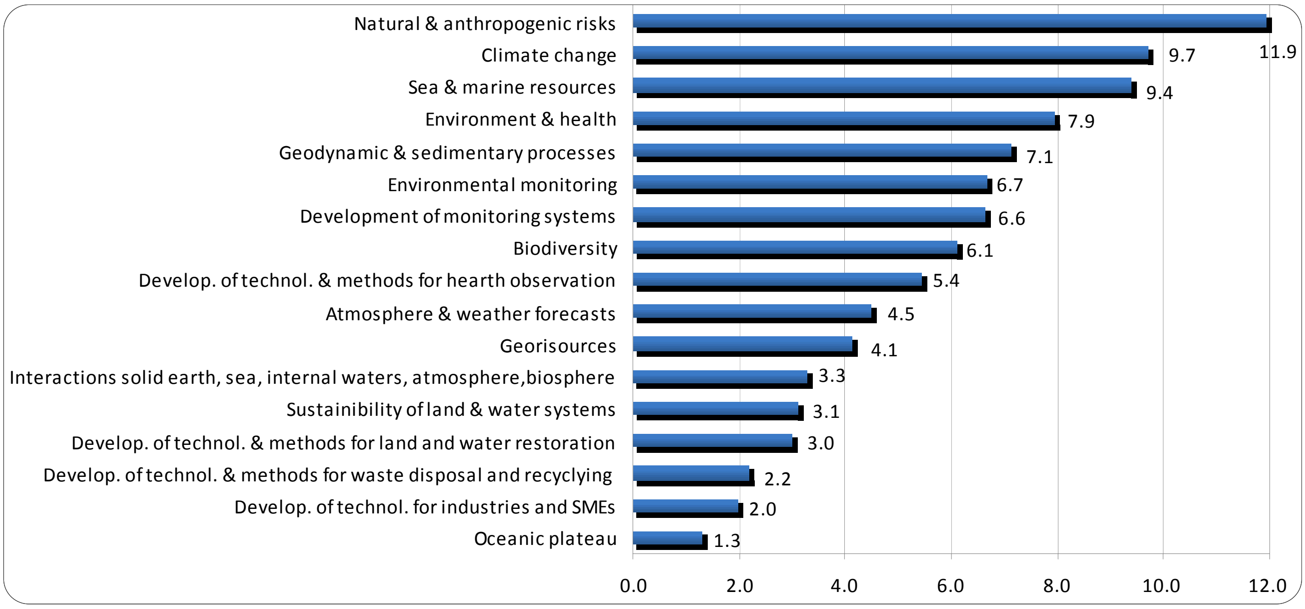
Data sharing does not simply represent an individual propensity, but it is influenced by socio-cultural, contextual and institutional factors. It has proved to vary from discipline to discipline and within disciplines, it depends on different factors: types of research and collaboration setting, types of data, modes of acquisition and handling strategies as well as human, technical and institutional support for long-term preservation, to mention but a few.

Therefore, the first part of the questionnaire is devoted to exploring general features of research practices to obtain a more detailed framework on how research activities are carried out in this area.

4.1. Research lines

Researchers were asked to provide a percentage of time dedicated to a set of research lines described in the website of the CNR Department of Earth and Environmental Sciences. Figure 1 shows the topics on which researchers concentrate their work (multiple answers were allowed). Researchers are generally involved in more than one research line, a relevant percentage on them deals with Natural and anthropogenic risks (11.9%), Climate change (9.7%) and Sea and marine resources (9.4%).

Fig. 1. - Distribution of the research lines carried out by CNR researchers



4.2. Funds

When asked to provide a percentage of funds received in carrying out their research activities, the majority of researchers (47.8%) reported that they receive national funds, 28.9% rely on EU and/or international projects and 14.9% on national and international private funds. On average, only 4.4% of researchers reported that their work is directly funded by CNR.

4.3. Collaboration

Two questions in the survey were focused on researchers' collaboration habits. The first one asked, whether they usually work as a single researcher, in small (max 3 persons), medium (from 3 to 7 persons) or in large groups (more than 8 persons). The majority of researchers work in a medium size (47.7%) group.

When asked how often and on which occasions they collaborate with multidisciplinary groups, 42% reported that they always do so in international projects and with colleagues of the same Institute (36.9%). Working in multidisciplinary groups occurs sometimes with other CNR institutes (56.2%) and with other Italian institutions and/or Universities (61.6%).

4.4. Data acquisition

As data sharing is part of data lifecycle, a set of questions was devoted to exploring types of data used, how they are acquired and managed. A prerequisite of data sharing is that data are acquired following defined procedures, is associated with proper metadata, so that data are interpretable and properly reusable. Therefore a set of questions was focused on the type of research carried out, types of data used, how data are acquired, as well as information on measurements and instrumentation.

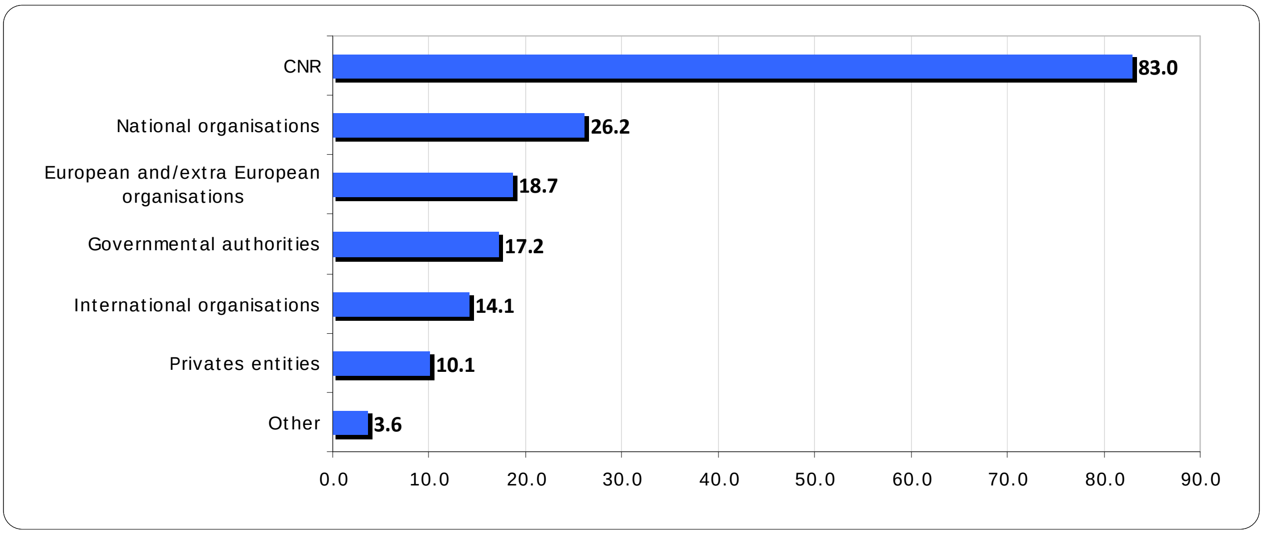
We first asked researchers to provide a percentage of time dedicated to theoretical and/or experimental research in order to gain insights into the type of research most frequently carried out in this field. On average the majority of CNR researchers (77%) carry out experimental research that generally implies the collection as well as an intensive use of data.

Tab. 3. - Type of data used in the analysis of land, sea, internal waters, atmosphere and biosphere

	Biological	*Chemical*	*Physical*	*Geological*
Land	18.5	31.0	31.9	45.1
Sea	27.0	30.0	30.8	26.2
Internal waters	22.6	36.5	30.2	27.3
Atmosphere	6.5	27.3	42.4	13.4
Biosphere	28.3	24.3	21.2	16.4

Researchers were asked to indicate the type of data used when they analyse phenomena related to land, sea, internal waters, atmosphere and biosphere. Multiple answers were allowed. Table 3 shows that CNR researchers more frequently use geological data related to the study of land (45.1%) as well as physical data related to the Atmosphere. Data gathered in the analysis of sea, internal waters and biosphere tend to be almost equally distributed among biological, chemical, physical and geological data. This multidisciplinary approach is confirmed by some researchers, who specified in the variable "other" that they use biogeochemical, geo-morphological, geophysical or geo-mechanical data. A small percentage of researchers (0.6%) indicate in "others" that they use remote sensing data.
Moreover, 21% of researchers also use demographic data to carry out their research activities.

Fig. 2. - Distribution of respondents to the question:
"The data you are working on come mainly from instrumentation managed directly by …"



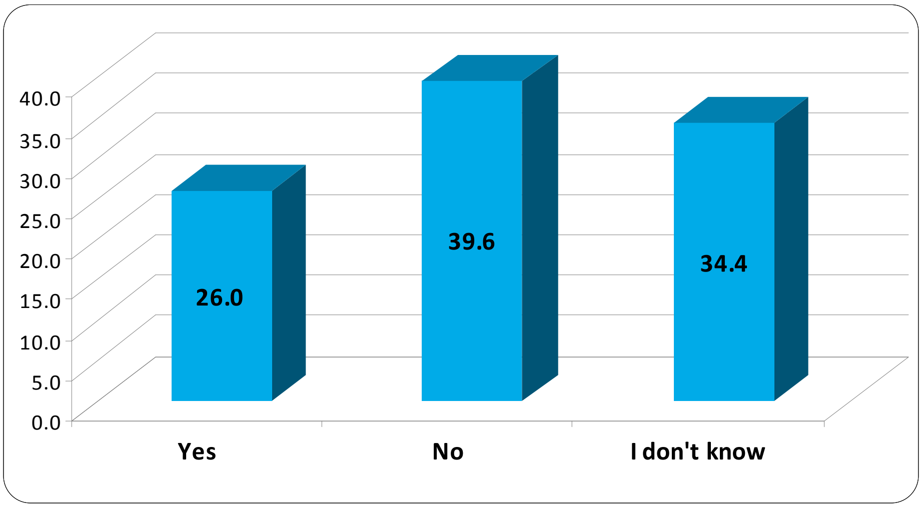
When asked whether they take measurements directly by themselves, or use measurements taken by others or alternatively use both, the majority of CNR researchers (53%) reported that they use measurements directly taken by themselves and/or by their research group, while 8.2% use measurements taken by others and 38.8% use both. Moreover, data are mainly acquired from both laboratory work and in the field (53.8%), while 32.5% of CNR researchers collect data from fieldwork alone.

A multiple answer was allowed to indicate who manages the instrumentation used. Figure 2 shows that an overwhelming majority of CNR researchers (83%) obtain data from instrumentation directly managed by CNR, while 26.2% of them also use data taken from instrumentation managed on the basis of agreements with other national organisations.

4.5. Data management

The use of standards facilitates data sharing, while re-use and evaluation of data also depends on the metadata associated to the data acquired. Therefore, a set of questions aimed to explore different aspects of data management, such as the availability of standard of researchers' community of reference, use of descriptive metadata in their current research practice, data management plan in place in their Institutes and presence of trained staff that may support data curation.

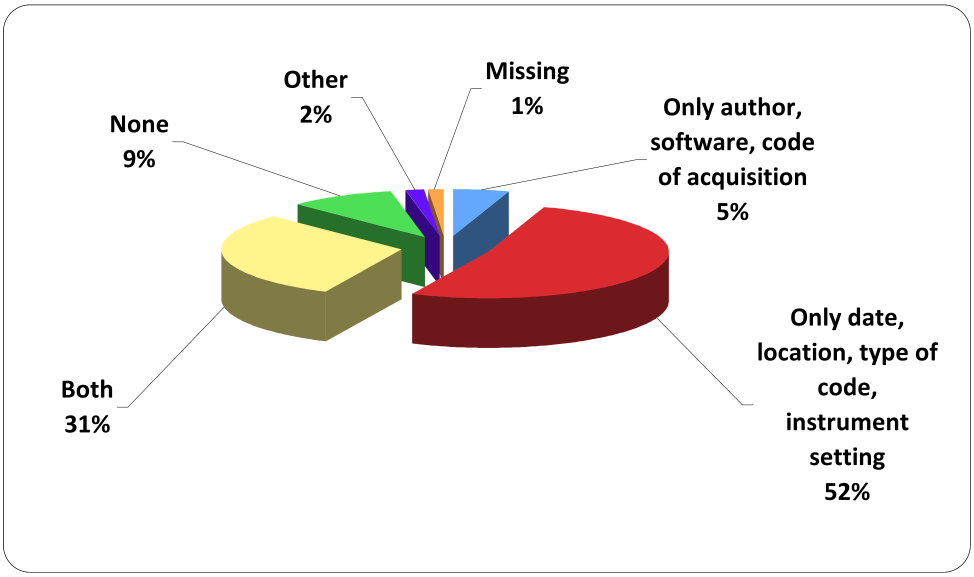
Fig. 3. - Distribution of respondents to the question: "Does your community of reference use standards to manage data?



When asked about the use of standards, a high percentage of researchers reported that their community of reference doesn't use standards (39.6%), while 26% of them don't know about the use of standards in their research field (fig. 3).

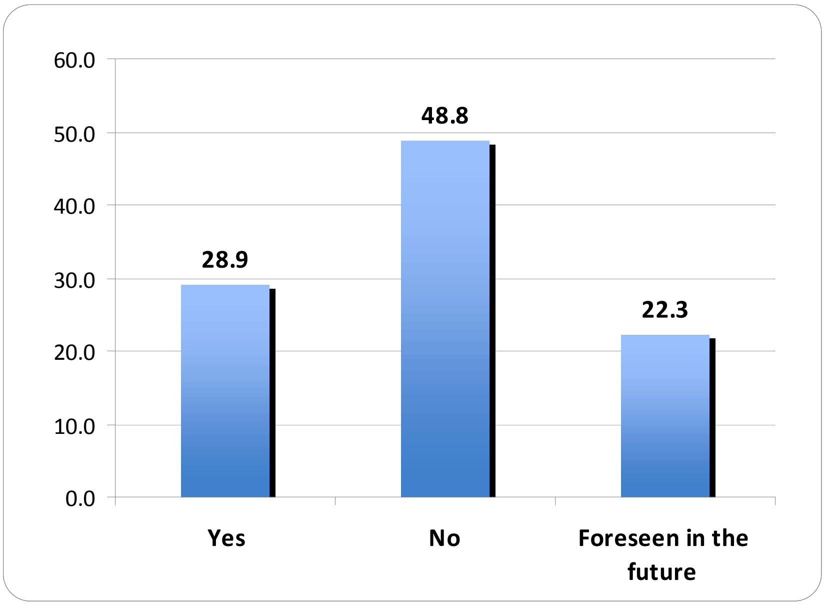
The remaining 26% of researchers that answered positively to this question also specified the standard they more frequently use. Many of them use a set of standards specific to the type of data and infrastructure of reference for their work. Here a brief overview of the standard more frequently indicated. Many mention the European initiative INSPIRE (Infrastructure for Spatial Information in the European Community) that established a general framework for Spatial Data Infrastructure (SDI) together with ISO19115 (Geospatial metadata) as well as the standard developed by the Open Geospatial Consortium (OGC). Others rely on the SEG Y standard file format developed by the Society of Exploration Geophysicists for storing geophysical data, or on NetCDF (Network Common Data Form), an open standard for sharing array-oriented scientific data, and/or on ISO/WMM (World Meteorological Organization) to standardize meteorological data.

Fig. 4. - Distribution of respondents to the question: "What type of additional information do you generally associate with data you have collected/analysed?"



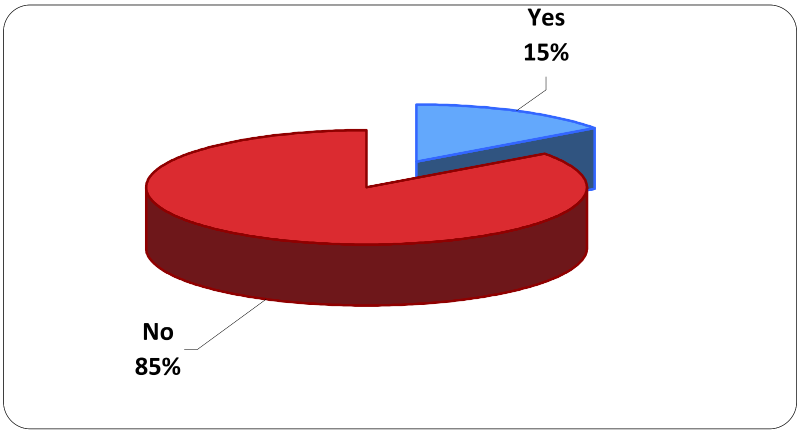
When data are collected and/or analysed, 52% of CNR researchers provide metadata related to the date of collection, information on location, type of code used and instrument setting. 9.4% of researchers associate data with additional information on the author, software, code of acquisition, while 30.6% associate both types of the above-mentioned metadata. Only 9.4% do not associate any type of metadata to data gathered or analysed (fig. 4). The addition of descriptive metadata is an encouraging result as it makes research data more easily interpretable and reusable, thus more accessible and better suited for preservation.

Fig. 5. - Distribution of respondents to the question:
"Does your Institute have specific procedures for data preservation in place?"



When asked whether specific procedures for preservation are set up by their institute, 28,9% of researchers reported that these procedures are in place in their institutes, while 22,3% reported that these procedures are going to be set up in the future (fig. 5).

Fig. 6. - Distribution of respondents to the question:
" Is there anyone in your Institute who is specifically trained to manage data?"



The presence of personnel specifically trained to manage data is reported by 15.4% of researchers, while the majority of them answered that there is no one in their institute that is in charge of this task (fig. 6). Out of 79 researchers that reported on the presence of personnel dedicated to data preservation, 60 indicated the type of personnel that carry out this task. Generally they are IT experts that manage local databases, GIS, digital images. Many researchers mention technicians or researchers that carry out this task, and only in few cases do respondents refer to a data manager, that is the emerging professional skill often mentioned in official documents on data management and preservation. One respondent reports that data preservation is carried out by the same person who manages the Institutes' publications, probably a librarian.

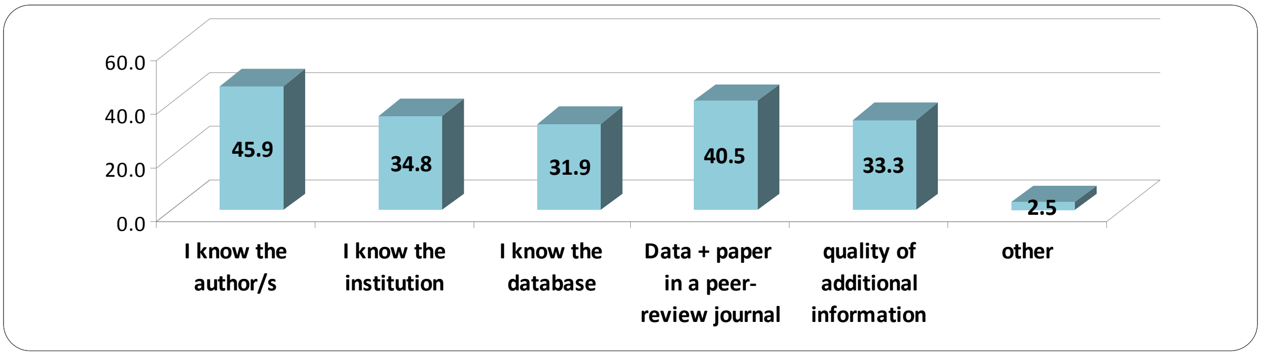
4.6. Data re-use and availability

This group of questions aims to ascertain whether researchers use data produced by others, in which field, along with the criteria they apply to consider data reliable. Generally the use of data generated by others is associated with the propensity of sharing researchers' own data, in the hypothesis that this could represent a mutually coherent behaviour. Results of other surveys (PARSE.Insight, 2009, Tenopir et al., 2011, Enke et. al., 2012) generally showed a lower percentage of data sharing when compared with the re-use of data generated by other researchers.

59% (= 307) of researchers indicate that they use data produced by others. Among them, 43% re-uses data in the same disciplinary field, while a similar percentage of researchers re-use data coming both from the same disciplinary and from cross-disciplinary fields.

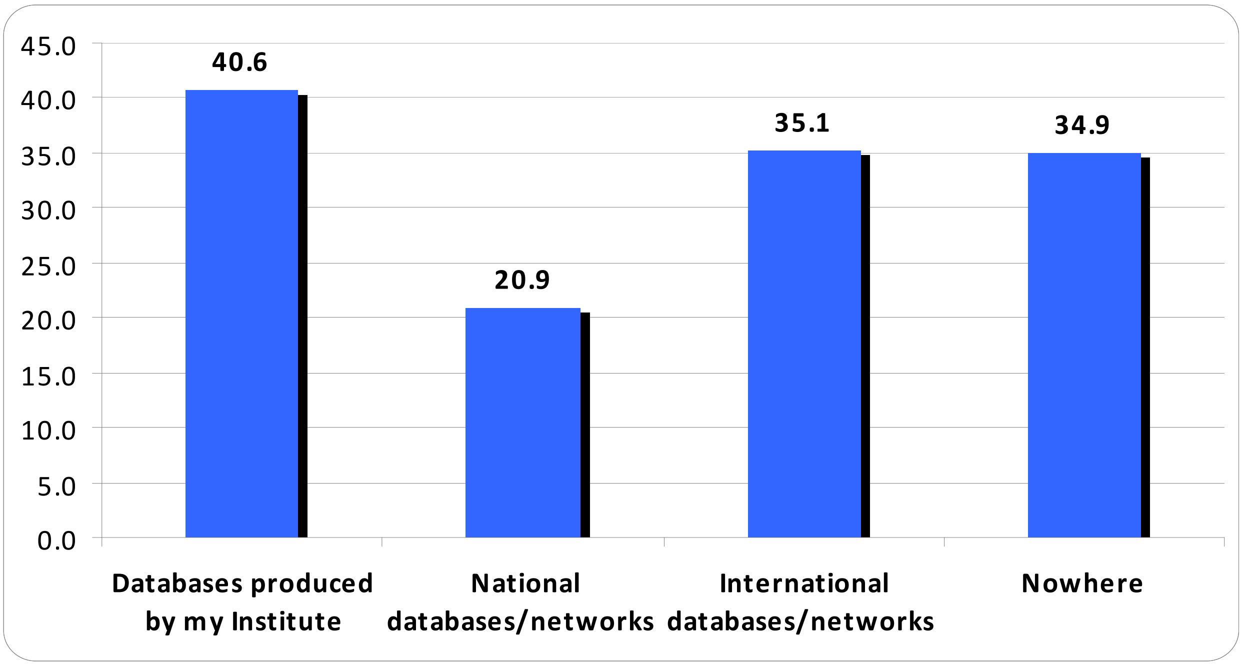
When researchers are re-using data produced by others they consider data reliable if they know the authors (45.9%) and when data are associated with peer-reviewed journals (40%) (fig. 7). Answers reported in the variable "Other" help to give a more complex picture of data reuse. Some researchers reported that they apply procedures of quality control and validation; others consider the experimental method adopted as well as methods and instrumentation used to collect data. This highlights that data reuse may not always be a straightforward process.

Fig. 7. - Distribution of respondents to the question:
"What reassures you that the data produced by others is reliable?"



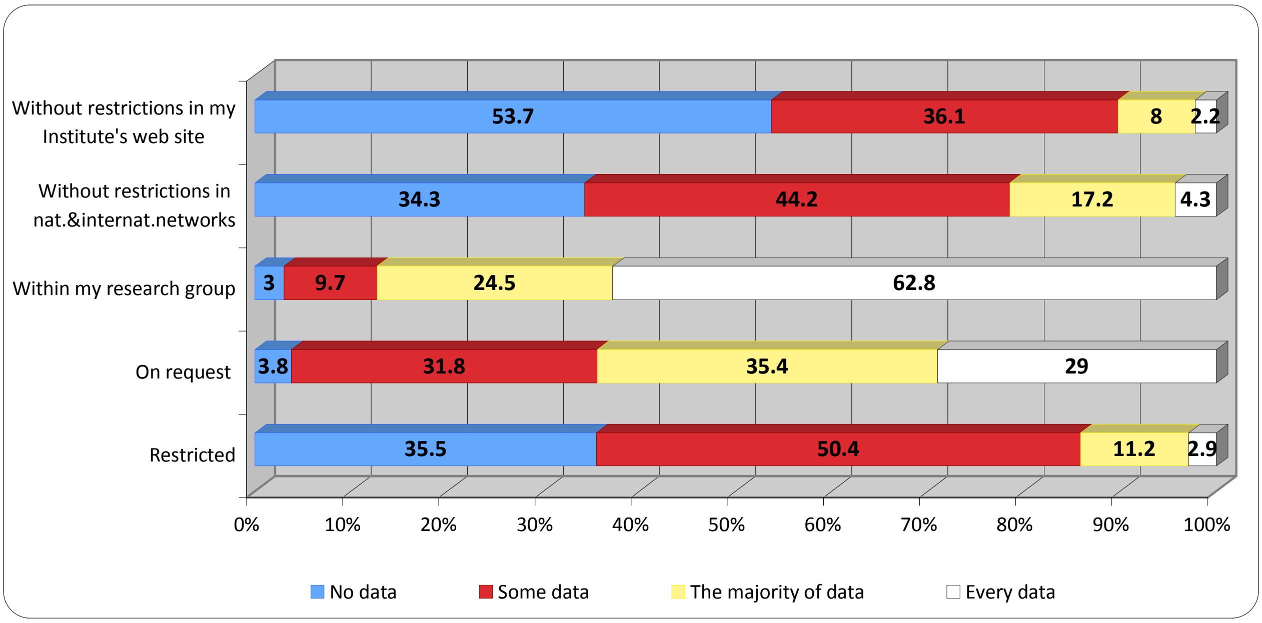
The willingness to share data also depends on the availability of databases or infrastructures where researchers can deposit their research data. For this reasons we asked whether there are databases or networks where they can deposit their data in their disciplinary field.

Fig. 8. - Distribution of respondents to the question:
"In your disciplinary field are there archives where your research data can be stored?"



More than 40% of researchers store their data in databases produced by their institutes, 35% in international databases and 20% in national databases, while for 34% of researchers there are no databases where their data can be submitted (fig. 8).

Fig. 9. - Distribution of respondents to the question:
" Data from your current research is available to everyone without restrictions "



Turning to data made available by CNR researchers, we can generally say that they make a selection of the data that they share (fig. 9). At least some data are available without restriction in the Institute's website (36.1%), or in national and international networks (44.2%). Of course all data are available within their research groups (62.8%). It is interesting to note that when data are requested, CNR researchers declare they do make them available, only 3.8% of them report that no data are available on request. A small percentage of researchers report that either all their data or the majority are restricted (2.8% and 11,2).

5. Researchers' attitude

This part of the questionnaire intends to explore researchers' opinions on the role played by research data, perceived obstacles and enablers to data sharing. As previously mentioned this part of the questionnaire is also based on other surveys carried out at international level, so that differences and communalities with the international context can be compared.

Before asking on data sharing practices and perceptions, we considered it important to let researchers express their opinions on reasons for the availability and preservation of data. We proposed a list of nine well-known statements (7 of which were taken from the Parse Insight survey) and asked whether they consider these reasons very important, important, slightly important or not important.

Almost all assertions of this self-evident list of reasons are considered very important or important by the majority of researchers. If we analyse how they ranked their importance, it emerges that researchers find that data availability and preservation foster the process of science (56.8%) and that it also enhances the transparency of research (53.9% very important and 40.7% important).

Table 3 - Distribution of respondents to the question:
"In your opinion for which reasons is it important to make research data available and preserve it?"

	Very important	*Important*	*Not very important*	*Not important at all*	*Missing*
The availability of data enhances the transparency of research results	53.9	40.7	3.8	0.6	1.0
When research is publicly funded, data should be available to anyone	50.7	38.6	7.5	2.1	1.1
The availability of data fosters the progress of science (new research is based on pre-existing knowledge)	56.8	38.2	3.6	0.4	1.0
It is a means to validate the results obtained	40.2	43.6	12.6	1.1	2.5
Existing results can be re examined	34.6	46.3	14.5	2.7	1.9
It can promote collaboration among different fields	39.6	45.5	12.8	0.8	1.3
It has a potential economic value	19.1	39.8	34.2	4.2	2.7
Research data are unique	19.9	40.0	26.8	9.8	3.6
The availability of data reduces the duplication of research efforts	36.9	35.9	18.9	6.3	1.9

Another reason to make data available and preserve them is that research is publicly funded and therefore should be made available to everyone (50.7% very important and 38.6% important). The economic value of data (4.2% not important at all) together with the assertion that data are unique (9.8%) is regarded as the least important reasons for availability and preservation. These two values are not surprising, as also in the Parse Insight project the survey obtained the same results. In the case of CNR researchers these values are balanced against the rate given as important (respectively 39.8% and 40%). Moreover, the Parse Insight survey found out that opinions on the very important and important reasons depended on the disciplinary field of the respondents. CNR researchers consider that data availability and preservation can stimulate the advancement of science like researchers in Humanities, Life sciences, Physical Sciences and Socio-cultural sciences.

5.1. Obstacles to data sharing

When asked on the obstacles of data sharing (table 4), we obtained a more homogenous distribution of responses, especially if we compare this question with the previous one. If we consider both the very important and important values we can notice a common agreement on some obstacles felt by the majority of CNR researchers (where sometimes the important value prevails on the very important one). These are: lack of technical support (41.9% important, 31.4% important) lack of standards (46.3% important, 25.8% very important), but also the fact that data are not evaluated like papers in scientific journals (37.5% very important, 31.5% important).

Table 4. - Distribution of respondents to the question:
"In your opinion what are the main obstacles to data sharing?"

	Very important	*Important*	*Not very important*	*Not important of all*	*Missing*
Lack of funds	31.4	30.6	27.9	5.4	4.8
Lack of standards	25.8	46.3	18.9	3.1	5.9
It requires too much time	16.1	38.0	32.7	7.8	5.4
Difficulties in adoption of standard	13.0	38.4	33.7	8.4	6.5
No technical support	31.4	41.9	16.4	3.6	6.7
There are no archives to submit to	23.3	37.3	23.9	9.2	6.3
Procedures of data sharing are too complicated	10.7	33.8	38.4	10.7	6.3
Loss of data control	19.9	31.4	30.4	12.6	5.7
Data may be misused and/or misinterpreted	22.8	35.6	25.4	10.5	5.7
Data are not evaluated like papers in scientific journals	37.5	31.5	20.8	5.0	5.2
Loss of exclusivity of the work	26.4	29.4	30.0	8.6	5.5

When we look at the least important perceived barriers, the statement related to the too complicated procedures receives the majority of responses, but the differences with the researchers that consider it important are not so high. At this stage of the analysis we could say that CNR researchers perceive a relevant number of different barriers as being rather important. It would be interesting to further analyse these perceptions combining these results with other variables of the questionnaire.

5.2. Enablers of data sharing

The last multiple-structured question summarizes some of the issues already investigated, but specifically asked on the conditions required to submit data to an open archive. A major consensus on very important facilitators of data deposit is evident here (table 5). The majority of researchers find very important to have the possibility to update data after submission (60.2%), to know who is using them, when and for which purpose (53.5%), to be contacted if data are used (52%). All these responses are related with a clear wish to keep control over their own data also after submission. Another very important factor that may encourage researchers to deposit is the availability of simple procedures for submission (52.6%) as well as receiving the same evaluation as in the case of publications.

Table 5. - Distribution of respondents to the question:
"What condition would you require to submit your research data to an open archive?"

	Very important	*Important*	*Not very important*	*Not important at all*	*Missing*
I will be able to update data after submission	60.2	30.8	4.2	1.7	3.1
I will be able to delete data	31.2	33.7	22.0	7.3	5.9
I know who is using data, when and for which purpose	53.5	27.5	11.7	3.6	3.6
Be contacted if someone wants to use my data	52.0	30.2	12.0	2.3	3.4
Receive a formal acknowledgment	35.4	36.1	20.3	4.0	4.2
Be reassured about long-term data preservation	38.6	39.8	13.4	3.3	5.0
Simple procedures to deposit data	52.6	37.1	5.5	0.6	4.2
Receive additional funds	24.7	39.2	27.5	4.4	4.2
Receive the same evaluation received for publications	41.1	37.5	13.0	3.8	4.6

6. Conclusions

Summarising some of the main results of the survey, CNR researchers in the field of environmental sciences tend to work in collaboration, often involved in multidisciplinary projects within the same institutes and with external organisations. They mainly carry out experimental research, use different types of data, gathered directly by themselves or by their research group in both laboratory and field work, using instrumentation directly managed by CNR.

There is not a diffuse use of standards, but researchers who use them apply different types of them, according to the data they are working on. Nevertheless, data collection is often associated with descriptive metadata that represent a pre-requisite for data reusability and interpretation as well as for preservation. It is also encouraging that a relevant number of researchers rely on procedures for data preservation already set up in their institutes or foreseen in the future. This process is generally carried out by the researchers themselves, as the majority of them do not have any support from specifically trained data managers.

Despite the use of data produced by others, CNR researchers tend to share only a fraction of data they produce. Generally they are more willing to share data on request, keeping control on whom is using their data and for which purposes.

A relevant number of obstacles are perceived by CNR researchers as rather important: lack of technical support, lack of standards, no formal recognition of practices of data sharing, but also lack of funds, fear of losing the exclusivity of their work. These perceptions are worth further analysis, combining these results with other variables of the questionnaire. Conditions required to submit research data to open archives concern both technical and policy-related aspects that confirm a clear wish to keep control over research data even after submission as well as the provision of simple procedures for submitting them. Doubtless a further motivation is that data sharing is evaluated the same way as publications are.

Generally the high rate of responses received to the questionnaire as well as researchers' opinions on the importance of research data indicate a high level of awareness and an encouraging willingness to share data that should be further strengthened by the introduction of policies and the development of infrastructures tailored to researchers' needs.

References

Bach Kerstin, Schäfer Daniel, Enke Neela, Seeger Bernhard, Gemeinholzer Birgit, Bendix Jörg (2012). A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. Ecological Informatics, 11: 16-24.

Bendix Jörg, Nieschulze Jens, Michener William K. (2012). Data platforms in integrative biodiversity research. Ecological Informatics, 11: 1-4.

Borgman Christine L. (2012). The Conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 64 (6): 1059-1078.

Dallmeier-Tiessen Sunje, Darby Robert, Gitmans Katrin, Lambert Simon, Suhonen Jari, Wilson Michael (2012). Compilation of results on drivers and barriers and new opportunities. ODE Project (Opportunity for Data Exchange) URL: www.ode-project.eu/ode-output

EU Directorate - General for Research and Innovation (2012) Online survey on scientific information in the digital age. URL: http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

Graaf Maurits van der, Waaijers Leo (2011). KE Knowledge Exchange Primary Research Data Working Group. A Surfboard for Riding the Wave: Towards a Four Country Action Programme on Research Data. URL: http://www.voced.edu.au/content/ngv48428>

Enke Neela, Thessen Anne, Bach Kerstin, Bendix Jörg, Seeger Bernhard, Gemeinholzer Birgit (2012). The user’s view on biodiversity data sharing – Investigating facts of acceptance and requirements to realize a sustainable use of research data. Ecological Informatics, 11, September 2012, pp. 25-33.

Hey Tony, Tansley Stewart, Tolle Kristin (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. s.l. Microsoft Cooperation. URL http://research.microsoft.com/en-us/collaboration/fourthparadigm/contents.aspx

Kim Youngseek, Stanton Jeffery M. (2012). Institutional and individual influences on scientists’ data sharing practices. Journal of Computational Science Education, 3 (1), June 2012

Kowalczyk Stacy, Shankar Kalpana (2011), Data sharing in the sciences. Ann. Rev. Info. Sci. Tech., 45: 247–294. doi: 10.1002/aris.2011.1440450113

Michener William K., Allard Suzie, Budden Amber, Cook Robert B., Douglass Kimberly, Frame Mike, Kelling Steve, Koskela Rebecca, Tenopir Carol, Vieglais David A. (2012). Participatory design of DataONE-Enabling cyberinfrastructure for the biological and environmental sciences, Ecological Informatics, 11, September 2012, pp. 5-15

Milia Nicola, Congiu Alessandra, Anagnostou Paolo, Montinaro Francesco, Capocasa Marco, Sanna Emanuele, Destro Bisol Giovanni (2012). Mine, yours, ours? Sharing data on human genetic variation. PLoS ONE 7(6): e37552.

National Science Foundation (2005). Long-lived Digital Data Collections: Enabling Research and Education in the 21st century. URL: http://www.nsf.gov/pubs/2005/nsb0540/start.jsp

PARSE.Insight (Insight into issues of Permanent Access to the Records of Science in Europe). D3.4 Survey Report. URL: http://www.parse-insight.eu/publications.php

Pinowar Heater A. (2011). Who shares? Who doesn’t? Factors associated with openly archiving raw research data. PLosONE, 6 (7) e18657.

Tjalsma Heiko, Rombouts Jeroen (2010). Selection of research data. Guidelines for appraising and selecting research data. Dans Studies in Digital archiving, 6. The Hague and Delft. SURFfoundation, Data Archiving and Networked Services (DANS), 3TU.Datacentrum. URL: http://www.dans.knaw.nl/en/content/categorieen/publicaties/dans-studies-digital-archiving-6

Tenopir Carol, Allard Suzie, Douglass Kimberly, Aydinoglu Arsev Umur, Wu Lei, Read Eleanor, Manoff Maribeth, Frame Mike (2011). Data sharing by scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. URL: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0021101

Acknowledgment

We thank Cristiana Crescimbene (CNR-IRPPS) for the technical support in the questionnaire submission and in the management of the LimeSurvey in all phases of the survey. We are also grateful to our CNR colleagues that voluntary answered this questionnaire.

Tracking the Influence of Grey Literature in Public Policy Contexts: The Necessity and Benefit of Interdisciplinary Research

**Bertrum H. MacDonald, Elizabeth M. De Santo,
Kevin Quigley, Suzuette S. Soomai, and Peter G. Wells**
Dalhousie University, Halifax, Nova Scotia, Canada

Abstract

Scientific information (much of it published as grey literature) can play a pivotal role in the search for solutions to serious global environmental problems. This fact is receiving growing attention by a diversity of researchers. How information functions within the interface between science and policy is only weakly understood, in part because most studies have been conducted through single disciplinary lenses. Moreover, determining the life cycles of scientific information and developing an understanding of the use and influence of this information are not trivial tasks. We believe that an appreciable increase in understanding can be achieved through an interdisciplinary perspective and a comparative approach employing a suite of research methodologies to document information pathways. In particular in our research (see www.eiui.ca), we contend that interdisciplinary research, drawing on "information science and management," "marine environmental science," "marine policy development," "fisheries science and management," and "public policy," can substantially increase understanding of the processes by which scientific information is incorporated into environmental policy decisions. This innovative, evolving interdisciplinary perspective enables addressing the question "what role and influence does grey literature have in marine environmental policy and decision-making processes" in an informative, holistic manner, otherwise unfeasible. As this paper shows, multidimensional thinking and analysis stimulated by an interdisciplinary perspective is essential for understanding the role of scientific information at the science-policy interface in marine environmental fields.

Introduction

In 2008, in his book entitled *Environmental Reform in the Information Age: The Contours of Informational Governance*, Arthur P. J. Mol, professor of environmental policy at Wageningen University in The Netherlands, stated forthrightly that "it is the production, the processing, the use and the flow of, as well as the access to and the control over, information that is increasingly becoming vital in environmental governance practices....and the motivations and sources for changing unsustainable behaviour are increasingly informational" (Mol, 2008). That environmental degradation is a serious global problem has been recognized for decades. September 2012, for example, marked the fiftieth anniversary of Rachel Carson's iconic *Silent Spring,* one of the most influential books of the twentieth century (Carson, 1962). Earlier in 2012, the United Nations Conference on Sustainable Development (the "Rio+20" conference) held in Rio de Janeiro, Brazil, forcefully highlighted the level of international commitment needed to halt environmental breakdown.

In the lead-up to the Rio+20 meeting, the Planet Under Pressure conference, held in London, England on 26-29 March 2012, stressed the seriousness of the matter:

> Research now demonstrates that the continued functioning of the Earth system as it has supported the well-being of human civilization in recent centuries is at risk. Without urgent action, we could face threats to water, food, biodiversity and other critical resources: these threats risk intensifying economic, ecological and social crises, creating the potential for a humanitarian emergency on a global scale. (Planet Under Pressure, 2012, p. 1)

Alarmist as this statement may seem, the London conference, attended by nearly 3,000 leading experts and decision-makers, sought a way forward while recognizing that new solutions would inevitably be required. The "State of the Planet Declaration," approved at the conference, boldly proclaimed:

> The challenges facing a planet under pressure demand a new approach to research that is more integrative, international and solutions-oriented. We need to link high-quality focused scientific research to new policy-relevant interdisciplinary efforts for global sustainability. This research must integrate across existing research programmes and disciplines, across all domains of research as well as local knowledge systems, across the North and South, and must be co-designed and implemented with input from governments, civil society, research funders, and the private sector. (Planet Under Pressure, 2012, p. 3)

Vast quantities of relevant scientific information have been generated, many solutions have been proposed, and some implemented in recent decades to address environmental problems. However, solutions can be slow in coming, limited in scope, or may even be thwarted by competing, sometimes opposing and fragmented views and an overemphasis on uncertainty rather than a need for precaution. Vociferous debate over the reality and effects of global climate change is one example, although increased understanding is winning out.

Delay in pursuit of solutions is no longer an acceptable strategy even though the hurdles are challenging. A new approach that links "high quality focused scientific research to new policy-relevant interdisciplinary efforts" may, in fact, achieve desirable results. Interdisciplinary research, which encompasses *all* relevant disciplines, is needed because as the "State of the Planet Declaration" noted, "The Earth system is a complex, interconnected system that includes the global economy and society, which are themselves highly **interconnected and interdependent**" (Planet Under Pressure, 2012, p. 2, emphasis in original), and such a system requires a holistic approach to research and understanding.

Interdisciplinary effort is already witnessed in global initiatives such as the Intergovernmental Panel on Climate Change (IPCC). Founded in 1988 by the World Meteorological Organization and the United Nations Environment Programme and later endorsed by the General Assembly of the United Nations, the IPCC produces comprehensive scientific assessments of current scientific, technical, and socio-economic information related to the risk of climate change (Bolin, 2007). The IPCC also played an instrumental role in the creation of the UN Framework Convention on Climate Change (UNFCCC), the main international treaty to address the causes and consequences of climate change. Drawing on the assistance of thousands of scientists and other experts, and obtaining the consensus of the more than 120 country signatories, the IPCC has produced four major climate assessment reports with the fifth planned for publication at the end of 2014 (IPCC, 2012a). These grey literature reports are the result of the review of a massive number of both primary and grey research literature publications. In addition to these periodic assessments, the IPCC also publishes special reports on subjects related to the implementation of the United Nations Framework Convention on Climate Change, the most recent of which is an almost 1,100 page report on *Renewable Energy Sources and Climate Change Mitigation* (IPCC, 2012b) released in advance of the meetings in Doha, Qatar in December 2012. During the preparation of this report, over 24,000 (24,766) comments were received as the text was being reviewed and finalized. This number alone highlights the extent of the effort to produce authoritative environmental assessment reports. Even though some aspects of the IPCC's work have proven controversial, including its use of grey literature to support some conclusions (see, Meyers & Petersen, 2010; Ravindranath, 2010), this impressive international initiative highlights the immense value and influence of interdisciplinarity in a key environmental field (Bjurström, & Polk 2011a; 2011b).

Commitment to an interdisciplinary perspective does not guarantee that all relevant disciplines have been brought to bear in the search for solutions, however. Information science (information studies, information management, informatics) is sometimes overlooked or is not at the table in some research initiatives. A case in point is the recent report of the Committee on the Use of Social Science Knowledge in Public Policy of the National Research Council in the United States, *Using Science as Evidence in Public Policy*, published by The National Academies Press (Prewitt, Schandt, & Straf, 2012). This report proposes "a framework for research on how policy makers make use of scientific knowledge and how the results of that research might lead to improved policy making and improved preparation of students in policy schools for careers in the policy world" (p. vii). The report makes no mention of the field of information behaviour, the knowledge and tools of which are surely important for developing an understanding of how people become aware of, use, and are influenced by information.

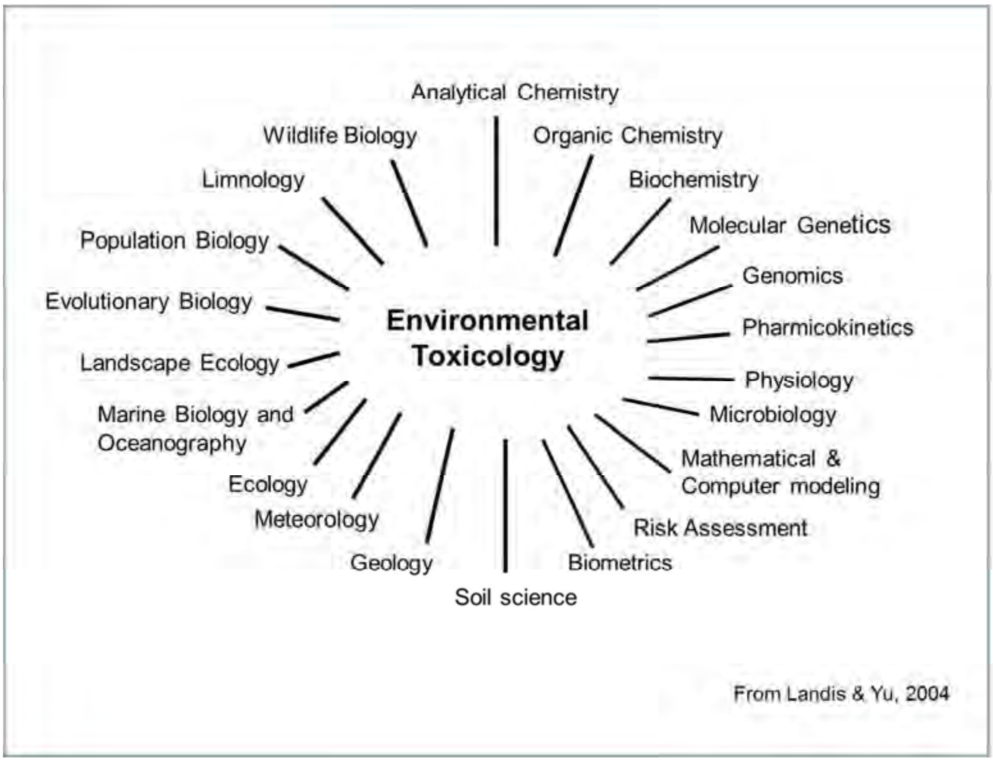
While an information studies perspective may be explicitly missing from some interdisciplinary undertakings, its absence may be more a matter of language or oversight than actual failure to recognize the contribution of this discipline. For example, the recent book, *Knowledge and Environmental Policy. Re-imagining the Boundaries of Science and Politics,* which draws on research based in political science, and environmental and natural resource policy, employs the term "knowledge" rather than "information" to convey research perspectives that govern information studies points of view (Ascher, Steelman, & Healy, 2010). An explanation of why some disciplines such as information science are overlooked or entirely missing from interdisciplinary research initiatives may be attributed to stereotypical misunderstanding of the potential contributions of such disciplines to such collaborations. Moreover, information specialists may overlook the benefits of working with researchers in other disciplines whose appreciation of the role of grey literature will likely be quite different from their own. These challenges notwithstanding, it is incumbent upon us to seize opportunities for interdisciplinary research along the lines of the urgent appeal of the 2012 "State of the Planet Declaration."

Interdisciplinarity

Interdisciplinary thinking is not new and has increasingly characterized some areas of scientific research for the past half century or more (e.g., toxicology, oceanography, ecology, biomedical sciences, and environmental sciences). Some might argue that twenty-first century interest in interdisciplinarity is a return to the perspective of Renaissance or Victorian scholars (see, for example, Lightman, 2012), except that today the vast and growing quantity of information and highly specialized techniques, methods and instrumentation, and knowledge have left "most scholars and artists stranded in ever-shrinking islands of competence" (Nissani, 1997). Even though research funding bodies in North America and Europe have given greater prominence to interdisciplinary research and interdisciplinary programs and administrative units have been established within universities (e.g., the Canadian Mountain Studies Initiative at the University of Alberta - www.mountains.ualberta.ca/en/ThinkingMountains.aspx), interdisciplinarity generates no shortage of debate. Moreover, a sizeable number of individuals within and outside "the academy" find it difficult to work at the intersections of their disciplinary boundaries with other researchers and practitioners who operate with different but complementary disciplinary points of view. According to Luhmann (1993), a paradox exists in modern society. The more systems evolve and specialize, the more critical it is for communication and coordination between these systems. Yet, at the same time these systems become more self-referential and unable to communicate between themselves.

Some fields of inquiry are inherently interdisciplinary, however. Take for example, environmental toxicology. This field "takes and assimilates from a variety of disciplines" (Landis & Yu, 2004, p. 1), as Figure 1 illustrates. Terrestrial and aquatic ecologists, chemists, molecular biologists, geneticists, and mathematicians all contribute to the evaluation of impacts of chemicals on biological systems (Landis & Yu, 2004). As the authors note in their *Introduction to Environmental Toxicology*, "biometrics...provides the tools for data analysis and hypothesis testing. Mathematical and computer modeling enables the researcher to predict effects and to increase the rigor of a hypothesis. Evolutionary biology provides the data for establishing comparisons from species to species and describes the adaptation of species to environmental change" (pp. 1-2). Even though this field is decidedly multi- and interdisciplinary, some disciplines are notably absent in Figure 1. Chemical toxicity can trigger far reaching effects in human society with social, economic, and/or political implications (e.g., the use of Agent Orange in Vietnam in the 1960s). As a consequence, social science disciplines could quite easily be included in the sizeable suite of largely natural and physical science disciplines populating this diagram. Although only one example, it is likely that even established and highly interdisciplinary fields of research and practice can benefit from broadening and illustrating the scope of their disciplinary perspectives.

Figure 1. Environmental Toxicology and Some of Its Components



From Landis & Yu, 2004

As researchers have pursued interdisciplinary work over recent decades, they have generated an extensive body of literature on the subject of interdisciplinarity and the flow of publications continues (e.g., Huutoniemi, Klein, Bruun, & Jukkinen, 2010). Since interdisciplinarity can be difficult to achieve in some contexts, interdisciplinary research or knowledge can be misunderstood of undervalued. This latter perspective emphasizes pitfalls to interdisciplinary study, and there are such; nonetheless, the potential and achievable benefits can be measurable (see Campbell, 2005; Elfner, et al., 2011; MacMynowski, 2007; Nuijten, 2011; Pickett, Burch, & Grove, 1999; Strang, 2009; Turner & Carpenter, 1999).

In the words of one scholar, “interdisciplinarity is best seen as bringing together distinctive components of two or more disciplines” (Nissani, 1997, p. 203). “In academic discourse,” Nissani has written, “interdisciplinarity typically applies to four realms: knowledge, research, education, and theory.”

> *Interdisciplinary knowledge* involves familiarity with components of two or more disciplines. *Interdisciplinary research* combines components of two or more disciplines in the search or creation of new knowledge, operations, or artistic expression. *Interdisciplinary education* merges components of two or more disciplines in a single program of instruction. *Interdisciplinary theory* takes interdisciplinary knowledge, research, or education as its main objects of study. (Nissani, 1997, p. 203)

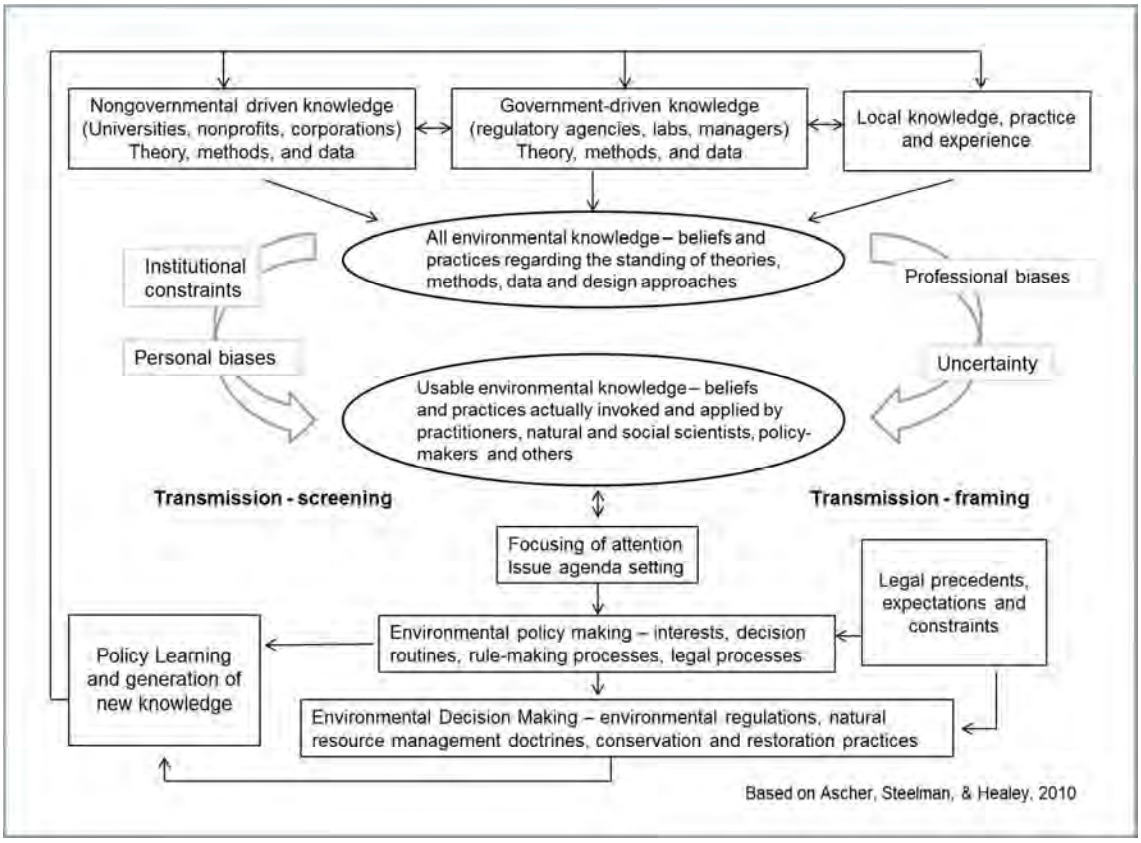
While we are interested in each of the four realms in this suite of options, we have chosen to collaborate and focus primarily on interdisciplinary research due to the benefits that arise when considering the multiple dimensions comprising the information-communication-policy interface or, more simply, the science-policy interface. We recognize that interdisciplinary research will result in unique and hopefully important interdisciplinary knowledge. Our research also flows into our educational work, which will see greater emphasis in a new graduate course entitled “The Role of Information in Public Policy and Decision Making,” being offered in 2013 at Dalhousie University.

The Challenge - The Science-Policy Interface

The necessity of interdisciplinary investigation becomes clear when the complexity of the science-policy interface is described. Tracking the movement and use of grey literature in this context poses challenges in framing research questions, determining what data to collect, deciding which methodologies or suite of methodologies must be applied or developed, and gaining access and establishing the trust of numerous stakeholders to undertake research within the ambit of governmental organizations, all the while appreciating that many dimensions of personality, culture, economics, politics, and social factors contribute to the processes of decision making and policy development. The magnitude and variation of these components outstrip the capacity of expert understanding of any single discipline. Quite simply, interdisciplinary is required. Figure 2 attempts to capture many of these dimensions beginning with knowledge generation through to policy formulation, decisions, and generation of new knowledge.

Figure 2 illustrates the complexity of the processes in which information is used (or not) in the activities within the science-policy interface. Two features should be noted. First, information comes from a variety of sources – non-governmental driven sources, governmental sources, and local knowledge. These three interact in interesting and complex ways to contribute to all environmental knowledge. Second, filters are typically applied to this body of all environmental knowledge resulting in usable knowledge. Those filters may be institutional cultural factors or individual biases, or they may involve professional biases and the uncertainty characteristic of scientific observations. This knowledge is then filtered even further due to attention and focus of particular agendas. This information may inform policy making in a context that is influenced by legal precedents and other constraints. Ultimately, the policymaking and decision making can stimulate new questions and new knowledge and a feedback loop will occur. This summation, which has simplified a complex amalgam of activities, shows that the scenarios encapsulated by the diagram are far from trivial processes (Nutley, Walter, & Davies, 2010).

Figure 2. Generation, Transmission, and Use of Environmental Information


Nongovernmental driven knowledge (Universities, nonprofits, corporations) Theory, methods, and data
Government-driven knowledge (regulatory agencies, labs, managers) Theory, methods, and data
Local knowledge, practice and experience
All environmental knowledge – beliefs and practices regarding the standing of theories, methods, data and design approaches
Institutional constraints
Professional biases
Personal biases
Uncertainty
Usable environmental knowledge – beliefs and practices actually invoked and applied by practitioners, natural and social scientists, policy-makers and others
Transmission - screening
Transmission - framing
Focusing of attention Issue agenda setting
Legal precedents, expectations and constraints
Policy Learning and generation of new knowledge
Environmental policy making – interests, decision routines, rule-making processes, legal processes
Environmental Decision Making – environmental regulations, natural resource management doctrines, conservation and restoration practices

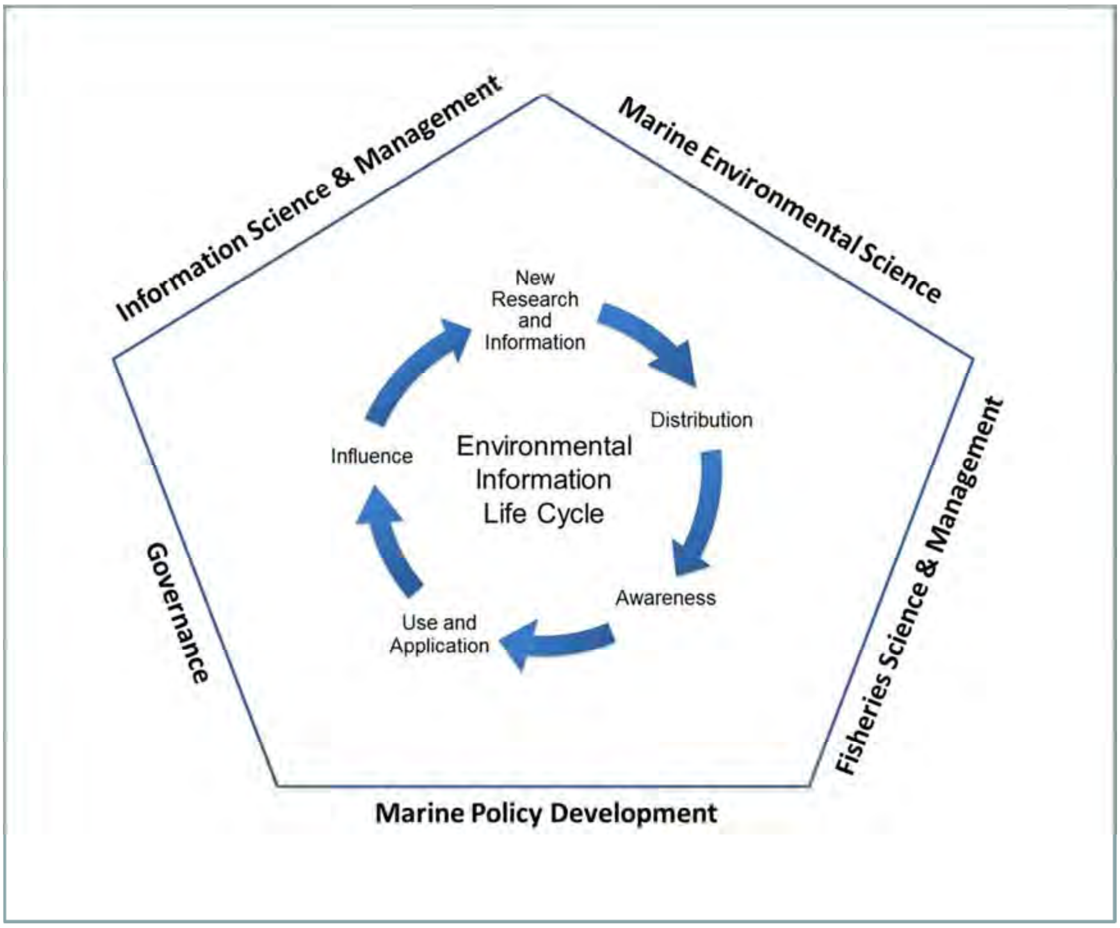
Based on Ascher, Steelman, & Healey, 2010

Environmental Information: Use and Influence (EIUI) Initiative as an Example of Interdisciplinary Research

In a paper entitled, "From research to policy and back," Aletha C. Huston stated that "policymakers are influenced by politics, ideology, interest groups, and the institutional rewards and pressures in governmental agencies. They sometimes appear indifferent to empirical evidence" (Huston, 2008, p.1). Some governments may ignore or discourage evidence-based policy making (Grisé, 2013). As a result, the "unpredictable and volatile world of ... policy has led some researchers to renounce efforts to inform it because they believe that decisions are entirely political and that data are invoked at best only to support a position that someone has already decided to endorse" (Huston, 2008, p. 1). Such pessimism, while understandable, may be a reflection of attempting to understand the activities of the science-policy interface through a single disciplinary lens, an approach that is bound to falter in gaining a meaningful appreciation of both visible and underlying dynamics of decision making work.

We are fortunate to work in an interdisciplinary team, which is making headway on opening new avenues of research to advance understanding of the use and influence of information published as grey literature in governmental and intergovernmental arenas. The composition of this team has been achieved in part because our multidisciplinary academic unit at Dalhousie University is staffed by faculty members of diverse disciplinary backgrounds, with interests in various aspects of management that demand and benefit from collaboration. Our team began with specialists in Information Management and Marine Environmental Science and expanded to include Governance and Public Administration, Marine Policy Development, and Fisheries Science and Management (Figure 3). More recently, additional expertise in Information Management, namely Social Network and Social Media Analysis, has been included and research students at the Masters and Doctoral levels complement the team. Individually, each of these disciplines tend to be outward looking, i.e., tend to draw on theories and methodologies of a variety of disciplines, which may contribute to shared interest in gaining greater understanding of the life cycle of environmental and fisheries information in policy development and decision making contexts (e.g., MacDonald, Cordes, & Wells, 2007; Soomai, Wells, & MacDonald, 2011; Soomai, MacDonald, & Wells, 2013).

Figure 3. Environmental Information: Use and Influence Interdisciplinary Research (from www.eiui.ca)


Information Science & Management
Marine Environmental Science
Fisheries Science & Management
Marine Policy Development
Governance
New Research and Information
Distribution
Awareness
Use and Application
Influence
Environmental Information Life Cycle

Benefits of Interdisciplinary Research

In a 1997 paper, Moti Nissani of Wayne State University in the United States suggested that the rewards of interdisciplinary knowledge and research fall within three overlapping categories: 1) growth of knowledge, 2) social benefits, and 3) personal rewards. Including the personal rewards (which are genuine and motivating), our EIUI research (see www.eiui.ca) has demonstrated the potential of substantial contributions to growth of knowledge with related social implications as noted in Table 1. The results of our efforts will have practical value in a world that is not demarcated by single disciplines.

Table 1 – Some Benefits of Interdisciplinary Research Identified by the EIUI Researchers

Category*	Benefits
# 1 - Growth of Knowledge	Big problems have many dimensions, requiring enquiry from multiple perspectives. Creativity in problem solving is facilitated and extends beyond the comfort zone(s) of disciplines. Greater capacity exists for determining core questions. New methods are developed since measuring use and influence of information is complex.
# 2 - Social benefits	Increased understanding occurs regarding different institutional cultures at the science-policy interface. Many types of stakeholders can be involved as the interdisciplinary perspective requires more comprehensive approaches and understanding.
# 3 - Personal rewards	Credibility increased with funding sources and partner organizations. Important societal issues are resolved.

* Based on Nissani (1997)

Conclusion

The influence of scientific information (much of which is published as grey literature) in the complex social contexts in which information is used can only be well understood if that context is comprehensively studied. Environmental problems and related policy decisions are multidimensional, as the *State of the Planet Declaration* emphasized (Planet under Pressure, 2012). "In one lifetime," the declaration stated, "our increasingly interconnected and interdependent economic, social, cultural, and political systems have come to place pressures on the environment that cause fundamental changes in the Earth system....But the same interconnectedness provides the potential for solution: new ideas form and spread quickly, creating the momentum for the major transformation required for a truly sustainable planet." Information published as grey literature will have an instrumental role to play in facilitating these solutions, as shown by the IPCC work on climate change. Thus, development of a clearer understanding of the function of this information and literature and its exchange at the science-policy interface, where these solutions will play out, is urgently needed.

Even if greater information exchange between communities is possible, problems persist. A cybernetic understanding of control points to three components to a control system, namely, information gathering, standard setting, and behaviour modification. Information exchange in the absence of common standards and behaviour modification will leave a system uncontrolled. In some respects, information sharing can be a "light touch" form of regulation. Problem solvers assume that, by sharing information, standards and behaviour modification will occur. Without an appropriate incentive structure, this latter scenario might be wishful thinking.

The challenge, therefore, will be to move beyond information exchange and into the role of establishing standards and changing behavior. This latter stage will be much harder to achieve. It requires power-sharing between disciplines and communities. In many respects, the result is a flat, leaderless approach to problem solving. Egalitarian structures, however, tend to be risk-averse; they strive towards consensus and in so doing potentially neglect and even undermine innovation and risk-taking. In this regard, grey literature may provide a half-way house: an opportunity for different communities to negotiate research findings, take risks, and propose alternatives. In some respects this form of publication lowers the constraints of the high academic standards of each individual discipline in order to allow something more innovative to emerge at the interface. An interdisciplinary research perspective, as we have discussed, is vital to achieve that outcome.

References

Ascher, W., Steelman, T. & Healy, R. (2010). *Knowledge and environmental policy. Re-imgaining the boundaries of science and politics*. Cambridge, MA: The MIT Press.

Bjurström, A., & Polk, M. (2011a). Climate change and interdisciplinarity: A co-citation analysis of IPCC Third Assessment Report. *Scientometrics, 87*, 525-550.

Bjurström, A., & Polk, M. (2011b). Physical and economic bias in climate change research: A scientometric study of IPCC Third Assessment Report. *Climate Change, 108,* 1-22.

Bolin, B. (2007). *A history of the science and politics of climate change. The role of the Intergovernmental Panel on Climate Change.* Cambridge: Cambridge University Press.

Campbell, L. M. (2005). Overcoming obstacles to interdisciplinary research. *Conservation Biology, 19*(2), 574-577.

Carson, R. (1962). *Silent spring*. Boston: Houghton Mifflin.

Elfner, L. E., Falk-Krzesinski, H. J., Sullivan, K.O., Velkey, A., Illman, D. L., Baker, J., & Pita-Szczesniewski, A. (2011). Team science: Heaving walls, melding silos. *American Scientist, 99*(6).

Grisé, Y. (2013, January 4). Unskakle government scientists and let them do their jobs. *The Globe and Mail*. Retrieved from http://www.theglobeandmail.com/commentary/unshackle-government-scientists-and-let-them-do-their-jobs/article6928508/

Huston, A. C. (2008). From research to policy and back. *Child Development, 29*(1), 1-12.

Huutoniemi, K., Klein, J. T., Bruun, H., & Jukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy, 39,* 79-88.

Intergovernmental Panel on Climate Change. (2012a). Key dates in the AR5 schedule. Retrieved from www.ipcc.org

Intergovernmental Panel on Climate Change. (2012b). *Renewable energy sources and climate change mitigation*. Cambridge: Cambridge University Press.

Landis, W. G., & Yu, M-H. (2004). *Introduction to environmental toxicology. Impacts of chemicals upon ecological systems*. Boca Raton: Lewis Publishers.

Lightman, B. (2012). Communicating knowledge to new audiences: Victorian popularizers of science. *Proceedings of the Nova Scotian Institute of Science, 47*(Part 1), 5-31.

Luhmann, N. (1993). *Risk: A sociological theory.* New York: Aldine de Bruyter.

MacDonald, B. H., Cordes, R. E., & Wells, P. G. (2007). Assessing the diffusion and impact of grey literature published by international intergovernmental scientific groups: The case of the Gulf of Maine Council on the Marine Environment. *Publishing Research Quarterly, 23*(1), 30-46.

MacMynowski, D. P. (2007). Pausing at the brink of interdisciplinarity: Power and knowledge at the meeting of social and physical science. *Ecology and Society, 12*(1). Retrieved from http://www.ecologyandsociety.org/vol12/issue1/art20/

Meyers, L. A., & Petersen, A. C. (Eds.). (2010). Assessing an IPCC assessment: An analysis of statements on project regional impacts in the 2007 report. The Hague: Netherlands Environmental Assessment Agency.

Mol, A. P. J. (2008). *Environmental reform in the information age. The contours of informational governance*. Cambridge: Cambridge University Press.

Nissani, M. (1997). Ten cheers for interdisciplinarity: The case for interdisciplinary knowledge and research. *The Social Science Journal, 34*(2), 201-216.

Nuijten, E. (2011). Combining research styles of the natural and social sciences in agricultural research. *NJAS-Wageningen Journal of Life Sciences, 57*, 197-205.

Nutley, S. M., Walter, I., & Davies, H. T. O. (2010). *Using evidence. How research can inform public services*. Bristol: Policy Press.

Pickett, S. T. A., Burch Jr., W. R., & Grove, J. M. (1999). Interdisciplinary research: Maintaining the constructive impulse in a culture of criticism. *Ecosystems, 2*, 302-307.

Planet Under Pressure. (2012). State of the planet declaration. Planet under pressure: New knowledge towards solutions. [London]: International Council of Science.

Prewitt, K., Schandt, T. A., & Straf, M. L. (Eds.). (2012). *Using science as evidence in public policy*. Washington, DC: The National Academies Press.

Ravindranath, N. H., (2010). IPCC: Accomplishments, controversies, and challenges. *Current Science, 99*(1), 26-35.

Soomai, S. S., Wells, P. G., & MacDonald, B. H. (2011). Multi-stakeholder perspectives on the use and influence of "grey" scientific information in fisheries management. *Marine Policy, 35*(1), 50-62.

Soomai, S. S., MacDonald, B. H., & Wells, P. G. (2013, in press). Communicating environmental information to stakeholders in coastal and marine policy-making: Case studies from Nova Scotia and the Gulf of Maine/Bay of Fundy region. *Marine Policy*.

Strang, V. (2009). Integrating social and natural sciences in environmental research: A discussion paper. *Environmental Development and Sustainability, 11,* 1-18.

Turner, M. G., & Carpenter, S. R. (1999). Tips and traps in interdisciplinary research. *Ecosystems, 2,* 275-276.

Grey communities
An empirical study on databases and repositories

Hélène Prost, Institute for Scientific and Technical Information, INIST-CNRS, France
Joachim Schöpfel, Charles de Gaulle University Lille 3, France

Abstract

The paper explores grey communities outside the Grey Literature Network Service (GreyNet) and identifies potential members for GreyNet. GreyNet can be compared to a Learned Society specialised in grey literature as a particular field of library and information sciences (LIS). Its relevance is related to its capacity to enforce the terminology and definition of grey literature in LIS research and publications, and its impact and outreach can be assessed through the proportion of experts dealing with grey literature and connected with GreyNet. From five databases (Web of Science, Scopus, LISTA, Pascal and Francis) and from open repositories we selected 2,440 papers on grey literature published between 2000 and 2012 by 5,490 authors. Publishing features, preferred journals and the number of publications per author are described for the whole sample. For a subsample of 433 authors strongly committed to grey literature, we present data on geographic origins, place of work, scientific domain and profession. We discuss the characteristics of grey communities in and outside of GreyNet and suggest strategies for the further development of the network.

Introduction

In 2012, the Grey Literature Network Service (GreyNet) celebrates its 20th anniversary, with nearly 300 contributors from 30 different countries. In recent years, GreyNet has directed its activities towards open access through the launching of OpenGrey, and the creation of the GreyNet LinkedIn group marked its entry into social networks. The activities of GreyNet, such as workshops, summer schools, curriculum development, discussion list, publications and conferences, contributed to the creation and development of a subject community by means of shared terminology, tools, events, experiences and topics.

GreyNet can be compared to a Learned Society specialised in grey literature as a particular field of library and information sciences (LIS). In fact, it is different from an academic group insofar as it is not confined to LIS but also reaches out to include information professionals and scientists from other disciplines (social sciences, computer sciences, law, economics...). GreyNet is more a kind of special interest group that recruits experts from the field of grey literature in order “to facilitate dialogue, research, and communication between persons and organisations (and to) to identify and distribute information on and about grey literature in networked environments” (GreyNet web site).

The relevance of GreyNet is related to its capacity to enforce the terminology and definition of grey literature in LIS research and publications, and its impact and outreach can be assessed through the proportion of experts dealing with grey literature and connected with GreyNet. In other words, if we want to evaluate the success of GreyNet as a community-creating structure and its potential for future development, we need to know the degree of GreyNet to attain and aggregate all (or at least a significant number of) scientists, academics and information professionals interested in the field of grey literature and contributing to its knowledge.

Our definition of community is pragmatic and follows the sociological approach to science (Kuhn 1962, Latour & Woolgar 1979). We consider a scientific and/or professional community as a social group with interaction and communication, common practice, identity and values, and shared interests, definitions and language (see Callon 1989, Schrecker 2006 or Paganelli 2012).

Our study builds on three other papers on scientific and professional members of GreyNet. In 2005, we analysed the citations of the first five conferences on grey literature (Schöpfel et al., 2005). This first paper defined the “stakeholders” as “those authors who focus their research and writing on the topic of grey literature (...) referred to as the meta-authors on grey literature” and identified 152 authors and co-authors of 139 papers. The citation analysis of 1344 records and 1721 authors or corporate authors revealed that roughly one quarter of these cited authors matched with the “meta-authors on grey literature” of GreyNet, and that another quarter “(...) deals (dealt) with grey literature, but does (did) not explicitly adhere to the term”.

Four years later, when the new *OpenGrey* service (formerly *OpenSIGLE*) was launched, Farace et al. (2009) defined the grey community as the “250 authors/researchers in the GL-Conference Series” and described the integrating role of GreyNet and the new open repository. The paper also mentioned two groups of potential interest for further development, i.e. the former EAGLE member institutions and “new stakeholders in Grey Literature” whoever this may be.

More recently, Marzi (2012) highlighted the particular relationship between terminology and community. She compared the usage of specific terms and the syntactic complexity of the proceedings of the GreyNet conference to social networks (Facebook) and subject-based communities (LinkedIn) and focused on “supporting relationships and content sharing”. Her conclusion was that subject-based collections such as the proceedings offer a more “coherent flow of shared and structured knowledge” than social networks but that subject-based communities such as the GreyNet group on LinkedIn can nevertheless contribute to “knowledge building and informative flow”, if there is a “strong interaction between medium and content (as in subject-based community exchanges)”.

Obviously, sharing common research interests, belonging to a group, attending the same events and speaking the same language are key factors for the definition of a community. GreyNet without a doubt is such a community, with an accepted terminology, reference definitions, social events and vectors of communication.

But if it is relatively easy to determine the core grey community based on membership, publishing behaviour and conference attendance, the frontier between inside and outside the community is hard to find. Surely, often enough we can read on Twitter or elsewhere questions like “what does grey literature really mean?” Often enough we can guess that outside of GreyNet the “inside-terminology” is not really in use and that grey literature remains a rather obscure topic, not quite clear, not well known.

On the other hand, we can also observe that other scientists and professionals from LIS or other domains publish about/on grey literature. Sometimes they apply the concept of grey literature together with the definition of GreyNet, sometimes they don’t for instance, when an article about PhD theses does not mention their grey character.

We were interested in the boundary (should we say continuum?) between inside and outside of the GreyNet community. This time we have not tried to define this boundary through citation analysis (= which authors are working together? Who cites whose publications?...) but through the analysis of usage of terminology and choice of topics. Who uses “grey literature” as an object of research and publication? Who works on documents belonging to grey literature without applying the term “grey literature”?

Our expectation is that this double approach may provide evidence of the outreach and impact of GreyNet beyond its community and of its potential for further development.

Methodology

The first step was a search for publications on grey literature in selected scientific databases. The search was conducted in March and April 2012 in five databases (table 1), applying three criteria:

Document type: The search was limited to published papers.

Time period: We considered documents published between 2000 and 2012.

Content: We searched for references that contain “grey literature” or variants in the title, abstract or keyword fields. Subsequently, we added references on PhD theses or Master dissertations.

The exact approach for each database is described in Appendix 1. The results were cleaned and consolidated. Cited publications and documents published by GreyNet (TextRelease) were discarded (Table 1).

Source	Owner	Nb of references	Search date
LISTA	Ebsco	465	4/10 2012
Scopus	Elsevier	1,412	3/31 2012
Web of Science	Thomson	1,206	4/1 2012
Pascal/Francis	INIST	129	3/31 2012

Table 1: Databases with search results

We then conducted the same research in different open access directories and search engines (DOAJ, OpenDOAR, ROAR, E-Lis, OAIster) but only the results from E-Lis were satisfying and relevant while the other tools were not specific enough (no limitation to the time period, to search on fields except full text and/or to published documents). Therefore, we only added the references from the E-Lis directory (Table 2). Again, these references were cleaned and consolidated, and GreyNet publications were discarded.

Source	Owner	Nb of references	Search date
E-Lis	RCLIS/CIEPI	124	4/4 2012

Table 2: Open access directory E-Lis with search results

The references from all sources were uploaded to a unique database. Again, double entries were eliminated, references were cleaned and consolidated. The final database contains 2,440 references and allows for three analyses:

- Study on the publication patterns: the study was conducted in order to know more about these references on grey literature – document types, publication years, preferred journal titles.
- Study on authors: the study was conducted in order to describe this community publishing on grey literature outside of the GreyNet, in particular their institutional affiliation, geographical origin, preferred journals and other vectors of communication.
- Comparison with GreyNet community: The GreyNet community (inside) is defined based on the authors who usually publish in the GreyNet newsletter, in *The Grey Journal*, or in the proceedings of the annual conferences on grey literature partially available in the *OpenGrey* repository. The list was downloaded in April 2012 from the TextRelease web page called "WHOIS in Grey Literature 2012"[1]. Together, this corpus (i.e. the GreyNet community) accounts for 296 members, more or less involved, active and publishing. Our comparison was conducted in order to better understand the specificity of the grey community, its boundaries, outreach and potential.

Results

The corpus of publications on grey literature and/or PhD theses or Master dissertations contains 2,440 references. As mentioned above, these references were retrieved from six different sources. No reference was available in all sources but some are listed in two or three sources; for instance, 762 references (30%) are indexed in both Web of Sciences and SCOPUS databases. All references from GreyNet sources (*The Grey Journal*, GL conferences) were discarded.

50% of the references contain "grey literature" in the title, keywords or abstract while another 49% of the references mention PhD or Master theses (dissertations etc.) in their metadata.

12% of all references can clearly be identified as belonging to scientometrics or bibliometrics, mostly citation analysis.

The results are presented in two stages. First we describe publishing features, preferred journals and the number of publications per author for the whole sample of 2,440 and 5,490 authors. This first part is followed by a more detailed analysis of geographic origins, place of work, domain and profession with a subsample of 433 authors who are strongly committed to grey literature.

Whole sample

Publishing features

82% of the references are from journal articles. Yet, about 18% are from other document types, such as theses or conference proceedings (Table 3).

Document type	Nb of references	in %
Articles	1,990	81.6%
Conference papers	201	8.2%
Theses, dissertations	4	0.2%
Books, sections or reviews	82	3.3%
Others, n/a	163	6.7%
	2,440	100%

Table 3: Document types (n=2,440)

Roughly 8,4% of the papers on grey literature and PhD theses are published and disseminated as grey literature while 82% are "mainstream" article publishing, probably mostly by commercial academic publishing houses.

The scope of our study was limited to papers published between 2000 and 2012. Figure 1 shows the distribution of the year of publication.

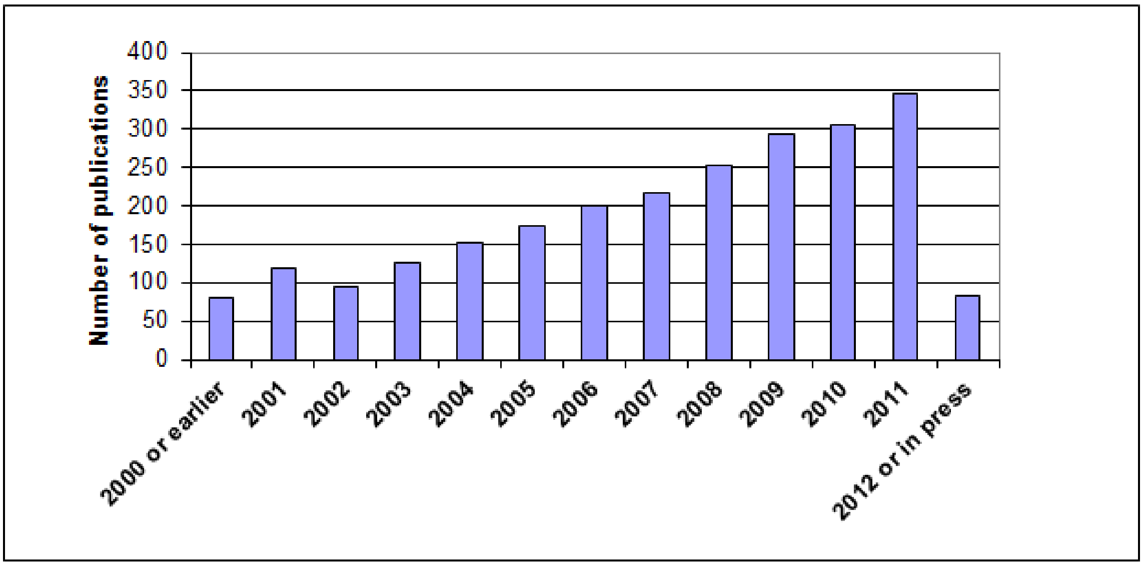


Figure 1: Year of publication (n=2,440)

The median age of publication is four years. 30% of the papers were published during the last two years (2010-2012) or are in press.

More than 36% of all papers were published in medical or public health journals while another 27% appeared in journals from library and information sciences or magazines from the publishing and book trade industry. The other papers were published in different disciplines, for instance in education, computers, biology or psychology.

Preferred journals

Academic journals are still the most important vector for scientific communication. Which titles do scientists and professionals prefer for the submission and publishing of their papers? Our sample provides two different answers.

Articles that mention grey literature in their titles are most often published in LIS journals (Table 7).

Publishing Research Quarterly
Archaeologies
Journal of the Medical Library Association
Interlending and Document Supply
Library Hi Tech News
Collection Building
Journal of Academic Librarianship
INSPEL
Science and Technology Libraries
Documentation et bibliothèques

Table 4: Ten preferred journals (grey literature in article title)

The high ranking of *Publishing Research Quarterly* can be explained by a former agreement with GreyNet that gave permission to PRQ to re-publish the best papers from the international conferences on grey literature.

In 2010 *Archaeologies* published a special issue on grey literature in archaeology that may at least partly explain the second place on the list.

Considering the whole sample and not only the articles with "grey literature" in the title, other titles appear to be as important if not more important than the cited LIS journals (Table 5).

Cochrane Database of Systematic Reviews
Publishing Research Quarterly
Journal of Advanced Nursing
Journal of Academic Librarianship
College & Research Libraries News
Health Policy
Library Hi Tech News
Technical Services Quarterly
Journal of English for Academic Purposes
Chinese Journal of Evidence-Based Medicine

Table 5: Ten preferred journals (all references)

Our data show that journals in medical science and public health (nursing) regularly publish papers that build on grey literature and/or theses and dissertations as valuable sources for literature reviews, scientometric analyses or state of the art contributions. Yet, these journals correspond to about 10% of all articles, and the whole/complete list of journals with articles on grey literature, theses and dissertations is much longer and contains nearly 1,300 different titles. This in other words, means that these publications are more or less scattered in a significant number of journals.

Number of authors
128 references (5%) are anonymous papers and/or without information (n/a) about the author(s). The remaining papers are signed by 5,490 different individual authors. Most of them signed only one paper (88%). The others published two (9%), three to five (3%) or up to fifteen papers (Figure 2).

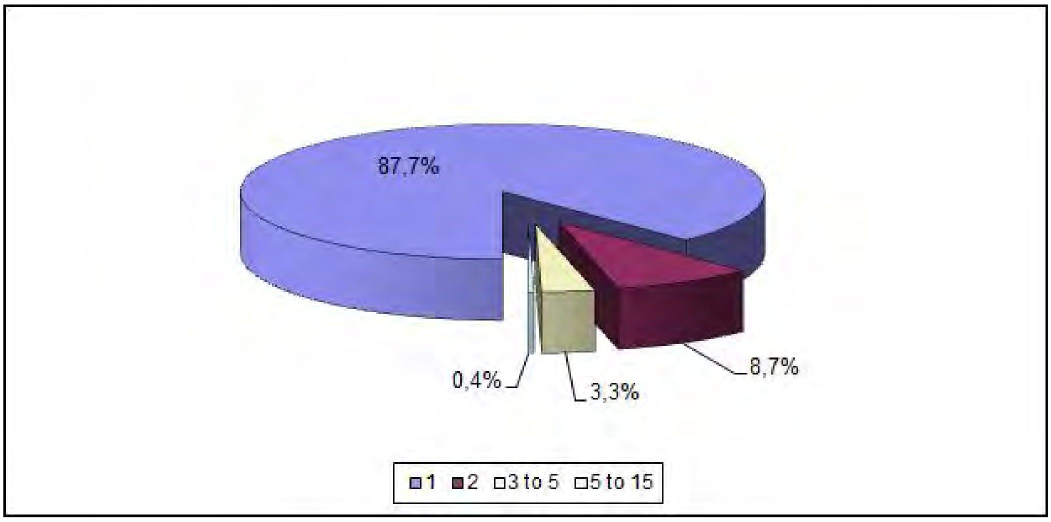


Figure 2: Part/proportion of authors with number of publications (all authors, n=5,490)

All these authors are in some way committed to grey literature or related subjects. Yet, as we were interested in those authors strongly committed to grey literature, we selected a subsample based on two criteria: authors with at least three papers (i.e. 199 authors or 3.6% of the whole sample), and those with at least one paper that mentions “grey literature” in the title. This subsample of authors strongly committed to grey literature contains 432 individuals and one committee of authors (GLISC), that is 8% of the initial sample.

Subsample of authors strongly committed to grey literature
Altogether, these 433 authors have published 550 papers on grey literature and/or PhD theses or Master dissertations.. One quarter of their papers mention “grey literature” in the title. Nearly 13% of the papers present results from scientometric studies.

Geographic origins
Half of these authors are from Europe; another third are from North America (Figure 3).

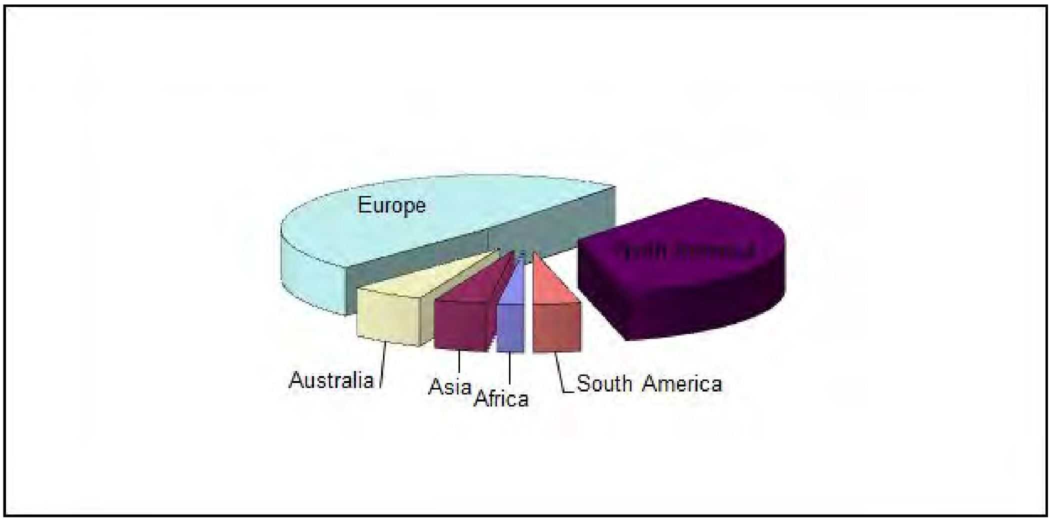


Figure 3: Geographic origins of authors (n=433)

Even if the authors are concentrated in Europe and North America, all continents are represented. Nevertheless, there is no BRICS effect, at least not in this sample.

Place of work and domain

Where do the authors work, which/what are their professional domains? Unsurprisingly, most of the authors (68%) come from Higher Education, mainly from universities, sometimes from schools. Another 17% are working in research organisations (laboratories, institutes...). The others are working in hospitals, government agencies, non-profit organisations or corporate companies. 4% are from more important and independent libraries, such as INIST, etc. (Figure 4).

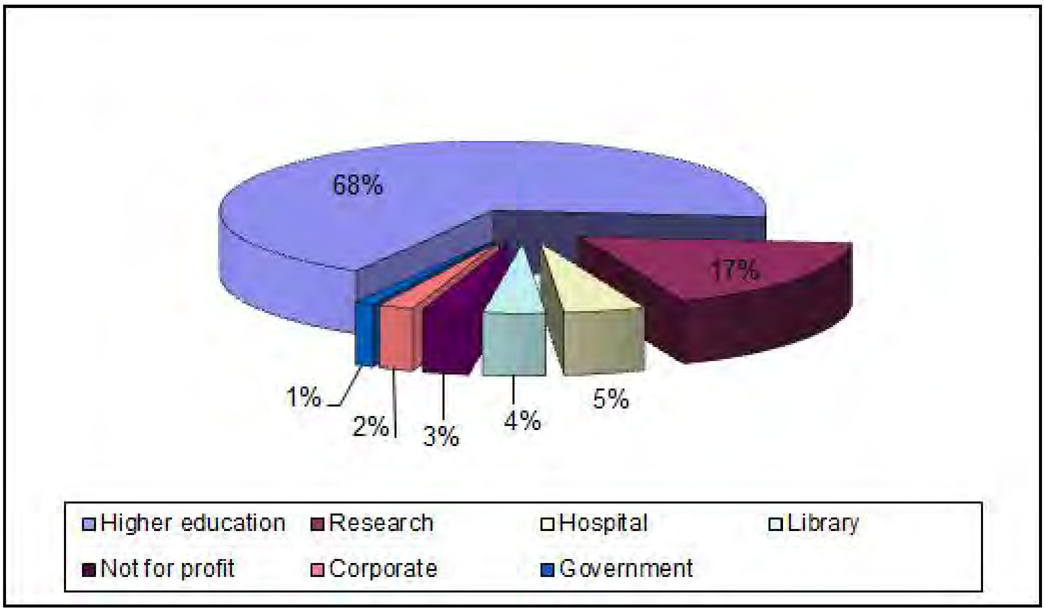


Figure 4: Place of work of authors (n=433)

In which domain are they working? We tried to identify their scientific or professional disciplines from their affiliation or other related information. 40% are working in structures of medical sciences and health, in hospitals or universities, followed by structures in library and information sciences (27%), often in universities (Figure 5).

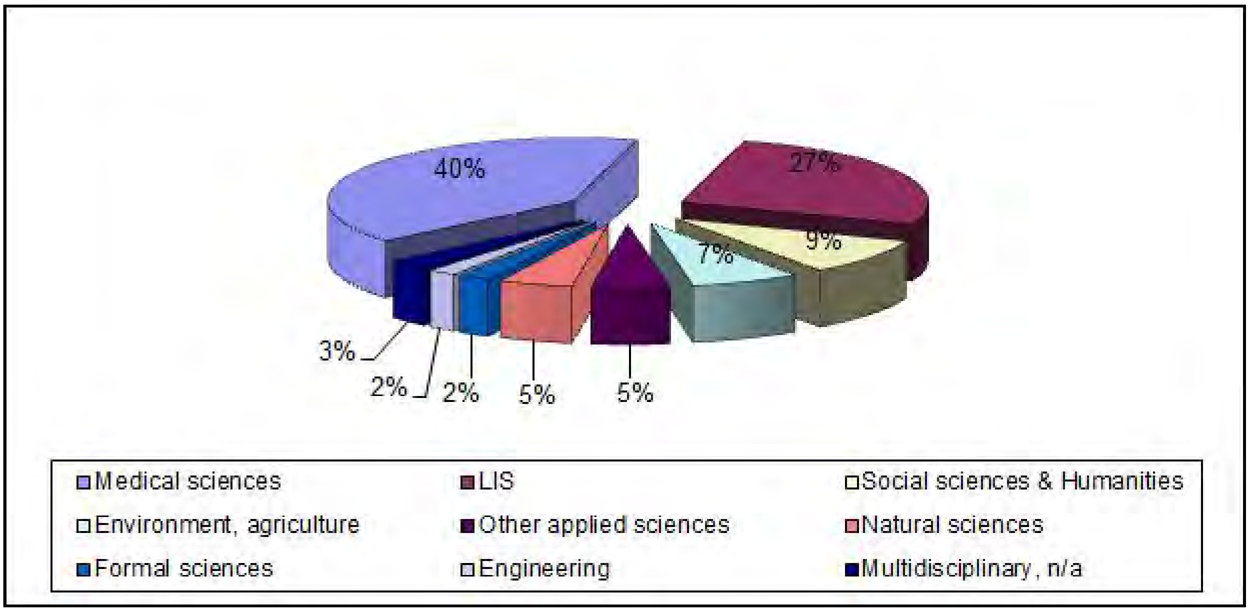


Figure 5: Scientific and professional domains of authors (n=433)

Other significant domains are social sciences and humanities (9%) and environment (7%) while only 5% of the authors are working in disciplines related to natural sciences.

Profession

What are the jobs of the authors publishing about grey literature? More than two thirds are librarians in different functions and settings. Another 25% are scholars teaching Library and Information Sciences at universities or working in research structures (see Figure 6).

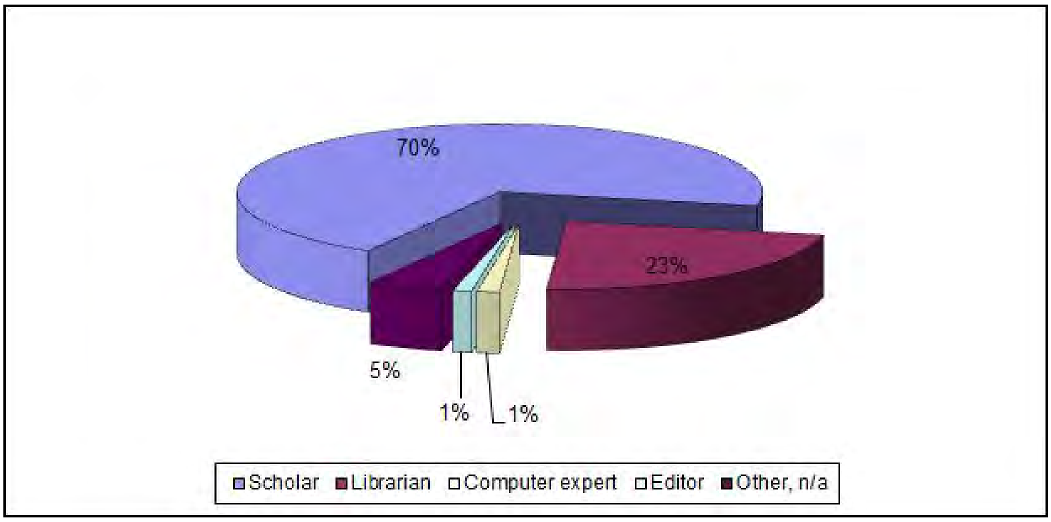


Figure 6: Profession of authors (n=433)

A very small number are computer experts, editors, consultants etc. For others, it was impossible to determine the profession. Most of the academics are teaching Library and Information Sciences while the scientists are often from medical research.

Overlap with GreyNet

Are these authors members of the GreyNet community? Are they listed in the WHOIS, members of the LinkedIn group, authors of papers communicated at a GL conference? We tried to find out by comparing the different samples.

First, we compared the GreyNet community (296 individuals) with our sample of 5,490 authors. 81 people belong to both. This means that 27% of the GreyNet community are represented as authors on grey literature in our initial large sample, but only 1.5% of the authors occasionally writing on grey literature belong to the GL network (sample 1 in figure 7).

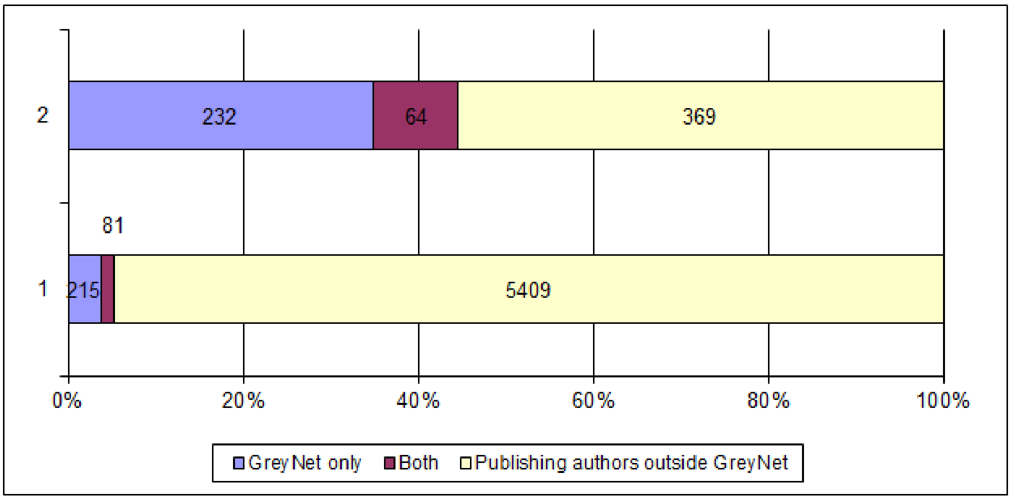


Figure 7: Publishing authors and GreyNet community (sample 1 n=5,490 authors ; sample2 n=433 authors)

Secondly, we limited our comparison to the smaller sample of 433 authors more strongly involved in publishing on grey literature. 64 people are in both communities (sample 2 in Figure 7). Again, this figure means that about 22% of the GreyNet community publish outside of the network and are engaged in grey literature while 15% of these authors are members of the GreyNet community.

The comparison between the two samples 1 and 2 shows that the more an author publishes on grey literature, especially when mentioning the term "grey literature" in the title, the more likely it is for him/her to be a member of the GreyNet community.

But these figures also mean that 70-80% of the members of GreyNet are not identified in databases and repositories as occasional or regular publishers on grey literature or related issues.

Outside of GreyNet - same or different?

85% of the authors strongly committed to grey literature (n=369 of 433) do not belong to the GreyNet community. Are they different?

The comparison of the two groups in our sample of 433 authors shows that the GreyNet authors are not represented in Africa, South America and Australia. Authors from outside of GreyNet represent 36 countries (Figure 7) while GreyNet authors are from only 12 countries.

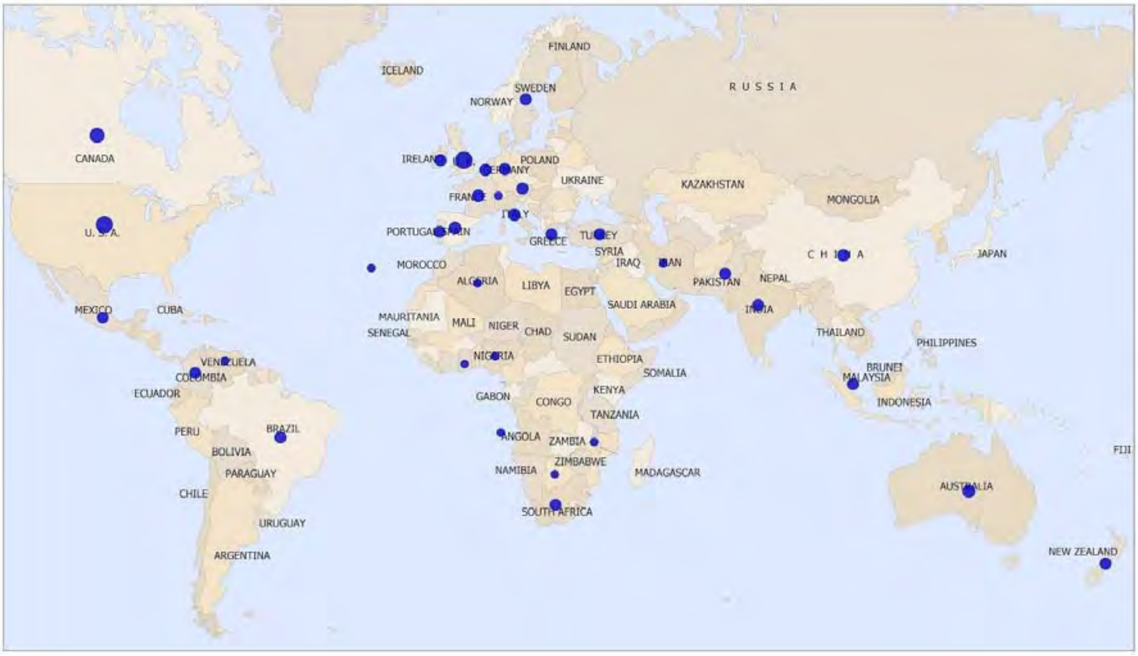


Figure 8: Geographical origin of authors outside of GreyNet (n=369)

Columbia, Mexico, Portugal, Greece, Austria, New Zealand, Pakistan and Turkey are countries so far beyond the reach of the GreyNet. At least this is so in this sample of authors strongly committed to grey literature and publishing more than others.

With regards to the other aspects, we can identify some other differences between GreyNet members and the other authors. For instance:

- GreyNet is less academic than the author sample.
- Research structures, non-profit organisations and libraries are better represented in GreyNet than outside.
- In GreyNet, LIS and natural sciences are better represented, while medical sciences, ecology, applied sciences and social sciences and humanities are underrepresented.
- The relationship between scholars and librarians is completely inverted. Publishing GreyNet members in our study are mostly information professionals while the majority of the publishing community outside of GreyNet is working in Higher Education or research organisations.

These results show two communities which, despite shared interests, are rather different, with a more consistent profile for the GreyNet community.

Discussion

Methodological shortfalls

Our choice to evaluate terminology (grey literature) and choice of topic (theses) as indicators for belongingness to a community is not exhaustive. Also, this kind of scientometric analysis depends partly on the quality and consistency of abstracting and indexing services of databases and repositories and also on the quality and relevance of their search tools and results. Missing search facilities for instance reduced the number of open access directories and most probably the number of identified publications and authors. Also, we had to apply different search strategies (see Appendix) that probably increased inconsistency in the corpus and eliminated relevant items.

Another problem shared with other scientometric studies is the author identification because of misspelling, changes or variants of the first and/or last name and so on. Because of missing person or author identifiers, the consolidation and validation were done manually, a procedure that decreases the reliability of the results.

The same remark applies to the analysis of geographical origins, affiliations and domains of activity. We tried to control this source of error through double-checking and, in some cases, by eliminating the reference or author from the sample. This may reduce the error rate but at the same time decrease the relevance of the overall results.

Grey community or grey communities?

A subject community, as defined above, shares terminology, tools, events, experiences and topics. The "grey community" should therefore share the basic terminology of grey literature, should attend the same events (conferences, workshops, other meetings), make use of the same tools of communication (journals, listservs...) and research (methodology), and discuss the same topics. But does it?

Drawing on our empirical evidence, we can distinguish different groups (Table 6).

Name	Number	Comments
GreyNet	**296**	Coordination and management by GreyNet Amsterdam and TextRelease.
GreyNet publishing outside	**81**	The "visible" part of GreyNet outside of the GreyNet events and communication vectors. More information professionals.
Occasionally publishing on GL	**5057**	One or two papers in ten years, no usage of grey literature terminology. All disciplines.
In 433 authors sample, Regularly publishing on GL with GL terminology	**149**	Three or more papers. SS&H, LIS. More scholars.
In 433 authors sample, Regularly publishing on GL without GL terminology	**31**	Three or more papers. Medical sciences. All disciplines. More scholars.

Table 6: Concentric and overlapping groups

These groups are not exhaustive, and we could add others such as:

- GreyNet members attending events,
- GreyNet authors publishing occasionally,
- Groups regularly publishing on grey literature with scientometric approach,
- Groups occasionally publishing on theses,
- Subgroups according to disciplines, organisation or profession, etc.

Some of these groups are defined mainly by practice (publishing features, attendance to events etc.), others by usage of terminology and methodology (citation analysis, grey literature definition etc.) or are mixed.

Probably, members of some groups would not have any problem to identify themselves as part of a "grey literature community" while others would probably prefer to define their affiliation by means of a scientific discipline or a profession.

In and outside of the community

One does not need to use Eskimo words to describe different kinds of snow. As Molière's *Bourgeois Gentilhomme* discovered with delight, one can speak prose without knowing it. And one can publish about grey literature without using the term, or even without awareness of it. We included in our sample papers on theses as a significant and central part of grey literature. We obtained two subsamples, one with references that explicitly mention "grey literature" in the title or in other header information; the others are papers on theses that may or may not mention "grey literature" in the body of the text.

In the whole/complete sample of references, the GreyNet authors represent 15%. But when we consider only the references with "grey literature" in the header information, this part/proportion increases up to more than 40%. On the other hand, when we consider the subsample on theses, the percentage of GreyNet authors decreases to 8%. This means that GreyNet authors when publishing outside of the GreyNet use the grey literature terminology more often but they do not do it in a consistent way. And this means also, that this terminology is known outside of GreyNet but not generally accepted or explicitly used when studying grey literature such as theses.

Comparison between the two samples of authors (the 433 authors and the others) against the two communities (GreyNet and outside) reveals another significant difference. The statistical distribution of the usage and non-usage of the grey literature terminology across these groups is significantly different from the expected values (Table 7).

	GreyNet	GreyNet	Outside	Outside	
	Terminology	No terminology	Terminology	No terminology	Total
433 authors	**64**	0	**336**	*33*	433
Others	*14*	3	3549	**1491**	5057
Total	78	3	3885	1524	5490

Table 7: Usage of grey literature terminology in header across the author samples

The figures in bold are higher than expected, and the figures in italics are lower than expected. In other words: Authors from GreyNet but also from outside who are highly committed to grey literature significantly use the term "grey literature" more often the others when describing their work. Compared to less committed authors, we can speak of a terminology-based community.

One reason for the differences may be that while GreyNet authors (mainly librarians from all disciplines and LIS academics) often choose grey literature as their object of research (Library and Information Sciences), authors outside of the GreyNet (mainly academics from different disciplines) include specific types of grey items in citation analysis or reviews (state of the art), in particular in medical or life sciences, without having a global concept of grey literature.

Our empirical data tend to confirm the reality of two grey communities, with relatively closed frontiers between inside and outside of Greynet - only 3.5% of all papers are co-authored with somebody "from the other side". Obviously in the whole/complete sample we cannot speak of some kind of freedom for the movement of ideas, projects and publications across this border. In so far as both are working with the same material, is this a problem? Perhaps not, as long as mutual understanding and exchange remain possible. Yet, the concept of grey literature draws a framework for research and practice and allows for a better understanding of a specific kind of scientific and technical communication. Therefore, the use and promotion of "grey literature" terminology is all but trivial, not only inside the GreyNet community but also and above all outside of the community.

Conclusion

As we stated at the beginning, a community builds on shared terminology, definitions (concepts) and practice, such as common methodology and events. Our empirical data indicate the existence of different shades of grey communities, with regards to GreyNet membership, publishing features and usage of terminology. Nine out of ten authors appear to be “bouncers” during the observed period 2000 to 2012, occasionally speaking of grey literature or using grey material. Only one out of ten authors can be considered as a kind of “returnee”, with a kind of loyalty and respectful behaviour regarding grey literature.

The analysis of 2,440 papers published by 5,490 authors confirms that the concept of grey literature remains more or less a professional affair applied by librarians and LIS academics. The papers can be mapped to/in three concentric circles: at the core, some studies on grey literature as an object of research, followed by a group of papers with a conscious and direct use of the concept. The third and largest circle contains papers that make usage of grey items, with or without awareness of the concept of grey literature.

Thus, the data show a potential for the development of the GreyNet community beyond the actual and often tight frontiers. We conclude with two possible strategies for the development of GreyNet, one based on proximity, the other on exploration.

Proximity

The first strategy is centred on authors outside of GreyNet with a profile adjacent to that of GreyNet members. This calls for contact as a priority:

- Authors who mention grey literature in the header information of their paper.
- Librarians and other information professionals, scholars from library and information sciences.
- Authors of papers published since 2008.

Contact could be established in order to invite publications for *The Grey Journal* or a monograph, to suggest communications for the GL conferences, and to invite to join the GreyNet listserv etc.

Exploration

The second strategy focuses on specific groups of authors that are not necessarily very near to GreyNet but nevertheless in (some kind of) grey literature. For instance:

- The Networked Digital Library of Theses and Dissertations (NDLTD), an international organization dedicated to promoting the adoption, creation, use, dissemination, and preservation of electronic theses and dissertations.
- Other library sections dedicated to special collections, acquisition policy and institutional repositories.
- Identified authors from Latin America, Sub-Saharan Africa, Australia or the BRICS countries.
- Identified authors of papers on grey literature (citation analysis, case studies etc.) in medical sciences.

Here, contact could be made in order to suggest joint publications or events, keynote addresses to GL conferences, or invitation to join the GreyNet listserv etc.

Beyond the community

These strategies could be implemented with the help of the GreyNet group on LinkedIn. This group already radiates beyond the traditional frontiers of GreyNet. Presently (October 2012) it has 271 members, and only around twenty of them usually publish in the GreyNet newsletter, in *The Grey Journal*, or in the proceedings of the annual conferences. Here, the GreyNet community clearly has the potential to expand and to enhance its impact.

However, radiating beyond the community can also mean publishing on grey literature in other media and products, editing special issues related to grey literature not only in LIS journals but also in selected journals from other disciplines that are usual “consumers” of grey literature (medical sciences, social sciences and humanities...), and fostering joint research and publishing projects between professionals and academics.

Integrating authors near to GreyNet, exploring other communities and reaching beyond the GreyNet community could be three different but complementary ways to promote and foster sustainable development of this network.

Bibliography

M. Callon (1989). *La Science et ses réseaux :genèse et circulation des faits scientifiques*. Paris, La Découverte.

Farace, D. (2011) What the Future holds in Store for GreyNet International, Business Report Amsterdam, TextRelease. http://www.greynet.org/images/Business_Report_2011.pdf

Farace, D., Frantzen, J., Stock, C. et al (2009) OpenSIGLE, Home to GreyNet's Research Community and its Grey Literature Collections: Initial Results and a Project Proposal, in *GL10 Tenth International Conference on Grey Literature: Designing the Grey Grid for Information Society*. Amsterdam, 8-9 December 2008. http://archivesic.ccsd.cnrs.fr/sic_00379643

T. S. Kuhn (1962). *The Structure of Scientific Revolutions*.University of Chicago Press, Chicago.

B. Latour& S. Woolgar (1979). *Laboratory life : the social construction of scientific facts*. Sage Publications, Beverly Hills.

Marzi C. (2012). `Knowledge Communities in Grey'.*The Grey Journal***8**(1):27-33.

Paganelli C. (2012). `Recherche en SI. Analyse des discours sur la notion d'« usage » dans deux revues en sciences de l'information : Doc-SI et BBF'. *Documentaliste*49(2):64-71.

Schöpfel, J., Stock, C., Farace, D. (2005) Citation Analysis and Grey Literature: Stakeholders in the grey circuit, in *GL6 Sixth International Conference on Grey Literature "Work on Grey in Progress"* New York Academy of Medicine Conference Center New York City, 6-7 December 2004. http://archivesic.ccsd.cnrs.fr/sic_00001534

C. Schrecker (2006). *La communauté : histoire critique d'un concept dans la sociologie anglo-saxonne*. L'Harmattan, Paris.

Appendix

For each data source, we preferred if available the advanced search interface, eliminated double entries and references from the GreyNet (The Grey Journal, GL conference series etc.), and limited the results to the time period mentioned above (2000-2012).

OAIster was accessed through the OCLC WorldCat but the results were not specific enough. We could not search in the LISA database because we had no access. We did not consider the results from ROAR and OpenDOAR because of missing search functionalities. The search results were much too large and without interest for our study.

The table shows the search strategies. After some exploratory tests, we decided to limit the search in the following way/manner:

- Grey (or gray) literature in the title of the publication.
- Grey (or gray) literature in the abstract.
- Grey (or gray) literature in the keywords.

In order to produce some information about authors dealing with grey literature but not using the concept of grey (or gray) literature, we also searched for papers on theses and dissertations, excluding biographical papers.

Corpus	Extraction
LISTA	all txt contains "grey literature" or "gray literature" or dissertation or "master thesis" or "doctoral thesis"
Scopus Grey	Title abstr keyword contains "grey literature" or "gray literature"
Scopus Thesis	KW contains thesis or Title contains doctoral dissertation
WoS Grey	Topic contains "grey literature" or "gray literature"
WoS Thesis	Title contains dissertation or "master thesis" or "doctoral thesis"
PASCAL and FRANCIS	All text contain "grey literature" or "gray literature"
E-LIS	All metadata & Full Text contain "grey literature" or "gray literature", or subject = "H. Information sources, supports, channels > HB. Gray literature" or "master thesis" or "doctoral thesis"

All references were cleaned and consolidated, and GreyNet publications were discarded.

[1] http://www.textrelease.com/whois2012.html

An Environment Supporting the Production of Live Research Objects

Massimiliano Assante, Leonardo Candela, and Pasquale Pagano
Istituto di Scienza e Tecnologie dell'Informazione – CNR, Italy

*Ab**stract***
Modern science communication requires innovative environment and means for providing stakeholders with scientific outcomes. Research objects are emerging as replacements of traditional "documents" in scientific communication. These objects are multi-media and multi-part objects that aggregate all the "pieces" that contribute to a research result. Supporting these objects has gone beyond the capacity of traditional technological approaches based on locally specialized data management facilities. In this article we present an environment for producing "live research objects" by exploiting the capabilities offered by a Data Infrastructure. Such environment includes: (i) a workspace where users can organize and share with their co-workers very different items in a file-system-like environment; (ii) an editing framework where users can define the structure of a live research object and compile objects that comply with one of the defined templates; and (iii) a workflow engine where users can define the workflow governing the production of a live research object by specifying the phases and the relative responsible actors(s).

1. Introduction

Scientific research is rapidly evolving in all fields, it is multidisciplinary, networked and driven by new patterns, e.g. data-intensive sciences [1]. In this complex scenario scientific communication must go well beyond traditional scholarly communication. Specifically, it requires accessing all the elements exploited and developed during the scientific workflow to achieve a result, e.g. datasets, analysis tools, and methods [2]. This wide corpus of primarily grey elements are at the moment mostly unavailable and, even when they are available, they are not linked to the scientific result. This makes difficult to completely understand the result and validate it.

To overcome this limitations many initiatives have been proposed in the literature as replacement of traditional "documents" in scientific communication, such as *Executable Papers* [16,17], *Enhanced Publications* [6,15], *Living Reports* [8,7]. Some of them were sponsored by major scientific publishers [12].

Lately, *Research Objects* [10] are emerging as an abstraction for communicating, sharing and reusing research results. These are multi-media and multi-part objects that aggregate all the "pieces" that contribute to a research result. Such elements, which may range from binary files to compound objects including maps, time series, and tabular data, are generally structured according to well-established templates and produced according to user-defined workflows. We extended the Research Object definition and included facilities to make some of these "pieces" contributing to a research result directly embedded in the document. The idea behind a Live Research Object is depicted in figure 1.

However, supporting Research Objects has gone beyond the capacity of traditional technological approaches based on locally specialized data management facilities. This paper discusses how an infrastructure-oriented approach aimed at promoting *sharing* and *re-use* of resources (including data, services and computational and storage resources) is an effective approach for producing Live Research Objects and poses the accent on the user-oriented facilities supporting the collaborative production of such research products.

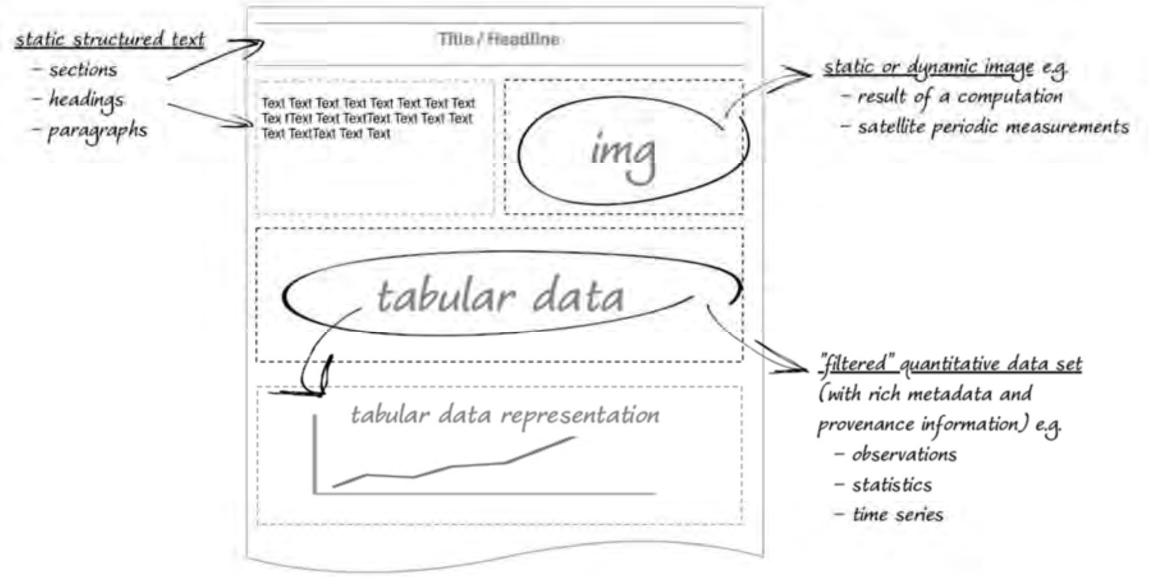

static structured text
– sections
– headings
– paragraphs
Title / Headline
Text Text Text Text Text Text Text Text Tex tText Text TextText Text Text Text Text TextText Text Text
img
static or dynamic image e.g.
– result of a computation
– satellite periodic measurements
tabular data
tabular data representation
"filtered" quantitative data set (with rich metadata and provenance information) e.g.
– observations
– statistics
– time series

Fig. 1. The Idea behind a Live Research Object

The production environment includes: (*i*) a *workspace* where users can organize and share very different items (from binary files to compound objects) in a file-system-like environment; (*ii*) an *editing framework* where users can define the structure of a live research object (a template indicating sections, layout, active elements) and compile objects that comply with one of the defined templates by entering content or taking it from the workspace via *drag and drop*; and (*iii*) a *workflow engine* where users can define the workflow governing the production of a live research object by specifying the phases and the relative responsible actors(s).

Accordingly, the article is structured as follows. Section 2 describes the requirements and the main design decisions driving the development of the proposed approach. Section 3 presents the environment developed to offer Live Research Objects production. Finally, Section 0 concludes the article and summarizes its results.

2. Motivations and Design Philosophy

We feel that the classic notion of documents viewed as static entities has to be abandoned. This notion should be moving towards one where documents are constantly evolving, by enriching them with components that yield seamless data access to contents, user cooperation, collaboration, interaction, and personalization.

However, this notion cannot evolve if not supported by progresses in the way documents are created, produced and made available too. In fact, thanks to the new technologies, scientists, researchers or experts in a field, can produce novel research outputs based on new resources (in terms of hardware, services and content) that were not available in the past.

Despite that, this production can still require a lot of work due to *(i)* the complexity of interfacing with different sources and tools, and *(ii)* the people involved in the task that may need coordination, concurrent and rule-based access to the same research object or part of it. In particular, for the research object instance case the resulting work may also not meet the requirements of its consumer, *i.e.*, the reader, since it could present a picture of the subject at the time of its production and not at the time in which the information produced is accessed and used.

The latter point illustrates the potential for new scientific advances. For example, the Food and Agriculture Organization of the United Nations (FAO) exploits its rich information sources, ranging from raw data sets to graphs and map archives, to periodically prepare reports on the status of the agriculture and fishery, per country. This activity often:

- make necessary to have access to a wealth of textual, graphical and tabular information located in several (external) data sources;
- requires several people working together in the collection, collation, drafting, and reviewing;
- demands for "freshness" of the information reported.

To overcome these problems we decided to experiment the construction of a workflow[1]-driven Live Research Objects production environment. This environment *(i)* exploits the facilities provided by an underlying *Hybrid Data Infrastructure* [18] for accessing and exploiting the entire spectrum of resources needed to achieve a research result including data, services and computing resources "as-a-Service", *(ii)* uses a component-oriented flexible document model for the representation of the research objects, *(iii)* benefits from a workflow driven mechanism to ensure concurrent and rule-based access to its users, and *(iv)* is accessible through a thin client (namely a web browser).

The solution for the described approach can be logically divided in two main modules: one module realizes an environment for producing (namely defining and editing) innovative research objects, one module realizes an environment driving users (namely assigning actions and notifying the right actor(s) when needed) while producing such research objects.

The first module is delegated to solve the issues related to the complexity of interfacing with different sources and tools, providing users with a familiar and easy-to-use environment to produce research objects. The following design decisions have been taken for the implementation of the proposed approach:

- The management of different data sources should be made transparent to the end-users (by the hybrid data infrastructure) and promote the seamless access and sharing of a rich array of information objects;
- The research objects production environment should enable a WYSIWYG[2] editor (similar to Google docs) to give users an immediate feeling on how the actual research object will look like;
- The production environment should enable the production of research objects sharing a common structure, when needed. Thus it supports a production based on two phases: *(i) template definition*, to define a basic research object structure (including sections and layout) to be adopted by research objects expected to be compliant with such a structure, and *(ii) reporting*, to produce the actual research object in compliancy with one of the defined templates;
- Supported Research Objects should enable the definition of living elements, i.e. Research Object elements that are potentially willing to evolve whenever an user accessed the object;
- The produced research object should be available in different exporting formats, including OpenXML, PDF, and HTML.
- The second module of the solution instead is in charge of coupling the research object with a workflow driven mechanism and ensuring through a series of steps, rule-based access to the same instance. For the implementation of this module the following design decisions have been taken:
- support for visual representations of the workflow, through workflow diagrams outlining the whole process;
- support for workflow states;
- support for manual and automatic (policy-based) routing;
- support for routing to individuals and to groups (set of users having the same role).

Accordingly, the next section presents the environment developed to offer workflow driven Live Research Objects production facilities.

3. The gCube Live Research Objects Environment

gCube[3] is software system enabling the building and operation of an Hybrid Data Infrastructure. In a nutshell, it offers a number of mediators for interfacing with (data) providers and a number of services for data management over a rich array of data types. The gCube Live Research Objects Environment is one of these data management services and it implements the environment envisaged in the previous sections to realize Live Research Objects. It is made of three main components: (i) the *Virtual Workspace*, that is a virtual file system promoting the sharing of a rich array of information objects, (ii) the *Research Object Editing Environment*, that is a graphical editor and renderer of a Live Research Object, and (iii) the *Research Object Workflow Manager*, that allows to create and associate workflows to a given Research Object and to manage and control its status.

3.1 Virtual Workspace

The Virtual Workspace (WS) is the core element of the cooperation environment. It is conceived to resemble a classical folder-based file system any user is familiar with. As a consequence, the operations it supports on the items are the expected ones, namely items creation, deletion and their organization in folders and subfolders. Thus every single user is free to organize its items.

However, the real added value of this file-system-like environment is represented by the types of items it can manage in a seamless way. They range from binary files to information objects representing tabular data, species distribution maps, and time series. Every item in the workspace is equipped with a rich metadata including bibliographic information like title and creator as well as lineage data.

Another distinguishing feature is represented by the sharing that is fundamental since users need to work collaboratively on the same research object and rely on common research materials. Sharing can be performed per single item as well as per folder and it is invite-based. Any item or folder and, in turn, its content can be shared with other users. The users involved in this sharing are alerted by a notification mechanism in charge of delivering the related invite.

The end-user interface

A web application has been developed to offer users the possibility of viewing and managing their research object instances and as well as any other information object. This web application offers a remote file system view similar to any Operating System over the content available within the data infrastructure along with files management facilities such as copy, delete and upload.

3.2 Research Object Editing Environment

Document templates motivate their uptake by easing the strain of repetitive manual work. It is quite common that a document's structure and style is maintained through time and used for multiple document instances. This is the reason why the Research Object Editing Environment, that is a web application developed by using the Google Webtool Kit framework [13], is logically divided in two phases: the Template Definition and the Research Object Editor. The first one is needed to define research object templates that, during this stage, are dynamically and statically completed. The second instead is capable of loading these templates to produce actual research objects by filling out their dynamic parts. In the following we explain our design approach from a functional description point of view together with actual examples for both phases.

Template definition

During this phase, document templates can be created through user interfaces by exploiting an extension of the Document Editor web application, called Template Creator.

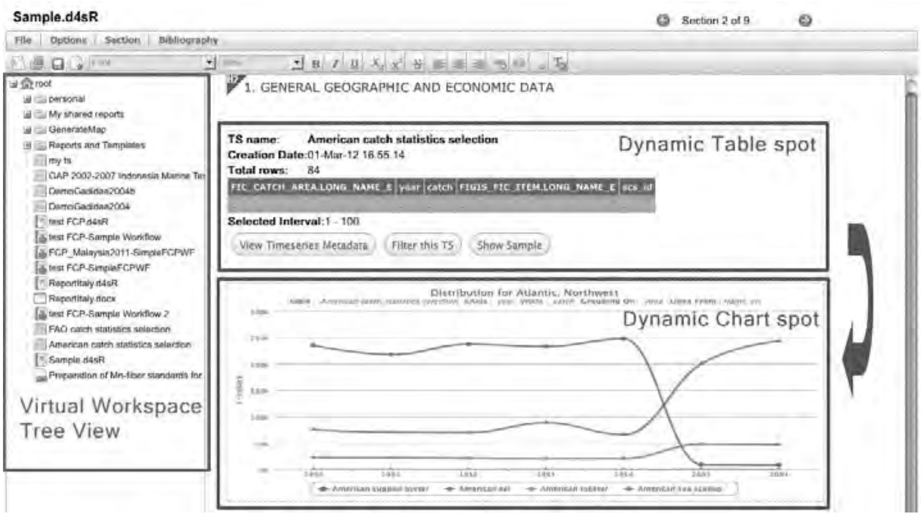
The Template Creator adopts a component oriented approach for templates composition and supports several component types such as structured and styled text areas (title, headings, body), images, tables, table of contents (ToC), bibliography, page breaks. Template components are divided in two classes: *static* and *dynamic*. A static component of a template is, as the name suggests, a part that is not meant to change across diverse research objects sharing the same template, such as filled-in texts or images. A dynamic component instead is meant to be completed in the second phase and, in turn, can belong to the following two categories:

- *dynamic text*: empty text areas, including headers and empty tables, fall into this category. An example would be having the same template for a set of research objects aiming at representing a study on a species. One would change species' specific data while the structure and the presentation would be uniform for all of them.
- *dynamic spot*: empty rectangular spots designed to host an image, a table, a diagram or a chart, to be instantiated using a given data source or data set. This is a key category as during the second phase these spots become configurable, providing users with the possibility of entering configuration parameters. Examples vary depending on the "hosting type" and will be explicated in the following.

The template components belonging to a template can be grouped into sections. It is a human-friendly interface assuming the form of an HTML page. This application provides users with toolbars and toolboxes from where they have the possibility of adding or removing template components, adding or removing sections, formatting text and saving their work into the Virtual Workspace.

Research Object Editor: The Document templates, created through the Template Creator extension, can be loaded by accessing the Virtual Workspace from within the Reporting Editor through user interfaces.

The Research Object Editor in this phase offers the possibility to complete or instantiate the dynamic components that might be present in a template. Depending on their type, this action can be performed in two ways: *(i)* by typing in, for components belonging to the *dynamic text category* (formatting text as one would in a word processor by using a formatting bar) or *(ii)* by providing a configuration for components belonging to the dynamic spot category. This configuration can vary depending on the *dynamic spot* type. For instance, users would specify which column and which row intervals to show for *tables* or, what data to be set on x and y axis for *charts* and so on.



TS name: American catch statistics selection
Creation Date:01-Mar-12 16.55.14
Total rows: 84

FIC_CATCH_AREA.LONG_NAME_E	year	catch	FIGIS_FIC_ITEM.LONG_NAME_E	scs_id

Selected Interval:1 - 100

Fig. 1. A partial view of a template being completed in the Research Object Editor

Figure 1 illustrates the Research Object Editor user interface, a tree view of the Virtual Workspace is presented next to the *working area*. This tree view is needed to connect dynamic spots to the actual data they need to display. This connection is made explicit through *Drag and Drop* operations (by dragging the desired item over the desired spot).

Figure 1 also shows an example of template section being completed next to the tree view. The section is composed by three components, a static heading component on the top and two dynamic spots below, the first of type Table and the second of type Chart. In the example, both spots have been connected to the same *quantitative dataset*[4], that is represented as an item in the user's Virtual Workspace, and describes a catch statistics *time series*[5] (reporting statistics on catches of marine species on a give geographic area, per year). The example shows the type Table being configured while the type Chart, that has been already, is actually displaying the related chart.

It is important to stress the fact that the *dynamic spot* component during this phase becomes *living, interactive* and capable of redisplaying itself depending on the parameters specified by the user.

Research Object instances can be successively exported into different formats by using functions provided by this editor, such as OpenXML (docx), PDF, HTML and saved into the user's personal Virtual Workspace.

3.3 Research Object Workflow Manager

The idea behind the Research Object Workflow Manager (ROW) is to work collaboratively to the creation of Research Objects. During this phase the Research Object is being created and is still incomplete. It is initiated by an actor, generally identified by a role, and then passed to other people involved in the realization. In the following, to distinguish this phase we will use the term Report instead of Research Object.

For example, one author is asked to start drafting the Report, successively several iterations can be possible between authors and editors, before the Report is sent for review and authorization. To support such scenarios, ROW is equipped with a *workflow roles editor* that, as the name suggests, allows to distinct the users involved in the creation of the Report into categories *e.g., author, editor, publisher.*

The ROW is also equipped with a *workflow templates editor*, a graphical editor and renderer of a workflow diagram. A *workflow template* (WT) is naturally represented as a digraph (directed graph) where vertices, *i.e.*, workflow steps, are connected by directed edges, or arcs. WTs support the possibility to apply *labels*, *i.e.*, *roles*, on edges to indicate the required role to perform the transition from one step to another and are always defined by an entry and an ending point, i.e., Start/End steps are mandatory in any WT. An example of a workflow template, loaded in this editor is depicted in Figure 2.

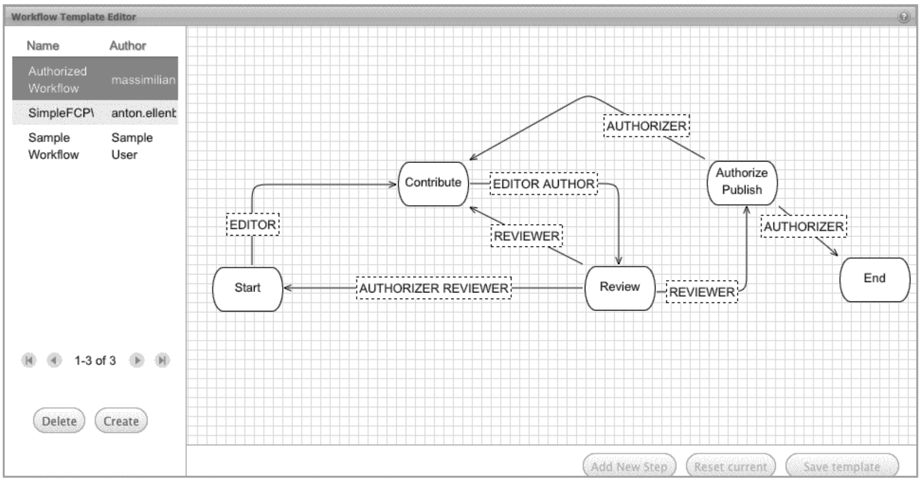


Fig. 2. A workflow template diagram loaded in the Workflow Templates Editor.

The presence of workflow templates is justified by the fact that reuse is highly desirable. Often one would like to reuse the same workflow diagram for multiple Reports, alternating the users involved but the process.

Roles and Templates are indeed complementary components for the ROW: they are required for the creation phase of a workflow document, where a Report's draft is linked to a Workflow Template, and for the functioning and monitoring of the reporting activity, *i.e.*, the step-by-step procedure needed to complete the job.

In particular, the creation phase of a new workflow document is performed by an entitled user and requires the three stages illustrated in Figure 3. In the first stage the user associates a Report draft to a Workflow Template: selects the Report to work with, created with the Research Object Editor by accessing the Virtual Workspace and couples it with an already available template (created previously through the workflow template editor). The second stage requires the user to specify, for each step, the associated roles and their relative permissions. It is possible to associate any Role to a given step, however Roles (labels) connected to a step's outgoing edges are mandatory and already present for permissions to be added. It is worth noting that the permissions attached to a role can vary from one step to another, *e.g.*, during the step "review" an author has read-only permission while during the step "contribute" an author has both read and update permissions. During the third stage instead the steps are not interested while roles, defined in the workflow report during the second stage are. In fact this is the part where these roles are linked to the actual users partaking in the workflow report production.

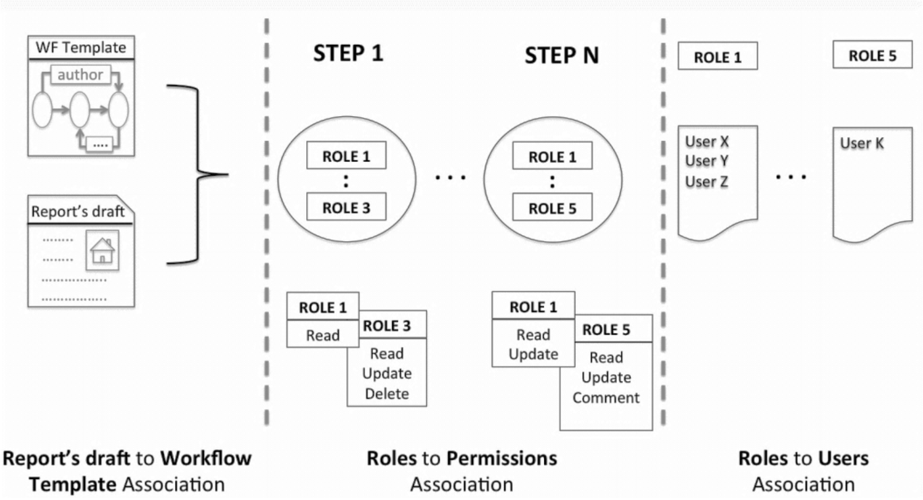


Fig. 3. The three stages needed for the creation of new documents workflows.

Regarding ROW *functioning*, manual routing and automatic routing are supported as by requirement of Section 2. The former is achieved by providing the workflow reports owner (WFO), *i.e.*, the user who created workflow reports through the three stages procedure previously described, the possibility to decide when and where performing step transitions, for the latter instead, the concept of forward action has been introduced. A *forward action* is a sort of green flag users make us of to indicate their work is completed for the current step (in which they are required to do a job). For instance, suppose there are three authors assigned to a step "contribute", each author completes his job and then performs a forward action towards the next step, *e.g.*, "review", ROW's logic recognizes it and consequently executes the transition (automatically) towards the expected step, *i.e.*, "review" for the sake of this example. It is worth noticing that the policy applied to automatic transitions is the one of executing it if, and only if, the totality of the actors involved in a job have forwarded. However, this can be changed to majority (instead of totality), by setting a parameter during the workflow report's creation phase.

Regarding ROW *monitoring*, the WFO can continuously monitor the users activities involved in a workflow. Clearly, a WFO can see and monitor only the workflow reports he owns. For each workflow report it is possible to *(i)* check the current status of the workflow, that is characterized by the current step and the forward actions performed until that time, *(ii)* check the workflow history, a chronological reversed list of user actions on the workflow.

The concurrent access, that is required since multiple users could work on the same report instance at the same time, is guaranteed by a locking mechanism provided by the ROW. When a user is editing a report instance a lock is acquired over it and no other user can access the report until the editing one finishes his work and commits the changes. The information a user has over any workflow report he needs to work with are the following: *name, current status, his role, date of creation, last action performed, grants (read, update etc.), whether the report is locked or not.*

4. Conclusion

Scientific research is nowadays multidisciplinary, networked and driven by new patterns, *e.g.*, data intensive sciences. This calls for innovative approaches and tools that are capable to cope with disciplines and tasks that can not be tackled by researches operating in isolation. Moreover, research products expected to be produced are richer than traditional research products, namely scholarly publications.

In this paper, an innovative environment for the production of live research objects was presented. In particular, a number of facilities supporting the whole lifecycle leading to the production of Live Research Objects was described. These facilities include: *(i)* a shared workspace resembling a classical file system and enabling users to store and organize any research artifact leading to a research product; *(ii)* an editor conceived to support the definition of research object structures, to promote the realization of research objects compliant with a given structure by implementing a number of user friendly features, *e.g.*, drag and drop of object constituents from the user workspace; and *(iii)* a workflow engine supporting the definition and operation of workflows driving the realization of a live research object in a collaborative and distributed approach.

Acknowledgments. The work reported has been partially supported by the D4Science-II project (FP7 of the European Commission, INFRA-2008-1.2.2, Contract No. 239019) and the iMarine project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644).

References

1. T. Hey, S. Tansley, and K. Tolle. The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research, 2009.
2. Boulton, G.; Campbell, P.; Collins, B.; Elias, P.; Hall, D. W.; Laurie, G.; O'Neill, O.; Rawlins, M.; Thornton, D. J.; Vallance, P. & Walport, M. Science as an Open Enterprise. *The Royal Society,* 2012
3. http://en.wikipedia.org/wiki/timeseries. TimeSeries Definition, 2012.
4. M. Altman and G. King. A Proposed Standard for the Scholarly Citation of Quantitative Data. 13(3/4), Apr. 2007.
5. T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni. Deploying general purpose virtual research environments for humanities research. Philosophical Transactions of the Royal Society A, 368:3813–3828, 2010.
6. OpenAIRE EU Project Website - What is an Enhanced Publication? http://www.openaire.eu/en/component/content/article/76-highlights/344-a-short-introduction-to-enhanced-publications 2012
7. L. Candela, F. Akal, H. Avancini, D. Castelli, L. Fusco, V. Guidetti, C. Langguth, A. Manzi, P. Pagano, H. Schuldt, M. Simi, M. Springmann, and L. Voicu. DILIGENT: integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure. International Journal on Digital Libraries, 7(1- 2):59–80, October 2007.
8. L. Candela, D. Castelli, P. Pagano, and M. Simi. From Heterogeneous Information Spaces to Virtual Documents. In E. A. Fox, E. J. Neuhold, P. Premsmit, and V. Wuwongse, editors, Digital Libraries: Implementing Strategies and Sharing Experiences, 8th International Conference on Asian Digital Libraries, ICADL 2005, Lecture Notes in Computer Science, pages 11–22, Bangkok, Thailand, December 2005. Springer
9. G. Crane, A. Babeu, and D. Bamman. eScience and the humanities. International Journal on Digital Libraries, 7(1-2):117–122, October 2007.
10. Belhajjame K., Goble C., & De Roure D. Research object management: opportunities and challenges. Data Intensive Collaboration in Science and Engineering (DISCOSE) workshop, collocated with ACM CSCW 2012.
11. G. Crane, A. Babeu, and D. Bamman. eScience and the humanities. International Journal on Digital Libraries, 7(1-2):117–122, October 2007.
12. The Executable Paper Grand Challenge, Elsevier http://www.executablepapers.com/
13. Google Inc. Google Webtool Kit. http://developers.google.com/web-toolkit/
14. J. Lave and Wenger. Situated Learning: Legitimate Peripheral Participation. Cam, 1991.
15. S. Woutersen-Windhouwer, R. Brandsma, P. Verhaar, A. Hogenaar, M. Hoogerwerf, P. Doorenbosch, E. Durr, J. Ludwig, B. Schmidt, B. Sierman, “Enhanced Publications”, edited by M. Vernooy-Gerritsen, SURF Foundation, Amsterdam University Press, 2009
16. Van Gorp, P. & Mazanek, S. SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science,* 2011*, 4*, 589-597
17. Nowakowski, P.; Ciepiela, E.; Harezlak, D.; Kocot, J.; Kasztelnik, M.; Bartynski, T.; Meizner, J.; Dyk, G. & Malawski, M. The Collage Authoring Environment. *Procedia Computer Science,* 2011*, 4*, 608-617
18. Candela, L.; Castelli, D. & Pagano, P. Managing Big Data through Hybrid Data Infrastructures. *ERCIM News,* 2012, 37-38

[1] Commonly agreed, workflow is the series of steps procedure taken to complete a given task or job.

[2] WYSIWYG is an acronym for “what you see is what you get”. A WYSIWYG editor is one that allows a user to see what the end result will look like while the document is being created.

[3] www.gcube-system.org

[4] A quantitative data set represents a systematic compilation of measurements intended to be machine readable. The measurements may be the intentional result of scientific research or information produced by governments or others for any purpose, so long as it is systematically organized and described [4].

[5] A time series is a sequence of data points, measured typically at successive time instants spaced at uniform time intervals [1].

Creating and Assessing a Subject-based Blog for Current Awareness within a Cancer Care Environment

Yongtao Lin and Marcus Vaska
Health Information Network Calgary, University of Calgary, Canada

Abstract

The Health Information Network Calgary (HINC) is comprised of a group of libraries providing information services and resources to urban and rural sites in the Calgary Zone of Alberta Health Services. Establishing a current awareness service is a necessity in any discipline, especially in health care. Web 2.0 and social networks have transformed how health care professionals and researchers create knowledge, access information, collaborate, and disseminate research.

One of the earliest forms of social media, blogging has taken the world by storm (1) . Although there is a wealth of literature on the use of blogs in providing current awareness services for libraries, there is a pronounced gap on how blogs are assessed or evaluated, especially for information alert purposes (2) .

Clients within the HINC subscribe to e-mail alerts and RSS feeds, a trend particularly evident within the Cancer Care environment where a number of researchers have already implemented feed readers to remain aware of current literature. However, they often comment on challenges associated not only with maintaining alerts and managing RSS feeds, but also in selecting and creating alerts for unpublished materials. The need for a librarian-facilitated current awareness strategy became more and more apparent. The literature reviewed addressed the value of an alert, namely to indicate a gap in the participant's knowledge, rather than to deliver content the librarians may have perceived as useful (3). The authors saw the creation of a subject-based blog as an opportunity to disseminate current awareness "grey" information to this specific research community.

The Grey Horizon Blog was created in April 2012 using Blogger. The selection and re-aggregation of information involves ongoing assessment of user needs and continuous work on the Blog. A weekly global email-digest listing of the postings will be distributed two months after the launch.

Several metrics will be employed in October 2012 to evaluate the Blog. Blogger itself tracks the number of page-views over time. Google Analytics was set up as it tracks additional information on access and use of the Blog. As clients may be using feed readers to read Blog entries and may thus not visit the Blog at all, Feedburner has also been incorporated to track the number of times that the Blog RSS is accessed, as well as calculating the number of subscribers.

A post-survey will be conducted in six months to complement the web statistics data. The additional feedback and comments will help us determine whether the Blog has successfully created an easy platform for users to keep current with unpublished literature, the type of resources found most important, and whether the amount of time spent maintaining the Blog met expectations.

It is anticipated that this case study will portray how to successfully plan a subject-based blog to meet users' current awareness information needs in grey literature. Further efforts will focus on targeting the Blog to the topic areas in grey literature where users feel more information is needed. The findings from this assessment will direct us to potential marketing opportunities and changing technology that haven't been fully utilized in our Grey Horizon Blog.

Introduction: HINC and the Role of Social Media and Grey Literature in Healthcare

The Health Information Network Calgary (HINC) is a strategic partnership between the University of Calgary and Alberta Health Services (AHS). As a network of libraries covering different disciplines, the HINC is currently in the midst of progressing and extending coverage to a provincial-based service, collectively referred to as Knowledge Resource Services (KRS). The vision for the Network grew from a 2004 report, *Making Information Count*, advocating that "a transformed healthcare system needs a transformed information delivery system, if its practitioners, researchers, patients and their families are to have adequate and timely access to the information essential to their needs" (4). It proposed a new model for library services in the local health organizations, showcasing the successful integration of more dynamic information services supporting patient care, teaching, research, and learning, thereby strengthening the bond between healthcare practitioner and information provider (5). Of the six sites comprising the HINC, two cancer facilities serve nearly 500 cancer care staff.

As has been discussed time and time again, the power of grey literature lies in the fact that it is information that is rapidly produced, available when the user needs it, bypassing the often longwinded peer-review process present in commercially published journals. In fact, this body of literature can perhaps best be equated with social media, in that it is "accessible to anyone at little or no cost…can have a very short time frame of production, and can be altered as needed" (1). With the advancement of technology, social media has become widely integrated in numerous aspects of information services. When considering the role of social media in the pursuit of grey literature, questions may arise as to whether or not healthcare professionals are prepared to accept this type of material as a means of keeping current and staying abreast of the latest information published in their disciplines. Social media engages readers with similar interests, fostering a sense of community development. In health care settings, social media has been widely adapted to "promote health, improve health care, and fight disease" (1). With *Grey Horizon* (http://grey-horizon.blogspot.com), social media has been adapted in our project as an effective platform for information and knowledge management. As such, it entails the provision of better communication, an alternative and expected social norm. As information professionals, we share our skills in retrieving and selecting appropriate grey literature material that will enhance our credibility in finding quality information.

One of the fundamental purposes of current awareness, exhibited in this paper by means of a blog, is to ensure the easy, convenient, and wherever possible, free access to information all in one place, offering support to the user every step of the way. As the discovery of grey literature material is heavily dependent upon information being made available online, engaging users to embrace social media components for their information-seeking pursuits is becoming crucial for health researchers. An interesting paradigm, considering that the blog, of which there are an estimated 70 million today, is often considered the earliest form of social media (1). According to a survey and research findings conducted in this field, 50% of physicians in Europe and 41% in the U.S. regularly engage and/or post in blogs (1). Further, 50% of medical organizations surveyed currently use blogs in their information pursuits, with another 80% planning to do so shortly (6). The cancer care librarians and creators of *Grey Horizon* thus echo Chu's notions that due to the interactive nature of the diary-like postings, "blogs are more dynamic and permit writers to engage in one to many conversations with their readers" (6).

Background: Current Awareness' Importance in Cancer Care Leads to Creation of the Blog

The notion of current awareness in the cancer care environment served by the *Grey Horizon* Blog is not new. Current Awareness services, "designed to keep users informed about recent developments" (7), have existed in Cancer Care since 2008. Librarians involved in this study have responded to user demand by spending countless hours over the years meeting with researchers and health care professionals on an individual basis to create current awareness guides or digests, providing instruction on how to set up a table of contents e-mail alert and/or an RSS feed for a particular journal or search strategy, catering to their clients' needs. While cancer researchers were keen, motivated, and engaged with this initiative at the beginning, increasing workloads, time constraints, and other pressing commitments soon put current awareness on the backburner. Many researchers echoed the sentiments of Attfield, Blandford, and Makri, as "participants frequently found attending to current awareness information overwhelming" (7). By sorting, selecting, and placing relevant information in a central location (i.e. blog) that all researchers could access on their own time (in lieu of filtering out a deluge of e-mails), the authors of this paper believe that issues of information overload and time constraints can be better managed.

For a librarian not to recognize his/her role as an information mediator and accept the direction that social media and the selective dissemination of information is taking, is to demonstrate that one's role in this profession is done. While the introduction of a blog in the specific cancer care environment discussed in this paper is a new initiative, it is not entirely unique, but rather merely goes with the flow of the connected health care professional. As a subject-based blog, *Grey Horizon* delivers timely and accurate information exclusively from grey literature resources, to the community in need.

Method: Designing and Creating the Grey Horizon Blog

Deciding on a subject-based blog

Findings from the literature support a strong association between current awareness and social media in healthcare. The HINC has and continues to integrate social media in an effort to reach out to the connected users served by this organization. Instant messaging chat reference, a *Twitte*r account, and a *YouTube* channel containing brief self-paced tutorials, all demonstrate the need of library services to recognize the importance of this communication medium. The nature of the cancer care environment within which the authors of this paper are employed is fast-paced with ongoing clinical trials, evidence-based guideline development, , grant proposal applications, physician meetings, and conference presentation opportunities. In fact, Health Information Network staff can be seen as information brokers, acting "as a layer of 'intelligent filters' sensitive to complex, local information needs; their decisions aim to optimize conflicting constraints of recall, precision, and information quantity" (7). Cancer Care librarians integrate current awareness in their provision of information, by offering in-person training, facilitating virtual interest groups, and creating subject-specific e-mail alert digests, emphasizing that librarianship most go forward and acknowledge the joint role of current awareness and grey literature, particularly in a 21st century technologically-enriched society.

Planning for the Blog

Despite the wealth of information available focussing on the need and importance of blogs, far less literature has been written on evaluating and assessing a blog (2). The authors of this paper have thus applied Blair and Level's guide of subject blogging etiquette while planning *Grey Horizon*. Before embarking on the actual creation of the Blog, the authors first established and planned the process and methodology dictating how *Grey* Horizon would come to fruition. Blogger (http://www.Blogger.com) was chosen as the software, due to the program's ease of posting information, tracking followers, and generating statistics. Further, the program is free anduser-friendly. Once the program of choice was decided on, appropriate material from which the Blog's postings would be generated was selected, based on the authors' past experience and knowledge of grey literature cancer resources. This included a wide array of grey literature cancer resources, such as Canadian Partnership Against Cancer, Cancer Trials Canada, New York Academy of Medicine, Canadian Cancer Society, Canadian Health News, NICE guidelines, American Society of Clinical Oncology, American Association for Cancer Research, the European Society of Medical Oncology, and several others.

Organization of content was considered an essential planning step, as the volume of posts in a Blog can become unwieldy if not appropriately managed. To manage the scope and breadth of content, a controlled list, consisting of a series of tags corresponding to grey literature types, was established at the outset. Each new item posted on the Blog would thus be labelled accordingly. The most common tags, many of which were formed according to the nature of the postings, were Cancer Research, News, Conferences, Drug Updates, Case Studies, Reports, Clinical Trials, and Guidelines. In addition to the tagging mechanism utilized, the Blog's interface also contained a Favourite Links section, as well as an area where the reader could find out additional information about grey literature, as well as instructions on how to set up an e-mail alert and RSS feed.

To keep track of all of the information and make it easier for the project team to oversee all information sources for *Grey* Horizon, a shared email account was created via Gmail for storing and monitoring all RSS feeds and e-mail alerts. Meanwhile, a Blog log was initiated to keep track of the progress of postings, complete with time spent and any comments the Blog creators had with regards to issues encountered with the postings (i.e. website down, inappropriate posts, etc.) Inclusion and exclusion criteria were established before the Blog went live, to ensure that postings covered numerous aspects of cancer care, including research findings, clinical trials, and conference proceedings, as well as to maintain consistency among blog postings. Criteria common across all document types, and adversely affecting the decision of whether or not to post a particular news item included current, unbiased information from selected sources, clinical trials not endorsed by pharmaceutical companies, recognizing the various formats of grey literature and posting accordingly, as well as adhering to the main subject areas/types of sources previously identified from the users. To attract and sustain reader attention, new postings must appear on a daily basis.

Creation of the Blog

Grey Horizon was created in April 2012, and officially launched on April 30, 2012. Prior to the official unveiling, a soft launch was held for two weeks to allow for pretesting with a few key users, measure workload involved, and finalize the interface design.

Bi-weekly digests

Three months after launching the Blog, the librarians decided to re-aggregate selected postings into a bi-weekly digest format. The digest was disseminated to all staff at both cancer sites via an e-mail listserv. This promotion effort, combined with a link to the Blog being placed on the HINC website, established multiple avenues by which the Blog could be accessed and postings read. Along with the posting's relevancy to the cancer community at large, presenting information via a digest format provides a brief and condensed way for users to access the same information. Due to the nature of grey literature and the importance of disseminating current information in cancer care to Blog readers, *Grey Horizon* is reviewed daily, not only in posting up the latest cancer care news, but also in allowing our readers determine the direction the Blog will take. While the primary evaluation period was not scheduled until six months after the Blog launch date, librarians review postings and listen to reader comments on a continuous basis, thus following "what users would change or like to see added" (2).

Monitoring and maintaining the Blog

Due to the elusive nature of grey literature, websites that served as the focal point one day can be moved or disappear. Thus, as Attfield, Blandford, and Makri comment, the importance of continuous monitoring, particularly with a form of social media such as the blog, cannot be overlooked: “monitoring is an ubiquitous activity and can take many forms, frequently combining both formal routes with less formal routes, such as the use of social networks” (7). As creators of the Blog, the two cancer care librarians applied Ellis’ model of information seeking (8-9); namely, appropriate cancer resources were identified and located (whether by conducting a search or implementing an e-mail alert or RSS feed), accessed, selected, and processed (i.e. posted on the Blog).

Methods: Analyzing and Evaluating the Blog

As there is no one standard metric for the evaluation of a blog (2), the authors decided to employ several metrics addressed in the literature to evaluate the usefulness and effectiveness of *Grey Horizon* . This included looking at participation through comments (10), webtrackers and other RSS feed reader tracking websites, in addition to monitoring blog traffic. Both qualitative and quantitative data from the post-survey and comments were gathered to assess how the blog meets successful criteria.

Blogger

Blogger, as an online platform for the creation of the blog, tracks the number of page-views over time. A total of 6806 page views on 463 postings were tracked from the Blogger from May to October 2012. As indicated from Figure 1, traffic has significantly increased during the six-month project phase. There were nearly three times (1835 page views) as many visitors in October as the first month in May (691 page views)

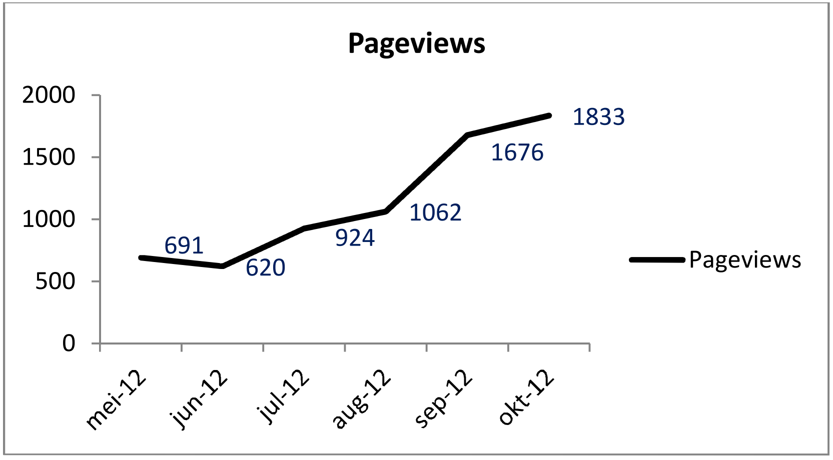


Figure 1: Number of page views by month from Blogger

The popularity of individual postings was also determined by the number of page views tracked, thus confirming the value of core types of grey literature for cancer researchers. Based on the total number of 463 postings as of October 22, 2012, the following was noted:

- 84 page views for news about a specific research **conference**
- 56 page views for a cancer **drug update** from Health Canada
- 53 page views for a **report** on cancer health services
- 51 page views for a clinical **guideline** from the National Institute for Health and Clinical Excellence (NICE)
- 46 page views for **statistics** on cancer incidence, survival and risk factors

Further, a definition of grey literature (66 page views), along with instructions on how to obtain full text papers from the references mentioned in the postings (23 page views) were among the most accessed pages within the More Information section of the Blog. Undoubtedly, users seek expert guidance in successfully retrieving, filtering, and understanding information.

Google Analytics

As a popular web tracking instrument, *Google Analytics* captures how visitors interact with a website. Therefore, this tool was set up to track additional information on access and use of the *Grey Horizon* Blog, including number of visits, number of hits, types of access, visitors from referring websites, and detailed user behaviour when visiting the Blog. The breakdown of new visitors and return visits was particularly helpful in evaluating the usefulness of the content and effectiveness of blog promoting strategies (Figure 2).

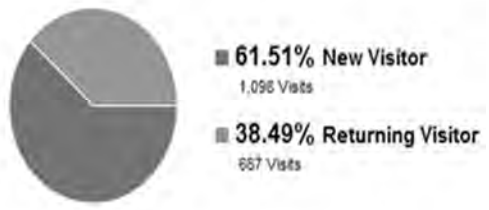


Figure 2: Percentage of new visitors vs. returning visitors of the Blog

According to *Google Analytics*, a total of 1785 people have visited *Grey Horizon*, as identified by unique IP address. An average visit duration of 4 minutes 14 seconds indicates that time is being taken to read and digest the information presented. In addition, it is helpful to examine how visitors navigate various pages, along with the different interactions undertaken.

Feedburner
Feedburner is a tracking service for RSS subscribers. As clients may be using feed readers to peruse blog entries and may thus not visit the Blog at all, by replacing the automatically generated RSS feed with one from *Feedburner*, *Feedburner* tracks the number of times that the Blog RSS is accessed, as well as calculating the number of subscribers. A total of 17 people subscribed to the *Grey Horizon* RSS feed over six months. On average, 12 out of 17 subscribers remained active by taking the actions of clicking the links to the postings, or viewing them in their feed readers. Responses from the post-survey questions on whether visitors subscribed to the Blog feed and their experience regarding use of an RSS feed reader confirmed users' education needs on using RSS feeds for current awareness purposes, a future instruction effort for librarians via more integrated alert services.

Reader feedback/comments
Having the ability and opportunity to comment based on postings is one of the most important features of a blog. The use of comments often foreshadow a blog's success. Since the "Comment" feature was turned on in September 2012, *Grey Horizon* has received seven comments from readers. Some echoed their own experiences with cancer, while others expressed their opinions on some of the research posted. Reader feedback and comments on our Blog were also observed from the connection activities with the Health Information Network Twitter account. The HINC Website links the *Grey Horizon* Blog in the current library news column and the tweets on individual postings were therefore tracked accordingly. As no single evaluation metric is able to present the full story, it was extremely helpful with ongoing service planning to have seen similar comments and feedback from different evaluation channels.

Post survey
A post-survey was distributed to the Cancer Care listserv in September 2012 to complement the web statistics data, with the purpose of helping the authors determine whether the Blog has successfully created an easy platform for users to keep current with unpublished literature, the type of resources found most important, and whether the amount of time staff spent maintaining the Blog met users' expectations. Fifty-one people completed the survey, comprised of questions in four sections:

- Users' familiarity with grey literature and current awareness practices
- Experience with the Blog
- Experience with bi-weekly digests
- Additional comments and suggestions on future current awareness services

43.5% of respondents mentioned unfamiliarity with the concept of grey literature prior to accessing the Blog. This finding was consistent with previous studies conducted by the authors indicating that most users successfully incorporated grey literature in their research despite being unaware of this term. (11) Google still appeared to be the most predominant source of information for grey literature searching for 21 out of 51 respondents. Despite an overwhelming number of people (67.3%) finding it challenging to keep abreast of current information in their fields, cancer researchers and healthcare professionals are still in the traditional mode of receiving and sharing new information. Attending workshops and conferences, setting up journal alerts via email, and communicating with colleagues are among the most frequent practices.

The post-survey has also reflected the success of the *Grey Horizon* Blog. 35 out of 51 respondents had accessed the Blog, a frequency ranging from a few times a month to every day. Among these, nearly all respondents (94.1%) found the postings useful. For those who hadn't accessed the Blog before, having a busy clinical/research schedule was the most noted barrier. The usefulness of the topics was rated by the readers, as shown in Figure 3. Cancer Research, Clinical Trials, Reports and Drug Updates were the top four categories, the findings being consistent with the number of page views tracked by Blogger.

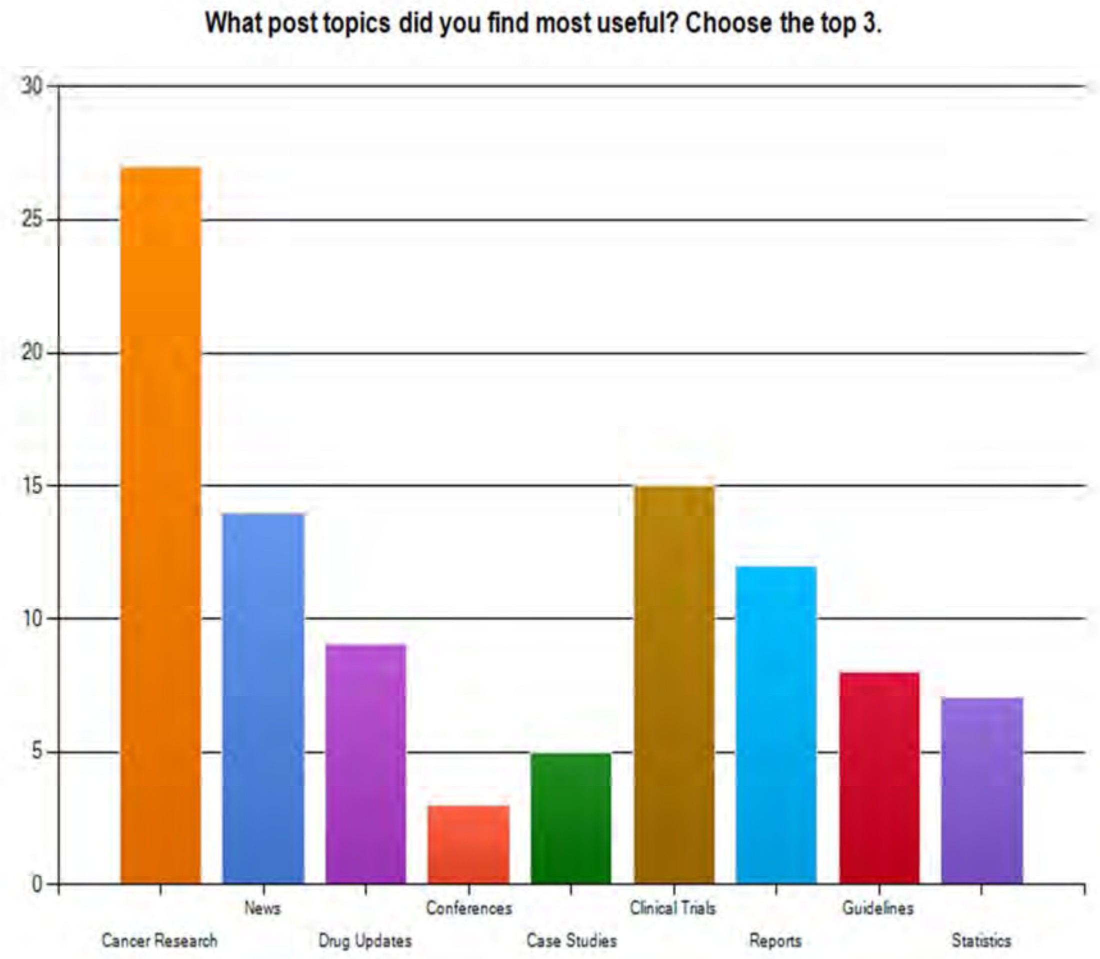


Figure 3. Usefulness of posting topics perceived from the Post-Survey

Additional features of the Blog, including providing links to other websites, pointing to the published studies mentioned in the post, and creating pages to inform about grey literature and how to access full-text articles were most favoured by the readers.

Bi-weekly digests appeared to be a very successful pursuit in re-aggregating the information in a user-friendly format, promoting the site to new staff as well as serving as reminders for existing readers to revisit the Blog. Ongoing assessment of the user needs and thorough planning of this current awareness project have become common themes the authors discovered when analyzing and evaluating the success of the Blog. When readers were asked to articulate their views on what may have contributed to the usefulness of the Blog and the postings, key features of a blog, including ease of interface navigation and information gathered in one place, as well as selection of various relevant sources and provision of current, pertinent information in one place were stated as the main criteria.

Discussion and Future Directions

Enhanced research interaction and collaboration

Grey literature's placement and availability on the World Wide Web, particularly the new Web 2.0 generation is essential to its success. These new services "encourage group interaction in a space where individuals can participate, socialize, and set social norms" (12). Such was the ideology behind the creation of *Grey Horizon*: to allow readers to not only learn about the latest cancer care news, but also to actively participate and interact with their colleagues, as well as with the cancer care librarians. The Blog has achieved a chain-reaction effect, demonstrating the power of collaborative learning and information sharing.

Librarians' integrated roles in current awareness services

Librarians in this project act as the conduits of information, browsing through RSS feeds and e-mail alerts to actively pull relevant information and responsively push this out to the users, the readers of the

Blog (7). The authors are thus the gatekeepers, controlling the flow of information to craft Cancer Care South into a well-informed and efficient organization (7). Although the creators of the *Grey Horizon* Blog identified resources that, in their opinion, were best suited to researchers within a cancer care research facility, it is important to note that the two librarians do not consider themselves as content experts in this field. This then emphasizes the collaboration needed between librarian and researcher, where the librarian seeks a better understanding of the information being requested is a form of contextual inquiry, playing a key role in the data that is gathered and eventually placed in the Blog (7).

Blog as a successful platform in promoting current awareness services

It is evident from this project that many researchers and healthcare professionals appreciated being “pushed” relevant information in their field, as they often did not have the time to go searching for this information themselves. Tailored information to different fields of interest in cancer was suggested as a future direction for a more dynamic and laddered current awareness service. The Blog has also become an invaluable tool that can aid in planning integrated information services, and promoting library instruction courses, special events, and new resources.

Becoming a source of grey literature

Although *Grey Horizon* has only been in existence for a mere 6 months, it is gaining ground and is serving as an important source of grey literature material in the cancer care environment in Alberta. This relative success within a short time frame demonstrates that subject-based blogs work well as a current awareness service, highlighting the importance of disseminating useful content on relevant topics to the intended audience. Readership is growing, requests from readers to link to external blogs is raising even further awareness.

Current awareness project planning

Grey Horizon reinforced the need for thorough project planning, achieving balance between staff time, workload, and user expectations. Additional features of the Blog are being explored, along with new methods of marketing and promotion. The authors are presently investigating the possibilities of linking and sharing with other blogs in the same subject field, as well as creating a mobile site for the Blog. . Despite making a few adjustments over the course of this pilot study, the creators of the Blog often heed the words of Chu et al. words that can be applied to the importance of intertwining current awareness with grey literature: “what attracts users is not technology but how you make use of the technology so that they can fully utilize the tools to accomplish things they want” (6).

References

(1) Graham DL. Social media and oncology: Opportunity with risk 2011:421-424.

(2) Blair J, Level AV. Creating and evaluating a subject-based blog: planning, implementation, and assessment. Reference Services Review 2008;36(2):156-166.

(3) Attfield S, Blandford A. Conceptual misfits in e-mail-based current-awareness interaction. Journal of Documentation 2011;67(1):33-55.

(4) Calgary Health Region, University of Calgary. Making Information Count: AN integrated knowledge service for healthcare practitioners, staff, patients and their families 2004.

(5) Reaume R, Aitken E, Powelson S. Library advocacy at the bedside and the boardroom: Calgary Health Information Network 2010; Available at: http://www.chla-absc.ca/2010/graphics/chla2010-poster17.pdf. Accessed October 29, 2012.

(6) Chu SKW, Woo M, King RB, Choi S, Cheng M, Koo P. Examining the application of Web 2.0 in medical-related organisations. Health Information & Libraries Journal 2011;29(1):47-60.

(7) Attfield S, Blandford A, Makri S. Social and interactional practices for disseminating current awareness information in an organisational setting. Information Processing and Management 2010;46(6):632-645.

(8) Ellis D. Modeling the information-seeking patterns of academic researchers: A grounded theory approach. The Library Quarterly 1993;63(4):469-486.

(9) Meho LI, Tibbo HR. Modeling the information-seeking behavior of social scientists: Ellis’s study revisited. Journal of the American Society for Information Science and Technology 2003;54(6):570-587.

(10) Jackson A, Yates J, Orlikowski W. Corporate blogging: Building community through persistent digital talk. Proceedings of the 40th Hawaii International Conference on System Science; Available at http://doi.ieeecomputersociety.org/10.1109/HICSS.2007.155. Accessed November 12, 2012.

(11) Lin Y, Vaska M. Raising awareness of grey literature in an academic community using the cognitive behavioral theory. Eleventh International Conference on Grey Literature: The Grey Mosaic, Piecing it all Together; Available at http://www.opengrey.eu/item/display/10068/698105. Accessed November 14, 2012.

(12) Cho A. AN introduction to mashups for health librarians. Journal of the Canadian Health Libraries Association 2007;28:19-22.

Conference Announcement

GL Conference Series 1993-2013

Fifteenth International Conference on Grey Literature

The Grey Audit: A Field Assessment of Grey Literature

Bratislava, Slovak Republic 2-3 December 2013

Slovak Centre of Scientific and Technical Information

CENTRUM VEDECKO-TECHNICKÝCH INFORMÁCIÍ • SR

CVTI SR

GL15 Program and Conference Bureau

TextRelease

Javastraat 194-HS, 1095 CP Amsterdam, The Netherlands
www.textrelease.com ▪ conference@textrelease.com
Tel/Fax +31-20-331.2420

Centralised National Corpus of Electronic Theses and Dissertations

Július Kravjar and Marta Dušková
Slovak Centre of Scientific and Technical Information

Abstract

ETDs are a significant source of grey literature and not only for the academic community. Slovakia has made a big step forward by implementation of the Centralised National Corpus of Electronic Theses and Dissertations in 2010. The national ETD corpus consists of bachelor's, master's, dissertation and habilitation theses. This implementation was coupled with the concurrent implementation of the National Plagiarism Detection System (aka the originality check or the anti-plagiarism system). Both systems have to be used by all higher education institutions operating under the Slovak legal order. The new theses and dissertations incoming into the national ETD corpus are compared to this corpus and to the selected internet resources. The higher education institutions pay no fee for the service, the system acquisition costs were covered by the Ministry of Education, Science, Research and Sport, and the operating costs are also paid by the Ministry. The formation of both systems and the first two years of its existence is analyzed.

The first signs of activities towards ETD in Slovakia were recorded on the threshold of the Millennium. March 2004 was to become a significant milestone: sixteen academic libraries of twelve Slovak universities decided to solve the ETD.SK project: "Building Digital Academic Libraries - Collecting and Providing Access to Full Texts of Slovak University Publications". The ETD.SK project marked the beginnings of cooperation on a national level in this area, with the effort to follow up international ETD activities. Unfortunately, the project was not sufficiently implemented due to the lack of financial and personnel resources, but mainly because of the lack of legislative support.

The ICT and internet penetration, low copyright awareness and the rapid growth in the number of higher education institutions and students in our country contributed to the expansion of plagiarism. There was also an inherent lack of systemic action that would act as a barrier for its future growth. The establishment of the nationwide electronic theses and dissertation repository and their originality check was considered as a perspective solution.

A significant step in this matter was made in 2008: the Ministry of Education decided to implement a comprehensive nationwide solution for the collection and processing of theses and dissertations produced at Slovak higher education institutions. The goal: creation of the national theses and dissertations repository, increase in the quality of theses by their originality check, copyright and intellectual rights protection. In 2009, the Higher Education Act was amended and the most relevant change was this: Before the defence of the thesis, the higher education institution forwards the thesis in the electronic form to the Central Repository of Theses and Dissertations (CRTD) and the originality check is performed.

During 2009, the CRTD was built and the whole system with the originality check became reality at the end of April 2010. The existence of such a system has a preventive effect, and not just in the student community. CRTD is now a publicly-available source of grey literature.

A yearly increase in CRTD is about 80 thousand items of bachelor's, master's, dissertation and habilitation theses.

The paper analyzes the creation and two years' operation of the national corpus of bachelor, master, diploma, dissertation and habilitation theses of Slovak higher education institutions and the follow-up plagiarism detection system. The national corpus is called The Central Repository of Theses and Dissertations (CRTD). Each thesis has to be entered in CRTD before defence and it is then checked for plagiarism.

Creating Collections of Higher Education Theses: First Steps

"Creating digital collections of own academic production and making them available as full text digital documents, accessible in the computer network, has been gradually taken up by Slovak academic libraries in the last ten years. Projects aimed at the electronic processing of publications of university employees and at the electronic processing of theses and dissertations were the pilot projects in this area. Academic libraries implemented the first projects already at the beginning of the millennium.

March 2004 was a significant milestone in the area of Slovak academic digital libraries when 12 Slovak Universities presented the central development IT project "Building Digital Academic Libraries – Collecting and Providing Access to Full Texts of Slovak University Publications (ETD.SK)". ETD is the internationally used abbreviation for the university graduation theses in the electronic form, i.e.

Electronic Theses and Dissertations. ETD.SK Project started the cooperation in this area at the national level with the effort to continue in international activities in the ETD area. The Project involved the specification of the organisational, technical and technological requirements for ETD collection, digitalisation workplaces were created in individual libraries and two hardware storage sites were built (in the cities of Košice and Prešov). Unfortunately, the Project was not implemented sufficiently in all libraries, mainly for insufficient financial and staff coverage; however, mainly for insufficient legislative support.“ (Haľko, 2011)

A positive effect was achieved as the universities started to make theses and dissertations accessible through online catalogues of academic libraries available on the Internet.

Plagiarism

Plagiarism existed in the past and will probably continue to exist in the future. In the Ancient Rome the word plagiarist (*plagiarius*) designated a kidnapper, mainly kidnapping children or slaves. Plagiarism action (*plagium, crimen plagii*) also meant offering refuge to an escaping strange slave.

The content of the word has been significantly modified and is kept only in a figurative meaning. Today ”kidnapping” is understood as a theft of ideas, opinions or other intangible property that are illegally “appropriated“ and presented to be original, authentic. Plagiarism can be defined as use of original ideas and creative formulations of another person with the aim to present them as one’s own ideas or formulations. (Szattler, 2007)

In the absence of suitable instruments for the prevention and suppression of plagiarism, this phenomenon may overgrow to unacceptable dimensions. It would probably not be possible to fully eliminate all kinds of plagiarism in the future; however, it is necessary to build barriers where possible. We must not ignore or tolerate it.

The first higher education institution in Slovakia – the “early bird“ – started using the system to reveal plagiarism in 2001 – it was a lonely runner during long seven years.

Background Situation

Slovakia with its 5.4 million inhabitants is confronted with plagiarism of higher education theses and dissertations like other countries. Plagiarism was growing at the higher education scene before 2010. There were no systemic measures preventing its growth. Rapidly increasing number of higher education institutions and students, growing Internet penetration, low understanding of copyright and intellectual property rights – this helped the growth of plagiarism. If we look at 1989 (year of Velvet Revolution[1]) as a basis – and compare it with 2011, the number of schools increased three times to 39 and the number of students increased four times to 250 000. In 1989, Internet penetration was zero; it reached 74.3% in 2010 and 79.2% in 2011.

How does the Internet help plagiarism? In two ways. The Internet offers free papers, final theses, dissertations and also expert literature. Insensitive use of the “copy and paste“ functions with no adequate quotations compound in the text signed by the “author“ – this is a typical example of plagiarism. Offers related to the preparation “supporting materials“ for various types of theses can be found on the Internet with little effort. There are web pages offering the preparation of a wide range of theses (seminar, graduate, bachelor, diploma, dissertation, MBA etc.) covering markets of more than ten countries, and there are also web pages aimed only at local markets or markets with similar language or history. It is necessary to realise that “publication of another person’s work as one’s own is plagiarism“. (SME.SK, Adamová, 2012)

Order and payment for the prepared bachelor, diploma or dissertation theses is not only the Slovak speciality, this is a global problem. If it is found before the graduation that the thesis was ordered, such student may be dismissed from his studies prematurely. If this is found after the graduation, nothing happens. Our Higher Education Act does not define the removal of a university degree. The degree will remain in the hands of the owner (Hospodárske noviny, 2012). It seems that this will change in the near future. The removal of fraudulently acquired academic degree should be defined in the Criminal Code as informed by the daily news in August 2012 (TASR, 2012).

The study of Králiková “Introduction of Rules of Academic Ethics at Slovak Higher Education Institutions“ summarises the state of academic ethics at Slovak higher education institutions and shows that the majority of inquired pedagogical workers directly experienced fraud of students. The most discussed topic in the media was the plagiarism of students, mainly the ways in which the students cheat and the instruments used by higher education institutions to eliminate plagiarism. Other themes included plagiarism of pedagogical workers or persons in high public positions. The absence of a wide discussion on academic ethics has also other consequences. One of them is that the academic environment members and the general public do not understand the importance of academic ethics and they are also less sensitive to a breach of it.“ (Králiková, 2009)

Issues related to the collection and processing of higher education institutions theses in electronic form and issues of plagiarism were often repeated in discussions in the academic area; however, with no significant progress. Seeds of future changes were seeded at the meeting of the Slovak Rectors' Conference in September 2006 (Slovak Rectors' Conference, 2006) when two documents on academic ethics were approved. These nationwide documents deal with ethics of pedagogical workers and students. However, the measures proposed to prevent plagiarism did not come to life (Králiková, 2009). In February 2008, the Slovak Rectors' Conference returned to the problems of plagiarism and asked The Ministry of Education, Science, Research and Sport of the Slovak Republic to coordinate activities connected with the acquisition of the plagiarism detection system. Higher education institutions were recommended to punish plagiarism and to create electronic archives of theses (Slovak Rectors Conference, 2009).

Although the plagiarism and the need to fight against it were being discussed a lot, the final decision was adopted by The Ministry of Education, Science, Research and Sport of the Slovak Republic in 2008 (in 2008, only two higher education institutions used the plagiarism detection system) for all higher education institutions operating under the Slovak legal order:

- To establish The Central Repository of Theses and Dissertations (CRTD), higher education institutions have to send any thesis or dissertation to CRTD;
- Any thesis and dissertation with the name and surname of the author and higher education institutions will be stored in CRTD for a determined period of time from the day of registration;
- Theses and dissertations must be checked for originality before defence, as proved by the originality check protocol;
- The originality check protocol is the result of the comparison with all the theses and dissertations in CRTD and to Internet sources and other available electronic sources;
- Higher education institutions‘ local repositories of theses and dissertations are the CRTD's communication partners.

The defined tasks were aimed to increase the quality of higher education institutions studies and also:

- Copyright and intellectual property rights protection, its better understanding;
- Theses and dissertations quality improved due to the originality check;
- The creation of The Central Repository of Theses and Dissertations;
- The creation of barriers to growing plagiarism.

Establishment of the Central Repository of Theses and Dissertations

In 2009, the Higher Education Act was amended and the most important modifications were:

1. The Ministry of Education, Science, Research and Sport of the Slovak Republic will administer the Central Repository of Theses and Dissertations
2. Before any person is allowed to defend his/her thesis/dissertation, the higher education institution sends his/her thesis/dissertation to the Central Repository of Theses and Dissertations in electronic form and undergoes the originality check.
3. Thesis/dissertation will be stored in the Central Repository of Theses and Dissertations with the name and surname of the author and the name of higher education institution for the period of 70 years from the day of registration. (Zbierka zákonov (Collection of Acts), 2009)

Theses and dissertations registered in the Central Repository of Theses and Dissertations after 31 August 2011 are publicly available – this was defined in the amendment to the Higher Education Act. (Zbierka zákonov, 2011)

The Central Repository of Theses and Dissertations of the Slovak higher education institutions and the Plagiarism Detection system became real at the end of April 2010. Higher education institutions operating under the Slovak legal order are obliged to use both systems. Approximately 80 000 theses are registered in the Central Repository of Theses and Dissertations yearly. As of 31 October 2012, there were 223,757 theses registered.

What did the implementation of the Central Repository of Theses and Dissertations and the plagiarism detection system (the originality check) bring?

Before the implementation of the project, it was necessary to consider and address:

- The centralised national corpus of electronic theses and dissertations;
- The growing plagiarism of theses and dissertations;
- The occasional efforts of higher education institutions to check originality of theses and dissertations;
- The absence of system instruments to fight against plagiarism at the national level;
- The public access to theses and dissertations from one spot;
- The increased understanding of copyright and intellectual property rights.

The published information on the prepared centralised national corpus of electronic theses and dissertations and the originality check brought preventive effect even before the implementation. Other effects of the implemented project were the following:

- The principal breakthrough in the fight against plagiarism in Slovakia, obligatory use of the Central Repository of Theses and Dissertations and the originality check (with the Central Repository of Theses and Dissertations, Internet sources and other electronic sources).
- The existence and real operation of the instrument for the protection of copyright and suppression of plagiarism.
- Unified methods for collecting theses and dissertations (creating centralised repository for all higher education institutions) and unified system for the originality check (Plagiarism Detection System).
- Automated collection of theses and dissertations, originality checks and distribution of originality check protocols.
- The existence of CRTD (as one of the sources of grey literature) and the system to detect plagiarism is preventive in the community of students and others. It increases the understanding of copyright and intellectual property rights at least in the academic area; it improves how students work with literature, Internet, quotations, and improves the quality of theses and dissertations.
- All higher education institutions in Slovakia operating under the legal order of the Slovak Republic are obliged to use the complex of the CRTD with the originality check at the national level. This applies to 35 higher education institutions of 39.
- All theses and dissertations are collected in the CRTD where they are stored for 70 years.
- The public may verify the suspected plagiarism on the web page where theses and dissertations are published: http://www.crzp.sk/crzpopac?fn=*searchform.
- The originality check protocol does not confirm that a thesis/dissertation is original or a plagiarism. The protocol is the basis for the examination committee’s decision, it helps – it informs about documents a supervisor or an opponent might have overlooked. The originality check protocol identifies the parts of the text of the presented thesis/dissertation identical with the parts of texts deposited in CRTD and with Internet sources. The originality check protocol is available to the examination committee for evaluation and it is a part of the final (national) exam records.
- Higher education institutions do not pay to use these systems; the procurement costs were covered by the Ministry of Education, Science, Research and Sport of the Slovak Republic and the operation costs are also paid by the Ministry.

In Conclusion

There are large reserves in the fight against plagiarism in the education of the young generation. Plagiarism should be prevented and the process of education to the non-cheating culture must start gradually and adequately from the earliest age, from the pre-school education level. (Skalka, et al., 2009).

A correctly oriented and properly timed educational process and the implementation of advanced technologies have high potential in the fight against plagiarism. Technologies, however, are not a panacea. Very important and irreplaceable is the role of education – from the beginning of the educational process – in close connection with the prevention and plagiarism uncovering with clearly defined rules and penalties, and with mutual interaction of all these parts. (Kravjar, 2011).

The implementation of the CRTD together with the originality check at the national level in everyday practice is a unique solution in Europe and probably in the world. The system has high potential to be implemented in many areas. The originality check can be done where a written work is the outcome.

The following may be considered:

- Seminar and other works at higher education institutions;
- Research reports;
- Applications for projects, grants, their outcomes;
- Final works for increasing qualification of pedagogical and other professions;
- Secondary school works within the curriculum;
- Secondary school works beyond the curriculum;
- General publication activities (the establishment of relationships with publishers);
- The repositories of grey literature.

The system is still being developed. Last year, the system supplier (a company called SVOP, Ltd, http://svop.sk/en/Default.aspx) won the first prize at the international competition of anti-plagiarism systems in Amsterdam "PAN 2011 Lab Uncovering Plagiarism, Authorship, and Social Software Misuse" held in conjunction with the "CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation", where their algorithm detecting plagiarism was the best in all four indicators (plagiarism detection, recall, precision, granularity). It is necessary to remind that the competition corpus did not include Slovak texts, only English, German and Spanish texts, and to point out that the system was also able to uncover the so-called "translated plagiarism", and that the detection is invariant against a change of word order, against the occurrence of changed words, against omission or additions of words in the text in a suspicious document.

The plagiarism detection system received a significant award at the international conference ITAPA 2011 in Bratislava (Information Technology and Public Administration). In the category New Services, the nationwide plagiarism detection system and CRTD ranked second.

This paper is a reworked and updated version of the conference paper Barrier to thriving plagiarism (Kravjar, 2012).

References

HAĽKO, P. (2011). *Elektronický zber záverečných prác na Prešovskej univerzite v Prešove*. [ONLINE] Available at: http://www.pulib.sk/elpub2/PU/Uninfos2011/data/ P11_Halko.pdf. [Accessed 26.6.2012].

KRÁLIKOVÁ, R. (2009). *Zavádzanie pravidiel akademickej etiky na slovenských vysokých školách.* [ONLINE] Available at: http://www.governance.sk/assets/files/publikacie/akademicka_etika .pdf. [Accessed 21 March 2012].

KRAVJAR, J. (2012). *Barrier to thriving plagiarism*. [ONLINE] Available at: http://www. plagiaris madvice.org/documents/conference2012/finalpapers/Kravjar_fullpaper.pdf. [Last Accessed 22.8.2012].

KRAVJAR, J. (2011). *Antiplagiátorský systém.* [ONLINE] Available at: http://www.inforum.cz/pdf/2011/kravjar-julius.pdf. [Accessed 23 March 2012].

SKALKA, J. et al. (2009). *Prevencia a odhaľovanie plagiátorstva.* [ONLINE] Available at: http://www.crzp.sk/dokumenty/prevencia_odhalovanie_plagiatorstva.pdf. [Accessed 21 March 2012].

SZATTLER, Eduard. *Právne a morálne aspekty plagiátorstva. Duševné vlastníctvo : Revue pre teóriu a prax v oblasti duševného vlastníctva* [online]. 2007, 1, [cit. 2011-04-06]. Dostupné z WWW: <http://www.solain.org/clanky/20070325192403.pdf>.

SME.SK (e.g. 2011). *Veľký biznis: Diplomovka na kľúč v akcii za pár stoviek eur* . [ONLINE] Available at: http://www.sme.sk/c/6329494/velky-biznis-diplomovka-na-kluc-v-akcii-za-par-stoviek-eur.html. [Accessed 28.6.2012].

HOSPODÁRSKE NOVINY (2012). *Napíšem vám diplomovku. Stačí zaplatiť.* [ONLINE] Available at: http://style.hnonline.sk/vikend/c1-55146090-napisem-vam-diplomovku-staci-zaplatit. [Accessed 23 March 2012]

SCHÖPFEL, J. (2010). *Towards a Prague Definition of Grey Literature*. [ONLINE] Available at: http://archivesic.ccsd.cnrs.fr/docs/00/58/15/70/PDF/GL_12_Schopfel_v5.2.pdf. [Accessed 22.8.2012].

TASR.SK (2012). *Podvodne získaný titul bude trestným činom* . [ONLINE] Available at: http://www.sme.sk/c/6511466/podvodne-ziskany-titul-bude-trestnym-cinom.html. [Accessed 30.8.2012].

ZBIERKA ZÁKONOV (2009). *Zbierka zákonov č. 496/2009.* [ONLINE] Available at: http://www.zbierka.sk/sk/predpisy/496-2009-z-z.p-33272.pdf. [Accessed 20 March 2012].

ZBIERKA ZÁKONOV (2011). *Zbierka zákonov č. 6/2011.* [ONLINE] Available at: http://www.zbierka.sk/sk/predpisy/6-2011-z-z.p-33957.pdf. [Accessed 20 March 2012].

[1] The Velvet Revolution (Czech: *sametová revoluce*) or Gentle Revolution (Slovak: *nežná revolúcia*) was a non-violent revolution in Czechoslovakia that took place from November 17 to December 29, 1989. Dominated by student and other popular demonstrations against the one-party government of the Communist Party of Czechoslovakia, it saw to the collapse of the party's control of the country, and the subsequent conversion to capitalism (Wikipedia).

A funder repository of heterogeneous grey literature material with advanced user interface and presentation features

Ioanna-Ourania Stathopoulou, Nikos Houssos, Panagiotis Stathopoulos, Despina Hardouveli, Alexandra Roubani, Ioanna Sarantopoulou, Alexandros Soumplis, Chrysostomos Nanakos
National Documentation Centre; National Hellenic Research Foundation, Greece

1. Introduction

The present contribution concerns the development of a funder repository aiming at the dissemination, reuse and preservation of mainly grey literature material of diverse types. This material which was produced under the auspices of large scale (multi-billion Euros) funding programmes of the Hellenic Ministry of Education, co-financed by the European Union. The project involved the handling of a wide range of content, like studies, reports, educational material, videos, theses, material from a range of conferences/events. The project has been successfully completed and the system is publicly available since spring 2011 at http://repository.edulll.gr.

In this repository creation use case, technical enhancements were used to provide the means to organise, process and import to the repository a wide range of heterogeneous material. Special facilities for the support of these workflows was included, along with enhanced user viewing capabilities for metadata, PDF files and video content, providing an end user experience suitable for the repositoty's users, which include the educational community, researchers and scientists and the general public The metadata schema used is an application profile using elements for Qualified Dublin core, LOM and PREMIS. The repository has been implemented using the DSpace platform, with significant extensions developed specifically for this project. Futhermore a supplement external lightweight CRIS system has being developed in internal beta version, in order to appropriately represent semantically rich information.

The rest of the article at first explains the overall project scope and the followed methodology and approach, including important issues encountered and relevant decisions that had to be made. Implementations aspects, particularly regarding metadata schema and workflows are presented in Section 3. Advanced repository features are described in Section 4 and the paper concludes with summary, lessons learnt and ideas for future work.

2. Project scope, methodology and approach

Since 1994, the Hellenic Ministry of Education has been executing a series of large scale (at the range of billion of Euros) funding programmes, co-financed by the European Union through Structural Funds, under the names Operational Programme for Education and Initial Vocational Training (EPEAEK I & II, 1994-1999 and 2000-2006, respectively) and Operational Programme for Education and Lifelong Learning (2007-2013). The entity responsible for the management of the programmes is a Special Managing Authority within the Ministry (www.edulll.gr). The programmes finance a variety of activities related to all level of education (i.e. K-12, Higher Education and Lifelong Learning) including, among many others, restructuring and enhancement of educational programmes, production of educational material, scholarships and fellowships for doctoral students and post-doctoral researchers, collaborative research projects with the participation of Greek Higher Education Institutions, as well as a wide range of studies and reports, conferences and events, seminars, etc.

In autumn 2010, the respective Special Managing Authority assigned to the Hellenic National Documentation Centre the task of organizing the digital material that has been produced in the course of the aforementioned funded activities into a publicly available repository that would provide a single point of access to the programmes results, ensure wide dissemination and reuse of content through enforcement of interoperability standards and also address certain long-term preservation aspects.

The input material available from the Special Authority had the form of digital files, arranged in folders mainly according to project, however without any metadata attached to them. A considerable portion of these primary sources concerned administrative reporting during project execution and related files (e.g. administrative reports, detailed schedules of courses and seminars, participants' lists for seminars / events, etc.) It was decided to ignore this kind of material so that the repository includes only items that contain content of archival value. It is worth noting that during this first-phase project also a lot of the digital material produced during these programmes was not included, simply due to non-availability in digital form.

Processing of digital files before publication via the repository was a challenging issue due to the inherent heterogeneity of the material and the non-uniformity in formats, naming and structure. A considerable part of the overall project concerned this aspect. An important decision was to store and preserve both initial and processed digital files within the repository so that potential problems in the sometimes non-trivial processing workflow could be corrected a posteriori. Notably, providing perpetual persistent access to the the digital material and handling certain preservation issues was an important goal of the project; among the main reasons for that was the fact that a significant part (probably the majority) of the content concerned grey literature that would not be handled through official publishing and preservation channels.

Regarding metadata schema, a custom application profile was created, based on existing common international standards. This has been the result of an iterative process. The first phase consisted of a detailed mapping of the input digital material during which the most common types of items were identified and candidate fields per type were defined. This activity was necessary due to the diversity of the material and the limited available information initially on types and formats of content. It has been a tedious – despite assisted to some extent by automated tools – process of going through the directories/files and capturing types of items, formats and doing an initial distinction of content candidate for inclusion to the repository and administrative documents. The first phase produced a first version of the metadata schema that was to a large extent stable for the most common document types and enabled the beginning of the cataloguing process early in the project, followed by subsequent phases of refinement. Further details on the metadata schema are provided in paragraph 0.

Certain important cataloguing decisions were made early on. First, there was a considerable allocation of effort to subject cataloguing and appropriate user interfaces for browsing by subject (see paragraphs 0 and 0), since the diversity of the material would make it inadequate to rely only on navigation/browsing based on fields like keywords, type, project or collection. Secondly the idea of self-archiving was abandoned, at least for this first phase project and metadata entry was assigned to experienced information scientists that could safely distinguish among archival and administrative material, identify and remove duplicates (probably scattered across different folders of primary material), perform subject cataloguing, produce abstracts of items (where appropriate) and provide clear guidelines to the digital files processing staff (e.g. for file merging/splitting).

Furthermore, the heterogeneity of the material created a requirement of various ways of browsing to be applied (e.g. not only listings but also tag clouds for presenting subjects and keywords; image-based browsing of material types) and also accessing the digital material itself (e.g. both the ability to download a PDF file and to stream it via an online reader; video material embedded into item pages).

3. Implementation

3.1 Metadata schema

Due to the heterogeneity of the material to be included in the repository, a custom application profile has been developed, utilizing elements from various metadata standards, in particular Dublin Core, Learning Object Metadata (LOM) and PREservation Metadata Implementation Strategy (PREMIS). Furthermore, a new EKT namespace/schema was created to capture custom elements that were required for our implementation but were not available in standard schemata.

The EuroVoc thesaurus [1] has been utilized for the thematic indexing of the material, as well as for spatial coverage. This particular vocabulary was selected due to the following properties:

(a) Breadth and depth of coverage and inter-disciplinarity. EuroVoc is a detailed thesaurus across scientific disciplines and domains, so it is perfectly suitable for the diverse material of our case.

(b) Multi-linguality. EuroVoc is available in 24 European languages. Therefore, the thematic index can be made automatically available in all these languages.

Furthermore, regarding education levels, the controlled vocabulary of the Eurydice network [2] was adopted.

3.2 Content submission and processing workflow

As mentioned above, the content of the repository includes diverse material in terms of their format and type, therefore it was necessary to implement a workflow which would include the required processing of the digital data depending on their type and format.

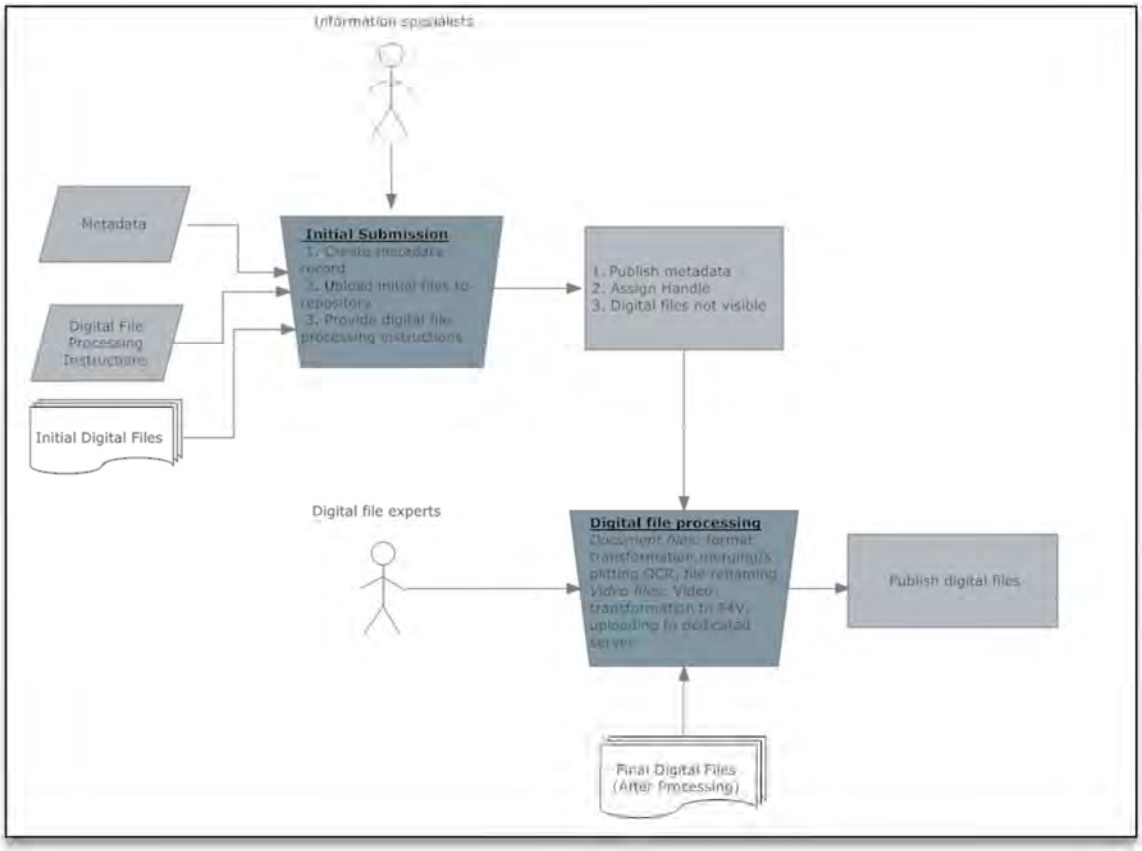

Information specialists
Metadata
Digital File Processing Instructions
Initial Digital Files
Initial Submission
1. Create metadata record
2. Upload initial files to repository
3. Provide digital file processing instructions
1. Publish metadata
2. Assign Handle
3. Digital files not visible
Digital file experts
Digital file processing
Document files: format transformation,merging/splitting OCR, file renaming
Video files: Video transformation to M4V, uploading to dedicated server
Publish digital files
Final Digital Files (After Processing)

Figure 3 Content submission and processing workflow

The workflow of the repository consists of the following steps, graphically depicted in the aforementioned Figure 3:

1. **Create metadata record:** Firstly, the item is submitted to the repository using an appropriate submit form. This step is performed by authorized users, specialized in the organization and management of information materials (in our case, information specialists who belong to the library staff of EKT/NHRF). In addition to providing the necessary metadata, the information specialists also upload the input digital files and provide, in separate fields and in a custom specification language, the necessary instructions for the processing of the digital file(s) by the digital material processing staff. Specifically:
   a. ***Digital Processing Instructions:*** necessary instructions for the processing of the digital file(s), such as merging sequence for multiple files, renaming instructions, etc.
   b. ***Managing metadata:*** Information regarding the filepath of the specific files (in the material provided from the Special Managing Authority within the Ministry of Education), which can be used as a reference of the cataloguing process status.
2. **Handle assigned to record:** Upon submission by information specialists, the item metadata record is published online and a persistent identifier is assigned to it. At this stage, the digital files are already within the repository, but not visible and available only to authorized personnel.
3. **Digital file processing:** The processing of the digital files from the respective expert staff takes place. It follows the directives of the information specialists for the production of final digital items for publication of each record. In case of text material, the final digital item must comply with the following rules:
   i. The final document must be a full-text indexable and searchable pdf file. Optical character recognition (OCR) might be needed to achieve that, for example when the input file(s) are in pdf format, and they are either image pdfs or have improper encoding or font support (e.g. for Greek content).
   ii. If specified by the information specialist, the final document(s) may be the outcome of merging or splitting the input data files.
   iii. The name of the final document should follow a particular pattern, which includes the item's handle in the repository.

In the case of video material, it has been selected to be hosted by an external FLV streaming video server and integrated to the repository using enhanced FLV players. Video has been received in various formats and the workflow is as follows:

i. The video is encoded in H.264 format and the files are saved in the standard file format.
ii. If specified by the cataloguer and/or considered necessary by the digital material staff, some video editing or enhancement takes place.
iii. The video is uploaded to a the dedicated video streaming server (WOOWZA) and the RTMP link is stored in a separate field in the repository.
iv. The embedded video player is generated dynamically according to the stored RTMP link, which includes also time related information, i.e. start and stop of each video segment corresponding to a serarate metadata entry.

4. **Publication of digital files:** After the processing of digital files the final files become publicly available. Furthermore, the respective staff member records in an appropriate metadata field the fact that the processing has been completed.

4. Advanced user interaction and presentation features

Several advanced user interaction and presentation features, not typically found in repositories, have been implemented to improve the user experience.

The DSpace platform was greatly customized and enhanced in order to accommodate the requirements of this specific digital repository implementation. Furthermore a number of additional software and infrastructure elements, namely video server, image backend server processing and delivery for achieving page by page reading of the digital items were developed. A number of the aforementioned DSpace extensions have been incorporated to the DSpace 3.0 Release. The main issues that have been addressed include modifications in the submission form in order to simplify the process, supporting multi-lingual controlled vocabularies both in metadata entry and presentation to end user and context-sensitive presentation of material as well as a range of modifications in order to encourage efficient and effective human-computer interactions and optimize the overall user experience in the repository.

4.1 Submission form modifications

The default submission process in DSpace platform comprises quite a few steps (collection selection, metadata entry, digital file upload, veriy and grant copyright license). In order to somewhat simplify the process, we changed the submission form so as the user would be able to provide the item metadata and upload one or more digital files in the same step, in a single web page/form. Thus, the submission process is more simple and straightforward, as it consists of three steps: (1) collection selection, (2) describe (metadata entry including copyright, upload one or more files), (3) verify.

Moreover, the following new input-types have been developed: *'advancedName', 'refinementdrop_advancedName', 'refinementdrop_value', 'multiTextArea', 'onebox_lang', 'textarea_lang', 'startHeading' and 'endHeading'*. These input-types include capabilities like dynamic processing of persons' names which may be entered in various patterns and storing the value as expected by the repository (e.g. *{Surname, First and Middle Names}*), further refine the values by using controlled vocabularies, handle multiple inputs from the same textarea, or allow the user to choose the language of each metadata value.

Finally, in order to improve the user experience and make the submission easier, the submission process was modified in order to be fully localised and support different languages. Specifically, all the labels and control vocabularies in submission form can be configured from the message properties and input-forms configuration file and be displayed in different languages. The item presentation page, also uses these configuration files in order to display the stored control vocabulary in the user's respective language. Also, help tips were implemented which are shown when the user clicks on one of the submission form fields and provide context-sensitive help.

4.2 Visual representation of browsing

The repository browsing functionality was greatly enhanced by applying visual representation mechanisms. The visitor can browse through the subject classification metadata using a tag cloud which provides a quick perception of the most frequently used EuroVoc terms. The tag cloud is fully parameterisable. Specifically, parameterisation concerns how the tag cloud is displayed (e.g. maximum number of display terms, order of appearance rules for terms, minimum required occurences of a term for showing up in the tag cloud, colour and size patterns for tags, etc.) and selecting which metadata fields this browsing should be applied on. Furthermore, a custom, more user-friendly browsing by type has been implemented using a characteristic image per type, in addition to string labels.

The repository homepage, which includes visual representation mechanisms for browsing has been implemented as an extension to DSpace Community. This extension allows the creation of a tag cloud in any repository page for any metadata field (i.e. subject, keyword). The tag cloud is fully configurable via the DSpace configuration files. The tag cloud is implemented as a custom Java Server Pages (JSP) tag, allowing it to be easily included in any page of the DSpace repository besides the homepage. An example fragment of a configuration file is shown below.

```
#######################################################
# TAG CLOUD configuration #
#######################################################
#
# Should display tag cloud in the home page?
# Possible values: true | false
webui.tagcloud.home.show = true
#
# Select the browse index to create a tag cloud for in the home page
# Possible values: any of the browse indices declared earlier in this conf file
webui.tagcloud.home.bindex = subject
#
# Select the total tags to show
# Possible values: any integer from 1 to infinity
webui.tagcloud.home.maxtags = 50
#
# Should display the score next to each tag?
# Possible values: true | false
webui.tagcloud.home.showscore = false
#
# The score that tags with lower than that will not appear in the rag cloud
# Possible values: any integer from 1 to infinity
webui.tagcloud.home.cutlevel = 5
#
# Should display the tag as center aligned in the page or left aligned?
# Possible values: true | false
webui.tagcloud.home.showcenter = true
#
# The font size (in em) for the tag with the lowest score
# Possible values: any decimal
webui.tagcloud.home.fontfrom = 1.3
#
# The font size (in em) for the tag with the highest score
# Possible values: any decimal
webui.tagcloud.home.fontto = 2.8
#
# The case of the tags
# Possible values: Case.LOWER | Case.UPPER | Case.CAPITALIZATION | Case.PRESERVE_CASE |
Case.CASE_SENSITIVE
webui.tagcloud.home.tagcase = Case.PRESERVE_CASE
#
# If the 3 colors of the tag cloud should be independent of score (random=yes) or based on the score
# Possible values: true | false
webui.tagcloud.home.randomcolors = true
#
# The ordering of the tags (based either on the name or the score of the tag)
# Possible values: Tag.NameComparatorAsc | Tag.NameComparatorDesc | Tag.ScoreComparatorAsc |
Tag.ScoreComparatorDesc
webui.tagcloud.home.tagorder = Tag.NameComparatorAsc
#
# The first color of the tags
# Possible values: hex value of the color (i.e. e3d67a)
webui.tagcloud.home.tagcolor1 = D96C27
#
# The second color of the tags
# Possible values: hex value of the color (i.e. e3d67a)
webui.tagcloud.home.tagcolor2 = 424242
#
# The third color of the tags
# Possible values: hex value of the color (i.e. e3d67a)
webui.tagcloud.home.tagcolor3 = 818183
```

4.3 Handling and presenting Controlled Vocabularies

A major feature, needed to accommodate the requirements of this specific digital repository implementation, concerned the support of user-friendly presentation and search for multi-lingual controlled vocabularies. Users need to view and search terms of controlled vocabularies in human-friendly forms and terms need to be available in different languages. At the system level, however, it is sometimes preferable to store encoded values for terms, for the sake of guaranteed unique identification and clarity. For example, in a controlled vocabulary for the “language” metadata field, it is better for the system to store for a language a standard encoded value according to the ISO 693-2 specification (e.g. “eng” for English, “gre” for Greek), while users wish to view the corresponding values as “English” or “Greek” in item metadata pages (or respectively, “Englisch” or “Griechisch” for German users) and get appropriate results both when searching the repository using the queries “English”, “Englisch” or “eng” for the language metadata field.

This functionality has been implemented in the presented repository in which a search for a controlled vocabulary term produces the same set of 1178 results either by searching for items with the English term “Greek” or the respective Greek term “Ελληνικά”, while in the repository database the respective value (“gre”) stored is the code name for the Greek language in the ISO 639-2/B standard.

To achieve this, enhancements to the DSpace repository platform have been necessary. During the submission process, DSpace supported the entry of controlled vocabularies having for each vocabulary term a separate “displayed-value”, a user-friendly one (e.g. “English”), and a respective "stored-value" which is stored in the database (e.g. “eng”). Nevertheless, the only use of a “displayed-value” is in drop-down lists in the submission form; this value would not appear in item metadata pages and would not be indexed for search.

Thus, improvements were applied in order to show the displayed values in the item page and index both stored and displayed values for search. Both extensions are fully configurable via the DSpace configuration files. In particular, the repository administrator can specify if it is desired to show the displayed-value in item page or if DSpace should index both the "stored-value" and the “displayed-value”. Both functionalities have been contributed to the DSpace open source platform and have been incorporated in the core DSpace platform, since version 3.0 [3][4].

4.4 Video integration and external video server system

As already stated, video content forms a considerable part of provided content in the presented repository, whose material includes full length documentaries, short stories for educational use, online courses, etc. Various file formats and encoders were used in the content, from DVD MPEG2 VOD files to proprietary video files and encoding formats (H.263, MPEG-1, MPEG-4, DIVX). On the other hand a unified, meaningfull and intuitive experience should be provided to the end user when accessing this content. Relevant requirements were:

- Easy to use and intuitive user interface for streaming video.
- The option for downloading the video file should not be given, due to IPR issues.
- Integration of a web based video player to the repository and not an external video player program.
- Based on open file standards and codecs, if possible.

In order to cope with the various file formats and encoders the following decision were made:

- Content was batch transcoded to H.264 enconding using a F4V video file container, selected based on the encoder properties and its open nature.
- The video stream was provided by an appropriate video server external to the core repository system. The Woowza video server was selected based on means of versatility in the protocols used (RTMP and RTSP) and efficiency.
- The flowplayer embedded video player was selected based on features and configuration options. The flowplayer configuration is dynamically generated based on the video repository metadata. E.g. in case of a single video stream, with different repository entries, e.g. a DVD with multiple episodes, it was decided that instead of splitting the video file, rich metadata, including episodes start and end time would be included. This approach enables the capability to easily rearrange the video player configuration, be independent of particular video player and streaming servers and provide future enhancements, according to available metadata.

Relevant examples are available at repository.edulll.gr on the video content category, demonstrating the dynamically generated embedded video player according to the video’s repository metadata.

4.5 Providing an ebook-like experience as an alternative to PDF viewer

The initial format of most of the material imported to the repository was PDF files. While PDF is a widespread format, a large percentage of the files included rich photographic material, multihundred pages books, or has been OCR processed, thus file sizes of tens, or even hundreds, of megabytes were common. Therefore an alternative was sought in order to not only make shorter the time required for opening the document, but also to provide a more intuitive user experience.

Initiatives such as the “Google Books” and “Google Art” project, the Internet Archive “Open Library” and advanced repository systems, e.g. Islandora, etc., have paved the way for novel online reading capabilities and experience, with features such as “page by page” reading of electronic resources and tile-based image viewing systems, exploiting advanced codecs such as JP2000.

In order to provide such a viewing experience the following backend components were integrated with DSpace:

- The online reader, based on the open source Internet Archive Open Library BookReader provides the end user interface. It features advanced end user capabilities such as interactive online reading with zooming, thumbnail view, full-text search and hit highlighting.

- A two tier backend for processing, generating and delivering image files, comprising:
  - A distributed multithreaded conversion management tool, in order to interface with DSpace and manage the batch conversion process from PDFs to page by page JP2000 files which wraps around existing standalone converters. This tool can batch convert selected PDF content from a DSpace repository to JP2000, with the conversion process transparently being initiated from the repository manager. It is available as open source software [5] and has been already been integrated also with Open Journal Systems e-publishing platform [6].
  - The highly scalable image delivery backend, based on the The Djatoka image server for providing the page by page content in openurl format and in various sizes. It comprises a Nginx/Varnish load balancer/caching frontend, connected to a 2 node tomcat/djatoka cluster processing on the fly JPEG2000 over a shared SAN storage and utilising a 3 node Postgres 9 cluster for handling file locations and file technical metadata.

    The overall system architecture is depicted in Figure 2.

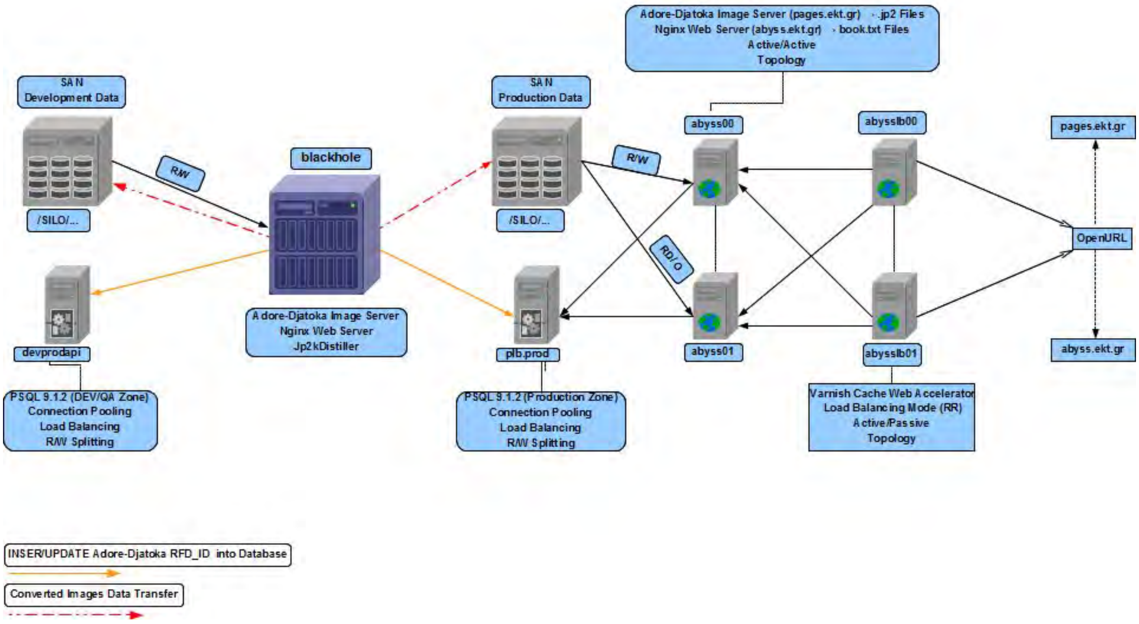


Figure 2 Architecture of the back-end image server infrastructure

The system provides a more natural way to open and view PDF based content, especially suitable for the case of multi-MB PDF files and tablet devices.

An architecture involving a CERIF/CRIS system

The metadata that is available for the digital material (e.g. documents, data sets, multimedia material) is to a considerable extent contextual; thus, it represents the context in which the digital content has been been created, including entities such as persons (e.g. authors, editors), funded projects and the corresponding funding programmes, organizations (e.g. funders, author organisations, project coordinators or supervisors). The context information comprises also the relationships among these entities for example project-person, project-organisation, publication-person, publication-organisation, and many others. Contextual metadata is extremely useful in the case of grey objects, where, compared with white literature, there is no publication venue (e.g. a journal or conference) that is known to the reader and provides some level of reliability guarantees for the content. Therefore, it is even more useful to become aware of provenance and other information providing answers to questions like “is this document produced by a reliable source – e.g. authored or supervised by a reputable person and/or organization?”, “is this document the result of a project in a well-known funding programme?”.

Flat metadata schemata commonly used in digital repositories are not since they do not represent contextual entities as first-class citizens and do not have the capabilities to represent rich semantics on relationships among entities. This kind of metadata is handled very well by Current Research Information Systems (CRIS) and the dominant metadata standard in this area, namely the Common European Research Information Format (CERIF) [7], whose suitability for grey literature has been well documented in other work [8][9].

Therefore, to appropriately represent contextual metadata, a CERIF/CRIS has been developed running as a separate system than the repository. It is being used to represent in CERIF metadata about persons, projects, funding programmes, organisations and links among those entities and grey objects. The relationship semantics are represented using the CERIF Semantic Layer [7]. Example relationships that need to be represented are project-organisation (indicative roles: funder, supervisor, contractor), person-project (indicative roles: coordinator, team leader, team member, evaluator), person-grey_object (author, numerous types of contributors) and recursive relationships, for example part-of or derived-from relationships among grey objects.

The system is currently being used as a private system that forms an authoritative source of the available contextual metadata, avoiding ambiguity and information loss. Public access is not provided to the CERIF/CRIS, due to considerations originating from the commissioning organization; however, the CERIF/CRIS is very valuable in ensuring the integrity of information that is provided publicly.

Conclusions - future work

In this repository development use case, technical enhancements were used to provide the means to organise, process and import to the repository a wide range of heterogeneous material. Special facilities for the support of these workflows was included, along with enhanced user viewing capabilities for metadata, PDF files and video content, providing an end user experience suitable for the repositoty's users, which include the educational community, researchers and scientists and the general public. Future work, possibly in the frame of a sequel project, includes the inclusion of further material from past funded projects, and the integration of a workflow for the quick inclusion of content produced by currently running projects.

References

[1] EuroVoc thesaurus, http://eurovoc.europa.eu.

[2] The Eurydice Network, http://eacea.ec.europa.eu/education/eurydice/index_en.php.

[3] Show display values for controlled vocabularies in Item Page, DSpace feature request DS-1125 (https://jira.duraspace.org/browse/DS-1225) and pull request #54 (https://github.com/DSpace/DSpace/pull/54).

[4] Index (for search) both display and stored values for fields using controlled vocabularies, DSpace feature request DS-1231 available at https://jira.duraspace.org/browse/DS-1231 and pull request #55 (https://github.com/DSpace/DSpace/pull/55).

[5] dPool Elastic Cluster - PDF/TIFF to JP2000 Distiller, Open source project available on Google Code, http://code.google.com/p/dpool/.

[6] Stathopoulos, P., Houssos, N., Stathopoulou, R., Stavrou, G., & Soumplis, A. (2011). Enhancing OJS journals with advanced online reading and viewing capabilities. Presented at the PKP Scholarly Publishing Conference 2011, http://pkp.sfu.ca/ocs/pkp/index.php/pkp2011/pkp2011/paper/view/319.

[7] Jörg, B. (2010). CERIF: The common European research information format model. Data Science Journal, Volume 9, pp. 24-31.

[8] Jeffery, KG, Asserson, A. (2009). Mosaic: Shades of Grey. Conference Proceedings on Grey Literature: GL-conference series, ISSN 1386-2316 ; No. 11.

[9] Jeffery, KG, Asserson, A. (2011). GL Transparency: Through a Glass Clearly. The Grey Journal (TGJ), ISSN 1574-1796, Volume 7, Number 2.

Working for an open e-publishing service to improve grey literature editorial quality

Rosa Di Cesare and Marianna Nobile,
Institute for Research on Population and Social Policies, IRPPS
Silvia Giannini, Institute of Information Science and Technology, ISTI, Italy

Abstract
A survey of in-house publishing practices at CNR Institutes is described. Fourth categories are introduced to measure the level of innovation in the management of in-house publications in order to identify the business model used by each CNR Institutes to manage their editorial products, especially digital products. Data used for this descriptive and quality study were obtained from CNR Institute websites.

1. Introduction

The widespread adoption and diffusion of the Internet as well as the increasing application of digital publishing technologies have modified and streamlined functions, processes and products in the scholarly communication chain. The consequences of the application of ICT in context of scientific research and education has been analyzed in many studies. Starting from the first contributions (Roosendaal 1997, 2004; Gierveld 2002) that were focused on the transformation of the linear scientific information chain, through to more recent studies on the changes in scholarly communication focused on e-science collaboration and information and data sharing (Borgman 2007; Tenopir et al. 2012), many analyses are now focused on added—value services embedded in the digital publishing technologies. The latter particularly point out the key role played by academic and research libraries in challenging the traditional business publishing model on in favour of a more sustainable economic model to produce and diffuse scholarly research outputs. Many of these studies refer to e-publishing library services development which *is rapidly becoming a norm for research libraries, particularly journal publishing services* (Mullins et. al. 2011; Iglezakis et al. 2011).

Moreover, the widespread diffusion of electronic publishing technologies is increasing digitization initiatives addressed to traditional libraries' print collections, which involve also Grey literature collections. This means that digital publishing technologies are creating a "second life" for traditional and valuable grey documents and this innovative way of managing grey contents can improve editorial quality, as well as, their diffusion and discovery.

Several studies and reports have demonstrated that moving from print to a digital publishing model reduces production costs through rationalization and automation of editorial procedures (Houghton 2011;Willinsky 2011; Crow et al. 2009). Within this context studies on innovation emphasise the combination of electronic publishing and open access as represent the major drivers of editorial changes.

2. Research hypothesis and aims

Generally an editorial activity is a component of scholarly communication that can be managed as an autonomous activity and/or integrated with the management activities of the scientific production of an academic or research institution.

These activities can be carried out in-house and/or in collaboration with commercial publishers.

Our basic premise is that the editorial activity depends on the organizational-productive context while the number and type of editorial products depends on the disciplinary area. Moreover it could be carried out in an innovative way in terms of process and/or products, to make the editorial process more efficient and effective.

The National Research Council (CNR), is one of the biggest Italian multidisciplinary research institutions and comprises a network of 109 Institutes geographically distributed Institutes, which have scientific and organizational autonomy. One of the institutional missions of CNR is the diffusion of scientific information in Italy and its editorial products reflect this institutional mission. At local level too, CNR institutes have always created their editorial products strictly connected with their studies and research interests, tailored to different target of users (general public and/or their research community), for their production and diffusion they have used and still use conventional and non conventional channels. Of course this activity may vary from Institute to Institute.

The survey analyses in-house production available at CNR Institute web sites to identify: a) type of products and use of bibliographic elements; b) technology used to manage in-house products and c) degree of innovation concerning the management of the contents; and finally d) access and discovery of

products. In other words the analysis concerns the description of the business model used in the editorial production and the policy adopted for the diffusion of its contents.
As the framework of our analysis are new publication models strictly connected with open access movement, the general aim of the survey is to obtain information on current publishing practices at CNR institutes as a means to improve editorial quality of in-house scientific publications and increase visibility of CNR scientific products.

3. Materials and Methods

3.1. Survey design

The object of our analysis is CNR institutes' editorial products published in-house and/or in collaboration with commercial publishers.

The survey was divided into two phases.
In the first phase, we checked each CNR Institute website to gather preliminary information about in-house production, identifying current and ceased editorial products directly produced and managed by CNR Institutes, in a stable standardized way and with continuity. Then we classified Institutes according to a set of criteria that measures the level of innovation in the management of their editorial products (see box).

I (*Innovative*) Institutes that manage in-house-publications applying an editorial control that includes at least an identifiable standardized series title and numbers; and using electronic-publishing systems to manage editorial processes.
A (*Traditional*) Institutes that manage in-house-publications applying an editorial control that includes at least an identifiable standardized series title and numbers.
Z (*No editorial control*). Institutes that produce GL without applying any editorial control or where we could not find any information on their production on their website
X (*No in-house publications).* Institutes that do not produce GL at all and/or produce it sporadically - <5 per year-.

In the second phase, we have chosen, among in-house publications, the editorial products, with a minimal set of editorial and bibliographic elements (i.e.: series title and/or number) produced by the Institutes belonging to Letter I and A, and we have analysed their products in terms of:

- Type of products and their publication frequency
- Type of production/diffusion
- Technology used
- Access policy

Data was collected from CNR Institute Websites, between June and September 2012.
For those websites with scarse information about their in-house production, we also checked the CNR central archive that collects both conventional and non conventional literature, and also includes the editorial products produced and managed by CNR institutes.
In addition we decided to conduct informal phone interviews with the manager of the CNR Institute library, to obtain further and more detailed information, in particular on the procedure used to manage the editorial products, critical issues connected with the editorial activity, such as budget and human resources as well as future editorial plans. The phone interview further considered the characteristics of the publishing systems used, focusing in particular on the technical requirements for locally developed systems.

2.2. The business model

An important part of our analysis aimed to identify the business model used to manage in-house publications by CNR Institutes that fall in the category I, namely the Institutes which have introduced changes to the editorial process and/or created a new end product, in other words those Institutes that make the editorial process more efficient. As mentioned before, digital publishing technologies have in addition led to many changes in the core functions of the libraries and publishers, as well as of scholars, affecting the general flow of an editorial processes.
In our view the main steps that are foreseen in an editorial process encompassing different phases.

Fig. 1. *Business process model*

Analysis of the business process

NO Linear process:
Depending on the business model adopted
on the organisational framework
on the type of products
on the access policies
on the evaluation strategies
on the technology used
…..

It starts with the production of content, that is the acquisition of contents that are going to be published. These contents can be managed as an internal or external activity and can be automated or not. The contents may be subject to peer review or other reviewing systems or not. Then there is the phase of copy-editing where the manuscript is submitted to improve editorial quality and ensure the content's bibliographic and textual style, while the proof reading checking for typos and layout transforms the editorial product for publication and distribution
Most of these activities can be carried out in-house or outsourced to external services providers or also to commercial publishers.
Along with the aforementioned issues the business model also depends on the policies adopted fo access to contents, ranging from subscription fees to full OA, but also on the products which can be peer reviewed or not and, of course, on the technology used. In addition the editorial process can be influenced by the organizational framework and human resources available, or by the type of products to be produced. (Journals, monographs or reports are very different in terms of cost. The cost is higher for a monograph compared to for example that of a reports.

To identify the business model we checked for each editorial products the type of business process carried out: a) how the Institute manages products in terms of production & diffusion and/or distribution and which phases were externally managed; b) if the Institute has introduced some changes to the editorial process management (we checked whether the contents of the editorial product were not just a version of the printed one) and, c) if they make use of an online content management and publishing platform or an electronic handling manuscript system (content management system).

4. Results

Table 1 shows the results of the first phase of the analysis that aims to classify the level of innovation of CNR Institutes in the management of in-house production.

Table 1. *Number of CNR Institutes by Department according to grouping criterion (see box**)*

DEPARTMENTS	*Number of Instistutes*	***I***	***A***	***Z***	***X***
Earth & Environment	13	2	3	8	0
Agricullture & Food	10	1	2	7	0
Biomedical Sciences	17	0	0	7	10
Chemistry & Materials Techn. Sciences	14	0	0	7	7
Physics Sciences	14	0	2	5	7
Engineering & ICT	21	1	7	12	1
Social Sciences & Humanities	20	6	8	5	1
Total	**109**	**10**	**22**	**51**	**26**

The majority Institutes that have a variety of in-house production, managed in different ways (ranging from the most innovative to traditional) are concentrated in the Departments of Earth and Environment, Agriculture and Food, Engineering and Social Sciences and Humanities.

Conversely, institutes belonging to the Department of Biomedical Sciences and Chemistry don't either produce GL or carry out any in-house editorial activity to manage and diffuse their products. They obviously use traditional channels to diffuse their research results.

However, as we can see in the table, there are many Institutes that are classified in category X (26), these are Institutes which produce GL sporadically and in a very limited number (> 5 x year).

The Institutes classified in category Z (51) produce a lot of GL documents, that are not organized in well-established, standardized series. This is also confirmed in a previous survey (Di Cesare, 2010). These are GL documents produced "ad hoc", such as project deliverables, conference proceedings and so on and this is the case especially for institutes belonging to the Engineering and Information Communication and Technologies Department (ICT).

They were excluded from the analysis because their products lacked the minimal set of editorial and bibliographic elements.

Further to the recent signing of the Berlin Declaration by the CNR, following the development of the CNR Institutional Repository (IR), a working group was established to elaborate specific guidelines for quality and metadata control of the CNR researchers' output, including current and back GL collections. So we hope in the near future to have more suitable procedure in the management of in-house publications, together with consistent editorial policies for all CNR Institutes.

4.1.CNR Institutes editorial products

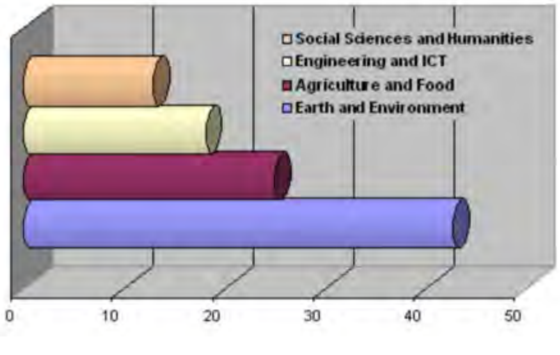
Table 2 shows the number of traditional and digital products broken by CNR Department. We found 106 editorial products with the minimal editorial set (i.e.: series title and/or number, corresponding to inclusion criteria) to be included in the analysis. 106 out of 19 – equal to 18% - are digital products.

Table 2. *Number of editorial products by Department*

Department	***Number of editorial products***	***Traditional product***	***Digital product***
Earth & Environment	14	8	6
Agriculture & Food	8	6	2
Biomedical Sciences	1	1	0
Chemistry & Materials Techn. Sciences	0	0	0
Physics Sciences	2	2	0
Engineering & ICT	11	9	2
Social Sciences & Humanities	70	61	9
Total	***106***	***87***	***19***

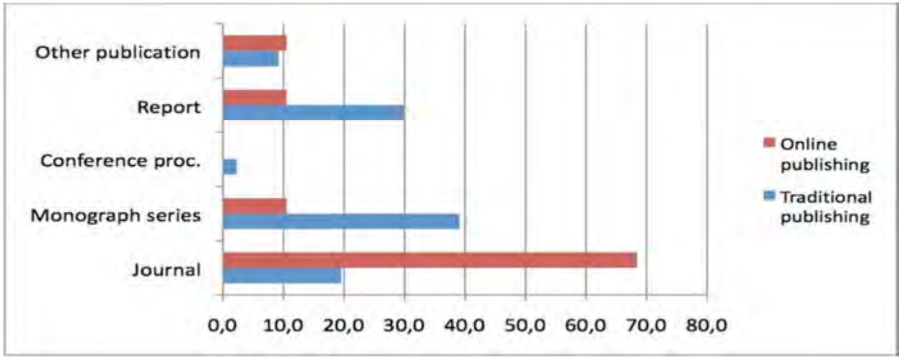
There is an evident concentration of traditional products in the Department of Humanities and Social Sciences, with 70 out of 106 and are predominantly only in print version. The digital products represent a small proportion (fig. 2) of the editorial products (17%). The Department of Earth and Environment accounts for 42% of digital products, a high value considering the total number of products produced and compared to the 12% of the Department of Humanities and Social Sciences.

Fig. 2. *Digital products by department (%)*



As expected (fig. 3) we found that the journals are the majority of document types that are published online. This is not surprising because usually the Journal product is the first type of document that was supported by electronic manuscript handling system and has now shifted to online publishing. In our sample, journals account for 13 out of 19 – equal to almost 70%, - 2 of which are digital-born and open access. Both are journals produced by the Department of Humanities and Social Sciences.

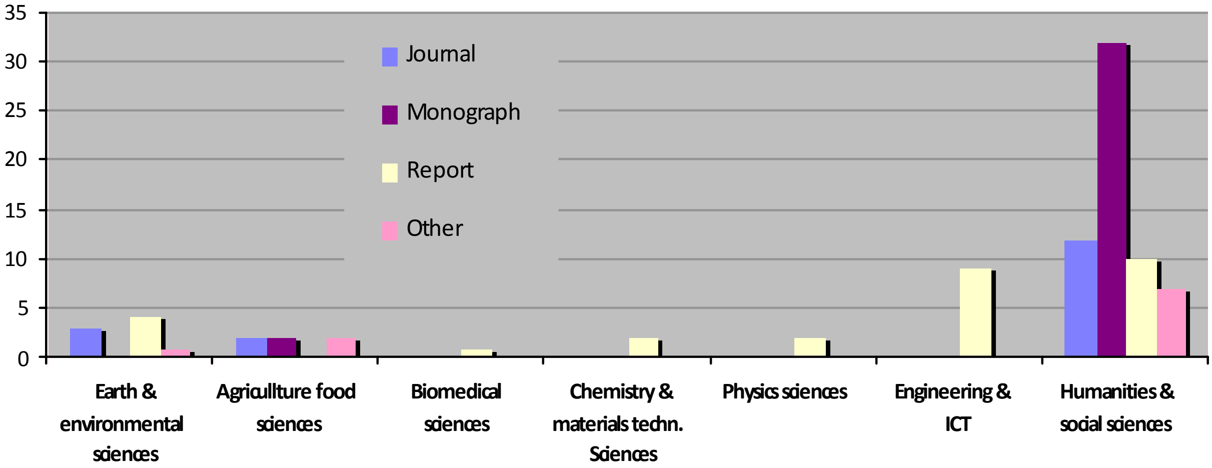
Fig. 3. *Traditional and online publishing by document type*



4.2 Traditional and online publishing

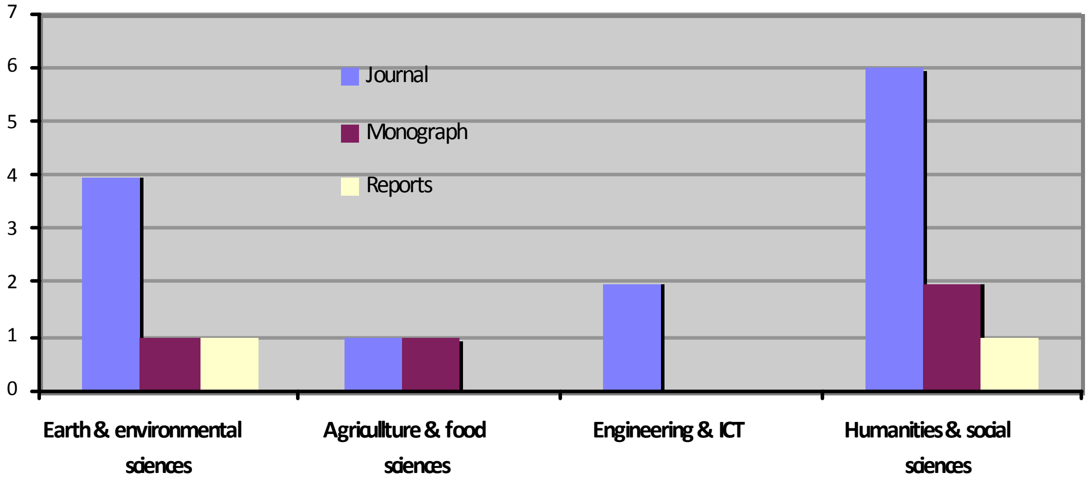
Figure 4 shows traditional editorial products produced by CNR Departments. Many of them are established and valuable collections and have been published without any interruption since the foundation of the Institute. They are products that are typical output of a scientific community, including journals, monograph series and reports. The Department of Humanities & Social Sciences produces the whole range of products, especially Monographs that are well known in the international scientific community for their high-profile. While reports are almost exclusively in-house productions of other Departments and equally quantitatively relevant.

Fig. 4. *Traditional editorial products by Department*



The figure below shows the digital products managed by electronic publishing systems. They are, as already mentioned, mainly Journals that more frequently shifted from printed to digital formats and adopted proprietary or open source systems to support the whole publishing process taking advantage of flexibility, easy reusability of the content and cost saving. The majority of online publishing journals are concentrated between the Department of Earth & Environment and of the Department of Humanities and Social Sciences, while lower numbers are found respectively for the Departments of Engineering and Information Communication Technology and only a single example for that of Agriculture & Food.

Fig. 5. *Digital editorial products*



Clear differences in the digital publishing can be noted for the other kind of documents such as reports and monographs. Of course we think that these documents too can benefits from the advantage of electronic publishing systems[1]. (Costigan 1999, 2004; Crow 2009).

As regard the reports series, there are not many of them that benefit from electronic publishing systems. The general practice is that "high quality printed report series" have been gradually replaced by digital formats, and sometimes back issues have been digitized and made freely available from the institutes' website. This conversion from paper to electronic formats has not changed the production process, even if there is a clear advantage for end users. However, the use of electronic publishing systems to manage also the production and distribution of reports could represent a step forward in terms of editorial quality and above it can introduce additional services such as peer review, indexing and abstracting in general and specialized search engines facilitating the web discovery.

In this context, it is well known that the CNR central library collects and manages valuable reports produced over time by CNR research Institutes. They could represent a starting point for a consistent digitations project within CNR. Moreover, it would be interesting to connect this initiative with the print collections stored in OPENSIGLE. This initiative would take into account Lynch's observation in reported in a recent study (Lynch, 2009; Hahn, 2008),.

In the case of monographs as we have seen previously, (fig. 4), a certain number of them are produced in-house following the traditional publishing process, while only a limited number is currently managed by e-publishing systems (fig. 5), because this would imply the change of the whole editorial process. In fact their shift to an innovative production model is more complex if compared, for instance, with the production of reports. However, considering that monographs, especially in Humanities and Social Sciences, are mainly scientific-based output and often a *niche product,* the use of e-publishing systems could be a valid alternative for cost containment, and we know that monographs in particular suffer budget constraints. Besides they are generally locally oriented, and often publish in their native language since they have greater difficulty finding commercial publishers. Last but not least, they are very expensive to produce and only few commercial publishers accept to publish in narrow and targeted study areas and even if they publish, they do not always guarantee a wide diffusion of contents or indexing and abstracting services.

In this regard many studies and reports stress these critical issues within Monographs. Some of them focus on the insustainability of traditional business models and report examples of successful experiences carried out by e-publishing library services as well as through alliance with more collaborative commercial publishers. (Hahn 2008; Alenius, 2007; Besen 2012, Ferwerda 2010; OAPEN project[2])

4.3. Some characteristics of the editorial business process

Table 3 provides an overview of the main characteristics of digital products management that we have analysed. For the majority (68%) of digital products all the activities connected with their production and diffusion are carried out in-house, from content acquisition (including submission and peer review) to the online publications of the content. For a smaller proportion (26%) of digital products, the Institutes externalize part of the editorial process, often the phases of copyediting and publishing. As can be noted paper and electronic diffusion is still the dominant solution and for these reasons some commercial publishers are in charge of providing the printed copy while in other cases print versions are available on demand directly from the producers. Sometimes commercial publishers are responsible for management of membership fees and subscriptions.

Tab. 3 *Profile of digital products (=19)*

Production & Diffusion	n.	%
In-house	13	*68,4*
Partially in-house	5	26,3
National commercial publisher (for print distribution)	*10*	*52,6*
International Commercial publisher	1	5,3
Access policies		
Full OA	15	78,9
Delayed OA	2	10,5
Open access online/Subscription for print	1	5,3
Subscription (online & print)	1	5,3
Technology used		
Content management system	12	63,2
Open source electronic publishing system	6	31,6
Publisher'platform	1	5,3
Copyright & Licensing		
Yes	*14*	73,7
Not available	*6*	31,6
Peer review		
Yes	*14*	*73,7*
Not available	*5*	*26,3*
Scientific committee & editorial board	**19**	
Yes	*14*	*73,7*
Not available	*5*	*26,3*
International standard codes		
Yes	*16*	*84,2*

Coming now to the technology used, the table shows that 5 out of 19 digital products are managed using Open Journal System (OJS). OJS is a journal management and publishing system that has been developed by the Public Knowledge Project. It manages every stage of the publishing process, from submissions through to online publication and indexing, including peer review process. It is currently the most suitable and widely used system to manage online publications. OJS it is also OAI-PMH compliant and supports interactive functionalities, such as reading and social network tools (Willinsky 2005; Brian & Willinsky 2010).

However, in our survey we found that the majority of digital products are managed using content management systems. It is interesting to note that in one case the Institute has developed locally an OAI compliant open source publishing system to manage its journal. This is the case of the journal "Archeologia e calcolatori" (Moscati 2009)

With regard to access policies to the content, the majority of products are open access (15 out of 19) and most of them provide information related to copyright and licensing. Peer review and scientific committee and/or an editorial board are also contemplated for the majority of them. Finally, almost all have International standard codes.

Summarizing, in our survey the hybrid business model is the dominant solution: it combines in-house editorial activities with partial externalization. Moreover, it mixes open and toll-access as well as print and electronic format to diffuse the contents. In general we can say that each CNR Institute has its own business model and even within the same Institute there are different management models depending on the type of product.

4.4. Examples

The following examples are representative of different business models adopted by CNR institutes.

1. ***Journal of limnology* -** This is an example of the evolution from traditional to innovative publishing. It is a journal directly published by CNR since 1942 in the very specialized field of limnology one. At that time it was a forerunner in environmental studies. Since 1999, it has been an electronic open access journal. The content acquisition and management together with the peer review process is still managed by the CNR institute, while all the activities related to copy-editing and publishing are outsourced to an external e-publishing service that uses OJS platform. If subscriptions for the printed version are required, there is a local commercial publisher that provides them.

2. ***Archeologia e calcolatori* -** This is an example of best practice. In 2005 the Institute followed the open access principles and developed an OAI-PMH compliant e-publishing system. The OAISISTEMA used a simplified solution to manage an OAI-PMH repository. In 2006 Archeologia e Calcolatori was indicated by Los Alamos National Laboratory study as an example of systems that enabled easy and efficient content discovery.

3. *Geothermics* - The journal Geothermics represents a different example compared to the previous ones. It was founded in 1972 by the CNR institute and appeared immediately as an international peer reviewed journal. At the beginning the editor in chief belonged to CNR while at the moment this journal has become one the Elsevier journals and the business process is completely managed by this international commercial publisher and recently the journal lost its CNR branding.

4. ***IRPPS Editoria Elettronica*** (e-Publishing service) **-** This last example is different from the examples described above. The project of introducing an e-publishing system in IRPPS institutes was designed and carried out by the library with the collaboration of internal researchers. We' should stress that the introduction of Open Journal System (OJS) in the publication process has varying aims depending on the types of products. IRPPS started with working papers and monograph series, introducing internal peer review for WPS and external peer reviews for monographs. The editorial staff also intend to re-publish old reports that represented a breakthrough in population studies, giving a second life to GL documents. The entire business process is carried out internally.

Conclusions

This preliminary survey focuses on well-established editorial products published by CNR Institutes, with the general aim to better understand to what extent use of new digital publishing technologies have innovated their editorial process and products. Despite a limited number of innovatively managed products, they are in line with scientific scholarly publishing connected with digital publishing technologies and on open access publishing models.

From the results of the survey, it is also clear that the well-established and standardized products, with a solid tradition in print publishing are concentrated in the Department of Humanities and Social Sciences, where the products are predominantly monographs. At the moment innovative products managed by online publishing systems are concentrated in the Department of Earth & Environment.

Concerning the business process, first of all we can say that disciplinary fields do not influence the business model, or the trends in adopting new technologies. What we discovered to be very important is the evolution towards innovative products generally based on products having a long and stable tradition, representing the history of the institute as well as the scientific achievement in a specific field. Of course many back issues aren't accessible online and we consider this the ongoing task in future CNR digitation projects, which we think should be planned and defined at departmental level.

We have seen that the business model adopted is not uniform in all institutes, each one having found its own solution. Sometimes the entire business process is managed in-house, but there is a widespread tendency to contract the publication and distribution of both the electronic and printed format to commercial publishers. From our survey some best practice examples of in-house publication management seem to emerge, especially those using electronic publishing systems with their value-added services. Certainly the use of e-publishing systems increase the quality of editorial products: additional services can increase the visibility; indexing and abstracting of products in search engines make them more easily retrievable, and the peer review process can be quicker.

Moreover, products that represent narrow and targeted study areas with a limited potential market and therefore encounter difficulty finding commercial publishers could benefit greatly from in-house publishing services. Of course, even if costs are reduced by the use of e-publishing systems, it implies the setting up of its organization, training, maintenance and updating. For these reasons a possible sustainable model could be adopted by CNR organizing it at a departmental level in order to achieve economies of scale and to optimize coordination actions.

Taking into account that e-publishing initiatives developed locally by CNR Institutes will grow in the near future, our study was an exploratory pilot study for long-term program publishing activities. In the future we will test the role of CNR libraries in the development of e-publishing services together with the CNR research community, that should be involved in founding the best innovative publishing practices suited to their needs.

In this context it is well known that academic and research libraries that have had a fundamental role in supporting Open access practices in the construction of Institutional repositories, and digitization programs are currently moving toward the development of additional services for their community scholars. This is in line with the onus on libraries to reshape their role in the digital age following changes in scholarly communication models. In this context library publishing services represent a new modality to diffuse scholarly research outputs, improve the quality of in-house published products and decrease costs of publication.

References

Adema Janneke, Schmidt Birgit (2010). From service providers to content producers: new opportunities for libraries in collaborative open access book publishing. *New review of academic librarianship* 16:S1.

Alenius Marianne (2007) Open access, University press, and editorial responsibility. Conference paper at the Workshop on Open access, 23-24 April 2007. URL: www.mtp.hum.ku.dk/cgibin/PDFmedopenaccess/Open_Access__Univer_0_0_9788763510868.pdf

Bareghed A., Rowley J., Sambrioik S. (2009). Towards a multidisciplinary definition of innovation. *Management decision*, 47 (8).

Besen Stanley M. Kirby Sheila Nataraj (2012). E-books and libraries: an economic perspective. American library association.

Borgman Christine L. (2007) Scholarship in the Digital Age: Information, infrastructure, and the Internet. Cambridge : MIT Press.

Brian D. Edgard; Willinsky J. (2010). A survey of the scholarly journal using Open Journal Systems. *Scholarly and research communication*. 1(2). URLS: http://pkp.sfu.ca/files/OJS%20Journal%20Survey.pdf

Costigan S. International Affairs Research and the Web (1999). In: Grey Literature GL99. Washington, D.C.

Cooke M., S. Costigan (2005). Making Your Way Through Grey: Metadata, MARC and User Tools. In: GL6 Conference Proceedings. Amsterdam : TextRelease, 2005.

Crow R. (2002). The case for institutional repositories: a SPARC position paper. Washington, D.C. : SPARC. URL: http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf

Crow R. (2009). Campus-based publishing partnership: a guide to critical issues. Washington, D.C. : SPARC. URL: http://www.arl.org/sparc/partnering/guide/

Ferwerda Eelco (2010). Open access monographic publishing in the humanities. *Information Services & Use*, 30 (3). URL: http://iospress.metapress.com/content/l6wg61l0mg6426w8/fulltext.pdf

Gierveld H. (2002). A conceptual analysis of functions, processes, and products in the scholarly communication chain. Elpub 2002 proceedings . URL: http://elpub.scix.net/data/works/att/02-16.content.pdf

Hahn, Karla L. (2008). Research library publishing services. New options for university publishing. American Research Library Report 258. URL: http://www.arl.org/bm~doc/research-library-publishing-services.pdf

Hougton John W. (2011). The costs and potential benefits of alternative scholarly publishing models. *Information Research*, 16 (1)

Iglezakis Ioannis, Synodinou Tatiana-Eleni (2011). E-publishing and digital libraries: legal and organizational issues. Hershey – New York : Information science reference

Lynch Clifford A. (2009). Special collections at the cusp of the digital age: a credo. Research library issues: a bimonthly report from ARL, CNI and SPARC 267. URL: http://www.arl.org/bm~doc/rli-267-lynch.pdf

Mill Fethy (2000). Trends in publishing academic grey literature: examples from economics. *The International journal on grey literature,* 1(4)

Moscati Paola (2009). Archeologia e calcolatori: le ragioni di una scelta. *Archeologia e calcolatori*, 20. URL: http://soi.cnr.it/archcalc/indice/PDF20/12_Moscati.pdf

Mullins, James L.; Murray-Rust, Catherine; Ogburn, Joyce; Crow, Raym; Ivins, October; Mower, Allyson; Newton, Mark P.; Nesdill, Daureen; Speer, Julie; and Watkinson, Charles (2012). Library publishing services: strategies for success. Research Report Version 1.0 Libraries Research Publications. Paper 136. URL: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1166&context=lib_research

OAPEN project. URL:http://www.oapen.org/home

Ridi R. (2001). Gli incerti confini dell'editoria digitale. URL: http://www.burioni.it/forum/ridi-confini.htm

Roosendaal Hans E. (2004) Driving change in the research and HE information market. *Learned publishing,* 17(1).

Roosendaal Hans E.(1998). Forces and functions in scientific communication: an analysis of their interplay. CRISP 97 Cooperative Research Information Systems in Physics. URL: http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html

Stempfhuber M. et. Al. (2007). Enhancing visibility: integrating Grey literature in the SOWIPORT cycle. In: 9th International Conference on Grey Literature (GL9). Tools for publishing, archiving and accessing GL. Amsterdam : TextRealase, 2007.

Tenopir Carol, Birch Ben, Allard Suzie (2012). Academic libraries and research data services. Current practice and plans for the future. An ACRL White paper. www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

Willinsky J. (2005). Open Journal Systems; an example of Open source software for journal management and publishing. *Library Hi-Tech*, 23 (4). URL: http://pkp.sfu.ca/files/Library_Hi_Tech_DRAFT.pdf

[1] CIAO (Columbia International Affairs Online) is one of the first digital publishing project to put online scholarly information in the field of international affairs. CIAO includes now full-text online books, journals, seminars and research projects, working papers, reports and conference proceedings. Launched in 1997, CIAO project was presented at a previous GL Conferences (1999, 2005). It is now managed by EPIC (Electronic Publishing Initiative at Columbia University), together with other two e-publishing projects developed at the Columbia University

[2] OAPEN project stands for Open access publishing in European Networks- OAPEN developed and implemented open access publication model for peer reviewed academic and research monographs in the Humanities and Social Sciences.

Grey literature in Australian education

Gerald White, Paul Weldon and Helen Galatis, Australian Council for Educational Research
Julian Thomas, Swinburne Institute for Social Research, SISR
Amanda Lawrence, Grey Literature Strategies Project
Jessica Tyndall, Gus Fraenkel Medical Library, Flinders University, Australia

Abstract

The prevalence of informal publishing or grey literature in education appears to have increased as digital technologies have become main-stream, educators have become more proficient and policies have moved increasingly towards supporting its use. In addition, the take up of social networking technologies and innovative methods of digital publishing have encouraged educators to produce, distribute and share content and commentary. Grey literature may make a substantial contribution to education even though issues such as credibility, access and a lack of standards can pose problems for producers and users. This paper begins by providing a context for the discussion of grey literature within the broader policy and education environment in Australia. An overview of grey literature as it appears in education in Australia introducing evidence of its usage, dissemination and application in Australian education then follows. Evidence about the access, dissemination and use of grey literature is drawn from an examination of the characteristics of a leading social networking and digital publishing service that was used by educators in schools, training institutes and teacher education faculties. This evidence is discussed in the context of influential national, state and institutional policies that address the use of digital technologies in education. As the take up of digital technologies in education increases, there is an expectation that the access to, dissemination of and use of digital publishing by and for educators will increase and have an impact on online professional learning and awareness of education research and practices.

Introduction

Education and training is one of Australia's largest enterprises employing hundreds of thousands of people and engaging millions of students. In schools alone, in 2011, there were 290,854 teachers and 3,541,809 students (ABS, 2011). In the main, education is funded by State Governments of which there are eight (six States and two Territories) and the national Australian Government.

The publication of most materials in education such as reports, policies, curriculum, research, guidelines, surveys, speeches and conference papers is undertaken by governments and educational agencies none of which have commercial publishing as a primary function. Although major policies and resources can be readily located during their time of implementation and prominent public discussion, some may fall into obscurity over time because they have not been commercially published and catalogued or preserved. One such example was a major national initiative called Education Network Australia (EdNA). EdNA and the resources developed during its lifetime will be discussed in some detail in this paper, in order to raise the issues associated with non-commercially published materials.

Non-commercially published resources and materials are often referred to as grey literature. Unlike commercially published materials, grey literature often does not go through a rigorous quality control and production cycle before being made available to its intended audience (Lawrence, 2012, p. 4). This raises a number of practical issues, especially for educators who seek evidence for supporting national projects of innovation, change and improvement.

This paper raises some of the issues associated with grey literature in within an educational context. EdNA is used as an example to demonstrate the impact of such information on education. Two possible solutions that seek to address issues about grey literature in education are then discussed suggesting ways that education might overcome some of the disadvantages of grey literature although issues associated with intellectual property are not canvassed. Finally, the paper advocates further actions for educators to maximise and build on the body of important knowledge published as grey literature in education.

Background

The Australian Research Council (ARC) is a national body that advises the Australian Government on research funding. Its focus is on research and innovation for the benefit of the community and globally (Australian Research Council, 2012). The ARC is responsible for the provision of significant research funding to Australian Universities.

In 2011, Swinburne University of Technology, in collaboration with four industry partners and the University of Victoria, was successful in receiving funding for an ARC research grant to be conducted over three years. The research program titled *Grey literature, innovation and access to knowledge: realizing the value of informal publishing* began in late 2011.

The 'aims of the Grey Literature Strategies project are to:

- Define the role and value of grey literature and establish ways in which its impact and value can be evaluated and measured.
- Improve the way grey literature is produced and published in Australia in order to maximise its quality, impact and use.
- Improve access, retrieval and preservation of grey literature by collecting institutions, universities and other organisations.
- Build networks of collaboration across sectors active in producing and/or managing policy-oriented grey literature in order to build capacity for shared administration and technological development.' (Swinburne University, 2012).

The project will review the production, dissemination and collection of grey literature in Australia and develop best practice guidelines to maximize its benefits and use in public interest research (Swinburne University, 2012).

The partners who have joined with Swinburne University on this project include the National Library of Australia (http://www.nla.gov.au); the Eidos Institute (http://www.eidos.org.au/), a think-tank that focusses on the work-force, skills and productivity; the State library of Victoria (http://www.slv.vic.gov.au/) representing the National and State Libraries Australasia (http://www.nsla.org.au/) and the Australian Council for Educational Research (ACER) (http://acer.edu.au). ACER is heavily involved in the production and dissemination of grey literature in education and training, including through digital research projects such as the Digital Education Research Network (DERN) (http://dern2.acer.edu.au).

Grey literature definitions

There are many definitions of grey literature. The most commonly accepted definitions were originally agreed at the Luxembourg Grey Literature Conference in 1997 and then expanded at the New York conference in 2004. Grey literature was defined as a collective noun to mean, 'information produced on all levels of government, academia, business and industry in electronic and print formats not controlled by commercial publishing i.e. where publishing is the not the primary activity of the producing body' (Farace, 1978; Farace & Frantzen, 2005).

There are those who would describe grey literature as information that is obscure, poorly distributed, of mixed value, ephemeral, of low value, less traded, less monetised and used in limited ways (Whitehead, 2012). Grey literature can be undated, the author not known and the body taking responsibility for the document obscure. In other words, grey literature is seen by some as difficult, disorganised and variable in quality.

The work of education and training in Australia is led and informed by the national and state governments through policies, white papers, discussion papers, inquiries, investigations, evaluations and guidelines all of which is grey literature. As such it is an important body of literature and there is a need for it to be readily locatable, accessible and retrievable from an organised source that has the capacity to manage and preserve electronic resources. Currently, in Australia, this is not the case and grey literature is dispersed, restricted and can give the impression of being hidden from public access. Much grey literature is lost to the education sector due to the discontinuation of services and a lack of policies to ensure its survival.

Australian education

Australian education is organised into a hierarchy of three tiers: schooling, training and higher education. Schooling operates from Kindergarten to year 12, training is vocationally and technically focussed for post-compulsory education, and higher education concentrates on studies for degrees and research.

In the school tier, there are three sectors that are comprised of: the public or state school sector catering for 66% of school students; the Catholic sector which educates 20% of students and the independent school sector with 14% of students (Australian Government, 2011, p. 4). In the training tier the Australian states manage 64 Technical and Further Education (TAFE) institutions on multiple campuses in parallel with over 400 Registered Training Organisations (RTOs). Australia has 39 universities, 38 of which are publicly funded.

Although the Australian states have a legislative mandate to manage education and training, the Australian Government provides the bulk of the funding from taxation for schooling in the three schooling sectors. Therefore, policies, research, strategies and priorities in education emanate from both the national and the state governments. These documents are usually stored electronically on websites and disseminated by notification, usually from government websites. They are then commented on in the news media, which monitors education and training, and also by bloggers and microbloggers (eg Twitter) involved in education. Therefore, finding, accessing, retrieval and use of documents about education produced by governments and research bodies constitutes an important body of knowledge for educational leaders, teacher educators and practicing teachers as well as researchers.

The proliferation of grey literature is underlined further by research that found that more archeological information is made available via grey literature than printed literature (Houghton, 2012). There is no reason to doubt that this situation is similar in education and training. As broadband is implemented throughout Australia, and particularly in services such as education and training, internet usage will increase. As the use of the internet increases so too will the production and use of blogs (Thomas, 2008) and other forms of grey literature publishing that will be at the forefront of information dissemination about educational change and innovation.

Educational innovation

National education policies, arrived at through cooperation between education Ministers and senior education officers of the national and state governments, drive educational innovation in Australia. Currently, national policy programs such as the supply of a computer to all year 9-12 students, the development of a national curriculum, the implementation of national teacher standards and teacher professional development, and the distribution of funds for special programs to improve equity for Aboriginal students, disabled students and students in poverty are central to educational improvement, innovation and change.

Each of the national programs mentioned above has produced a raft of documents from inquiries, policies, research, reports, speeches, guidelines, articles and scholarly papers. Researchers wishing to locate these documents need to be aware of the national priorities and where documents for these are stored. However, a number of issues associated with this tranche of documents become apparent in their production, dissemination and access. These issues can be seen more clearly by examining one such national project that has now been completed: EdNA[1].

EdNA was a national project that ran for fifteen years from 1995. It was managed by a national agency charged with the responsibility for its maintenance and development, and consulting with the national government and the state governments about its provision of services and strategic direction. The national consultative processes were undertaken through a range of collaborative electronic and physical means such as digital services, websites, blogs, group spaces, listservs, emails, video-conferencing meetings, distribution of papers and the like. Each of these processes produced a vast amount of important educational and technological literature as the project progressed.

The EdNA website in 2009 had a database of 43,368 items that had been evaluated against developed content guidelines and manually added metadata to improve the speed and accuracy of document locations. These items were aimed at supporting the use of digital content and services in education and training. There were three million additional linked resources accessible via a real-time federated search function of significant international databases of educational resources including research. EdNA also supported 35,349 educators engaged with collaborative professional groups and email distribution services. The production of knowledge was substantial by the users of these EdNA services. This knowledge included not only the policies, research and strategies related to the project and the many national high level groups that collaborated on the project but also advice from expert groups, international bodies and teacher support materials.

In 2010 the EdNA service was closed by senior education officials following a stakeholder review which did not take into account or consult with EdNA users. The EdNA users were the people who had engaged with and benefitted from the service. None of the aforementioned electronic resources and materials is officially available today and cannot be accessed. However, a subjective selection of resources including some research and reports were unofficially stored on Australia's electronic archive called Pandora[2]. In addition, the research, strategy, consultation and management documents associated with the project and developed by the national technology agency responsible for the management of EdNA are also not available. All in all, the pioneering work of a large number of professional educators, educational leaders, researchers and experts has been lost to posterity, research and historical analysis.

The question that then emerges is, 'Why was this information lost?' In a small and limited enterprise losing information may be understandable. However, education and training in Australia is a very large enterprise that in 2011 consumed 7.1% (Australian Bureau of Statistics, 2012) of the national income; that is $94 billion. An innovative and ground breaking national project such as EdNA can have a lasting impact on future policy and innovation upon which education and training is advanced. Today, the national collaborative physical and online processes that were conceived and tested by EdNA are used in the development of the national curriculum. However, the EdNA resources and materials were lost because they were not archived and preserved. Instead, the EdNA online service was shut down and most of the grey literature information that was in electronic formats vanished.

The capacity to bring important information together into one accessible database or multiple databases connected in real-time, and to archive and preserve important information should be a central function of all online services that produce and disseminate important literature. The next two sections discuss ways in which grey literature can be harnessed to enable searching, accessing, retrieving, usage, archiving and preservation.

Integrators or grey literature curators

A number of online websites provide integration services within a specific field. An integration service can be a web service where a group of discipline experts aggregate and curate information in one place for access by interested users. An example of an integration service is the Digital Education Research Network (DERN) (http://dern2.acer.edu.au) managed by the ACER (http://www.acer.edu.au). DERN is a research service that concentrates on and specialises in the use of digital technologies in education for the purpose of improving learning. DERN provides a research news and review service that also facilitates access to research archives and a database of digital education research. It is focused on servicing teachers, educators, researchers, policy makers and digital education commentators.

Each week DERN distributes an email notifying users about research news and also a review of a significant research article, paper or report. Predominantly, the news and research reviews that are provided on DERN are about works that have been openly published and are freely accessible for viewing and often for downloading. News and research reviews of articles and papers available in commercially published journals are rarely cited on DERN because they are inaccessible to the general user. DERN is an aggregator and curator of grey literature in the field of learning using digital technologies and digital media.

Of note is the observation that it would appear that the bulk of research, scholarly papers, policy statements and strategic documents about the use of digital technologies in education and training are openly available and published as grey literature. Integration services such as DERN (http://dern2.acer.edu.au) support educators and educational researchers to locate, access, disseminate and use grey literature in the field of digital technology. Integrators such as DERN bring highly skilled expertise to concentrate on the provision of a service in a specific field of inquiry that removes many of the barriers inherent in grey literature. However, if web services that contain significant research that has been reviewed by DERN and linked to DERN is not preserved and archived then, it may also be lost to research and knowledge building.

A framework for harnessing grey literature

Another method for systematically harnessing grey literature in a specific field of research or scholarly endeavour was developed by Jessica Tyndall, Medical Librarian at the Flinders University in South Australia. Tyndall (2008) developed a framework in order to systematically collect and evaluate grey literature for use by students, researchers and scholars. Tyndall (2008) listed the types of grey literature that may be encountered by scholars and researchers, and then went on to propose that grey literature could 'be critically appraised for strength and validity using a simple approach … [by] marrying the concept of "expert opinion/insider knowledge" with the general principles used to evaluate web resources' (p. 5).

Tyndall (2008) identified the essential characteristics of grey literature and argued that the following criteria checklist 'has the flexibility to be applied to the widest range of resources: from models of primary healthcare to dissertations, maps, diaries, podcasts, blogs and so on' (p. 6). The criteria in the Tyndall checklist include:

- Authority,
- Accuracy,
- Coverage,
- Objectivity,
- Date, and
- Significance.

Tyndall (2008) abbreviated these grey literature characteristics as AACODS for brevity and as a memory aide. Each criterion is explained in some detail and the researcher is able to use them to evaluate resources from a grid of inquiry locations based on possible sources of information.

An example using Tyndall's (2008) framework to systemically build a search schema for a discipline can be seen in Table 1. In Table 1, the AACODS criteria have been listed in the centre with the types of possible information sources shown on the left and levels of search on the right. Each located item can then be evaluated by using Tyndall's criteria and sorted for sources and levels of information.

Tyndall AACODS framework			
Possible sources	**Criteria for each item**	**Levels**	
Govt policy/strategy reports	*Authority*	Open	Local
Blogs	*Accuracy*		National
Projects	*Coverage*		
Conferences	*Objectivity*		International
Newsletters	*Date*		
Web services	*Significance*		

Table 1: AACODS framework

A researcher can develop a schema for research by adapting the Tyndall (2008) AACODS framework for application in a specified field of inquiry. Although this framework is not exhaustive, it is sufficiently general to be applied to wide variety of disciplines and provides an excellent starting point for locating and evaluating grey literature in a specific discipline.

Conclusion

Grey literature as defined in this paper is being relentlessly produced by a range of government, businesses and research bodies. Education and training in Australia is a large and ever changing enterprise that is dependent on government funding. Change and innovation often occurs in education and training at the national level as a result of policy, strategy and funding shifts. The documentation produced by local, state and national education and training projects in the course of such shifts in policy and priorities is published by governments, national agencies and research institutions. Information about innovations also emerges from conferences, workshops, newsletters and blogs.

The increasing proportion of grey literature emanating from education and training innovations although important is obscure, difficult to locate and access, of mixed value and used in limited ways. However, major national projects have compounded these ephemeral characteristics of educational grey literature through a lack of attention to archiving and systemic preservation. Grey literature that is not archived contributes to the loss of an important body of material on which education could build and advance.

The importance of grey literature in education and training cannot be underestimated. As the take up of digital technologies in education escalates, there is an expectation that the access to, dissemination of and use of digital publishing by educators and for educators will increase and have an impact on online professional learning and awareness of education research and practices.

There are ways in which grey literature can be preserved which include access to expert integration online services such as DERN (http://dern2.acer.edu.au) where the searching, curation and aggregation are done for the user. A grey literature search framework is another way that can empower researchers to build a search schema for systematically locating, accessing, retrieving, using and applying open information in a specific scholarly discipline. The Tyndall (2008) AACOD grey literature framework is one such method for developing a schema for systematically researching and evaluating grey literature.

In order to maximise the use of grey literature in education and training, there is a need to systematically collect, store, archive and preserve important information as a body of knowledge for posterity, research and historical analysis. Further research is warranted into the role of grey literature in order to improve access, retrieval and preservation of important information produced as grey literature.

The Australian Council (ARC) Linkage Project 2012-2014 titled *Grey Literature Strategies: Enhancing the value of informally published research and information* led by Swinburne University in collaboration with five partners which includes the Australian Council for Educational Research (ACER) (http://www.acer.edu.au) is one such grey literature research effort. Progress on the research project can be seen at http://greylitstrategies.info/.

Bibliography

Australian Bureau of Statistics. (2011). *4221.0 – Schools, Australia*, 2011. Canberra, Australian Bureau of Statistics. Retrieved November, 12, 2012, from http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/4221.0Main%20Features12011?opendocument&tabname=Summary&prodno=4221.0&issue=2011&num=&view=.

Australian Bureau of Statistics. (2012). *Education and training. 1301.0- Year Book Australia, 2012. Canberra, Australian Bureau of Statistics*. Retrieved November, 12, 2012, from http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Financing%20education~111.

Australian Government. (2011). *Review of Funding for Schooling: Final Report*. Canberra; Australian Government. Retrieved November 3, 2012, from http://www.deewr.gov.au/Schooling/ReviewofFunding/Documents/Review-of-Funding-for-Schooling-Final-Report-Dec-2011.pdf.

Australian Government. (2012). *Australian Research Council*. Retrieved November 4, 2012, from http://www.arc.gov.au/.

Ewing, S. & Thomas, J. (2008). Broadband and the 'Creative Internet': Australians as consumers and producers of cultural content online. *Observatorio Journal*. Vol. 6, pp. 187-208.

Farace, D. (1998). Foreword. In *Third International Conference on Grey Literature: Perspectives on the Design and Transfer of Scientific and technical Information, GL 2004 Conference Proceedings*. Luxembourg, 13-14 November, 1997. Amsterdam, GreyNet/TransAtlantic.

Farace, D. and Frantzen, J. (2005). (Eds.). *Sixth International Conference on Grey Literature: Work on Grey in Progress, GL 2004 Conference Proceedings, New York, 6-7 December 2004*. Amsterdam, TextRelease.

Lawrence, A. (1202). Electronic documents in a print world: grey literature and the internet. *Media International Australia*, Issue 143, pp. 122-131.

Swinburne University. (2012). *Grey Literature Strategies*. Retrieved November 12, 2012, from http://greylitstrategies.info/about.

Tyndall, J. (2008). *How low can you go?: Toward a hierarchy of grey literature*. Retrieved July 12, 2012, from https://dspace.flinders.edu.au/jspui/bitstream/2328/3326/1/Tyndall.pdf.

Whitehead, D. (2012). Grey Literature is easily orphaned. Presentation at the *Where is the Evidence Conference, National Library of Australia, Canberra, 10 October, 2012*. Retrieved November 2, 2012, from http://greylitstrategies.info/presentations.

[1] A snapshot of a number of top-level layers of EdNA can be seen at http://apps-new.edna.edu.au/edna_retired/edna/go.html

[2] Pandora can be accessed at: http://pandora.nla.gov.au/

Research Life Cycle: Exploring Credibility of Metrics and Value in a New Era of eScholarship that Supports Grey Literature

Julia Gelfand, University of California, Irvine (UCI)
Anthony Lin, Irvine Valley College, United States

Abstract

The fundamental components of the research process are defined by academic tradition, discipline and its participants. Traditional scholarship has now evolved into eScholarship with emerging technologies providing new methods of innovation and new ways of handling classical research processes. This revised research life cycle not only incorporates the established parts of the research chain, from discovery, gathering, and creating, but now has added phases of citing, sharing, preserving and archiving. There are quantifiable elements that help describe unique elements depending on sector and subject matter and format. Previously defined barriers such as geographical, institutional, digital and domain boundaries that had earlier existed can now be transcended to either accelerate or retard the research lifecycle in amazing and innovative ways. This new paradigm shift of current practices or activities today include the range of literacies that must be demonstrated and include information literacy, visual literacy, financial literacy, and increasingly data literacy. The role of the academic library has become increasingly visible as scholars and scientists seek support in managing their research lifecycle components. Librarians are now managers and curators of the scholarly research lifecycle by protecting, harvesting, and promoting reuse of content for new and unprecedented purposes. In a similar fashion, Grey Literature has previously followed the lifespan of more traditional output. Now new technologies exist to extract value metrics that compare favorably with other information products. This paper will explore how Grey Literature matures through different pathways or life cycles as the new grey becomes less grey with metrics of increasing value to support and describe it. Also, the world of publishing has become increasingly accessible to a new population of scholars to release new information and ideas, contribute to emerging fields and frontiers without the barriers or requirements of following a specific trajectory of traditional publication processes. Examples will be shared about how the research life cycle has evolved with new tools to support Grey Literature from the life cycle management (LCM) and life cycle assessment (LCA) models to determine impacts and drive future directions concerning options for actions like open access, intellectual property and other forms of rights management.

Background

Innovation is the catalyst for positive change and grey literature is the measure of benchmarks in the further process of research and development. Innovation and grey literature share parallel life cycles in which early growth is relatively slow until their use and application become recognized both within and later beyond their community of origin. Expected top-line growth and increased bottom-line results are achieved in part through new technologies, through redeployment and enhancement of existing products and services, which at times are unachieved. Nevertheless, the process shared by innovation and grey literature carries on with analogies to product life cycle management strategies. If grey literature is considered a product, and business or industry the context, then the conditions in which it is developed, tested, applied and sold becomes its lifecycle and changes over time. Product life-cycle management (PLC), often known as marketing include stages such as analyzing the time to market, improving product quality; and considering ways to penetrate new markets, reach new customers and develop new applications. In our contemporary environment, software and technology play important roles in determining the lifespan and obsolescence of any product.

Research is an abstract environment and has many different components but studied in its most benign and natural state, one can assume that research generates products, births new knowledge and engages in a re-engineering process while confirming the following assumptions:

- Every product has a defined life cycle and a limited life
- Sales and marketing require different passages, and both challenges and opportunities are presented to the distributor or seller
- Each life cycle stage demands attention to marketing, financing, manufacturing, purchasing and other strategies to support it successfully

This can be demonstrated by the main stages of a product's life cycle:[1]

1. Market Development – full of unknowns, uncertainties; trying not to prematurely fold
2. Growth – responds to consumer demand/interest; develop brand loyalty over other products; establish pricing
3. Maturity – responds to competitive intelligence; requires creative marketing, communication with users or customers
4. Decline (saturation) – over capacity is usually the result of this stage but it can be a positive outcome with prices and margins reduced diverting the decline with a more creative force.

Again, we confirm that innovation is never easy – the population must perceive that it needs and wants a product or service. The more unclear or new it is, the more challenging it is to enter the mainstream.

Research Life Cycle

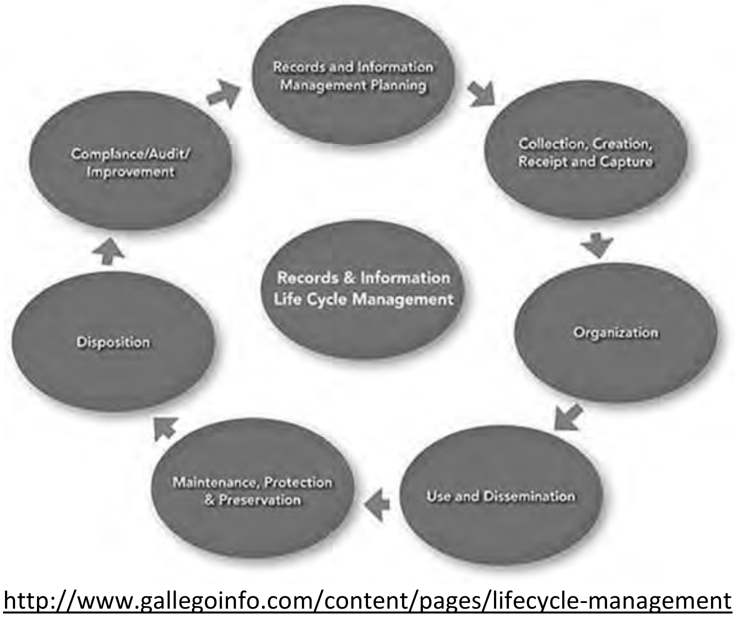
The research lifecycle appears to be measured in different ways. It also refers to different elements in the research chain. In the framework of assessing outputs, one traditionally uses metrics that reflect the activity of the author or contributor, what the product is and how it is cited to determine the value it has to a user, subscriber, or to a more general audience. Libraries have increasingly become engaged in assessing the value of their collections and the source of their resources. Research generates many different products and services, but in library-speak, they tend to be universally described as information sources.

The Library and especially the digital library environment that organizes information have become increasingly central to the research process. The scholarly communication pipeline as it has been referred to at conferences, especially the future of scientific communication has been the focus of all stakeholders representing scientists and scholars, librarians, publishers, technologists who are exploring "how to create continuity across the entire ecosystem and give the scientific community the ability to observe or participate wherever they need to in that ecosystem, …by filling the need for better filtering."[2] The added value of providing a structure to find and store research products or outputs, and as a source for discovering new research trends or evaluating research excellence and thus improve the process of how research is conducted. These values are increasingly part of a new research landscape covered by meetings now held worldwide to help advance publishing, research sharing, informatics, data analysis and hopefully develop new technologies and products to expand the research of research.[3]

"Life cycle management (LCM) is a business management approach that can be used by businesses and organizations to improve products and performance of those companies or institutions involved in production and thus can be used to target, organize, analyze and manage product related information and activities towards continuous improvement along the life cycle."[4] Also it can be simply described as "a framework to analyze and manage the sustainability performance of goods and services."[5]

An example, or one model that diagrams this for the information industry may be something like this:

Lifecycle Management


Records and Information Management Planning
Collection, Creation, Receipt and Capture
Organization
Use and Dissemination
Maintenance, Protection & Preservation
Disposition
Compliance/Audit/ Improvement
Records & Information Life Cycle Management

http://www.gallegoinfo.com/content/pages/lifecycle-management

Life cycle assessment (LCA) involves "implementing standards, metrics, procedures and inventory data for the social dimension of sustainability"[6] creating a community that can vet and compare data elements or social indicators as part of this process. With roots in environmental standards, it allows providers a means to improve resource efficiency. It can be extended to information products or resources by adding a social life cycle assessment, (S-LCA) that "can be used to assess the social aspects of products and their potential positive and negative impacts along the life cycle."[7]

What appears to be critical is to know that in addition to LCM and LCA approaches, there are a wide variety of other tools that are used to understand the transition from LCM to LCA or management to assessment. This chain includes the actions of discovery, gathering, and creating, but now has added phases of citing, sharing, curating, preserving and archiving the content whether it be text, images, data, 3D objects, media, etc. This can correspond in the information industry and work of librarians to be part of the scholarly communication model, building out current service emphases of information literacy, outreach, instruction, digitization and the most recently launched data management/curation programs.

Research and scholarship are independent activities that define the academic experience and contribute to creativity and innovation. In recent years, the concept of "knowledge generation" has been applied to the outputs and the goals of research.

When libraries are trying to forge ahead increasing the digital footprint and aligning services to support eResources, the quandary became more complicated this year, as publishers raised the price of eBooks to libraries, and consortia struggle to find ways to reduce unit costs by multi-institution buying plans. This is a serious detour in the research lifespan.

Definition of Innovation

Standard dictionary definitions of the verb innovate include: "make changes in something established, especially by introducing new methods," or "something different that has impact."[8] In order to study or evaluate innovation, some business executives or analysts may consider the context as being one of "perpetual change," and the "imperative of today's times;" and ways to categorize innovation as by the "innovation's strategic intent" or examining the "type of innovation."[9]

Innovation comes in many shapes and sizes with different forms of influence. It is never easy and the population or users must perceive that it needs and wants a product. The more unclear or new it is, the more challenging it is to enter the mainstream and be accepted. The research environment attempts to consider things in a transformational way, again by weighing the impact over time. To understand innovation and see how it plays out, it is critical to determine and anticipate the desired outcome. This can be done by considering three types of innovation:

- Continuous innovation – suggests incremental changes or improvements and is "a common way to satisfy existing customers while grabbing new users."
- Dynamically continuous innovation – represents a change in the way we use a product without changing the technology behind the product all that much
- Discontinuous innovation – requires a significant behavioral change but is not synonymous with disruptive innovation because that causes immediate changes while discontinuous innovation may take significantly longer to influence change.[10]

Certain types of innovation require more resources, time and energy from an organization than others and will determine how to select the right type of innovation to fit the appropriate commitment level.[11]

"The world no longer cares about what you know; the world only cares about what you can do with what you know," writes Tony Wager, Innovation Education Fellow at Harvard's Technology and Entrepreneurship Center, and author of the recent release, *Creating Innovators: The Making of Young People Who Will Change the World*.[12] This book is innovative in itself as its promotion is available in multiple formats inviting discussion and commentary. (see http://creatinginnovators.com/). It states that innovation is dependent upon a skillset, learning style, and keeping up with change. It is inevitable that subsequent generations of students and thus workers will have a different skillset than current employees and the transition to a different work life and organizational culture will continue to be challenging for some time, as new members of the workforce appear to have a different, more technology intensive background.

Innovation in Higher Education – with extensions to libraries

With the recent budgetary challenges, the field of higher education needs to invest in innovation to find solutions to long standing problems. Never easy, the word "innovation" can be used in many different ways. In the *Future of University Libraries: 2012 Midwinter Report*, citations to several examples of entrepreneurial action point to how libraries are:

- demonstrating value
- rethinking library services
- reconfiguring library spaces
- preparing psychologically for the future.[13]

James Neal, Vice-President for Information Services and University Librarian at Columbia University suggests that research libraries of today and the future are “more entrepreneurial organizations, more concerned with innovation, business planning, competition and risk, leveraging assets through new partnerships to create new financial resources.”[14]

If innovation can be classified into two basic types, administrative and technical as it describes innovation in organizations such as libraries, then two stages may emerge, 1) initiation and 2) implementation each with its own substages that form what Jantz calls an innovation process model with four major construct groups:

1. leadership
2. new knowledge
3. organizational structure
4. perceived innovation attributes.

The substages for initiation are:

- knowledge awareness substage
- attitude formation substage
- the decision substage

and the substages for implementation:

- initial implementation substage
- continued sustained implementation substage[15]

Perhaps this sets us up for the prediction made by Daniel Greenstein in 2009, “The university library of the future will be sparsely staffed, highly decentralized, and have a physical plant consisting of little more than special collections and study areas.”[16] This scenario he describes is already a truism as budgets decline to support library collections and associated services are reconfigured to offer greater access via more remote connections and library reorganizations tackle how best to meet ongoing service and information needs. Extending the reference to a research context, Christine Borgman has stated, “the role of libraries in research institutions is evolving from a focus on reader services to a focus on author services.”[17]

Unique to the United States: The Community College Experience

In a one-year study completed by O’Banion, T. & Weidner, L. (2010), *The nature of innovation in the community college,* the authors examine community college views of innovation in higher education. The community college movement got its stride in the 1960s in the United States when divisions between higher education sectors were defined. Community colleges allowed students to fulfill their core competencies with an A.A. degree before matriculating in a baccalaureate degree program, or gain vocational or specialized expertise and credentials. Today, community colleges are providing a critical and necessary function in higher education as students and their families are facing increased tuition and fees, are less mobile, where the emphasis is on teaching and learning outcomes, the greater demand for new skills in the workforce and lifelong learning is at an all time peak. O’Banion and Weidner found after examining hundreds of statements on the definition of innovation in both higher education and business, all definitions of innovation led to this fundamental concept, “a creative process that results in an improved product.”[18]

Institutional barriers to innovation

Institutional support from a college administration is the most important source for responding to institutional needs. It is logical to conclude that at the research university level, that these more complex institutions would place a much higher emphasis on innovation than their community college counterparts. However, evidence largely parallels their community college findings. In a report issued by the American Society for Engineering Education (ASEE), fiscal resources to sustain innovation and policies and practices to support innovation were shockingly low. The report also states that “This is mostly good news in that the faculty committees and administrators generally want to increase their involvement in educational innovation. They just do not feel that they have the policies, practices, resources, and infrastructure they need to be more successful.”[19]

In contrast to O’Banion and Weidner, the ASEE report tends to take a more holistic approach to creating a culture of innovation in their engineering institutions. Both groups stress the need for partnerships and teamwork, securing administrative support for innovation, and assessment. The ASEE takes their recommendations one step further by integrating them into the "pre-professional, professional, and continuing education of engineers."[20]

Interestingly enough, lack of resources was not the main source of barriers for innovation at the community college level. The main road block was "lack of time," and is understandable due to declining budgets, decreased staffing levels, and an increase of duties for the remaining faculty who have to teach larger enrollments, more sections with less support. Often innovation is thought of as an investment in time when the institution can dedicate resources to innovate. However, in a time strapped environment, people are focused in a reactive planning mode versus a proactive strategy.

Comparing higher education realms

Paradoxically, at the four-year research university level, the main barrier to innovation was "resources." This was consistently cited by faculty and academic leaders. Workload, the main barrier for community colleges, was listed towards the middle of the five challenges among this group.

The community colleges and research universities both have different approaches to foster innovation in higher education. The community colleges tend to focus on a narrower approach that is defined within the community college framework:

- Demonstrate a need.
- Develop a vision and a plan.
- Put the plan into action.
- Talk with colleagues.
- Build a team.
- Secure administrative support.
- Dedicate the required time and effort.
- Evaluate the innovation's effectiveness.
- Tie the innovation to the college mission, values, and goals.
- Take risks.
- Plan for sustainability of the innovation. [21]

Promoting innovation through assessment

By far the most surprising finding in the O'Banion and Weidner study is the question "How do you know that the innovation had an impact?" The most popular response was "Faculty/staff testimonies or anecdotes" followed by "Student testimonies or anecdotes". This is rather disappointing considering how in today's movement of student learning outcomes and evaluation, there is little to no formal planning or assessment to measure the success or failure of a particular innovation project. At the same time, this could also be explained by the "lack of time" previously mentioned in the survey since community colleges are many times in a reactive "firefighting" mode that will leave little to no time left for proactive, evidenced based planning and outcomes.

Innovations in Higher Education – Distance Education

With the nature of higher education changing from the physical classroom to the online environment, the popularity of distance education has increased in recent years not only in the United States, but globally as students enroll in courses and programs at institutions around the world without ever leaving home. According to the most recent report by the US National Center for Education Statistics, the percentage of undergraduates enrolled in a distance education course has grown from eight percent in 1999 to almost 20 percent in 2008, with variances across the higher education spectrum.[22]

Furthermore, the breakdown of students enrolled in distance education courses from a US public 2 year institutions was at a surprisingly high 22 percent compared to their public 4 year university counterparts at 16 percent. The private 4 year non-profit universities were not as inclined to have distance education courses at this time since only 12 percent was reported. This gap may be narrowing as new and emerging technologies have become more evident in the past 3-4 years; it is safe to assume that from 2008 that these figures have increased overall for all categories of distance education. [23]

There are a variety of reasons why Distance Education has enjoyed an exponential increase in popularity over the past decade. The most salient reason is cost, as higher education institutions grapple with leaner budgets, increasing staff costs, and decreasing support from public funding sources. Providing distance education is one way to trim costs as Distance Education courses require no physical space to teach courses; a higher number of students enroll without impacting space constraints on campus facilities; in extreme circumstances the professor can telecommute from her/his home negating the need for a physical office on-campus, and the asynchronous learning environment of Distance Education courses allow colleges to decrease the number of on-campus courses while providing the maximum flexibility for both students and faculty.

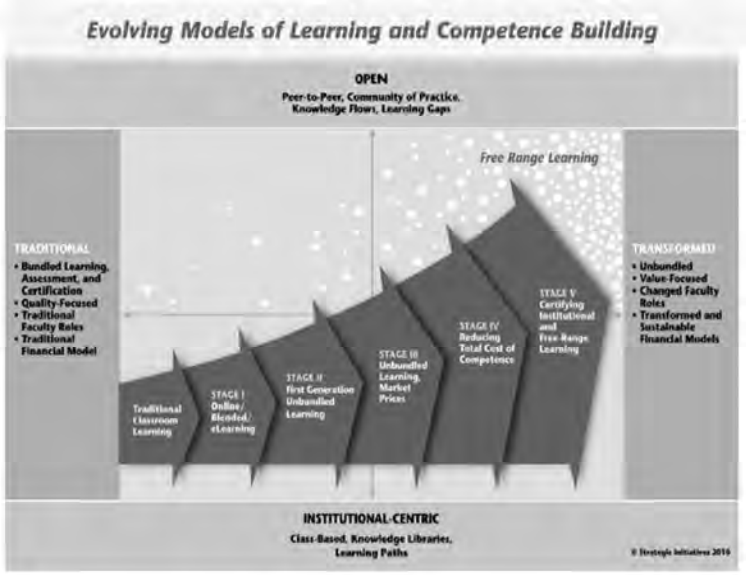
For students, the benefits and personal convenience of Distance Education are equally numerous as the asynchronous nature of Distance Education allow students to take courses and study at their own pace. This is especially attractive for students who work full-time and can take advantage of Distance

Education courses by studying and completing the coursework before and after their jobs. According to the National Center for Education Statistics, nearly 27 percent of undergraduates who were employed full-time took Distance Education courses compared with 17 percent for part-time workers and only 16 percent for those not employed. [24]

Learning online

Today we are in a transitional phase in higher education, a shift away from traditional paradigms of brick and mortar learning and breaking through to online learning environments that have virtually no boundaries. Norris and Lefrere identified five phases of Distance Education environments that lead to an evolution of online learning from the traditional physical classroom setting to a peer-to-peer learning system that is accredited by specialized bodies. [25]

As shown in the figure, there are four quadrants: on the left represents the traditional learning environment of bundled learning and assessment, on the right represents the "transformed" learning environment of unbundled learning and value focused coursework. On the bottom we see the "Institutional-centric" learning environment of traditional class-based learning vs. the open learning environment of peer-to-peer learning. As we move from the right to the left, we note that in Stage I is the start of most Distance Education programs that "digitize" the traditional learning environment through hybrid classes (a continuum of traditional classroom and online coursework), as we move towards the right Norris and Lefrere state that the Distance Education environment, specifically online learning, will evolve into a system where peer-to-peer learning is commonplace with institutional certification boards providing the levels of certification demonstrating competence in a particular field of study.[26]


Evolving Models of Learning and Competence Building
OPEN
Peer-to-Peer, Community of Practice, Knowledge Flows, Learning Gaps
Free Range Learning
TRADITIONAL
• Bundled Learning, Assessment, and Certification
• Quality-Focused
• Traditional Faculty Roles
• Traditional Financial Model
TRANSFORMED
• Unbundled
• Value-Focused
• Changed Faculty Roles
• Transformed and Sustainable Financial Models
Traditional Classroom Learning
STAGE I Online/ Blended/ eLearning
STAGE II First Generation Unbundled Learning
STAGE III Unbundled Learning, Market Prices
STAGE IV Reducing Total Cost of Competence
STAGE V Certifying Institutional and Free-Range Learning
INSTITUTIONAL-CENTRIC
Class-Based, Knowledge Libraries, Learning Paths
© Strategic Initiatives 2010

Source: Norris and Lefrere, 2011

Norris and Lefrere note the following:

Universal web access supports a shift toward greater reliance on informal local communities of practice and web-based and cross-border knowledge-sharing, and a shift away from peer-reviewed articles as the definitive source of accounts of experiences of successes in online learning and in the early adoption of transformational innovations at the course or subject level.

In other words, as we transition from Stage I towards Stage V, there will be a greater reliance on grey literature rather than the traditional peer-reviewed articles that are pervasive in traditional academic settings. Rather than relying on editorial boards and publishers to evaluate the level of scholarly information, the future will rely on practice-based and collaborative knowledge communities that transcend borders and the traditional academic boundaries of scholarly information sources. As time goes on, grey literature will become more important and proliferate in the Distance Education environment due to the universal access to of publication, the speed of producing a finished product, and harnessing the practice-based knowledge of experts in the field that are not traditionally tied to the peer-review environment. However, it is possible that peer-reviewed literature will continue to evolve into a format that is certifiable for the accreditation institutions to base the levels of competence in Norris and Lefrere, Stage V Distance Education institutions.

Grey Literature as distance education matures

Five years ago I presented at the Grey Literature conference how technology and other elements of cyberinfrasture, specifically distance education, impacted the maturity of grey literature.[27] With more ubiquity, a growing dependence on networking, and an infrastructure now to support a full industry of home schooling for the K-12 cohort, the latest evolution is the massive open online course or MOOC. 2012, known to be the "Year of the MOOC" is where we see traditional online courses where tuition and fees were charged for credits that translated into degrees, give way to the usually free, credit-less and massive enrollments of tens or hundreds of thousands of students taking a given class. All that is required is an internet connection and due to the large enrollments relationships with instructors are indeed different. The course delivery, methods of interaction, and student engagement are what contribute to the success of the course, and communities evolve to foster that with the support of social media directing a framework of options for students to collaborate and congregate to share the learning experience. Lectures still are the central core but the rollout will take elements from entertainment, especially gaming and social networking to facilitate the conversations, exchanges and dialogues course content should stimulate. Many questions remain about the efficacy and how to handle cheating and plagiarism, how to scale learning this way, whether it can translate into degree programs, global registration and how credits will be earned and students will record such registrations and experiences. Currently three major platforms exist for massive open online courses, edX, Udacity and Coursera[28] and each of them has their own "flavor" indicating subject orientation, technology focus and directives. Some top, prestigious academic institutions continue to partner with innovators in industry to bring this form of instruction and learning to the mainstream, and since it remains a "work in progress," with "many kinks" it is premature to suggest what role it will play in coming years in higher education. We can speculate that it may have a role for the Emeriti Colleges, for lifelong learning and other continuing education roles but will it assume a place in post-secondary higher education as we currently know it?

Assessing consumer needs: the academic experience

With the current generation of students as digital natives, their customs and expectations, according to a new edition of, *Generation on a Tightrope: A Portrait of Today's College Student*, are grounded in technology, they appear pragmatic, concerned about their future and jobs, have a more global interest, but act locally and deal with diversity by claiming to want to live in a networked world but many of their associates, including parents and teachers are analog or digital immigrants. Students struggle with face-to-face relationships and prefer communicating via texting and choose to "hook up" but not with a lot of talking, yet a more open communication system exists but the bond with parents is tighter than they want to admit, suggesting less independence, probably due to longer financial and residential dependency. They are part of a transformation to a new and diverse digital economy, where students appear to want convenience, service, quality and lower prices in all aspects of their daily lives, including their education.[29]

Students as consumers are an important population demographic thing to focus on. Not only by age, but it is taking longer to matriculate and graduate, students are attending multiple institutions in seeking credits for their degrees, and in several different educational settings and domains as the undergraduate career can be spent at all of the forms of higher education, due to transfer status and concurrent registrations, community college, a liberal arts institution, a research university and a distance education option from yet another discrete campus.

The 18-25 age cohort may be the largest concentration of any student group, but there are many older students returning to university, continuing their studies, beginning new programs, and seeking new academic training. Serving the needs of military veterans and more international students are examples of the extended services campuses must offer today.

Examining the Beloit College Mindset data, demonstrates how important it is for faculty to know who their audience is. Conducted annually since 1998, this review is a "look at the cultural touchstones that shape the lives of students entering college each fall."[30] The takeaways are how connected this generation is with their peers and family, how important hand-held devices and technologies are and how committed they are to cultural diversions, a healthy lifestyle and good eco-values.

The student community is a combination of millennials, boomers, Gen Xers and others surveyed by the Pew Research Center trying to better under understand how generations differ. According to Pew, "age group differences can be the result of three overlapping processes:

MILLENNIALS

A PORTRAIT OF GENERATION NEXT

Confident.
Connected.
Open to Change.

1. Life cycle effects – becoming more like their parents once they themselves age
2. Period effects – affected by major lifetime events, catastrophes and breakthroughs
3. Cohort effects – how period events and trends leave specific impressions as youth are still developing core values[31]

The influence of social media on organizations and research

It has been stated that using "a defined approach to manage social media can stifle innovation and creativity."[32] Some key factors that Bradley and MacDonald introduce to make mass collaboration an organizational capability include the premise that, "an organization becomes a social organization when it discovers the power of mass collaboration and develops the necessary corporate skills to address challenges by readily and repeatedly creating collaborative communities."[33]

- Understand when community collaboration is appropriate
- Know where community collaboration is more likely to deliver value
- Apply an understanding of your organizations goals and culture
- Craft an organizational vision for community collaboration[34]

The principles of mass collaboration and social organizations and the management guidance each requires are:

- Participation – encourage contributions from across community and make it safe by discouraging destructive and dysfunctional behaviors and promoting productive ones
- Collective – ensure results by reaching consensus and taking action together
- Transparency – use most accurate and appropriate information; encourage openness and inclusivity
- Independence – encourage and facilitate multiple viewpoints and broader perspectives
- Persistence – keep collaborative content, contributions, feedback and decisions with the social media platform and easily available to community members
- Emergence – concentrate on community results rather than controlling the means of producing those results. Defining terms of engagement may compromise community contributions.[35]

Social media practices - innovation strategies

Collaboration and social networking reduce a range of geographical, institutional, hierarchical, and digital barriers and promote implementing technology to bridge those structural silos. Within the information and higher education sectors, the more technology centric publishing, libraries and information distribution can become; the more likely that the research lifecycle will be impacted by improved access. Still a business enterprise, these services will benefit from more Open Access / Open Source content and the promotion of information sharing, repurposing and reuse of content.

On a global scale the ICT environment that is defined by "Information and Communications Technology" reaffirms that there are internal stakeholders of leaders, customers and support staff plus the network of external sources, vendors, customers, officials, lenders and media who also contribute to the success of innovation. With the flattening of the world, global geography is smaller and reachable via same time communication technologies. World events are shared within moments across the globe, introducing and displacing innovation all at the same time. In higher education, the customer may be the student, the innovator may be the faculty or scholar, the product is the learning outcomes, teaching and scientific methods that are advanced.

Many organizations are now using and relying upon social networking tools to reach out to all constituencies and solicit innovative ideas from them. The literature and media refer to ICTs as the method by which groups create and vent ideas – without a culture that supports ideation, innovation is challenged. Social ideation is an extension of engagement that encourages the use of social media internally so that colleagues can share ideas with each other and also can communicate externally with outsiders for more input and follow through.

More techniques and tools – leading to greyness

Social media can also encourage the use of crowdsourcing, the outsourcing of tasks to the external masses that are typically performed by internal employees or those close to the job.[36] This may be simpler than engaging research firms to conduct surveys, monitor and track consumer behavior. It also builds on the semantic enrichment of scientific publications and efforts of text-mining.[37]

Another method of gaining insight into the innovation process is the more academic version of storyboarding, a technique that has its roots in filmmaking and allows for groups of people to discuss sequencing and the narration of events. Somewhat related to the ethnographic analysis an anthropologist or social scientist may pursue when they study specific populations and cultural norms, these techniques utilize questions and answers, observation, surveying and other information gathering methods to gain insights about ideas, actions and activities. This can be applied to organizational settings to outline an acquisition, define an entry into a new market or movement of people, or to measure change.

Mind mapping is a tool that explores relationships, and ideas or connections are strung together by the power of the relationship. This visible interconnectedness illustrates potentially innovative ideas that can fix, remedy, re-engineer, and stimulate new ways to go forward and contribute a product or methodology.

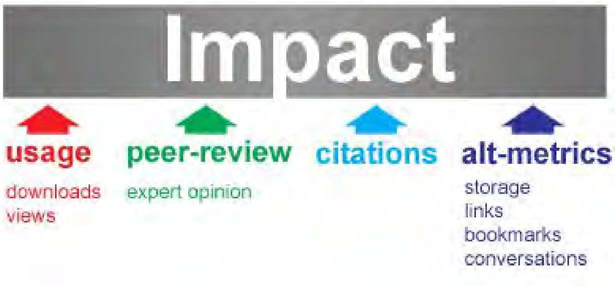
A wordle has become an accepted simplified, randomly and spontaneously software generated "mindmap" of concepts minus the relationships but releases a summary of covered ideas.

Byproducts of the research lifecycle

An environmental scan of products that document the research lifecycle suggests that there are a growing number of resource formats. Organizing and measuring various impacts are a common theme. Market penetration and other business and management themes are reflected in each of these byproducts of the research lifecycle. There are various forms of intellectual property. The release of constant new products measures productivity, commercial success and ultimately, innovation. The international stage promotes the global importance that each of these play flattening the world and reducing deficits of time. Examples are:

- Intellectual Property & Patents – greater global contributions
- Industry Standards – by certifying organizations - in the information industry with examples from ISO, NISO, OASIS, ANSI, and closed and emerging standards from the W3C (World Wide Web Consortium) and those maintained by the Society of Scholarly Publishing[38]
- Library Standards – based on literacies and created by professional societies (ACRL Professional Standards)[39]
- Benchmarking
- Social media – multiple ways to connect, view, contribute, participate, respond to issues of the day
- ICT channels – communications, news outlets, visual content, archiving
- New & multiple formats responding to user preferences and marketplace shifts focusing on digitization
- Publishing and Usage Metrics, Bibliometrics and Altmetrics

The academic community has always evaluated its members, both institutions and individuals by determining the rankings, reputation and success of its creators and increasingly on the commercial potential for those ideas and products. In their own way, each of these establishes new separate communities where there is an element of competition, sharing, informing, educating, that takes place.



(http://altmetrics.org)

Within the spectrum of innovation, several metrics are critical to the research lifecycle. Most scholars are familiar with the impact factor, coined by Eugene Garfield in the early 1960s when he founded the Institute of Scientific Information in Philadelphia and birthed the products, *Science Citation Index, Journal Citation Reports* morphed into other larger and more interdisciplinary products that tracked impact by individual article, author and publication. The measurement of overall scholarly impact, bibliometrics as a tool and science was defined by Alan Pritchard in 1969 as, "the application of mathematics and statistical methods to books and other media of communication."[40]

The expanding Altmetrics movement, defined as "the creation of new metrics based on the social web for analyzing and informing scholarship"[41] is just a few years old and gaining momentum with activists Jason Priem and Healther Piwowar directing much of the development, especially with her development of ImpactStory and her blog, ResearchRemix.[42] Altmetrics provides not just counts and cumulative totals, but rich metadata.[43] Citation analysis has driven the method of how the scholarly community attributes value to information. We can conduct literature searches by not only examining retrospective contributions but forward tracking by citation in different formats. Increasingly indexing now covers data elements, illustrative content as well as print. With the scholarly community focused on journal literature, the article and conference paper/proceeding has always been vital to track and follow. This gives insights into what defines competitive journals that publish the most important content critical to interested readers; and determines the lifeline of the journal, its cost and role in a subject. In addition to metrics associated with authors, content, institutional rankings, benchmarking and reputation, there are also methods for assessing impact for research groups and interdisciplinary intersections. Utilizing data mining and emerging technologies with many methods of reviewing the changing scholarly landscape new products and metrics are introduced.

Bibliographic management software is yet another example of managing references and organizing retrieval of multiple sources in a highly personal and customized way that can accommodate a range of information sources. Many are free software applications that can be downloaded, some have web-based functionalities and others are offered as fee-based subscription models.

Brand new applications of how to manage pdfs of journal articles or conference papers or book chapters allows one to build on products such as Mendeley or Zotero with the latest rollout of ReadCube by the Nature Publishing Group[44], which builds on providing access to fulltext articles. Priem states, “Bibliometrics mined impact on the first scholarly web, altmetrics mines impact on the next one.”[45]

Examples of different impacts are:

- Impact factor
- H-index
- Times cited (different variations but counts times cited in primarily journal articles)
- i10 Index (articles with 10+ citations)
- Highly cited (usually relates to authors)
- Eigenfactor
- Source Normalized Impact per Paper (SNIP)
- Google Scholar Citations
- Microsoft Academic Search
- Publish or Perish (PoP
- Altmetric for Scopus (tracks mentions of papers across social media sites, blogs and reference managers)
- Academia.edu

Innovation is right, front and center in the research lifecycle. In recent years, there have been many new examples of products that are offering new ways to analyze data. Most libraries have a subject guide that demonstrates and describes the processes and products. My LibGuide[46] is but one example.

Conclusion

When we consider what makes people, companies, organizations or products innovative, we often conclude that they are different in some special, unique way. Innovation is often a catalyst for the mindset needed to achieve success and the result of a journey in which discovery plays a large part. Influenced by the recent book, *The Innovator’s DNA*, we share how those authors provide five steps that they suggest leads to innovation where a bigger and better impact can be achieved:

1. Review and establish priorities
2. Assess discovery skills systematically
3. Identify a compelling innovation challenge that matters
4. Practice discovery skills (association, questioning, observing, networking, experimenting, skills)
5. Be coached to support development efforts[47]

Within the research community, innovation will be tested as new products are being released to manage the range of information and new knowledge generated by sponsored research.

Data is probably the most critical of the new forms of grey literature that libraries and librarians need to address. Issues related to eScience and grey literature were more fully addressed in a paper delivered three years ago at this conference.[48] Occasionally libraries have described the new research landscape as “data deluge” as they grapple with how to manage such resources so that they can be consulted, archived and reused. This is the new landscape of eResearch and eScholarship that has extended the boundaries and intersections of library collections and services. If libraries fail to embrace treating data, the risk of losing credibility and remaining relevant to the academic research and institutional landscape is huge. Activity at research libraries around the world now demonstrate how new structures are evolving to support institutional metrics that build on data management service models that include repositories, data curation and data management plans and extend to a range of other scholarly communication related services. The European Association of Research Libraries has documented ten recommendations for libraries to get started with data management plans[49] and in the US over the last year the Association of Research Libraries and the Digital Library Federation hosted a series of institutes for their members to respond to the eScience and data agendas on their campuses. [50] A recent conference in China highlighted such activity by Pacific Rim universities.[51]

Data and research impacts including the new bibliometrics are clearly examples of the next generation of grey literature that will pave the changing times of higher education. They are an important indicator of how scholarship is viewed from the outside looking in at a given point of time. These are times of great change and fluidity where coping with ambiguity is the new norm. Michael M. Crow, President of Arizona State University and known as perhaps the most visible innovator currently leading a major research institution in the US, in the thriving metropolitan area of Phoenix, where he has created, an "unusual academic structure with 'fused intellectual disciplines' meant to reflect the way knowledge is developed and applied today and a culture deliberately focused on admitting and graduating a student body that is ethnically and economically representative of the community."[52] Thus, innovation, grey literature and new paradigms in higher education fuel the research lifecycle.

Acknowledgements: Appreciation to SJSU Intern, Laine Thielstrom for her comments and insights on earlier drafts of the paper, and to Dan Tsang

[1] Levitt, Theodore (1965). Exploit the product life cycle. *Harvard Business Review* 43 (6): 82-84, November-December.

[2] Colloquium on Rethinking the Future of Scientific Communication (2012). Stanford University Libraries, March 8-9, 2. https://lib.stanford.edu/files/Colloquium.Summary.Final_.pdf Retrieved November 1, 2012.

[3] Scholarship 2.0: An idea whose time has come (2012). http://scholarship20.blogspot.com/2012/04/1st-international-workshop-on-mining.html Retrieved May 20, 2012.

[4] *Life Cycle Management: How business uses it to decrease footprint, create opportunities and make value chains more sustainable* (2009). UNEP/SETAC: vii. (http://www.unep.fr/scp/publications/details.asp?id=DTI/1208/PA - Retrieved October 20, 2012.

[5] Ibid., 4.

[6] Hunkeler, David, and Ribetzer, Gerald, (2005). The future of Life Cycle Assessment. *International Journal of Life Cycle Assessment* 10, (5): 307.

[7] *Life Cycle Management:* 5.

[8] Anthony, Scott D., (2012). *The little black book of innovation: How it works, how to do it*. Cambridge, MA: Harvard Business Review Press: 16.

[9] *Ibid*, 30, 27, 22, 24.

[10] Goldsmith, David. (2012). *Paid to think: A leader's toolkit for redefining your future*. Dallas, TX: Ben Bella Books, 432-433.

[11] Ibid.

[12] Wagner, Tony (2012). *Creating innovators: The making of young people who will change the world.* New York: Scribners. (see http://dailyedventures.com/index.php/2012/05/30/tony-wagner-usa/

[13] Dunaway, Michelle, (2012). *The Future of University Libraries: 2012 Midwinter Report*, 2.

[14] Neal, James (2011). Advancing from kumbaya to radical collaboration: Redefining the future research library. *Journal of Library Administration*, 51 (2): 67.

[15] Jantz, Ronald C. (2012). A framework for studying organization innovation in research libraries. *College & Research Libraries*, 73 (6): 528-529.

[16] Greenstein, Daniel (2009). Libraries of the future. *Inside Higher Ed*, available online at http://www.insidehighered.com/news/2009/09/24/libraries. Retrieved November 13, 2012.

[17] Borgman, CL (2010). Research Data: Who will share what, with whom and why? Fifth China – North America Library Conference, September 8-10, 2010: 13. http://dx.doi.org/10.2139/ssrn.1714427. Retrieved November 1, 2012.

[18] O'Banion, T. & Weidner, L. (2010). *The nature of innovation in the community college*. Phoenix, AZ: League for Innovation in the Community College. Retrieved November 11, 2012, from http://www.league.org/league/projects/nature_of_innovation/

[19] Innovation with impact, creating a culture for scholarly and systematic innovation in engineering education. Washington D.C. American Society for Engineering Education (ASEE). Retrieved October 1, 2012 from http://www.asee.org/about-us/the-organization/advisory-committees/Innovation-with-Impactj

[20] (2012). Innovation with impact, creating a culture for scholarly and systematic innovation in engineering education. Washington D.C. American Society for Engineering Education. http://www.asee.org/about- us/the-organization/advisory-committees/Innovation-with-Impact (accessed November 1, 2012)

[21] O'Banion, T. & Weidner, L. (2010). *The nature of innovation in the community college*. Phoenix, AZ: League for Innovation in the Community College. Retrieved November 11, 2012, from http://www.league.org/league/projects/nature_of_innovation/

[22] Radford, A. (2011). U.S. Department of Education, National Center for Education Statistics. Learning at a distance undergraduate enrollment in distance education courses and degree programs. *Stats in Brief*, (NCES 2012-154). Retrieved from website: http://nces.ed.gov/pubs2012/2012154.pdf

[23] *Ibid*

[24] *Ibid*

[25] Norris, D., & Lefrere, P. (2011). Transformation through expeditionary change using online learning and competence-building technologies. *Research in Learning Technology*, *19*(1), 61-72. doi: 10.1080/09687769.2010.549205

[26] *Ibid*

[27] Gelfand, Julia (2007). Updating Grey Literature as distance education matures. Paper presented at the Ninth International Conference on Grey Literature, Antwerp, Belgium, December 11, 2007.
[28] See https://www.**edx**.org/, www.**udacity**.com/, and https://www.**coursera**.org/
[29] Levine, Arthur and Deane, Diane R. (2012). *Generation on a tightrope: A portrait of today's college student.* San Francisco, CA: Jossey-Bass, 3rd ed.
[30] Beloit Mindset List 2016, http://www.beloit.edu/mindset/2016/ (accessed November 1, 2012)
[31] Millennials: A portrait of generation next, 2009. Pew Research Center. http://pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf (accessed November 17, 2012)
[32] Bradley, Anthony J., and McDonald, Mark P., (2011). *The social organization: How to use social media to tap the collective genius of your customers and employees.* Boston, MA: Gartner & Harvard Business Review Press: 29.
[33] *Ibid*, 23.
[34] *Ibid*, 31-32.
[35] *Ibid*, 148.
[36] Goldsmith, David. (2012). *Paid to Think: A leader's toolkit for redefining your future*. Dallas, TX: Ben Bella Books, 445.
[37] First international workshop on mining scientific publications (2012). http://core-project.kmi.open.ac.uk/jdcl2012
[38] http://sspnet.org/Publications_and_Links/Standards/spage.aspx (retrieved November 15, 2012)
[39] http://www.ala.org/acrl/standards (retrieved November 15, 2012)
[40] Pritchard, Alan (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 25(4). http://www.academia.edu/598618/Statistical_bibliography_or_bibliometrics (retrieved November 1, 2012)
[41] http://altmetrics.org (accessed November 17, 2012)
[42] ResearchRemix http://researchremix.wordpress.com/ (accessed November 17, 2012)
[43] Priem, Jason (2012). **Toward a second revolution: altmetrics, total-impact, and the decoupled journal.** Invited talk to Purdue University Libraries. Lafayette, IN, February 14, 2012. https://docs.google.com/present/view?id=ddfg787c_362f465q2g5
[44] http://www.readcube.com/ (accessed November 17, 2012)
[45] Priem.
[46] LibGuide, Research Impact Using Citation Metrics - http://libguides.lib.uci.edu/researchimpact-metrics (accessed Nov. 17, 2012)
[47] Dyer, Jeff, Gregersen, Hal, and Christensen, Clayton M., (2011). *The Innovator's DNA*. Cambridge, MA: Harvard Business Review Press, 259
[48] Gelfand, Julia, New shades of grey: The emergence of e-science, scientific data and challenges for research libraries. Paper presented at the 11th International Conference on Grey Literature, Washington, DC, December 14, 2009.
[49] LIBER's Ten Recommendations (2012). http://www.libereurope.eu/news/ten-recommendations-for-libraries-to-get-started-with-research-data-management (accessed October 10, 2011)
[50] ARL DLF E-Science Institutes 2011-2012. www.arl.org/rtl/eresearch/escien/escieninstitute/index.shtml (accessed Oct. 11, 2012)
[51] Todd, H. (2012). A partnership to support the research lifecycle: A case study from the University of Queensland Library. Presentation made at the Pacific Rim Digital Library Alliance (PRDLA) Conference, Beijing, China. November 4, 2012.
[52] Bluemenstyk, Goldie (2012). Change takes root in the desert: Embracing inclusiveness, Arizona State University pursues transformation on a grand scale. *Chronicle of Higher Education*, November 19, 2012.

Use Pattern of Archives on the History Of Mysore

Dr. N. Chowdappa, BMS College of Engineering, Bangalore, India
L. Usha Devi, Bangalore University, Bangalore, India
Dr. C.P. Ramasesh, University of Mysore, Mysore, India

Abstract

Records on the Administration of the Princely State of Mysore and Mysore History form rare collections for historians who venture to study the history of Mysore State under the rule of the Wadiyars, the rulers of Mysore State. These rare materials and archives have been carefully preserved at the Archival Section of the University of Mysore and also at other libraries in Mysore city. The present study furnishes the type of archival materials available at the University Library, Oriental Research Institute and the Karnataka State Archives, Mysore Division. Further, the study depicts the purposes of accessing archives and the use pattern of these rare archival sources on Mysore history by the research scholars, students and teachers in the discipline of Karnataka/Mysore History and allied fields. The present study also projects the extent of dependency of scholars from various professional fields, for information sources on Mysore History. The study also projects the rare collections of manuscripts and books of Tipu's Library.

Key Words: Archives, Mysore History ; Administrative Records, Mysore History ; Wadiyars, Mysore State; Tipu's Library Collection.

Introduction

Mysore was under the rule of Wadiyars for many centuries. For a brief period of time it was ruled by Hyder Ali and his son Tipu Sultan. Historians in their writings cover the most prominent aspects of Mysore : Dynasty of Wadiyars, Rule under Hyder Ali and Tipu Sultan, account of four Mysore wars, culture and philosophy, education and literature and also the art, architecture and music. Under the royal patronage, there was encouragement for the development and of art, music and education. Even today many researchers tend to concentrate on these major issues of Mysore State. Several publications are also being brought out on Mysore.

The present study aims to study the various archival collections available in the holdings of important libraries of Mysore city. The related objectives are to study the use pattern of archival collections on history of Mysore. The investigators have visited the libraries under the study and obtained data and information through the questionnaire. The available records pertaining to the holdings of archival materials and the annual reports of these libraries have also been consulted. The libraries which possess rich archives on history of Mysore are : Mysore University Library, Oriental Research Institute and State Archives, Mysore Region. Archives on history of Mysore including rare works cover the following types of documents: manuscripts, collection of letters, maps and charts, monographs, books, directories, handbooks, report literature and serials. The libraries which are consulted by the scholars include Department of Archaeology and Museums, Government of Karnataka, Mysore University Library, Oriental Research Institute, State Archives of Mysore Region. There are also few libraries which maintain archival materials and rare works under closed access. These are the libraries of Mysore Palace, Sarasvathi Bhandar, and Special collections of Tipu Sultan. The collection of Tipu Sultan's personal library was studied in detail by George Stewart during 1800 and 1805. He made a descriptive bibliography on the holdings of Tipu's Library and subsequently, the descriptive bibliography was published by Cambridge University Press in 1809. Most of the important collections of Tipu's library were religious texts and literature in Arabic and Persian. The manuscripts were also preserved with minor repairs and rebinding. However, of late, the collections are scattered and difficult to get access for reference and reading. The collections of Tipu Sultan has also gained importance on account of the fact that Tipu was a brave king and fought in the 3rd and 4th Mysore wars against the British Rule. He was a heroic figure and was physically fighting against tigers.

Important Library Holdings

SL. No.	Libraries in Mysore	Total Collection
01	Tipu's Library Collection (TLC)	2715
02	Mysore Palace Library (MPL)	16550
03	Sarasvathi Bhandar (SBM)	4120
04	Department of Archaeology and Museum (DAM)	5620

HOLDINGS ON HISTORY OF MYSORE
(In the Libraries Selected for the Study)

Sl. No.	Selected Libraries in Mysore	Archival Collection
01	Mysore University Library (MUL)	8095 (1.3%)
02	Oriental Research Institute (ORI)	3018 (2.7%)
03	State Archives, Mysore (SAM)	16506 (43.0%)
	TOTAL COLLECTION	**27619 (3.6%)**

The subject-wise number of rare works in the holdings of Tipu's Library have been depicted in the below table. Tipu was a lover of works on religion, philosophy, history and literature including works in science and technology. There were 1112 rare collections in his personal library. It is said that Tipu was keen in collecting rare manuscripts on religious texts, particularly works on the religious text, Koran.

RARE COLLECTIONS IN TIPU'S LIBRARY : 1799 A.D.
(Languages : Persian, Arabic, Turkish, Hindi and Kannada)

Sl. No.	Subject Areas	Number
01	Philosophy and Religion	239
02	Koran and Commentaries	120
03	Linguistics	74
04	Literature	237
05	Culture and Tradition	46
06	Collection of Letters	53
07	Jurisprudence	95
08	Science and Technology	108
09	Astronomy	22
10	History	118
	Total Rare Works	**1112**
	Total Library Collection	**2715**

Purposes of the Use of Archives

The below table projects the percentage of the use of archival documents available in the holdings of three prominent libraries in Mysore

Sl. No	Purposes of the Use of Archives	Percentage of Use		
		MUL	ORI	SAM
01	Teaching	14%	12%	22%
02	Research	21%	23%	34%
03	Students' projects	48%	41%	13%
04	Publish News Items	17%	24%	31%

It is clear from the table that a large majority of the users access archival collection for the purposes of research work as well as project works of master's degree programme. However, quite a number of users access materials at the State Archives for the purpose of writing articles and news items for newspapers and popular magazines. Among the purposes of use, teaching happens to be the low priority; comparatively, users seek archives for the purpose of teaching to a lesser extent.

Extent of Use of Archives

The extent of use depends on the availability of needed sources or facts. Secondly, the extent relies upon the facilities extended by these libraries which include the working hours, the extent of open access and the time required to retrieve the sources by the staff who work in these libraries.

Use of Archives at MUL

Sl. No.	Category of Users	Year-wise Visitors		
		2009	**2010**	**2011**
01	Teachers	409	419	426
02	Research Scholars	601	586	616
03	Students	1410	1316	1418
04	Public and Journalists	399	447	489
	Total	**2819**	**2768**	**2949**

The table depicts the use of archival collections at Mysore University Library (MUL). All the category of members makes use of the archives to a considerable extent. However, the students who approach the library for the purpose of preparing project reports are more in number when compared to the other categories of users, i.e., teachers, researchers and public, including journalists. Therefore, it can be deduced that all the categories of users access archives and rare works at MUL and the number of students who seek archives on Mysore history is comparatively high.

Use of Archives at ORI

Sl. No.	Category of Users	Year-wise Visitors		
		2009	**2010**	**2011**
01	Teachers	101	112	118
02	Research Scholars	196	210	228
03	Students	356	359	413
04	Public and Journalists	211	247	243
	Total	**864**	**928**	**1002**

It is evident from the table that students and public avail the facility at Oriental Research Institute to a great extent when compared to the teachers and research scholars. It is also true that there is considerable increase in the number of users every year. The students who visit ORI seeking archival collections, including manuscripts and rare books are master's degree students who compile and write dissertations or project reports as part fulfillment of master's degree programme. A large segment of the collection at ORI is in Sanskrit language, and the materials relating to history to some extent are in kannada and English.

USE OF ARCHIVES AT SAM

Sl. No.	Category of Users	Year-wise Visitors		
		2009	**2010**	**2011**
01	Teachers	816	793	803
02	Research Scholars	1212	1256	1229
03	Students	427	439	471
04	Public and Journalists	805	917	1107
	Total	**3260**	**3405**	**3610**

The above table depicts the use pattern of archival materials at State Archives, Mysore Regional Office, (SAM). Students who visit SAM are mainly for academic pursuits and the number is comparatively low. Whereas, teachers, research scholars and public who rely upon SAM are comparatively high in number. It is evident that a large segment of research scholars in the field of Mysore history rely upon State Archives for reference as the collection here is directly relevant and comprehensive. The statistics in the table also project that there is increase in the number of visitors every year.

Findings ans Recommendations.

- Archives and rare materials, including manuscripts on the history of Msyore is being used for research endeavour and quite a large number of students rely upon local libraries for their project works of the master's degree programme.
- The study on the use pattern of archives on Mysore history reveals that there is gradual increasing trend in the use of archives.
- The archival collections of utmost importance are related to the subject areas of ;
  - Wadiyars of Mysore : life and contributions
  - Education and cultural heritage of Mysore
- The archival collections on thrust subject fields like "Mysore Wadiyars and Mysore Cultural Heritage available in the holdings of the State Archives are being frequently used by scholars and public at large.
- Rare manuscripts and report literature on Mysore history have to be protected using modern preservation techniques at Mysore University Library and Oriental Research Institute. Quite a number of records have become brittle and susceptible for damage in these two libraries.
- It is recommended that the digitization of materials of historical importance has to be taken up and completed on priority in these two libraries. A major portion of materials of historical importance has been digitized at the State Archives and the process has to be continued covering manuscripts as well.

Bibliography

1. Rukminamma, P (2012). Mysore University Library : Evaluation of Facilities and Services. Tamil Nadu, Alagappa University (M.Phil., dissertation)
2. Leili Seifi and C.P. Ramasesh (2012). Digital and preservation of cultural heritage collection among Libraries of India and Iran, Germany, Lambert Academic Publishing.
3. Ramasesh, C.P. (1995). Record of 75 years of Mysore University Library. March of Karnataka. Feb. 1995: pp. 2-5.
4. Stuvert, George (1809): Tipu's Library. London, Cambridge University Press.
5. Ramasesh, C.P. (2009). Mysore : A Brief Note on its Culture. Silver Unifest, University of Mysore, pp. 26-28.
6. Ramasesh, C.P. (2002). Saga of Library Resource Sharing in Mysore City. In : National Conference on Consortia Approach for Content Sharing. Mangalore University. Pp. 185- 206.

Data Analytics: The next big thing in information

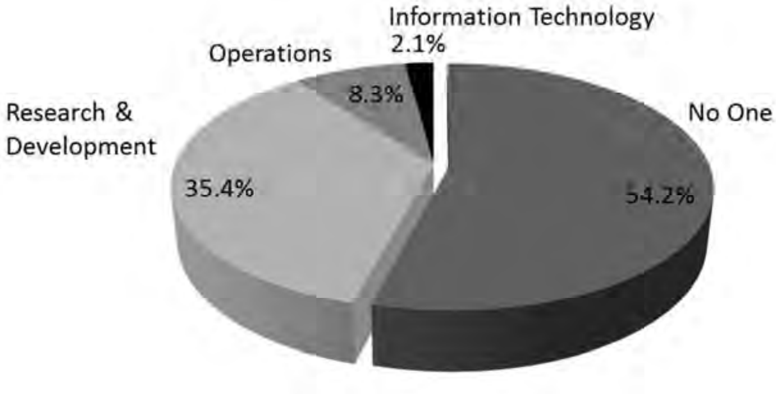
June Crowe and Joseph R. Candlish
Information International Associates, Inc. (IIa), United States

Abstract
Information is now available in an overabundance, so much so, that distinguishing the noise from the signal has become very problematic. In the past, the collection and storage of information was the primary issue. Currently, there are massive amounts of data both structured and unstructured, that need to be analyzed in an iterative, as well as in a time sensitive manner. In response to this need, data analytical tools and services have emerged as a means to solve this problem.
Grey literature repositories, libraries, and information centers are well positioned to take advantage of these new tools and services. The current trend is to make grey literature more easily discoverable, accessible, and with the new data analytical tools and services, more easily analyzed.
The intent of our survey of the Grey Literature community was to provide a snapshot of the Community's use, planned use, and knowledge of data analytical tools/services for big data as it affects grey literature. The survey summary that follows indicates where the Community currently stands in regards to the use of data analytical tools and services.

Represented Industries
From September 13 through October 31, 2012, an online survey was conducted and made available through two internet vehicles from: (1) the GreyNet Group on LinkedIn®, and (2) the GreyNet listserv. Forty eight responses scattered across South America (1), New Zealand (1), Africa (2), Asia (4), Australia (8), Europe (16) and North America (16), yielded insight into the Grey Literature Communitys' knowledge of the Big Data construct. Overall, academia represented 50% of responses, with government and private industry composing the remaining half. Within these industries, nearly 42% of the respondents were at the staff level, indicating that there's great understanding of the Big Data landscape, especially in academia.

The Current Landscape
The current landscape of Big Data products and services revealed several key significant points. First and foremost, a large majority (73%) of respondents indicated that their organization does not currently use Big Data products and services. This situation is partly due to the lack of drivers/champions to adopt them (>54%). However, in the area of Research and Development, there proved to be a significant contributing driver (35%) among the survey population that did denote the current existence and utilization of Big Data discovery and analytical tools (Figure 1).

Figure 1: Survey responses of current drivers for the adoption of Big Data services/products. (N=48)



Since such a large percentage of the Community has not adopted Big Data capabilities, it was not surprising to see that the majority (74%) indicated a novice expertise level. Moreover, 80% of the respondents had not seen any data analytical products/services demonstrated but were planning to use such products for web analytics, predictive analytics, and real-time analytics. For those who were familiar with existing analytical tools, SAP, SAS, and Google BigQuery were among the most popular. The survey question concerning the near future impact of big data analytical platforms, databases, services, and data analytical tools on grey literature (some impact-27%, moderate impact-19%, high impact-33%) clearly shows that the Grey Literature Community is expecting these services/products to provide some solution to the problem of Big Data.

Importance of Big Data

Big data is important primarily because it is growing at an exponential rate. Over five exabytes is created every two days. The problem with Big Data is not just data analysis, but with discovering, harvesting, curating, storing and its management. Steve Pederson, CEO of BrightPlanet Corporation, stated that 90% of Big Data content lies in the expanding universe of unstructured content; the vast majority of that information is hidden and unknown in the Deep Web segment of the Internet (Pederson, 2012). Not surprisingly, much of grey literature is found in the Deep Web.

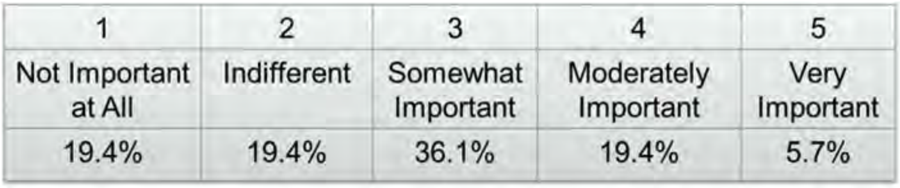
The Grey Literature community strongly felt that Big Data will be a huge positive for society just as it will be for science (56%). George Strawn of the Networking and Information Technology Research and Development (NITRD) Program identified four trends in Big Data in science and business:

(1) bigger data;
(2) increase in unstructured data;
(3) increase in distributed data; and
(4) increase in distributed computing.

These trends will spawn new tools and services for data sharing and collaboration, for data analytics, and for the management of data (Strawn, 2012).

Mobile devices are quickly becoming a primary means of accessing data. However, the survey results indicated that less than half of the respondents responded that it was only somewhat important (36%), on a 1-5 scale, as a method of accessing Big Data results. Figure 2 outlines levels of importance on a Likert scale of 1-5.

Figure 2: Likert scale based on levels of importance.



1	2	3	4	5
Not Important at All	Indifferent	Somewhat Important	Moderately Important	Very Important
19.4%	19.4%	36.1%	19.4%	5.7%

This result indicated that many in the Community do not see the mobile phone as a key tool for accessing data. This was another surprise because Juniper Research is predicting a greater demand in 2013 for data analytics solutions across mobile devices (Juniper Research, 2013).

Barriers and Potential Concerns to Adopting Big Data Products/Services

The two primary barriers to adopting big data services/products are the lack of skilled personnel and/or the lack of sufficient resources. In the Grey Literature community, the lack of sufficient resources proved to be the greatest barrier at 45.7%, with the lack of skilled personnel being a secondary barrier at 33.3%. Currently there is a shortage of personnel skilled in data curation, data integration and re-use, as well as data analysts (Schindel, 2012).

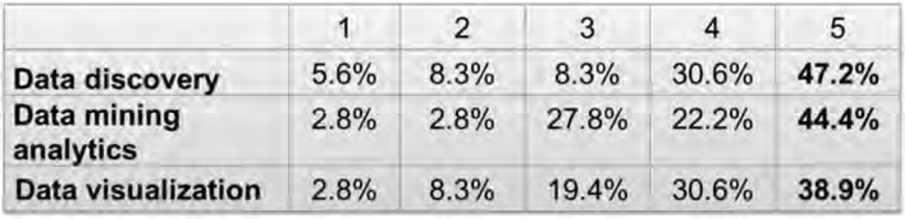
The barriers could conceivably be overcome if there were drivers/champions to led the cause of adopting these technologies/services, and to clearly identify the return on investments to management. Dennis Gannon, Director of Cloud Research Strategy pointed out that every area of science is now engaged in data-intensive research. Thus, the need for these technologies, products and services are not going away but will continue to increase (Gannon, 2012). As the volume of digital data continues to grow exponentially, there will be a continued need for skilled personnel and a need for adequate financial support. Both will be challenges the Grey Literature Community face in order to reap the benefits of big data analytical products and services.

Biased reporting has proved to be a leading issue to consider as well. Over 90% of respondents were concerned that biased reporting will be a cause of concern across multiple sectors (economic, political, health, scientific, social, etc.). From an historical perspective, an abundant amount of evidence supports this concern. For example, outcome biased reporting within the medical community has been a legitimate pubic concern when newly developed pharmaceuticals are trial tested then submitted to the Food and Drug Administration (FDA) for approval via New Drug Application (NDA) (Rising et al., 2008).

Big Data Goals

Results from the survey ranked the overall goals of utilizing Big Data products and services. The Grey Literature community favored data discovery (47.2%) with data mining analytics (44.4%) and data visualization (38.9%). The majority of responses emphasized the importance of these three categories by denoting a five on a 1-5 scale (Table 1).

Table 1: Survey responses on a 1-5 scale (1 being least important, 5 being most important) of which Big Data products/services would be most relevant to your organization's data goals. (N=36)



	1	2	3	4	5
Data discovery	5.6%	8.3%	8.3%	30.6%	**47.2%**
Data mining analytics	2.8%	2.8%	27.8%	22.2%	**44.4%**
Data visualization	2.8%	8.3%	19.4%	30.6%	**38.9%**

Summary of Survey and Future Considerations

Of the 48 survey takers, a trend in the responses revealed that respondents became increasingly impatient/distracted as the survey progressed. In the first third of the survey, responses were nearly 100% participation. Responses decreased by an average of 10 throughout the second third and plummeted by nearly 20 according to the last third of the survey. This indicated that the survey should have been shorter and the options for answers more limited. The survey administrators will take these findings into consideration for future surveys.

Overall, the Grey Literature respondents are keenly aware of the benefits of using Big Data services and products but have yet to identify people within their organizations as drivers/champions to make it reality. The lack of identified backers who can clearly and consistently make the case to show the immediate benefits and ultimate return on investments in these technologies and services to management has impacted their adoption or hindered their implementation. As indicated from the survey, the Grey Literature community is not using these products in substantial numbers nor have they seen these products/services demonstrated. Yet, the Community sees great value in these products/services for their local economy (>68% of survey takers), and they are planning to use these tools for web analytics, predictive analytics, and real-time analytics. Additionally, if the Community could select big data products/services for common data goals, they would select them first of all for data discovery and then for data mining analytics. Lastly, the lack of adequate financial resources is the greatest barrier to adopting these products/services.

In terms of future considerations, re-distributing the survey in three to five years may yield interesting responses as Big Data initiatives are readily explored. Additionally, as Big Data products and services mature, a better understanding of the developing landscape may reveal insight into trends that cannot yet be foreseen.

References

Gannon, D. (2012), Science as a service: Data Analytics and Data Mining - The Approaching Tidal Wave. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Gannon_Data_Analytics.pdf)

Juniper Research. 2012. Juniper Research's Top 10 Mobile Trends for 2013. (http://www.juniperresearch.com/shop/download_whitepaper.php?whitepaper=198)

Pederson, S (2012), Exploiting Big Data from the Deep Web: The new frontier for creating intelligence. BrightPlanet, Sioux Falls, South Dakota. White paper available (http://www.brightplanet.com/2012/07/creating-intelligence-from-big-data-whitepaper/)

Schindel, D.E. (2012), Data Curation: Skill-sets and Workforce Needs. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Schindel_Workforce_Needs.pdf)

Strawn, G.O. (2012), Big Data. – In: Proceedings of a CENDI/NFAIS/FEDLINK conference held the Library of Congress, Washington DC, Dec. 11, 2012 (http://cendi.gov/presentations/12_11_12_Strawn_Big_Data_Overview.pdf)

Rising K., P. Bacchetti, and L. Bero (2008), Reporting Bias in Drug Trials Submitted to the Food and Drug Administration: Review of Publication and Presentation. PLoS Med 5(11): e217.
– doi:10.1371/journal.pmed.0050217

Collection of Conference Proceedings and Improving Access to the Full Text of Proceedings

Misa Hayakawa, Shun Nagaya, Mayuki Gonda, Takeyasu Fukazawa, Minoru Yonezawa, and Keizo Itabashi, Japan Atomic Energy Agency, JAEA, Japan

Abstract

Conference Proceedings are "grey literature" due to the fact that they are not made commercially available frequently. While many Proceedings are published on the Internet, there are specific issues that can affect access, such as changes in the URLs. This paper introduces the case of the Japan Atomic Energy Agency (JAEA) library as an example of efforts to improve access of Proceedings using the Internet. JAEA Library uses the Internet to make available presentations by JAEA researchers. The paper notes that the conference secretariats tend to be temporary bodies and the links to conference websites are not permanent. The paper reports our investigation into these problems, and we introduce a new approach to provide access to these Proceedings.

1. Introduction

1.1 Conference Proceedings, a type of "grey literature"

Conference Proceedings are records of the papers and data which researchers submit for a specific theme at a conference or meeting. Proceedings contain information on the latest research trends, therefore, it is important information resources in the fields of science and technology.

These Proceedings papers are characterized by a shorter period between the writing and publication than the journal papers. In recent years, the period between writing and publication has been shortened even further by the Internet [1]. Due to this, the importance of Proceedings is growing.

Many researchers assume that the reliability of Proceedings is lower than journal articles. Nonetheless, many researchers use Proceedings as (1) points of reference in developing research topics, and (2) to investigate the latest research trends [2].

However, Proceedings are often distributed to the participants at a conference, and are often not commercially available. It is difficult for libraries to ensure a complete collection of Proceedings.

Proceedings have traditionally been published as conventional books. However, forms of publication have changed and many Proceedings are published as digital media, such as CD-ROM or Flash Memory, or are published only on the Internet.

Proceedings on the Internet may seem like an easy to collect, but there are in fact many challenges in terms of collection and use. For example, the websites on which Proceedings are usually published are owned and operated independently, and cross-searching these sites is difficult. Additionally, there is a problem of visibility as Proceedings on the Internet cannot be accessed when users do not know if the sites exist [3]. Furthermore, conference secretariats are often temporary bodies and the links to conference websites are not permanent. Changes to URLs happen frequently; a specific challenge relating to access to information on the Internet.

Proceedings are, therefore, still "grey literature" because they are difficult to search and use.

1.2 Purpose of this study

This paper discusses ways to improve access of Proceedings especially on the Internet.

The forms of publishing Proceedings are changing, and number of Proceedings available on the Internet is increasing. They are important as a means of providing the full text of papers to researchers. However, Proceedings published on the Internet cannot be accessed if their existence is not known, and currently it is difficult to search and use Proceedings papers.

We would like to introduce the case of the JAEA library as an example of improving access of full text Proceedings on the Internet.

We also address how the many sources of information about a meeting such as the conference website and papers on institutional repositories can be used to improve the access of Proceedings.

We introduce a new approach in providing conference information and linking to the full text Proceedings on the Internet.

2. Publication forms of Proceedings and library collections

2.1 Survey of the forms of Proceedings publications

We begin with an survey of the current forms of publishing Proceedings.

Fukazawa conducted a survey on the various forms of publishing Proceedings in 1982 and 1992. The survey classified Proceedings by publication forms such as books, journals, reports, and others [4]. 64% of Proceedings were published in journals or books.

We conducted our own survey to ascertain whether there had been any changes in the publication forms. We examined the Proceedings of international conferences in which researchers from the Japan Atomic Energy Agency (JAEA) participated in FY2010 (212 conferences). JAEA researchers mainly participate in conferences in the field of nuclear science.

We classified Proceedings forms as books, journals, reports, and other publications (Figure 1). Then, we classified books by mediums such as book form and CD-ROM (Figure 2).

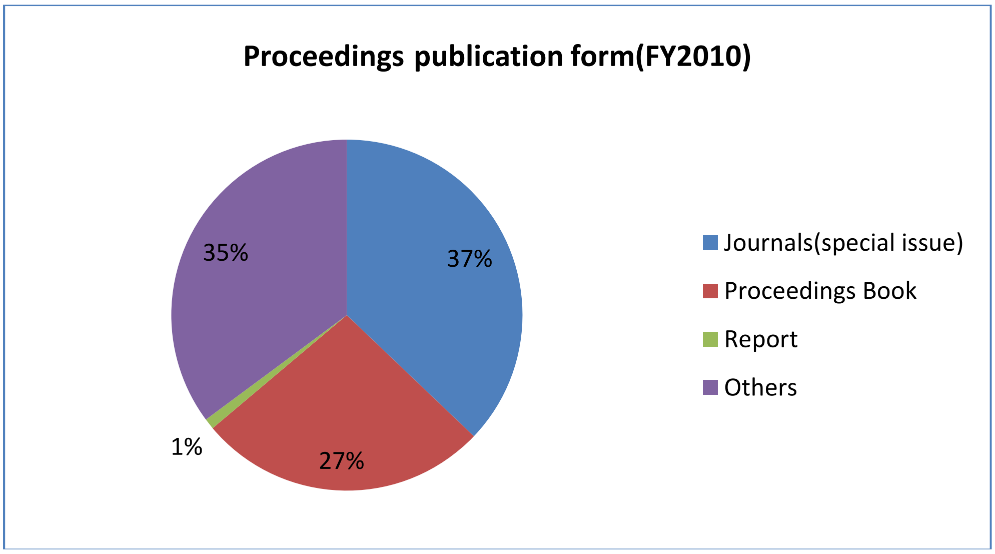


Figure 1: Proceedings publication form

We found that Proceedings are mainly published as journals (in special issues or as conference series). Next, Proceedings were commonly published in books. This is similar to the findings of Fukazawa's survey.

Incidentally, the "others" category includes cases where there was no form of publication of Proceedings (where, for example, only an abstract or program book were made available), and where there was an unknown publication form.

Proceedings published in journals are often commercially available, making them easy to access.

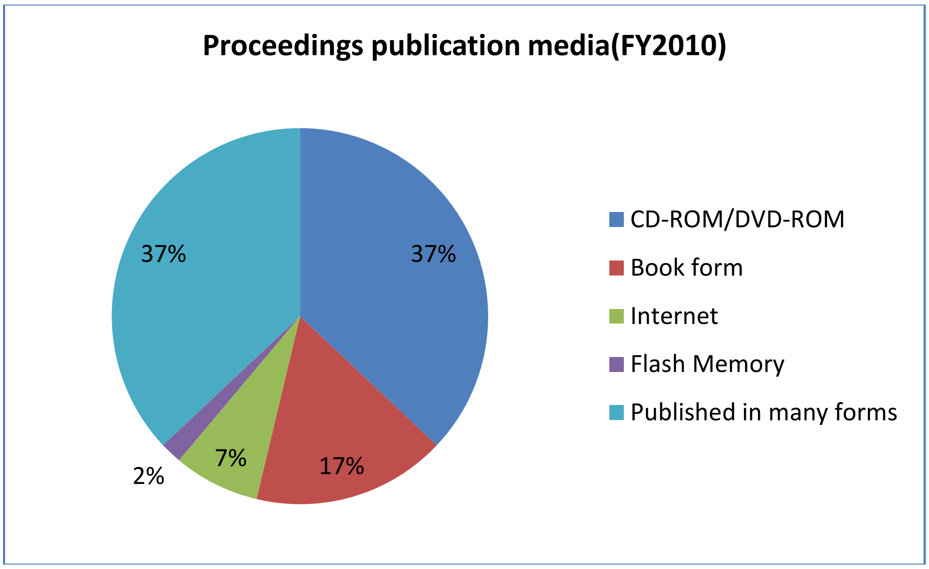


Figure 2: Proceedings publication media

We classified Proceedings by mediums, such as traditional books, CD-ROM/DVD-ROM, Internet sites, and Flash Memory. Proceedings published in two or more forms were classified as "Published in many forms" and duplication was not counted.

In all, 17% of the Proceedings were published as books, much less than the other forms. In comparison, 37% of the Proceedings were published as CD/DVD-ROMs.

Additionally, this survey shows many Proceedings are published more than one publication form. Many of these cases include Proceedings published on the Internet. Proceedings books or CDs were distributed to conference participants, and after the conference they were published on the Internet where they were available to buy. Published in a variety of forms in this way improves the availability of Proceedings.

Only 7% of the Proceedings were published solely on the Internet. In some cases, this publication form was open to the public, while some were only made available to the conference participants. This was, therefore, not necessarily a widely accessible form of providing access to the full text.

A few conference Proceedings were published as flash memory, although we do not dwell on this form of publication in this paper. Flash memory Proceedings were often distributed to conference participants at the conference venue, and pose a particular challenge in terms of collecting Proceedings by libraries. Additionally, it is difficult to preserve flash memory readability because they deteriorate and standards change. Where Proceedings published as flash memory do not conform to a meta-data system, it is also necessary to consider how Proceedings are made available by libraries.JAEA Library currently does provide a flash memory. However, to collect and provide flash memory still remains a challenging issue.

Proceedings published as CD-ROMs present similar problems to those associated with flash memory. For example, a Proceeding's CD-ROM published in 1996 created with Windows 95.CD-ROM also cannot be read when the operating systems or standards have changed. It is necessary to convert CD-ROM / DVD-ROM to up-to-date forms or to print.

2.2 Collection of Proceedings at the Library

We surveyed the library's collection of papers presented by JAEA researchers in FY2010 to confirm access of the Proceedings (not including those published as journals or reports) in libraries. We inspected the collections at the JAEA Library, the National Diet Library (NDL) [5], and CiNii Books [6] (to look for information about books and journals held by other university libraries in Japan). Additionally, we also investigated the availability of Proceedings on the Internet (Figure 3).

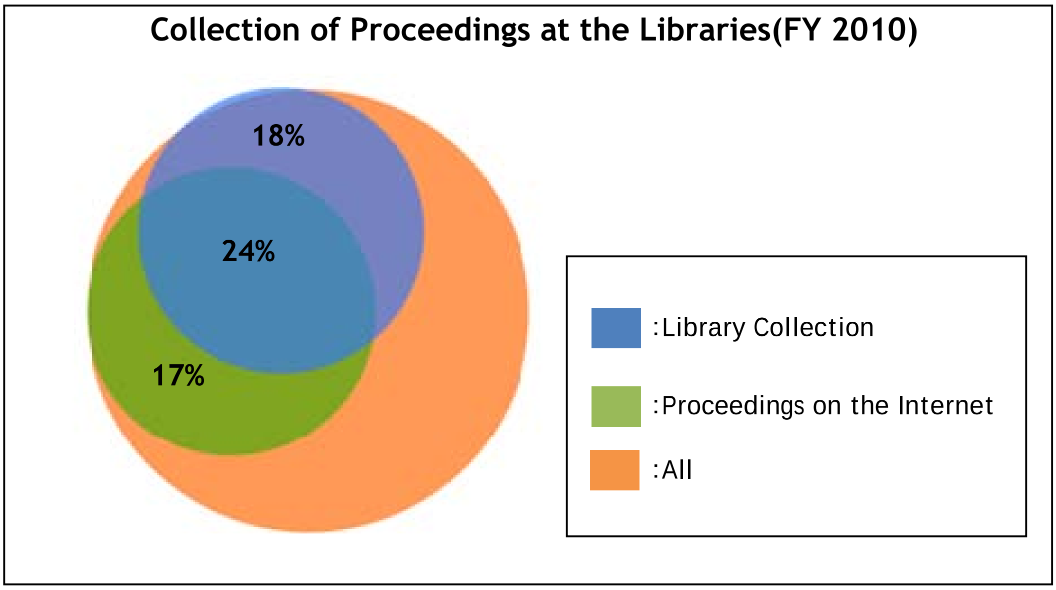


Figure 3: Collection of Proceedings at Libraries

The JAEA Library is one of the largest nuclear information centers in Japan, and we were collecting information on Proceedings specifically related to the field of nuclear science. We collected Proceedings from a variety of sources, including conference websites and from the researchers themselves.

However, we found that the JAEA Library together with the other libraries could provide only 42% of the Proceedings. As already mentioned, 17% of the Proceedings appear only on the Internet. This shows the difficulty of making Proceedings available in library collections. On the other hand, 41% of Proceedings were available as an Internet version. Providing links to the Proceedings on the Internet as well as the library's collections is important to improve access to Proceedings.

3. Management & providing links at JAEA

3.1 Providing links to Proceedings on the Internet within JAEA

As mentioned above, Proceedings on the Internet is increasing. Providing links to Internet Proceedings will improve access of Proceedings. We also advocated the management and provision of a system of the R&D results in JAEA as well as the links to the full text of the Proceedings.

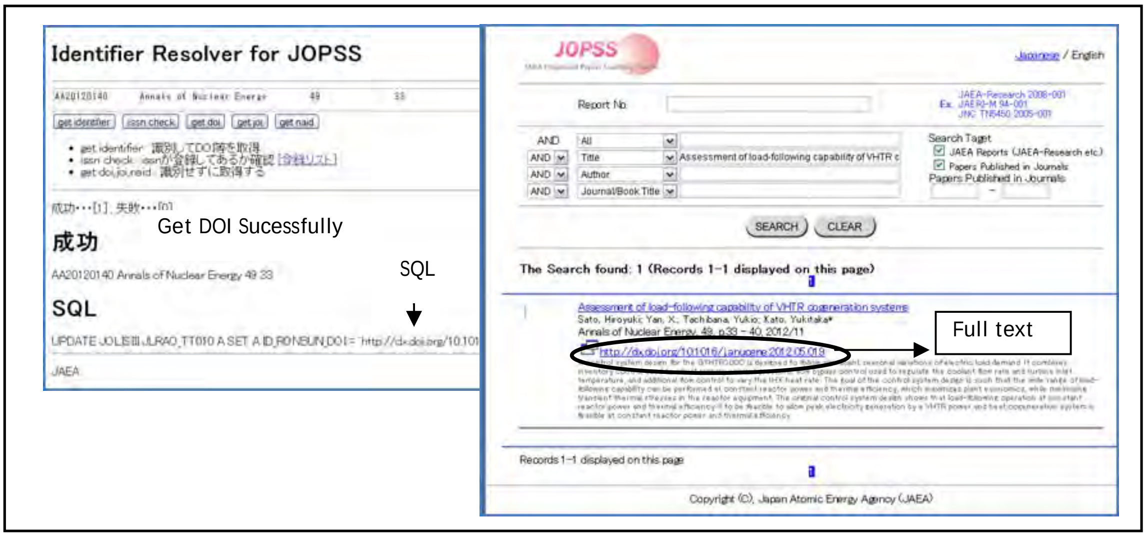
We manage the publication and presentation of information by JAEA researchers at the JAEA Library. At JAEA, researchers are requested to submit bibliographic data (title, author, journal/meeting name, etc.) via web forms before they undertake presentations or submit articles. Through this procedure, the library staff manage the information and produce an authority file of author names, conference names, and journal/book names.

The ability of this form of unified management by the library is particularly important where conference names have various notations.

Once the information is collated, we disseminate it via the Internet using the JAEA Originated Papers Searching System (JOPSS) [7].

Additionally, since July 2011 we have provided hyperlinks on JOPSS for the full-text versions of articles submitted by JAEA researchers. Providing links via JOPSS allows direct access to the Proceedings on the Internet.

The registration of the DOI/URLs is performed every month. Information is sent in the form of a hyperlink from a JAEA R&D result management system. The search and input of the DOI are then carried out using the Web system (Figure 4).



Identifier Resolver for JOPSS

AA20120140 Annals of Nuclear Energy 49 33

get identifier | issn check | get doi | get joi | get naid

- get identifier 識別してDOI等を取得
- issn check issnが登録してあるか確認 [登録リスト]
- get doi,joi,naid 識別せずに取得する

成功・・・[1], 失敗・・・[0]

成功

AA20120140 Annals of Nuclear Energy 49 33

SQL

UPDATE JOLIS Ⅲ.JLRAO_TT010 A SET A.ID_RONBUN_DOI = 'http://dx.doi.org/10.101

JAEA

JOPSS

Japanese / English

Report No

Ex. JAEA-Research 2008-001
JAERI-M 94-001
JNC TN8450 2005-001

AND	All	
AND	Title	Assessment of load-following capability of VHTR c
AND	Author	
AND	Journal/Book Title	

Search Taget
☑ JAEA Reports (JAEA-Research etc.)
☑ Papers Published in Journals
Papers Published in Journals
-

SEARCH CLEAR

The Search found: 1 (Records 1-1 displayed on this page)

Assessment of load-following capability of VHTR cogeneration systems
Sato, Hiroyuki; Yan, X.; Tachibana, Yukio; Kato, Yukitaka*
Annals of Nuclear Energy, 49, p.33 - 40, 2012/11
http://dx.doi.org/10.1016/j.anucene.2012.05.019

Records 1-1 displayed on this page

Copyright (C), Japan Atomic Energy Agency (JAEA)

Figure 4: Providing hyperlinks to full text

We currently provide access to 14,459 papers from conference. Among them, 3,137 are hyperlinked to the full-text version or abstract (as of October 2012). Past paper/presentation data will also include DOIs/ URLs. Additionally, conference information (conference name, venue, and date) also provide from JOPSS since September 2012.

However, few Proceedings papers published on the Internet have DOIs (except for those published as journal papers). For example, Proceedings papers published on conference websites almost never have DOIs. In these cases, we provided a URL for the paper (such as a PDF etc.). Additionally, even when full-text versions could not be accessed from the Internet, many conference websites provided abstracts of conference papers. We therefore provided links to the abstracts of conference papers when full-text versions of the Proceedings could not be accessed from the Internet.

3.2 Surveys of Proceedings on Internet/Conference websites

In recent years, many conferences have had dedicated websites which provide conference information (dates, venues, programs, abstracts) to the public and the conference participants. However, this information faces a challenge that is specific to the information on the Internet, i.e., the possibility that the access links to the URLs are either changed or the websites are not permanent.

We also, therefore, investigated access of the conference websites and the contents of websites for the conferences JAEA researchers participated in FY2010.

Access of conference websites with the URL registered before the meeting (Figure 5).

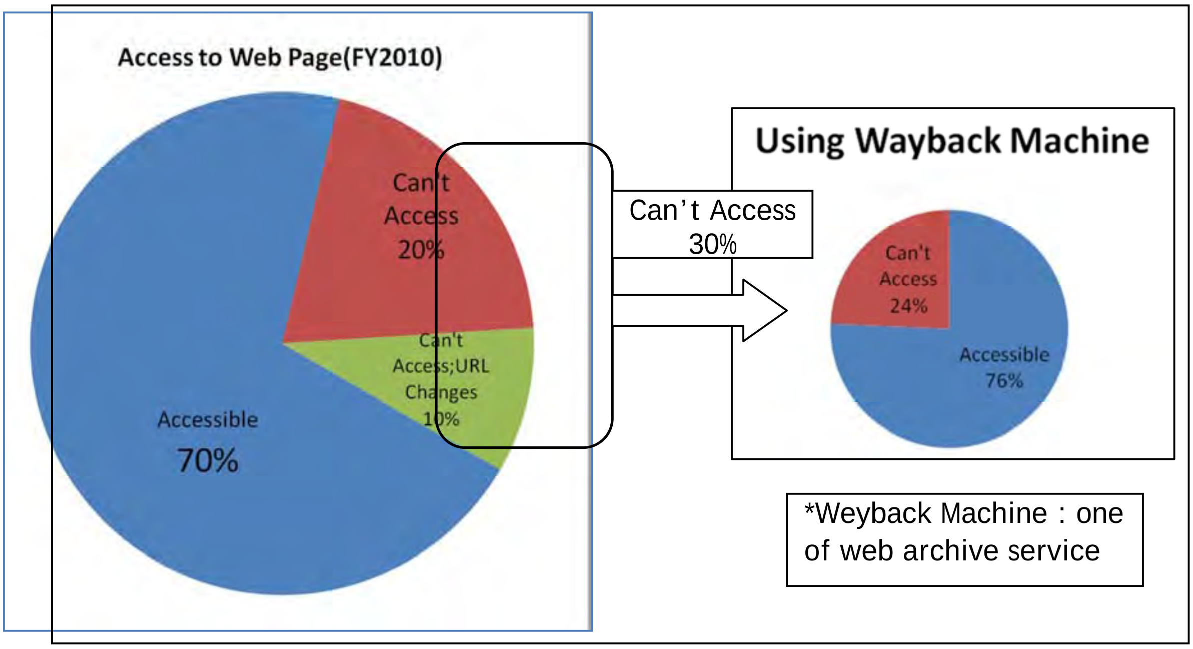


Figure 5: Access to conference webpage

Overall, 30% of the conference websites can no longer be accessed within 3 years. In all, 20% of the websites are no longer accessible because the website no longer exists, while 10% of the websites are not accessible because the URL has been changed. But 76% of these websites can access when use the web archive service.

Our investigation of website content showed that 10% of the websites provided the full text of the Proceedings, while 23% provided the abstracts. Some websites also provided presentation files and movie files, etc.

Some Proceedings available on the Internet were published on conference websites, while some were published on websites dealing with Proceedings in specific fields such as JACoW [8]. Moreover, some Proceedings available on the Internet were published on publisher websites.

Proceedings available on conference websites were published in a variety of formats, such as PDFs of each paper or PDFs of the whole Proceedings combined. Most of them were open to the public for no charge. Some, however, which were accessible only to conference participants. It required the ID and password to access details. These accounted for very few of the Proceedings, but did make it difficult to obtain full-text versions of the papers. Proceedings made available by publishers generally have DOIs, making them more accessible than other forms.

3.3 A new approach to providing Proceedings information

A large amount of conference information and a great number of Proceedings are available to the public on the Internet. The links to Proceedings and conference information on the Internet as well as in the library's collection is useful to improve access to the Proceedings. We, therefore, developed a prototype webpage that provides conference information, library collections of Proceedings, and Proceedings availability on the Internet using Google Calendar. Google Calendar is a free online calendar, and it is useful for information sharing and possible to integrate with other systems (Figure 6).

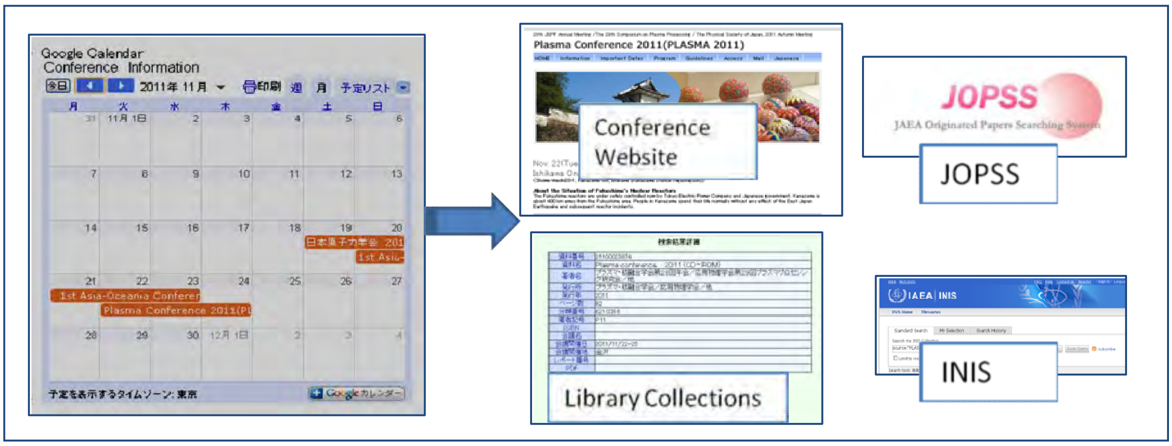

Google Calendar
Conference Information
2011年 11月
1st Asia-Oceania Conferen
Plasma Conference 2011(PL
Plasma Conference 2011(PLASMA 2011)
Conference Website
JOPSS
JAEA Originated Papers Searching System
JOPSS
IAEA INIS
INIS
Library Collections

Figure 6: Google Calendar to Internet sources

One of our initial difficulties was in providing conference information where the websites had been deleted within a few years of the conference. We found that it was possible to solve this problem by using the web archives that has developed in recent years. The Web archives collect Internet web pages, archives them, and makes them available to the public. We used the Internet Archive's Wayback Machine [9] to provide conference information from the deleted conference websites. Wayback Machine is a major Internet archive, offering access to Web pages from 1996. As mentioned above, web archives will improve access to deleted conference web pages.

Additionally, we use library collections information from the JAEA Library/NDL/CiNii Books. This covers most of the information on Proceedings in Japan.

Some of the Proceedings papers were included in institutional repositories by the authors. In recent years, institutional repositories have become popular, and some institutional repositories include Proceedings not available anywhere else. For this reason we also provided the hyperlink to the search results of Japanese Institutional Repositories Online(JAIRO) as a way of increasing the availability of conference Proceedings papers. JAIRO includes 11,888 of presentation files (2010/10).

We also provide hyperlinks to the search results of the International Nuclear Information System (INIS), which includes information from conference Proceedings in the fields of nuclear science and technology. INIS hosts the world's largest collection of published information on nuclear science and technology [10], and is appropriate for conference information relating to the participation of JAEA researchers.

We have registered conference information from FY2010.

This conference information on Google calendar is not yet open to the public, but we want to improve the resource and make it publicly available.

4. Summary

There are a variety of forms of publishing Proceedings and the number of Proceedings available on the Internet is increasing. Using links to Proceedings on the Internet as well as library collections can improve the access of the Proceedings. However, websites are not permanent and changes in URLs are frequent. The URLs of Proceedings on the Internet, especially those on conference websites, are particularly susceptible to change. When we provide links to Proceedings, we need to, therefore, undertake periodic checks on the links. Web archive also improve access to web pages when URL changes or deleted pages.

Additionally, some of the Proceedings available on the Internet are not open to the public, making it difficult to access the full text. These accounted for very few of the Proceedings, but we have to consider how to provide these Proceedings.

In recent years, some of the Proceedings papers were included in institutional repositories by the authors. Institutional repositories improve access to full-text Proceedings that are not available anywhere else.

There are a large number of information sources on the Internet. We aim to improve the access of the Proceedings and aim to provide access to this information.

References

[1] Tsuda, Yoshiomi. Conference Proceedings and materials. The present state of confrence Proceedings and materials. Jyoho no Kagaku to Gijyutu. 2005, 55(5), p.208-213. (Japanese)

[2] Toyoda, Yuji. Possibilities for effective use and distribution of preprints. Jyoho no Kagaku to Gijyutu. 1998, 48(6), p.336-343. (Japanese)

[3] Ikeda, Kiyoshi. Grey literature. Overview: rethinking the grey literature's definition. Jyoho no Kagaku to Gijyutu. 2012, 62(2), p.50-54. (Japanese)

[4] Fukazawa, Takeyasu. Analysis of the published Proceedings of international conference-Scientific conference held in Japan. Jyoho no Kagaku to Gijyutu. 1993, 43(10), p.913-915. (Japanese)

[5]NDL OPAC. https://ndlopac.ndl.go.jp/

[6]CiNii Books. http://ci.nii.ac.jp/books/

[7]JOPSS. http://jolissrch-inter.tokai-sc.jaea.go.jp/search/servlet/interSearch

[8]JACoW. http://www.jacow.org/

[9] Wayback Machine. http://archive.org/web/web.php

[10]INIS. http://www.iaea.org/inis/

Grey Literature, E-Repositories, and Evaluation of Academic and Research Institutes: The case study of BPI e-repository

Maria V. Kitsiou and Vasileios Souvlidis
Benaki Phytopathological Institute, Greece

Abstract

E-repositories are internet databases, in which the whole intellectual property produced by an Educational Foundation or a Research Institute can be gathered, classified, reserved and, of course, disseminated.

It is known that e-repositories are based on Open Access and Knowledge Dissemination concept, providing access-without restriction- to scientific information. But, it is not known that e-repositories can be proved useful and usable "tools" for the evaluation of Universities and Research Institutes.

In this study, we present the case of the BPI e-repository. The Library of the Benaki Phytopathological Institute (BPI) realizing the importance of the evaluation – necessary for funded projects- developed the BPI e-repository.

BPI e-repository has been planned and structured so in order to achieve the aims of the evaluation simultaneously in an "institution level", in "scientific department level" (common intellectual responsibility) and in "researcher level" (individual intellectual responsibility). The scientific material contained in it has been categorized so in order to responds to the organizational structure and function of BPI and, also, to highlights in the best way the research activities of the BPI scientific community members. Its implementation based on DSpace 1.7.0, open source software suitable for digital archives management, that uses OAI-PHM (Open Archives Initiative Protocol for Metadata Harvesting) and Dublin Core Standard, that is suitable for the documents description. The flexibility of DSpace software allowed us to implement changes into the repository system, like fields addition. So, these changes made the description of all documents in different types (books or book chapters, papers, dissertations, BPI publications, patents, conferences, technical reports etc) more efficient and complete.

Navigation and search function in an user-friendly interface using five diverse ways, i.e. type of material, title, author, year and subject. We note that we implement a simple bilingual subject standardization system. The keywords used by the authors themselves in their papers or conferences proceedings make the subject standardization procedure and indexing difficult. Nevertheless, we are aiming at the implementation of an optimized bilingual subject index adopting and implementing standards such as NLGSH or/and LC.

The appropriate use and exploitation of the search results via e-repository – according to the experience having been acquired by this e-repository development process and function- can lead certainly to faithful conclusions concerning the evaluation of an Educational Foundation or a Research Institute.

Keywords: Grey Literature, E-Repositories, Open Access, Evaluation, Universities, Research Institutes

Introduction

The Library of the Benaki Phytopathological Institute[1] (BPI) has been in operation since 1930. It is a specialized scientific Library dedicated to the subject of agriculture. In particular, it covers subjects of phytopathology, plant protection, plant health, botany, entomology, agricultural zoology, weed science, phytopharmacy, pesticides control, eco-toxicology, environmental protection, crops, food technology, food quality and integrated management systems of rural production. It intends not only to support and enhance the research activities of both the BPI community members and the external users, but also to exceed their expectations.

Its collection consists of books, printed and electronic journals, maps, grey literature, databases and audiovisual material.

As regard as grey literature, and realizing the importance of the internal evaluation of the Institute, the Library of BPI has developed the BPI e-repository[2].

[1] Established in 1929, Benaki Phytopathological Institute was the first Greek research institute to have a broad scientific basis in plant health, plant protection, risk assessment and safe usage of agricultural chemicals with regards to the protection of human health and that of the environment more generally. BPI, also, performs analytical work and mandated checks to ensure the safety of foodstuffs and other agricultural products. It is a Legal Entity of the Public Sector, donated by the National Benefactor Emmanouel Benakis and it operates under the supervision of the Hellenic Ministry of Rural Development and Food. Since January 2008, BPI has been publishing the semi-annual scientific journal, the Hellenic Plant Protection Journal (HPPJ), formerly Annals of the Benaki Phytopathological Institute.

[2] BPI e-repository is accessible via both website of BPI Library http://83.235.16.144:8080/jspuien and www.openarchives.gr

But, how BPI e-repository has been planned and structured? What is the type of material contained in it? What are the diverse *communities* and *collections*? What are its technical characteristics? How many persons have been working for its implementation? and – the most important- how is it related to the internal evaluation of BPI? How the e-repository can be used as a useful and usable "tool" for the internal evaluation of an Institute?

BPI e-repository

BPI e-repository has been in operation since January 2012. It has been planned and structured so in order to achieve a twofold aim, i.e. the *preservation* and *dissemination* of the BPI intellectual property and the *internal evaluation* of BPI, simultaneously in an "institution level", in a "scientific department level" (common intellectual responsibility) and in a "researcher level" (individual intellectual responsibility). Benaki Phytopathological Institute consists of three (four in the past) scientific departments "Department of Phytopathology, "Department of Entomology & Agricultural Zoology" and "Department of Pesticides Control & Phytopharmacy", and these are obvious if somebody choose "Author" from the main menu at the home page and then select the letter "D" from the index "browsing by author". In this way, the intellectual property of each scientific department can be easily evaluated. Also, the same results can be retrieved, if the options "search" and "advanced search" are used.

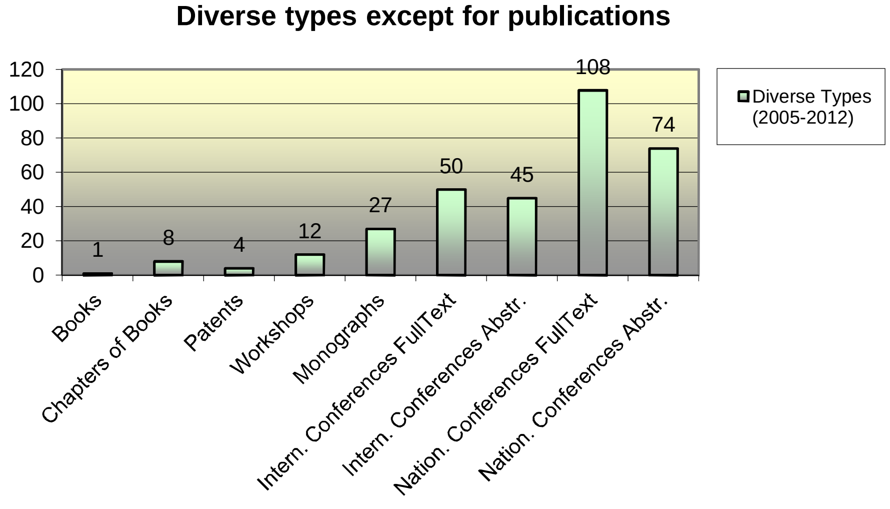
The scientific material contained in the e-repository has been *categorized* so in order to respond to the organizational structure and function of BPI and, also, to highlight in the best way the research activities of the BPI scientific community members. More specific, there are -up to present- eleven (11) *communities* and thirty two (32) *collections*:

- Books
- BPI Annual Reports
- BPI Editions
- Conferences
- Funded Projects
- Meetings/ Seminars
- Monographs
- MSc, Phd, Post Doc
- Patents
- Publications
- Technical Reports

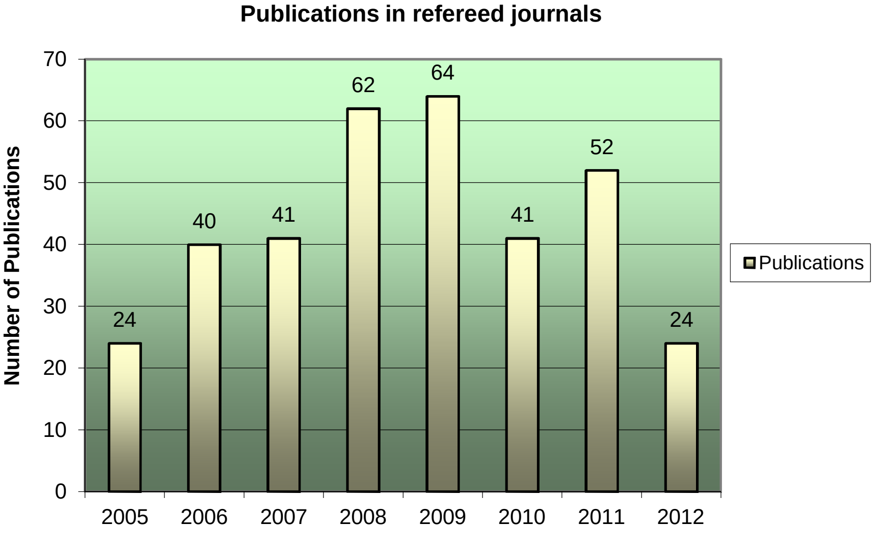
For example, the *community* "Books" consists of two (2) *collections* "Book Chapters" and "Books". Also, the *community* "Publications" consists of four (4) *collections* 1."Publications in foreign scientific journals (non refereed)", 2. "Publications in foreign scientific journals (refereed)", 3."Publications in greek non-scientific journals" and 4."Publications in greek scientific journals". Additionally, several *collections* are contained in the *community* "BPI Editions", like BPI books, Speeches, Guidelines, Newsletters and Technical Bulletins. As regard as the *community* "Conferences", there are six (6) *collections*, i.e. 1. "International Conferences-fulltext", 2. "International Conferences-abstract only", 3. "International Conferences-posters", 4. "National Conferences-fulltext", 5. "National Conferences-abstract only" and 6. "National Conferences-posters".

Shortly, one more *community* "Photographs" or "Photographic Material" is predicted to be added. Its *collections* ("photos of insects", "photos of plants", "photos of plant diseases" etc) will be included too. It is notable that this photographic material is original and it has been produced by the researchers of BPI during their research activities.

Up to now, there are totally 516 *records*. The graphs below represent the numbers of the *records* in some *communities* and *collections*. The numbers –as the note shows- concern only the period 2005 – 2012.



Graph1: Diverse types except for publications 2005-2012



Graph 2: Publications in refereed journals 2005-2012

As regard as the technical characteristics, the implementation of the BPI e-repository is based on *DSpace 1.7.0,* an open source software suitable for digital archives management, that uses the *OAI-PHM* (Open Archives Initiative Protocol for Metadata Harvesting) and the *Dublin Core Standard*, that is suitable for the documents description. Owing to the inefficiency of the *Dublin Core*, concerning the number and the diversity of the fields, it was needed to be several modifications, like fields addition. So, these changes made the description of all documents in the different types (books or book chapters, publications, dissertations, BPI editions, BPI annual reports, patents, conferences, technical reports etc.) more efficient and complete.

A common *record* consists of the following *fields*:

- author(s)
- title
- date of issue
- publisher
- citation
- identifiers (ISSN, ISBN etc.)
- type
- language
- subjects
- abstract
- description
- file attachment

With respect to the *field* “Author”, 650 authors names have already been recorded. During the recording, it was realized that the name of an author could be appeared in diverse types in different issues.

So, it was decided to be implemented a standardization system (*authorities*) in order the double records to be avoided. This means that only one name type is used for each author, the authorized one. For example “*Kitsiou, M. V.*”

Concerning the *field* “Subject”, there are almost 1500 subjects headings. It is notable that the bilingual (greek-english) subject standardization system that used, is so simple. The keywords used by the authors themselves in their papers or conferences proceedings make the subject standardization procedure and indexing difficult. Nevertheless, we are aiming at the implementation of an optimized bilingual subject index adopting and implementing *standards such as NLGSH (National Library of Greece Subject Headings), LCSH (Library of Congress Subject Headings)* or *Agrovoc thesaurus.*

The *"Browse"* and "*Search"* function in an user-friendly interface in which five diverse options, i.e. *communities/collections*, *issue date*, *author*, *title* and *subject* are available. At this point, we should refer that due to a technical problem –at the first stages- there was a quantitative difference between the results retrieved via *Browse* and the results retrieved via *Search* concerning the same keyword. We realized that *DSpace* uses two different indexes. Although, the problem was easy to be solved.

A librarian and a computer technician were the only persons needed in order the BPI e-repository be structured.

Conclusion

The appropriate use and exploitation of the search results via e-repository – according to the experience having been acquired by this e-repository development process and function- can lead certainly to faithful conclusions concerning the evaluation of an Educational Foundation or a Research Institute

Acknowledgment: The authors would like to express their gratitude to all contributors.

Bibliography

- Dietz, P. (2011) Dspace 1.7.0 System Documentation. Manual
- Hussos, N. K. et al. (2010) "Successful interoperability case studies in greek repositories and relative technological tools", Proceedings of the 19th National Conference of Greek Academic Libraries, pp 87-105, (3-5 November 2010, Panteion University, Athens, Greece)
- Kounoudes A. et al. (2010) "The way to the open access through Creative Commons. The case study of Ktisis" Proceedings of the 19th National Conference of Greek Academic Libraries, pp 319- 334), (3-5 November 2010, Panteion University, Athens, Greece)
- http://dspace.mit.edu/ (last access date January 9, 2013)
- http://ktisis.cut.ac.cy/ (last access date January 9, 2013)
- http://dspace.lib.uom.gr/ (last access date January 9, 2013)

Innovation, language, and the web

Claudia Marzi
Institute for Computational Linguistics, “Antonio Zampolli”,
CNR, National Research Council, Italy

Abstract

Language and innovation are inseparable. Language conveys ideas which are essential in innovation, establishes the most immediate connections with our conceptualisation of the outside world, and provides the building blocks for communication. Every linguistic choice is necessarily meaningful, and it involves the parallel construction of form and meaning. From this perspective, language is a dynamic knowledge construction process. In this article, emphasis will be laid on investigating how words are used to describe innovation, and how innovation topics can influence word usage and collocational behaviour. The lexical representation of innovative knowledge in a context-based approach is closely related to the representation of knowledge itself, and gives the opportunity to reduce the gap between knowledge representation and knowledge understanding. This will bring into focus the dynamic interplay between lexical creativity and innovative pragmatic contexts, and the necessity for a dynamic semantic shift from context-driven vagueness to domain-driven specialisation.

Keywords: *Lexical productivity, Language technologies, Web corpora, Grey Literature.*

1. Introduction

Understanding the relationship between language and innovation is connected with understanding that language determines what can and what cannot be talked about, and therefore what can be achieved and what cannot.

Language conveys innovative ideas and gives body to knowledge itself. Language is the common ability of our species that makes us shaping and referring to things, events, and concepts with remarkable precision. This common communicative ability connects people into an information-sharing network, and information allows people to expand their knowledge.

The importance of efficiently deploying knowledge for a complete and successful exchange is easily understandable: through a better understanding of information new ideas can be captured and exploited.

Knowledge transfer and innovation transfer are nowadays ubiquitous processes. The entire system of knowledge refers to knowledge creation and application, by defining a process going form acquisition and sharing to transfer and application. Knowledge extraction involves heterogeneous tasks related to the acquisition, from unstructured textual data in digital format, of structured and classified information relating to research topics. Nonetheless, far from being readily or easily transferred from the originator to the user of a technology, knowledge faces barriers, such as ambiguity, difficulty to be interpreted and absorbed, difficulty to be retrieved.

The spread of Internet has enabled development of better bibliographic scientific databases with significantly improved capacity for storage and retrieval. In recent years, web searching has become the default mode of highly innovative information retrieval, though the main sources of digital information are unstructured or semi-structured documents. Information relating to developments in scientific research is collected in the form of abstracts or full publications, in large and growing bibliographic repositories.

Knowledge ambiguity may also depend on language ambiguity; a shared language is part of a related language, and “for knowledge to be exchanged and combined, there has to be a shared medium of communication” (Hedlund, 1999:11). Language has effects at all stages of knowledge transfer. Language generates on-going impacts beyond a simple knowledge transfer act, and it is simultaneously an active agent in the knowledge transfer process itself.

Our goal is to focus on how words and language structures become vehicle for knowledge generation, and in particular for innovation transfer, and what kind of infrastructural support can enhance innovative knowledge transfer.

2. Background

2.1. Language of innovation and linguistic innovation

In considering the language of innovation, particular emphasis is laid on overlap and dissonance in terminology and word formation processes associated with innovation. The language of innovation suffers from a problem of lexical overabundance. Not only many words are offered, but different authors define or use these words in different ways; and this because the challenge is the complexity of categorising innovation.

Many terms are used to describe innovation (Linton, 2009, for an overview), and innovation itself offers new collocations and extension of use. Most of the differences in terminology can be accounted for by differences in perspective and domain.

Language change is a fundamental evolutionary phenomenon due to many natural, cultural and historical factors. In particular, we are interested here in shedding light on the phenomenon of terminology innovation and propagation of those novel forms across domains.

Through language cultural novelty can be transmitted vertically (from parents to children), horizontally (from peer to peer) as well as across generational gaps. Although many linguistic innovations fall into the category of what Andersen (1989) has called "fortuitous innovations" (i.e. spontaneous and purposeless innovations as the results of non-functional, non-intentional copying errors), in many cases the very structure of the innovation can be explained with reference to the speakers' (synchronic) perception and meaningful re-interpretation of linguistic surface forms within the pragmatics of the situation. Meaning itself is a consequence of interaction and context.

The meaning of words can change over time and discourse and, in particular, words can take on new senses when used in novel contexts. Words with emergent novel senses often reflect an extension of use from one domain to another. In this sense, linguistic innovations arise in the context of existing rules which they modify.

2.2. Language ambiguity

Many words have more than one sense, and each of their sense is reflected by their distribution across contexts. Lexical semanticists make a classical distinction between semantically ambiguous words (e.g. *bank*), whose meanings can vary unsystematically, and polysemous words (e.g. *school*), where different senses exhibit a predictable relationship. Both ambiguous words and polysemous words can be disambiguated by defining their context of use. Polysemous words, in particular, can shape their meaning as a function of their context of use. As a consequence of this context-sensitivity, if a polysemous word-type appears more times in the same text (e.g. a single document), it is extremely likely that its different tokens will share the same sense. Although people do not need too much context to perform a disambiguation task, in Natural Language Processing (NLP) and Information Retrieval (IR) larger contexts make the task easier. Methods for computing relational similarities and disambiguating polysemous words, based on large text corpora, can make rough sense distinctions. Although this is far from reaching the sophistication of human judgement, the field is making considerable progress in context-sensitive word sense disambiguation and in the identification of conceptual relations between words.

In the following sections, corpus-based investigations are analysed in the perspective of a lexical representation of innovative knowledge. In the field of corpus linguistics advantages and limits of various corpora are analysed, depending on both linguistic and innovative knowledge research questions.

2.3. Corpus linguistics

Corpus-based linguistics is the study of language on the basis of large text samples – the corpora.

A corpus, as defined by Sinclair (1999: 171), "is a collection of naturally occurring language text, chosen to characterize a state of variety of a language. In modern computational linguistics, a corpus typically contains many millions of words: this is because it is recognized that the creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are clues to the lexical structure of the language". Although it must be considered that an appropriate size of corpus is strongly dependent on the phenomenon to investigate and the purpose itself. Another factor influencing the size of corpora relates to the degree of internal variation in the language or genre under study (Meyer, 2002).

In any case, corpora are incomplete. Rather, the issue is whether they are representative of the inquired phenomena; in other words, a corpus should be large enough to give an adequate representation of the language and more occurrences of the elements under investigation.

The importance of findings, either quantitative or qualitative, depends on the representativeness of the selected corpus for the research question.

Dealing with machine-readable texts offers the basis for purpose-specific research questions.

Occurrence, distribution, and importance, are different issues to be taken into account.

Salient domain-specific concepts and relations are most often conveyed in text through statistically significant terms. Rare words often denote the most salient pieces of content information of a document together with its level of subject-specificity. Recurrent word combinations are defined as COLLOCATIONS.

In corpus linguistics, collocation defines a sequence of words or terms that co-occur more often than would be expected by chance. An example of a phraseological (multi-word expression) collocation is the expression *strong tea*. While the same meaning could be conveyed by the roughly equivalent **powerful tea*, this expression is considered incorrect by English speakers. Conversely, the corresponding expression for computer, *powerful computers* is preferred over **strong computers*. Unlike idioms, collocations have a rather transparent meaning and are easy to decode; yet they are difficult to encode – like idioms – since they are unpredictable for non-native speakers, and moreover they do not preserve the meaning of all their components across languages. Collocations are flexible and they can involve two, three or more words in various ways.

In NLP collocational information derived from corpora is useful in the perspective of text analysis; for instance, in word sense disambiguation collocations are used to discriminate between senses of polysemous words. Frequently occurring collocates give the idea of semantic preference.

Corpus data can be considered as very useful for revealing typically lexico-grammatical patterns and functional aspects of language. Corpus-based studies on word formation show that productivity of derivational suffixes are more pertinent in certain kinds of texts than others. In other words, register variation plays a salient role in word formation. It's, however, essential to state that register distinctions are not defined in linguistic terms, but rest on context, domain, and purpose; and that contextual knowledge allows to support knowledge processes and to better access them.

2.4. The Web as a corpus

The World Wide Web has become a primary meeting place for information and communication, and it provides texts to be mined for lexicographers and linguists. As the web is constantly expanding, it represents an unlimited universe of information and data, and offers ubiquitous accessible information, and large volumes of information are available. Increasingly, corpus linguists have begun using the World Wide Web as a corpus for linguistic analyses.

The Web as a linguistic corpus makes it possible to investigate how words are used to describe innovation, and how innovation topics can influence word usage and collocational behaviour.

As a source of machine readable texts for corpus linguists and researchers in the fields of NLP, IR and Text Mining, the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness. The Web and associated technologies have been both the catalyst for much linguistic creativity and the main vehicle for its dissemination. In contrast, any static corpus is cut off at the moment of its compilation.

However, the web is a particular kind of corpus, as an estimation of its size and especially its composition cannot always be assessed. Moreover, it must be seriously kept into consideration that investigation of selective corpora is better concerned with the description of use and structure of domain-specific language, by inquiring linguistic phenomena, such as co-occurrence distributions, collocational variability, derivational productivity, neologism coinage. While the notion of linguistic corpus as a body of texts rests on some related issues such as finite size, balance, permanence, the very idea of a web of texts brings about notions not only of flexibility but even of non-finiteness and provisionality.

How can data gathered from the web provide new insight into language usage? The Web as a corpus is a rich source of freely available linguistic data covering a lot of topics (Fellbaum 2005), but despite the great advantages in quantity and accessibility, there is no control for example on web posting, and especially concerning English data, a lot of data are posted by non-native speakers. In this sense, the language used on the Web does not represent thoroughly the standard usage. While statistically robust analysis of Web data to discover collocations can give a flavour of what Manning and Schütze (1999) defined as "a conventional way of saying things" by marking the most frequent expressions, high frequent occurrences cannot give a disambiguation of context usage and sense. Moreover, in using the web as a corpus especially when it is accessed through generic search engine, it is virtually impossible to replicate a test on the same data.

In short, the Web offers a huge repository of documents written in a multitude of – more or less standard – languages, of different types or genre, and constantly changing over time, though often helpless in telling intended purposes and in offering background and contextual knowledge.

In what follows we propose to approach the tight inter-relationship between sense extension, context and innovation, by exploring the usage of words in contexts with NLP technologies. In Computational Linguistics and Computational Lexicology, sense identification and words sense disambiguation are commonly modelled by focusing on the distributional similarity of word usages in context. These techniques provide a key to a deeper understanding of a constructive view of lexical meaning as the by-product of the interaction of a word with its surrounding context (i.e. its collocates), and represent the methodological basis of the ensuing analysis.

3. Methodology and experimental evidence

3.1. Method and materials

The challenge of identifying changes in word sense has only recently been considered in Computational Linguistics.

To investigate the themes discussed in the previous sections genre-oriented and stylistically heterogeneous English texts are analysed, with the support of SKETCH ENGINE (Kilgarriff et al., 2004), which is a corpus query tool, based on a distributed infrastructure, that generates *word sketches* and *thesauri* which specify similarities and differences between near-synonyms. By selecting a collocate of interest in a *sketched* word, the user is taken to a concordance of the corpus evidence giving rise to that collocate.

Ambiguous and polysemous words have been selected with particular reference to innovative domains, and their collocations are analysed. In particular, we considered the domain of brain sciences and new technologies of brain functional imaging, the domain of knowledge management processes, and the field of information technologies, by mainly focusing on the following test words: IMAGING, RETENTION, STORAGE, CORPUS, NETWORK, GRID.

The selected words present a potentially high degree of semantic ambiguity or polysemy and different degrees of semantic specialisation, which can be analysed objectively by studying their context collocations.

For a terminology exploration, both domain-specific and general-purpose texts materials are selected by using generic search web engine queries (www.google.com by using seed words), domain-specific databases and type coherent multidisciplinary large corpora (e.g. www.opengrey.eu, www.ncbi.nlm.nih.gov/pubmed by selecting the domain). Collocations and concordances are then compared with large balanced corpora (e.g. the British National Corpus, British Academic Written English, New Model Corpus, and the like, whose size ranges between 8 M and 12 G tokens).

3.2. Results

By comparing in different contexts of use collocates and keywords selected from reference corpora – both specific and generic – with simple keywords search in web context, we investigate the ambiguity vs. polysemy gradient, showing how dynamically word meanings are adjusted to novel usage, and how difficult it could be to disambiguate polysemy words without a predefinition of the specific context.

All six terms exhibit distinct senses when they are used in different domains/discourse contexts. However, the extent to which different senses of the same term are mutually related differs considerably from one term to another. The two different usages of Latinate CORPUS as referring to the medical domain, and in particular to brain areas (e.g. CORPUS CALLOSUM, CORPUS STRIATUM) as opposed to large collection of items in a general sense and to large collections of texts/specimens in the Humanities are related only etymologically, with no systematic sense shift or extension. Moreover, the use of CORPUS as 'collection' is by far more widely-spread than its (neuro)- anatomical sense. If we are not in a position to constrain automatic word search within particular text domains, the medical usage of CORPUS is likely to be severely under-represented, flooded by the vast majority of examples of the more generic sense.

In the case of IMAGING on the other hand, the prevalent use of the polysemous term in connection with the medical domain, as referred to specific diagnostic technology, is the result of the application of a general-purpose technology to a specific domain. Technical and scientific bibliographic databases present only this collocate, while by selecting it as a seed word in a web engine search, the more frequent collocation is the one referred to generic visual representations. The selected balanced corpora, on the other hand, exhibit both of them, with a higher frequency of occurrences in the medical domain, in particular in the brain sciences.

NETWORK and GRID represent somewhat extreme cases of such domain-sensitive specialisation, to the point that they appear to be overwhelmingly used in their specialised senses only. NETWORK and GRID are expression of the very popular domain of information technology, and even though related to innovation, they are identified as related to this specific context, exhibiting a very coherent collocation behaviour.

Finally, RETENTION and STORAGE appear to oscillate between their proper and extended senses interchangeably, thus witnessing a paradigmatic case of systematic, context-sensitive polysemy.

RETENTION can select both material and immaterial items, but in the bibliographical references system on Grey Literature presents collocations in the domain of information and knowledge retention, as part of the whole process of acquisition, storage and retrieval, whereas in biomedical scientific database is specialised for fluid retention. STORAGE can make reference to containing units, either physical or computational, to long term memory capacity, related to either computational or cognitive processing. Moreover, the storage process is part of the above mentioned knowledge process. Thus, reference to the storage process by itself does not disambiguate the object to be stored.

In figure 1, logarithmic relative frequency distributions across domains are plotted for the test words. Firstly, for each of the selected words, frequencies – expressed as a relative percentage of occurrences in a balanced corpus - are taken and ranked for the diverse domains. Secondly, the top ranked domains are selected and presented together.

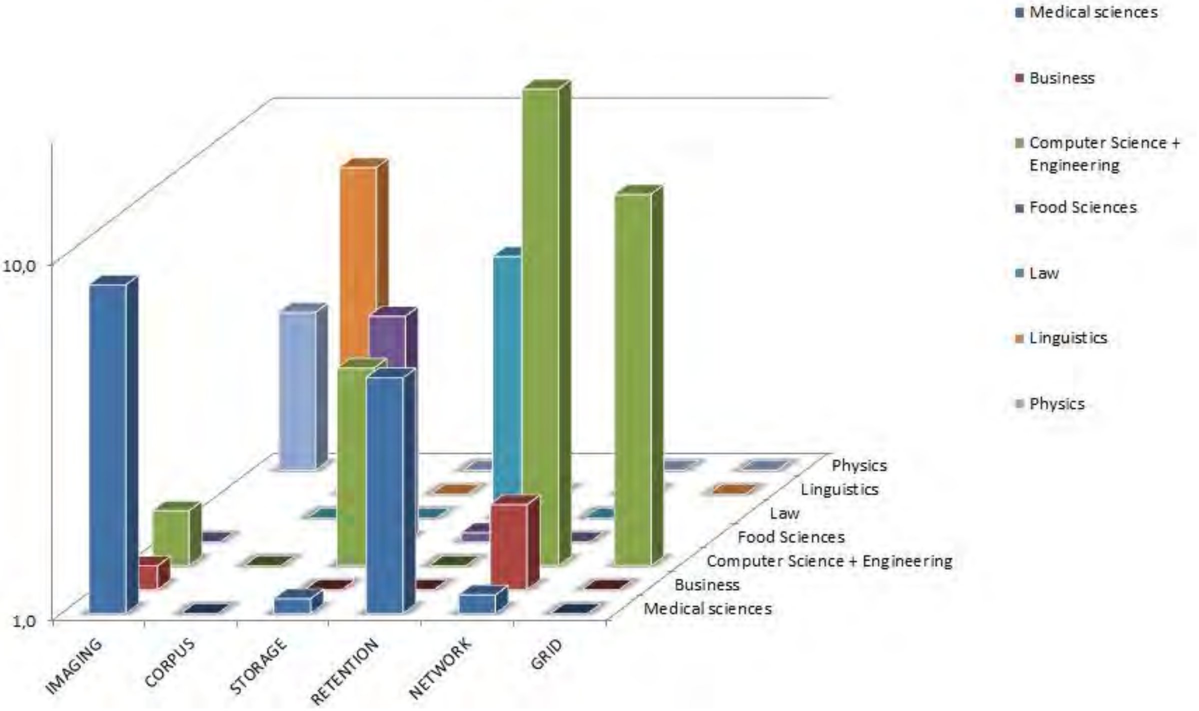
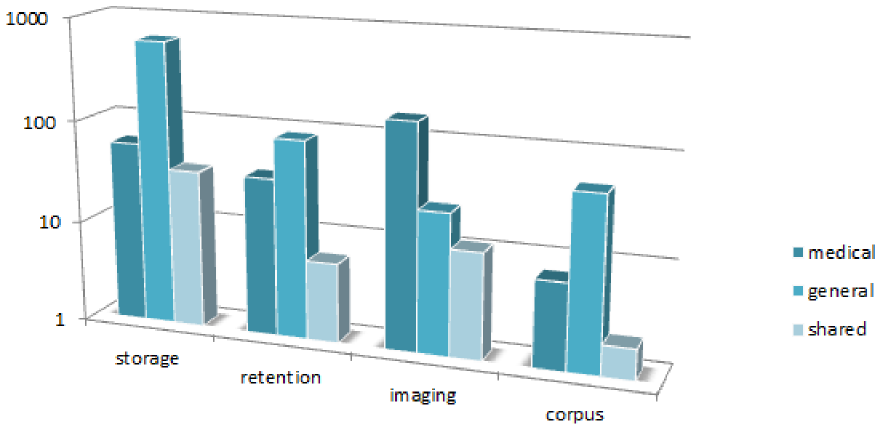


Figure 1- Test words log relative frequency distributions across domains.

Figure 1 provides a snapshot of the usage of our test words across domains, showing that some words are chiefly used in some domains only; however, it does not lay emphasis on the fact that all six terms exhibit multiple senses but relate them differently. The inter-sense relationship can be described in terms of i) AMBIGUITY, when multiple senses show no common schema/relation (as in the case of CORPUS), ii) POLYSEMY, when senses are specialisations/extensions of core meanings (see for example IMAGING); and iii) VAGUENESS, when words can interchangeably be used in both generic and domain-specific contexts with no appreciable sense shift (as with RETENTION and STORAGE). There are words (NETWORK and GRID) which show domain-sensitive specialisation and are chiefly used in their specialised sense.

To further investigate the relationship between different senses, we measured to what extent words which are used both in a specialised context and in a general context tend to co-occur with the same collocates, i.e. tend to be used in similar contexts. This is illustrated in Figure 2, where, for each test word, we counted the number of top-ranked collocates for each of two domains (medical vs. general), and then the number of collocates present in both ranks. As expected, the term CORPUS exhibits the lowest number of common collocates, due to the fact that the two senses of the word share no common meaning core.



	storage	retention	imaging	corpus
■ medical	59	35	159	7
■ general	603	87	24	49
■ shared	35	6	11	2

Figure 2 – Medical sciences and general domain shared collocates

It is important to note, however, that words like STORAGE, RETENTION and IMAGING present more overlapping contexts and different degrees of domain specialisation. In particular, IMAGING is by far the most frequent term in its medical usage, even in a non-specialised domain, as shown by the considerable number of collocates that are common to both medical and general domains. Moreover, it is often very difficult to distinguish two different senses of the same term on the basis of the observation of its contextual behaviour only. It turns out that terminological innovation – as expressing conceptual innovation - often implies a gradient extension of a core meaning rather that a radically different usage. We doubt that distributional analyses of contexts can provide a fully automatic basis to tackle sense distinctions at this level of granularity and flexibility. Nonetheless, innovation and domain-driven specialisation can be characterised distributionally in terms of a correlation between typical contexts of usage and knowledge domains.

To sum up, IMAGING, RETENTION, STORAGE, and CORPUS versus NETWORK and GRID are examples of linguistic knowledge in a very specific context-based approach, and therefore examples of phenomena such as ambiguity and polysemy. They are closely related to a representation of innovative knowledge, though still representing an existing gap between the semantic, context and domain dimensions. On the other hand, the relationship between meanings and reality, and lexical semantics are bridged in linguistic data related to concepts of the field of information technology, which is widely known, identified, spread and popular so to enhance a common and coherent understanding.

4. Discussion and concluding remarks

The spread of internet has enhanced the development of better bibliographic scientific databases with a significantly improved capacity for storage, access and retrieval. In recent years, automated search of electronic database has become the default mode of scientific information retrieval.

As an answer to the need for having better domain-specific databases available for improved storage and retrieval of scientific content, bibliographic domain-specific and type-oriented databases are developed, and emerged over the last two decades to offer materials to a specialised readership, thus providing highly-selected, well-targeted documents.

Such specific infrastructures define a communication network, where writers and readers – like speakers – are in a sufficient proximity to each other to have very high probability of communication with each other, by sharing the same “language", intended as the population of utterances in a speech community (Croft, 2000). The informative purposes are defined by the context, in line with Jakobson’s model of communication (1960), where six constituents of the communication act are modelled as functional roles: the addresser or encoder, a message or a signifier, the addressee or decoder, a context or the signified – where by context Jakobson means referent, a code or shared mode of discourse, and a contact or channel.

Analysis of word usages in large corpora offers the opportunity to investigate how words and language structures become vehicle for innovative knowledge generation and transfer. Lexical co-occurrences and collocations can be of considerable help in retrieving text materials which are relevant to a specific domain of interest, but they can be of little help in distinguishing one particular sense of a polysemous word from its other senses. This is especially true of innovative usages of existing terms, which appear to transfer and adapt their original and more general collocates to one or more specialised domains. This is the reason why automated sense disambiguation based on the distributional analysis of words in context proves to be less effective especially in distinguishing word usages which are most strongly related to innovation.

Due to ambiguity, polysemy and homography, most terms are multi-referential or multi-contextual (Renouf, 1993) in use. Low-precision results could be improved by restricting the contextual domain. A more supervised approach to the problem, where domains and texts are classified preliminarily by domain experts, rather than being bootstrapped from patterns of word distribution only, promises to be more successful in this task.

We investigated the hypothesis that extension of usage process and polysemous disambiguation correlate significantly with genre- and domain- oriented texts and intended readership, thus providing a convenient way to track down well-targeted, highly technical repositories of openly available text materials.

In requiring a dynamic shift from context-driven vagueness – in term of semantic polymorphism - to domain-driven specialisation – in term of terminological usage, the lexical representation of innovative knowledge in a context-based approach is closely related to the representation of knowledge itself, and represents the opportunity to reduce the gap between knowledge representation and knowledge understanding.

Our main emphasis is laid on the importance of effective information technology repositories and distributed infrastructures for the implementation of knowledge processes, and for an efficiently and widely distributed dissemination of research and innovation results, so to enhance future research.

References

Andersen H. (1989). Understanding linguistics innovations. In Breivik, L.E., Jahr, E.H. (eds.), *Language Change: Contributions to the Study of its Causes.* Mouton de Gruyter, Berlin. 5-27.

Baeza-Yates R., Ribeiro-Neto B. (1999). Modern Information Retrieval. Addison Wesley, ACM Press, New York.

Biber D. (1989). A typology of English text. *Linguistics,* 27, 3-43.

Buitelaar P., Cimiano P., Magnini B. (2005). Ontology learning form text. IOS Press, Amsterdam.

Church K., Hanks P. (1989). Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.

Croft W. (2000). Explaining Language Change. An Evolutionary Approach. Longman, London.

Deumert A. (2003). Bringing speakers back in? Epistemological reflections on speaker-oriented explanations of language change. *Language Sciences*, 25, 15-76.

Fellbaum C. (2005). Examining the constraint on the benefactive alternation by using the World Wide Web as a corpus. In Kepser S., Reis M. (eds.) *Linguistic evidence: empirical, theoretical, and computational perspectives.* Walter de Gruyter, Berlin. 209-238.

Fletcher W.H. (2011). Corpus Analysis of the World Wide Web. In *Encyclopedia of Applied Linguistics.* Wiley-Blackwell.

Frantzi K. T., Ananiadou S., Mima H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), 115-130.

Hedlund G. (1999). The Multinational Corporation as a Neatly Recomposable System. *Management International Review* 39,1. 5-44.

Jakobson R. (1960). Linguistics and Poetics: Closing Statement. *Style in Language*. MIT Press, Cambridge, MA.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. (2004) The Sketch Engine. *Proceedings EURALEX* 2004, Lorient, France. 105-116.

Linton J. (2009). De-babelizing the language of innovation. *Technovation*, 29, 729-737.

Lüdeling A., Evert S., Baroni M. (2006). Using web data for linguistic purposes. In Hundt M., Nesselhauf N., Biewer C. (eds.) *Language and Computers, Corpus Linguistics and the Web* 18, 7-24.

Manning C. D., Schütze H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Marzi C. (2012). Knowledge Communities in Grey. *Proceedings of the GL13 – International conference on Grey Literature: the Grey Circuit.* TextRelease, Amsterdam. 34-40.

Meyer C.F. (2002). English Corpus Linguistics: An Introduction. Cambridge University Press, Cambridge.

Nooteboom B. (2000). Learning and innovation in organizations and economics. Oxford University Press, Oxford.

Pangaro P. (2008). Innovation, Language, and Organizations. *Continuum Itaú Cultural magazine*. 7.

Seretan, V. (2011). Syntax-Based Collocation Extraction. Springer, Heidelberg-London-New York.

Sinclair J. (1991). Corpus, concordance, collocation: Describing English language. Oxford University Press, Oxford.

Smadja, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*, MIT, Cambridge MA, USA, 19(1), 143–177.

Stubbs, M. (2001). On inference theories and code theories: Corpus evidence for semantic schemas. *Text*, 21 (3), 437-465.

Renouf, A. (1993). What the Linguist has to say to the Information Scientist. In Forbes, G. (ed.) *The Journal of Document and Text Management* vol. 1/2, 173-190.

Vetere G., Oltramari A., Chiari I., Jezek E., Vieu L., Zanzotto F.M. (2011). Senso Comune, an Open Knowledge Base for Italian. *TAL*, 52 (3), 217-243.

Welch D., Welch L. (2008). The importance of Language in International Knowledge Transfer. *MIR*, 48(3), 339-360.

The Sketch Engine http://www.sketchengine.co.uk

Open Grey for Natural Language Processing: A ride on the network

Gabriella Pardelli, Sara Goggi, Manuela Sassi
Istituto di Linguistica Computazionale, "Antonio Zampolli", ILC-CNR, Italy

Abstract

The aim of this paper is to introduce the Open Access movement for Natural Language Processing (NLP) by means of a wide range of open access Grey Literature documentation available on the web. In 2008 Robert Dale, in the last issue of volume 35 of Computational Linguistics said: "There are a number of definitions of the term 'open access' in circulation, but almost all share the key principle that scientific literature should be freely available for all to read, download, copy, distribute, and use (with appropriate attribution) without restriction". At first glance it might seem that the Open Access movement has gradually become more influential in the field of language technology by building repositories accessible through the network. Today's digital archives are niches of intellectual production spread by means of a wide range of documents (such as journal articles and proceedings) which, paradoxically, the search engines do not always reach. The use of inappropriate terms in the formulation of queries and the fragmentation of repositories in this area of investigation does not allow to retrieve information on a large scale. The full paper, after a first introductory section, will be organized in two sections: 1) the first dedicated to the methodology for searching and tracing open access resources and to the criteria for analyzing and selecting the online documentation; 2) the second devoted to a description of the state-of-the-art of Open Access Grey Literature material in a statistical and thematic scenario.

As things stand, standardization of computational systems interconnected by links and tools of various nature allowing Internet users to easily retrieve the information that the web naturally makes available would then be essential.

Keywords: Open Access Movement, Natural Language Processing

1. Introduction

Open Access is the key in the development of Information Society (IS), a new method for sharing scientific resources which influences the dynamics of creation and dissemination of knowledge. In order to share and spread this knowledge ever more sophisticated digital devices are tuned up while scientific institutions and associations are lately committed to the creation of dedicated repositories with the intent of giving wide visibility to their resources.

There is more: sharing open access information does not only mean retrieving objects of digital nature from the origins but also digitally reproducing source material from the far and recent past.

The definition of Open Access is rather tricky, as Merkel-Sobotta says in 2005: " 'Open access' means many different things to many different people. To use an example from US politics: it is as difficult to be anti-choice as it is to be anti-life. In the flux of ideas generated by the new and rapidly developing phenomenon of web publishing, open access proponents were able to convince others that traditional publishers were "anti-open access" or even anti-access, period. Tested against the realities of e-publishing, this did not last very long".

The following are a few examples of significant Open Access repositories of our field:

I. the Association for Computational Linguistics (ACL) built a rich repository called *ACL Anthology*, a digital archive of research papers in Computational Linguistics. This archive traces down the history of Computational Linguistics from first research of the '60s by retrieving and putting on the web the articles published in the proceedings of the most important international conferences of the field (i.e. COLING series).

II. since some time several conferences publish their contributions as open access documents: the Global Wordnet Conference and the Language Resources and Evaluation Conference (LREC), just to make a couple of examples.

III. *Machine Translation Archive* is an electronic repository and bibliography of articles, books and papers on several topics in the field of machine translation, computer translation systems and computer-based translation tools. This archive contains knowledge: its documents, accessible by everyone, provide an historical overview of automatic translation which might turn out to be very useful both for experts and non-experts of the field.

Notwithstanding the fact that conferences and workshops scientific material is widely spread on the web, the available search engines – though very sophisticated – are not nowadays able to provide a comprehensive plan of open access resources for Language Technology.

As a matter of fact, the first decade of the new millennium has suddenly witnessed first the growth and then the rapid increase of the so-called "social networks" which totally transformed the way information is transmitted: nowadays the World Wide Web looks like an enormous collection of documents inter-connected and linked to the various search engines by sharing the same paradigm (Web 2.0).

The selection, conservation and storage of digital content apparently makes the users' fruition easier : but is this assumption really true?

To formulate appropriate and effective queries for a search is a difficult task for users and requires a careful terminological selection for obtaining the most from an Information Retrieval system: *Information Retrieval* is the academic discipline which studies the methodologies, tools, techniques and languages for searching and retrieving relevant data for an information need.

2. Web search technology

The term Information Retrieval was introduced by Calvin Mooers in 1951, who defined it in this way: "Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information. It is another, more general, name for the production of a demand bibliography. Information retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation. Information retrieval is crucial to documentation and organization of knowledge". (Mooers, 1951, p. 25).

The continuous development of the Internet-related technology makes available to the user a huge quantity of information perpetually increasing: "The Web is immense, free, and available by mouse click", as Adam Kilgarriff and Gregory Grefenstette said in 2003 (Kilgarriff and Grefenstette, 2003, p. 333). But though Internet unsettled the traditional scientific communication channels because publications on the web do not have (and do not need) any preliminary filter and sharing documents on the net becomes knowledge open to everyone, it is sometimes difficult to assess the quality of a document as well as to retrieve its semantics if the address of the portal is unknown to the user. Users are therefore often in difficulty when they submit a query and the web answers with a wide range of documents, most of them lacking any identification feature (such as place and time of creation) apart from the topic.

Today scientific documentation produced by the academia is transmitted thanks to the respective institutional web sites by means of ever more sophisticated software platforms for managing documentation; but, on the other side, there is a lack for what concerns a pertinent information retrieval: the answers to users' queries are usually not thorough and precise enough. When a user enters a query into a search engine (typically by using keywords), the engine examines the pre-existing indexes and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text; afterwards, documents are ranked according to their relevance probability and shown to the user in such an order. Given the results, the user can decide to refine the query and the whole cycle starts once again.

In this scenario, filtering the results of a query for selecting those valuable from a qualitative point of view amidst the great amount of useless information is a task which requires experience and patience. The classification system of the Web makes available to users several sophisticated search algorithms which do not have, though, an immediate impact on the human cognitive process of identification of the most significant and useful links for a given query. Therefore, for getting to what is really needed users have to start from a generic resource and afterwards follow the links generated by the first resource: search engines sound the web for capturing new pages by means of the URLs and then index them.

"Information retrieval today operates in a number of different contexts: full-text, digital libraries, the Web, online catalogs, and networked applications. IR utilizes a number of conceptual models such as: algorithmic, probabilistic Boolean, semantic, fuzzy logic, and vector space. As a result, IR has supported the creation of a number of different applications among them: latent semantic analysis, vector space analysis, information filtering, data mining, automatic indexing and classification, along with a number of paradigms for query formulation and information visualization" (Baeza-Yates & Ribeiro-Neto, 1999).

Whilst nowadays it is taken for granted that science concerns everyone thanks to the global net and in particular to the open access archives, it is also true that three centuries and a half have passed since the establishment of the journal called *Philosophical Transactions of the Royal Society of London* (1665), founded for the dissemination of scientific contributions.

3. Where to search for Open Access Language Technology documents?

Open Access documentation in Natural Language Processing, Computational Linguistics and Human Language Technology domains is not always easily retrievable although a massive amount of precious information has been published on the web by academia, association and private companies since many years.

In the '60s, the research in the field of natural language processing consolidated and consequently new associations were born and international journals were founded: here below a mention of some associations and the respective scientific open access production available on their portals:

≈ In the past: in 1959 the *Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée* (ATALA) was born; [in 1965, ATALA becomes the *Association pour le Traitement Automatique des Langues*]; today: http://www.atala.org/
Grey Literature for Language Technology: online proceedings of the TALN conference (*Traitement Automatique des Langues Naturelles*) to be found at http://www.atala.org/-Conference-TALN-RECITAL

≈ In the past: in 1962 the *Association for Machine Translation and Computational Linguistics (AMTCL)* was founded [in 1968 it becomes the *Association for Computational Linguistics (ACL)]*; today: http://www.aclweb.org/index.php?option=com_frontpage&Itemid=1
Grey Literature for Language Technology: online proceedings of the ACL conference series can be found on the ACL Anthology ("A Digital Archive of Research Papers in Computational Linguistics") website at http://aclweb.org/anthology-new/

≈ In the past: in1965 the *Association Internationale de Linguistique Appliquée* or *International Association of Applied Linguistics (AILA)* was established; today: http://www.aila.info/
Grey Literature for Language Technology: NO open access material.

≈ In the past: in 1973 the *Association for Literary and Linguistic Computing (ALLC)* was born; today: it has a new name: *The European Association for Digital Humanities* (http://www.allc.org/)
Grey Literature for Language Technology: online proceedings of the ALLC conferences can be found at: http://www.allc.org/conferences

≈ In the past: in 1978 the *Association for Computers and the Humanities (ACH*) was founded*;* today: http://www.ach.org/
Grey Literature for Language Technology: the collection includes abstracts from the "Joint International Conference of the Association for Literary and Linguistic Computing and Association for Computers and the Humanities" and covers the years 1996, 1997, and 2000-2003: http://67.207.129.15:8080/dh-abstracts/search

≈ In the past: in 1991 the *European Association for Machine Translation (EAMT*) was born; today: http://www.eamt.org
Grey Literature for Language Technology: 1) "Archive de la Traduction Automatique - Index des organisations": http://www.mt-archive.info/foreign/organisations-french.htm; 2) electronic repository and bibliography of articles, books and papers on topics in machine translation, computer translation systems, and computer-based translation tools: http://mt-archive.info/

≈ In the past: in 1995 the *European Language Resources Association (ELRA)* was established; today: http://www.elra.info/
Grey Literature for Language Technology: online proceedings of the LREC (*Language Resources and Evaluation Conference*) series are at: http://www.elra.info/LREC-Conference.html

≈ In the past: in 2000 *The Global WordNet Association (GWA*) was born:
today: http://www.globalwordnet.org/
Grey Literature for Language Technology: online proceedings of the GWA conferences can be found at: http://www.globalwordnet.org/gwa/gwa_conferences.html

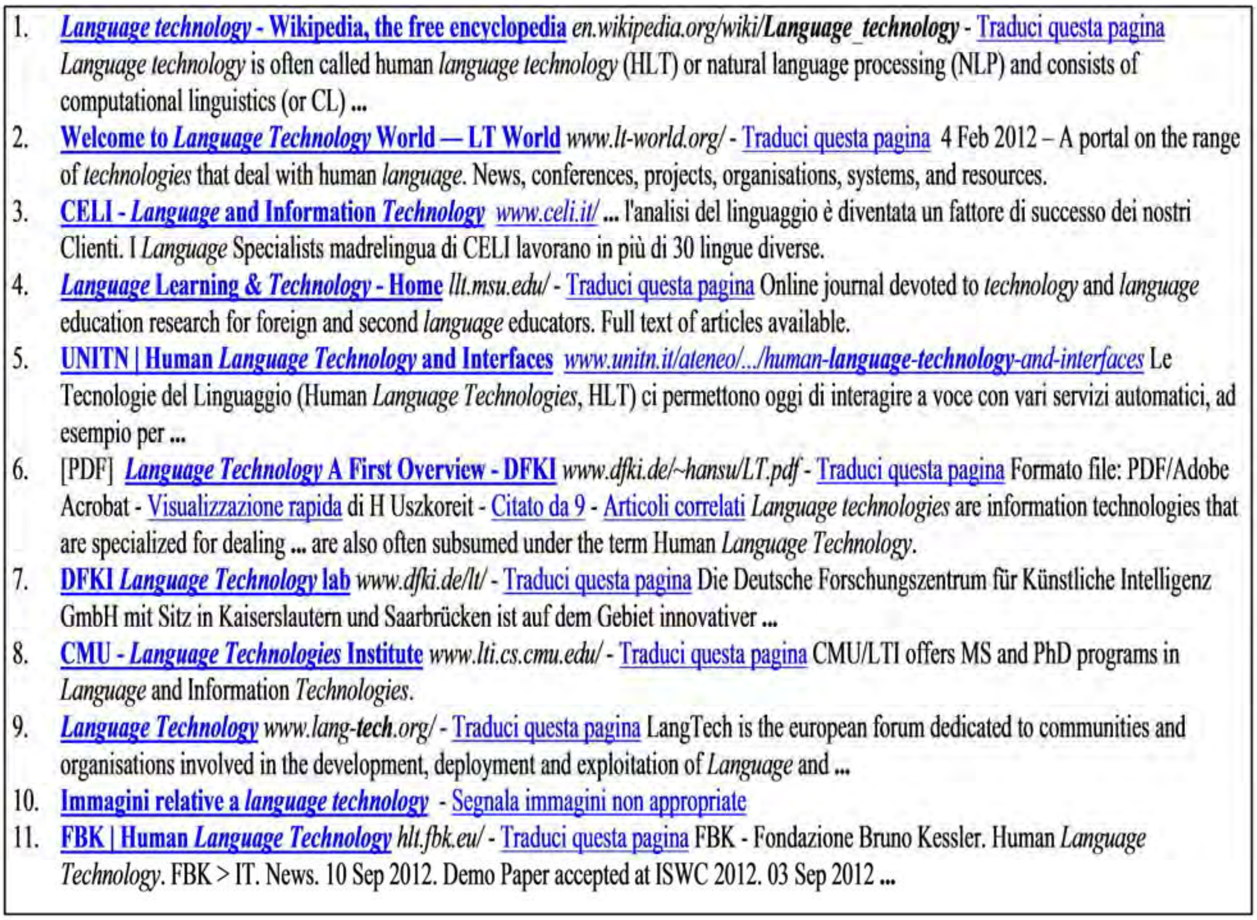
At last, a mention to the Grey Literature production of our Institute of Computational Linguistics stored on the PUMA "PUblication MAnagement System" Repository: http://puma.isti.cnr.it

3.1 The blind search

When a user tries to retrieve information on a given topic from online repositories there are several possibilities to formulate a query; for instance, given the query "Language Technology", the web replies with about 878.000.000 results in 0,26 seconds (September 25, 2012 at 4.30 p.m.).

Academic articles for language technology:

- ❑ of the state of the art in human language technology - Cole – Cited by 577
- ❑ Stirring up trouble about language, technology and …-Postman–Cited by 158
- ❑ Information extraction as a core language technology - Wilks – Cited by 69



1. Language technology - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Language_technology - Traduci questa pagina *Language technology* is often called human *language technology* (HLT) or natural language processing (NLP) and consists of computational linguistics (or CL) ...
2. Welcome to Language Technology World — LT World www.lt-world.org/ - Traduci questa pagina 4 Feb 2012 – A portal on the range of *technologies* that deal with human *language*. News, conferences, projects, organisations, systems, and resources.
3. CELI - Language and Information Technology www.celi.it/ ... l'analisi del linguaggio è diventata un fattore di successo dei nostri Clienti. I *Language* Specialists madrelingua di CELI lavorano in più di 30 lingue diverse.
4. Language Learning & Technology - Home llt.msu.edu/ - Traduci questa pagina Online journal devoted to *technology* and *language* education research for foreign and second *language* educators. Full text of articles available.
5. UNITN | Human Language Technology and Interfaces www.unitn.it/ateneo/.../human-language-technology-and-interfaces Le Tecnologie del Linguaggio (Human *Language Technologies*, HLT) ci permettono oggi di interagire a voce con vari servizi automatici, ad esempio per ...
6. [PDF] Language Technology A First Overview - DFKI www.dfki.de/~hansu/LT.pdf - Traduci questa pagina Formato file: PDF/Adobe Acrobat - Visualizzazione rapida di H Uszkoreit - Citato da 9 - Articoli correlati *Language technologies* are information technologies that are specialized for dealing ... are also often subsumed under the term Human *Language Technology*.
7. DFKI Language Technology lab www.dfki.de/lt/ - Traduci questa pagina Die Deutsche Forschungszentrum für Künstliche Intelligenz GmbH mit Sitz in Kaiserslautern und Saarbrücken ist auf dem Gebiet innovativer ...
8. CMU - Language Technologies Institute www.lti.cs.cmu.edu/ - Traduci questa pagina CMU/LTI offers MS and PhD programs in *Language* and Information *Technologies*.
9. Language Technology www.lang-tech.org/ - Traduci questa pagina LangTech is the european forum dedicated to communities and organisations involved in the development, deployment and exploitation of *Language* and ...
10. Immagini relative a language technology - Segnala immagini non appropriate
11. FBK | Human Language Technology hlt.fbk.eu/ - Traduci questa pagina FBK - Fondazione Bruno Kessler. Human *Language Technology*. FBK > IT. News. 10 Sep 2012. Demo Paper accepted at ISWC 2012. 03 Sep 2012 ...

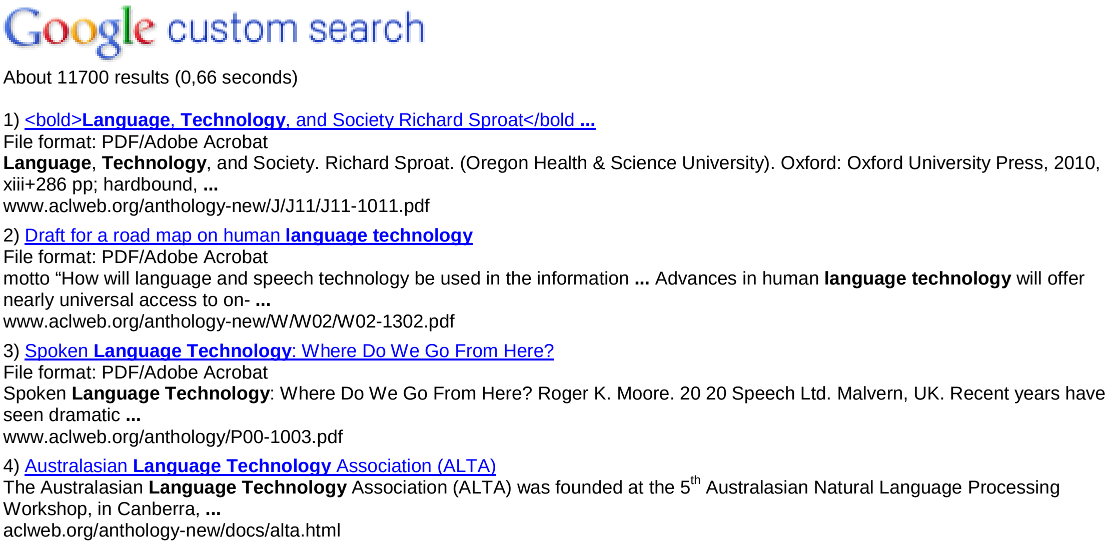
From this generic query it is possible to retrieve 9 portals, 2 open access publications (Cole and Wilks), 1 review (Portman). But only 2 documents from this set satisfy the user……

3.2 The conscious search

Let's try with queries formulated by an expert user:

1 Query: ACL Anthology

- ♦ Query: Language Technology



Google custom search

About 11700 results (0,66 seconds)

1) <bold>**Language**, **Technology**, and Society Richard Sproat</bold **...**
File format: PDF/Adobe Acrobat
Language, **Technology**, and Society. Richard Sproat. (Oregon Health & Science University). Oxford: Oxford University Press, 2010, xiii+286 pp; hardbound, **...**
www.aclweb.org/anthology-new/J/J11/J11-1011.pdf

2) Draft for a road map on human **language technology**
File format: PDF/Adobe Acrobat
motto "How will language and speech technology be used in the information **...** Advances in human **language technology** will offer nearly universal access to on- **...**
www.aclweb.org/anthology-new/W/W02/W02-1302.pdf

3) Spoken **Language Technology**: Where Do We Go From Here?
File format: PDF/Adobe Acrobat
Spoken **Language Technology**: Where Do We Go From Here? Roger K. Moore. 20 20 Speech Ltd. Malvern, UK. Recent years have seen dramatic **...**
www.aclweb.org/anthology/P00-1003.pdf

4) Australasian **Language Technology** Association (ALTA)
The Australasian **Language Technology** Association (ALTA) was founded at the 5th Australasian Natural Language Processing Workshop, in Canberra, **...**
aclweb.org/anthology-new/docs/alta.html

5) Evangelising **Language Technology**: A Practically-Focussed **...**
File format: PDF/Adobe Acrobat
Evangelising **Language Technology**: A Practically-Focussed Undergraduate Program. Robert Dale, Diego Mollá Aliod and Rolf Schwitter. Centre for Language **...**
www.aclweb.org/anthology-new/W/W02/W02-0104.pdf

6) Letter to the Editor: **Language Technology** for Beginners
File format: PDF/Adobe Acrobat
Language Technology for Beginners. Ronald A. Cole 1. (University of Colorado). I am writing in response to Varol Akman's review (Computational Linguistics, **...**
www.aclweb.org/anthology/J99-4012

7) Does **Language Technology** Offer Anything to Small Languages?
File format: PDF/Adobe Acrobat
Does **Language Technology** Offer Anything to Small. Languages? Nick Thieberger. PARADISEC, University of Melbourne/. University of Hawai'i at Manoa **...**
aclweb.org/anthology-new/U/U07/U07-1002.pdf

8) PROJECTED GOVERNMENT NEEDS IN HUMAN **LANGUAGE ...**
File format: PDF/Adobe Acrobat
for human **language technology**, this paper will discuss the uses which will probably **... language technology** a suitable solution to maximize the effectiveness in **...**
aclweb.org/anthology-new/H/H93/H93-1056.pdf

9) Proceedings of the Australasian **Language Technology** Summer **...**
File format: PDF/Adobe Acrobat
and Australasian **Language Technology** Workshop (ALTW). 2003 **...** The Australasian **Language Technology** Association is proud to present its inaugural **...**
aclweb.org/anthology-new/U/U03/U03-1000.pdf

10) Human **Language Technology** can modernize writing and grammar **...**
File format: PDF/Adobe Acrobat
Human **Language Technology** can modernize writing and grammar instruction. Gerard Kempen. University of Leiden. P.O. Box 9555, 2300 RB Leiden, The **...**
aclweb.org/anthology-new/C/C96/C96-2172.pdf

Results: 10 open access Grey Literature documents found!

Given the fact that the enormous amount of data available on the web is difficult to query from a semantic point of view, the human interpretation is always needed - - but which are the assumptions/conditions for making an effective query?

Nowadays knowledge extraction can be performed in a satisfactory way if:

i) the know-how of the state-of-the-art is updated; ii) there is a good **skillfulness** in navigating on the web portals; iii) there is the ability to interpret the data.

4. Conclusions

The web should be considered both as a knowledge repository and a knowledge dispenser: from this perspective, there is the need to create innovative paradigms for information retrieval, to establish features for semantic search on web portals as well as to achieve a certain degree of precision and recall, which are the coefficients measuring the performance of an information retrieval system:

- ✓ *Precision*: proportion of relevant data retrieved from the total data retrieved
- ✓ *Recall*: extent of relevant data retrieved from the total data relevant in the database.

These coefficients measure two different factors:

- ☐ *Noise* = non-relevant data retrieved;
- ☐ *Silence* = relevant data that have not been retrieved from the data base.

Retrieval models compute the degree to which certain elements answer to a query: a good model should be able to maximize recall and precision and minimize, respectively, "silence" and "noise".

References

Baeza-Yates R., Ribeiro-Neto B. (1999). *Modern information retrieval*. Reading, MA: Addison-Wesley

Dale R., LastWords What's the Future for Computational Linguistics? *International Journal for Computational Linguistics*, Volume 34, Number 4, 2008.

Kilgarriff A., Grefenstette G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics,* Volume 29, 3, Association for Computational Linguistics, 333-347.

Merkel-Sobotta E., Elsevier and Open Access, *Neuroinformatics*, 2005, Volume 3, Pages 5-9.
http://www.springerlink.com/content/x66772m265840111/fulltext.pdf

Mooers, C. N. (1951). *Making information retrieval pay*. Boston, Zator Co.

Ranger Sara L., Grey Literature in Special Library: Access and the Use*, Publishing Research Quarterly,* 2005, Volume 21, Number 1, Pages 53-63.

http://www.aclweb.org/anthology/J/J08/J08-4008.pdf
http://www.ling.ohio-state.edu/acl08/cfp.html
http://www.mt-archive.info/
http://www.aclweb.org/anthology/
http://www.cfilt.iitb.ac.in/gwc2010/
http://www.lrec-conf.org/proceedings/lrec2010/index.html
http://puma.isti.cnr.it//index.php?langver=it
http://www.greynet.org/greytextarchive.html
http://www.regione.emilia-romagna.it/wcm/LineeGuida/sezioni/generali/motori/documento_motori_di_ricerca.pdf
http://www.iva.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/information_retrieval.htm

IRPPS Editoria Elettronica: An electronic publishing web portal based on Open Journal Systems (OJS)

Marianna Nobile and Fabrizio Pecoraro
Institute for Research on Population and Social Policies, IRPPS, Italy

Abstract
This paper presents IRPPS Editoria Elettronica, an e-publishing service developed by the Institute for Research on Population and Social Policies (IRPPS) of the Italian National Research Council (CNR). Its aim is reorganize the Institute scientific editorial activity, manage its in-house publications and diffuse its scientific results. In particular this paper focuses on: the IRPPS editorial activities, the platform used to develop the service, the publishing process and the web portal developed.

Introduction.

In the last few years the development of sustainable economic models based on open source technologies for the management and dissemination of publishing activities represents a great opportunity for research institutions to improve the diffusion of their scientific information [1]. The wide diffusion of *"electronic publishing"* in the scholarly community is leading libraries to play a key role in the improvement of innovative systems for the dissemination of scientific research results, with the aim of increasing the quality of products and reduce cost of publications. This is particularly evident considering the widespread diffusion of open access journals published by academic and research Institutions.

In Italy, the Institute for Research on Population and Social Policies (IRPPS) of the National Research Council (CNR) is carrying out a project with the aim of developing an e-publishing service based on Open Journal Systems (OJS) with the aim of managing the GL collections. To its development an analysis of IRPPS current practices of publications was carried out considering types of documents and contents to be selected for future e-publications as well as monographs and/or digitization of previously published or unpublished works that have represented important achievements of IRPPS research results. A new editorial plan was designed in collaboration with the internal scientific community to define roles of the changed publishing process as well as editorial policies aimed to improve the scientific quality and visibility of IRPPS research products.

At the moment the IRPPS editorial group has developed a prototype customizing the OJS features to the Institute needs with the aim of: a) reorganizing its scientific editorial activity; b) managing its in-house publications; c) diffusing its scientific results. In this paper, after a brief description of the products edited by the Institute, the attention is focused on the platform used to develop the web portal, on the publishing process based on OJS and on the web portal developed.

IRPPS editorial products.

Table 1 shows IRPPS editorial activity since its foundation (i.e. 1981), highlighting for each scientific editorial product the number of publications in Italian and English, the years during which the product has been published, the frequency of publications and the ISBN or ISSN code if applicable. The Institute has planned and carried out its editorial activities, closely connected to the diffusion of its scientific results, as well as open to the external research community. All of these were "in-house publication" or grey literature. Some of them (Working Papers and Monographs) were published with continuity, while others (for example Conference proceedings series) were published not regularly or for specific scientific events organized by the Institute.

Table 1. IRPPS in-house publications (1981-2012)

Title	#	Language		Published		ISBN/ ISSN	Freq.
		EN	IT	From	To		
Monograph	12	4	8	1982	2012	Yes	Ad-hoc
Working paper	118	28	90	1981	2012	Yes	Ad-hoc
Report on the demographic situation in Italy	5	5	0	1983	1994	No	
Conference proceedings	3	0	3	1984	1992	No	
Series documents and reprints	3	0	3	1984	1986		
Demotrends journal	21	11	9	1997	2005	Yes	Quarterly
Quaderni di Demotrends *(Suppl. to Demotrends journal*)	7	0	7	1997	2007	No	
Report on Welfare state in Italy	11	0	11	1995	2012	Yes	Annual
Other	27	7	15	1981	2003	No	
Total	**207**	**55**	**146**				

IRPPS e-publishing service

Products

At the moment the “IRPPS Editoria Elettronica” portal publishes two series: *IRPPS Monographs* and *IRPPS Working Papers*. In particular:

- *IRPPS Monographs*, already published with irregular periodicity in a paper format from 1984 to 2002, restarted its publication in 2011 on an electronic and open access format. Monographs are released through the Institute web site and printed upon request. Concerning the content, *IRPPS Monographs* publishes essays, proceedings of conference organized by the Institute as well as digital versions of volumes already published.
- *IRPPS Working Papers* have a similar publishing process. A paper based format distribution started in 1981 and migrate on the Institute web site in 2002. Aim of this product is to examine emerging topics and to diffuse projects results and ongoing researches. Moreover, papers are published on an open access format and subjected to an internal evaluation process. *IRPPS Working Papers* are reserved to researchers of the Institute, sometime in collaboration with external researches.

Both Monographs and Working Papers are freely accessible and apply the creative commons version 3.0 licensing [3].

Open Journal Systems (OJS)

Open Journal Systems (OJS) [4-6] is a suitable software that manages both editorial activities and open access scientific publications. In particular it supports the editorial workflow, from manuscript submission to publication including the peer-review process, it facilitates the e-publishing of different types of editorial products, from serials to monographs, thus allowing the reduction of journal’s operating costs.

Editorial workflow

The whole editorial workflow, based on the OJS one, is depicted in figure 1.

Every author can upload his/her contribution directly through the IRPPS Editoria Elettronica portal. The author is requested to register both as author and/or reader (in order to be informed on the publication of new reviews). Once the registration is done, the author is directed to the Home page, where he can find the different roles of the user in the review (e.g. Author, reviewer, etc). By accessing as “author”, the user is directed to the submission page, where he can upload his article. The process consists of 5 steps and it is entirely guided. IRPPS Working Paper and IRPPS Monographs are associated with guidelines for authors concerning bibliographical and formatting standards to follow.

To make sure that propositions are correctly submitted, and that their format is appropriate at the moment of publication, the administrator and the editor of the review have conceived a checklist with submission requirements to be accepted by the author before moving towards the next step. If the author cannot satisfy one of the requirements for some reasons, he can accept the request and give explanations to the editor in the comments section. Furthermore, if he/she encounters any difficulty during the submission, he/she can still contact by email the system administrator or the editor.

For every single proposition, the editorial process is made of a submission process and a revision process, followed by the decision of the editor of accepting (or refusing) the publication. If it is accepted during the process of revision, the submission moves through the editing phase, which consists of copyediting, layout creating and draft correcting. After that, the publications of working papers or monographs in one of the proposed files are programmed.

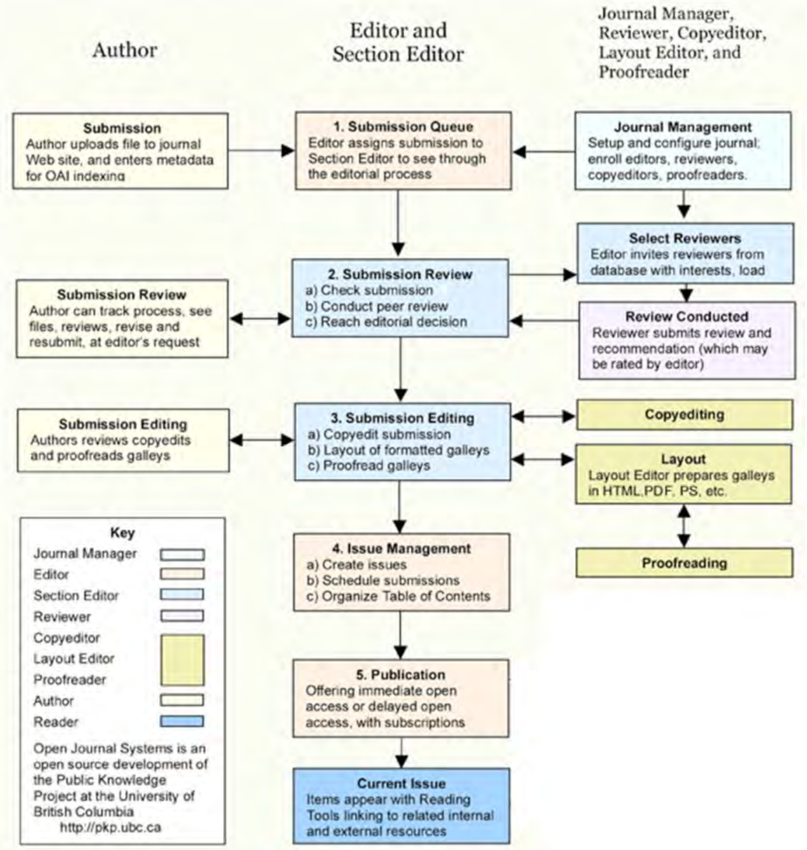

Author
Editor and Section Editor
Journal Manager, Reviewer, Copyeditor, Layout Editor, and Proofreader
Submission
Author uploads file to journal Web site, and enters metadata for OAI indexing
1. Submission Queue
Editor assigns submission to Section Editor to see through the editorial process
Journal Management
Setup and configure journal; enroll editors, reviewers, copyeditors, proofreaders.
Select Reviewers
Editor invites reviewers from database with interests, load
2. Submission Review
a) Check submission
b) Conduct peer review
c) Reach editorial decision
Submission Review
Author can track process, see files, reviews, revise and resubmit, at editor's request
Review Conducted
Reviewer submits review and recommendation (which may be rated by editor)
Submission Editing
Authors reviews copyedits and proofreads galleys
3. Submission Editing
a) Copyedit submission
b) Layout of formatted galleys
c) Proofread galleys
Copyediting
Layout
Layout Editor prepares galleys in HTML,PDF, PS, etc.
Proofreading
Key
Journal Manager
Editor
Section Editor
Reviewer
Copyeditor
Layout Editor
Proofreader
Author
Reader
Open Journal Systems is an open source development of the Public Knowledge Project at the University of British Columbia
http://pkp.ubc.ca
4. Issue Management
a) Create issues
b) Schedule submissions
c) Organize Table of Contents
5. Publication
Offering immediate open access or delayed open access, with subscriptions
Current Issue
Items appear with Reading Tools linking to related internal and external resources

Figure 1. IRPPS Editoria Elettronica editorial workflow based on OJS one.

The task of the editor, who can access all the submissions, is to oversee the editorial process and to take care of the section that controls accepted submissions and those referring to the editing phase (in the editorial process). Most particularly, the editor manages the submission, opens the submitted file, makes sure that the text is appropriate for the submission and that it has been proposed for the appropriate section (otherwise, the editor can send an email to the author to communicate that the contribution cannot be reviewed).

Moreover, the editor controls the contributions, in order to verify the presence of mistakes and make sure that the propositions are edited along with the guidelines. The editor is also the draft corrector. He can modify or update data insert by the author during the submission process. Data insertion and revision are important for the index of working papers or monographs.

The layout editor has the role of formatting the contribution (create the title page and the back-title page), and uploading on the portal the final version. When the layout curator uploads a file in the draft format, the system identifies the file extension (e.g. PDF; HTML), and gives information on its weight and original name. The label appearing in the summary of IRPPS Monographs or IRPPS Working Papers gives the reader the types of format available for the contribution. The layout editor can upload one or more files for each draft format, delete un-uploaded files, and delete information with the file name.

Future plans

In the future, IRPPS will continue to involve internal scientific community. The new service can represent an alternative especially for authors looking for some publication channels that can increase their visibility and impact, especially on the web.

IRPPS is thinking of providing a new online edition of Demotrends, a journal published by the Institute for 7 years (from 1997 to 2005). Demotrends was focused on the dissemination of high quality information on the topics in which the Institute is involved. It was indexed by the ACNP, the Italian Union catalogue of serials. The journal closed in 2005 due to budget restrictions.

Another editorial plan concerns the digitization of previously published or unpublished works that have represented important achievements for IRPPS research results. For instance one project deals with the reprinting of the First report on the demographic situation in Italy, published for the first time by the

Institute in 1985. Another important product was the Atlas on population aging in Italy published in 1991 and never reprinted. The e-publishing service can be an occasion to digitalize these publications and make them available in the portal.

Moreover, other plans concern the implementation of PKP (Public knowledge project) module for conferences organized or hosted by the Institute (Open Conference Systems) and the adoption of the new software recently released by PKP to manage monographs now published through OJS (Open Monograph Systems.

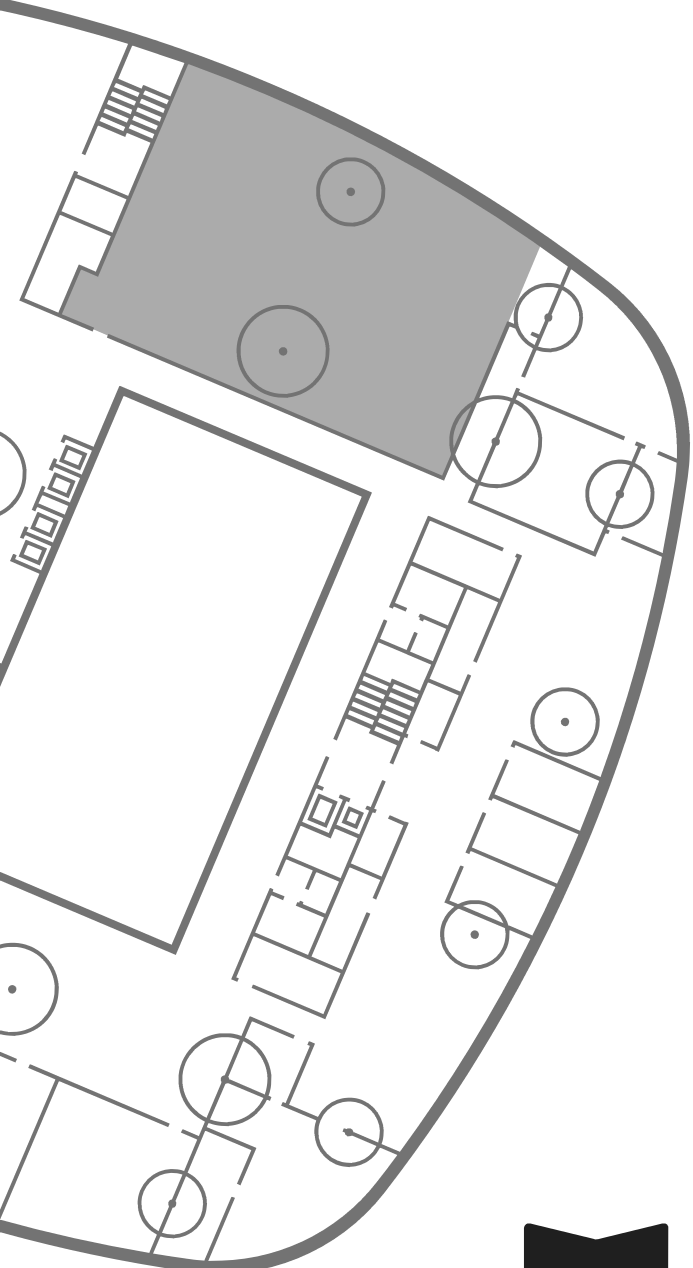
Finally, to improve the quality and visibility of the IRPPS scientific research various indexing procedure requests has been carried out in more closely disciplinary archives, such as REPEC (Research Papers in Economics) [7], for *IRPPS Working Papers*, and OAPEN (Open Access Publishing in European Networks) [8] for *IRPPS Monographs*. At the moment both Monographs and Working Papers are indexed by Google and Google Scholar portals.

Conclusion

The e-publishing services represent a sustainable -but not totally free- model: they need qualified personnel to be involved in various stages of project design and management. These services can improve the editorial quality, and also contribute to a larger diffusion and visibility of the scientific research products. For this purpose, the involvement of the internal scientific community is important, in particular the young researchers, who have been growing in a technological and rapidly changing context, can benefit from these new publishing services. Moreover, libraries can also become sites where a research can still be open to new perspectives, where their role is not limited to the management of list of references, but have a key role in supporting the scholarly communication chain. In this context, a redefinition of the role of libraries arises. Thanks to the progress in technology and open access, they can nowadays take advantage from new services and instruments allowing a better answering to users' needs (e-publishing services, projects of digitalization, etc.).

References.

1. Willinsky, J. (2006). The access principle: The case for open access to research and scholarship. Cambridge, MA: MIT Press.
2. IRPPS Editoria Elettronica. Available from: www.irpps.cnr.it/e-pub
3. Creative Commons 3.0. Available from: www.creativecommons.it/3.0
4. Open Journal Systems (OJS). Available from: http://pkp.sfu.ca/ojs
5. Willinsky J. (2005). Open Journal Systems; an example of Open source software for journal management and publishing. Library Hi-Tech, 23 (4).
6. Brian D. Edgard; Willinsky J. (2010). A survey of the scholarly journal using Open Journal Systems. Scholarly and research communication. 1(2)
7. REPEC (Research Papers in Economics). Available from: www.repec.org
8. OAPEN (Open Access Publishing in European Networks). Available from:www.oapen.org

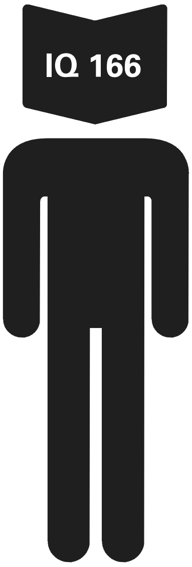


National Technical Library

National Technical Library (hereafter referred to as "**NTK**") is a central professional library open to public, which offers a unique collection of 250 thousand publications freely accessible in open circulation. Its holdings form the largest collection of Czech and foreign documents from technology and applied natural sciences as well as associated social sciences. It contains a total of 1,2 Mil. volumes of books, journals and newspapers, theses, reports, standards, and trade literature in both printed and electronic forms. Besides its own collection, parts of Central Library of the CTU in Prague and Central Library of the ICT holdings are accessible in NTK.

For detailed information on the National Technical Library visit **http://www.techlib.cz/en/**

94 mm



As corresponds to its statutes, NTK manages – among others – the project of building the **National Repository of Grey Literature.**
The project aims at gathering metadata and possibly full texts of grey documents in the fields of education, science and research.
The NTK supports an education in the field of grey literature through annual seminars in the Czech Republic.

For more information on the National Repository of Grey Literature visit our project Web site **http://nrgl.techlib.cz/** and for a search **http://www.nusl.cz/**

NTK Univers

Národní technická knihovna
National Technical Library

NU¶
· ŠL

Grey literature partnership network in the Czech Republic

Petra Pejšová and Hana Vyčítalová
National Technical Library, Czech Republic

Abstract

After the dissolution of initial partnership network, involved into the international SIGLE cooperation, at the beginning of 2005, new and broader partnership network was founded at the end of 2009. In November 2012, there are 91 data producers from the fields of research and science, education, culture and also enterprise, namely research institutes, universities, libraries etc. Partnership network has been built to support National Repository of Grey Literature (NRGL).

Grey literature data producers can choose from three types of cooperation with NRGL. All data are available through the Central Search Interface at http://www.nusl.cz in both Czech and English versions. National Technical Library (NTK) as a promoter supports partnership network in many ways. NTK runs informative webpages, e-mail conference, publishes manuals and methodologies, performs presentations and trainings and organizes full day conference every year.

Keywords: grey literature, data producers, digital repository, partnership network, legal framework

The starting point for building a grey literature partnership network in the Czech Republic was the state ascertained at the turn of 2007/2008, when the results of research, development and education paid for from public funds were often only available at the research workplaces (universities, research institutions) or through subsidiary information systems of subsidy providers. This meant that these valuable resources of information were often completely inaccessible to the expert public.

The National Technical Library (hereinafter only the "NTK"), which has systematically focused on collecting grey literature since the 1990s when it became a member of EAGLE (European Association for Grey Literature Exploitation), has sought to address this situation. The NTK has created the Grey Literature Cooperation System (hereinafter only the "GLCS"). This system was established on a contractual basis and accepted bibliographical records on grey literature (on dissertations) from cooperating Czech universities. Some universities also supplied the NTK with the actual printed dissertations so that the NTK could ensure access to them. This work always bore the marks of pragmatic access to information sources as carriers of information useful for education and research, and not merely the accumulation of library items without any direct relationship with their use. It was also for this reason that the building of a collection of printed dissertations in the NTK was ended in 2007, as it was already clear that other collections of these works would be available for study purposes in electronic form.

In 2009 the NTK began building a new, much more extensive partnership network of grey literature producers, within the framework of which it is attempting to capture grey literature from research, development and education fields such as grant agencies, state administration, research organisations, education and the commercial sphere. For this purpose the NTK, in cooperation with University of Economics, Prague (hereinafter only the "UEP"), created the National Repository of Grey Literature[1] environment (hereinafter only the "NRGL"), through which it is possible to search for documents, make them accessible or mediate access to them. Great emphasis is placed on the possibility of acquiring a document, and so in cases where there is no other option than to acquire a document at its storage location, the document records contain contact information for a service that will mediate the delivery of that document.

In view of legislative and organisational restrictions the NTK is creating a network of cooperating institutions on a voluntary basis. The first to be contacted, in 2009, were the institutes of the Academy of Sciences of the CR, followed by Czech state-run universities and grant agencies in 2010. In 2011 public research institutions, ministries and other selected organisations were contacted. In 2012 private universities, private research organisations, and selected museums and galleries were contacted.

Thanks to the commencement of cooperation with institutes of the Academy of Sciences of the CR, the NTK determined the diversity of institutions' access to grey literature. In the case of the production of grey literature by institutes of the Academy of Sciences of the CR, a connection was established with the Library of the Academy of Sciences, which is also interested in archiving grey literature at the NRGL, and this resulted in an agreement on cooperation. The Library of the Academy of Sciences performs the role of central administrator of the collection of grey literature for institutes of the Academy of Sciences of the CR. Initially the system for collecting grey literature mainly contained only bibliographic data, but in 2011 the Academy of Sciences commenced the collection of electronic documents and the retroactive digitisation of selected documents.

At universities university works are collected in particular on the basis of obligations laid down in the Act on Higher Education Institutions. A significant problem with Czech universities is that in the majority of cases they refuse to make public the actual works and only provide descriptive information about such works. The Act on Higher Education Institutions does not specify the method or the scope of the publication of information about university works. This means that each school stipulates the scope and method itself, and the form of the database of university works in the Czech Republic also reflects this. Some universities only provide public access to records of university works, while others also publish full texts. 11 state-run universities have so far joined the NRGL partnership network, and either their repositories have been connected to the NRGL repository, or they directly use the NRGL repository as their storage location. Discussions are under way with others. 5 private universities thus joined in 2012. Private universities and state-run universities generally do not have their own repositories and do not publish the full texts of university works. If they participate in NRGL activities, they often only provide other grey literature.

We have been unsuccessful in terms of contacting grant agencies with proposals for cooperation with the NRGL. Reports from projects that a support beneficiary must prepare and submit to the grant agency contain valuable expert information, but are a type of grey literature document that is almost inaccessible to experts. The majority of support beneficiaries submit working and final reports to the applicable grant agency, which subsequently does not make them public. In general there is a fear of publishing data about research itself, but also about drawn funds. This, however, does not contribute towards transparency and credibility in the field of the provision of support from public funds for science and research in the Czech Republic. Experts are more interested in the part of the report with an analysis of the solution for the grant project, the general summary and the complete listing of all the results achieved through the grant project. The Ministry of Agriculture of the CR, for example, chose a good solution. It publishes so-called editorially modified final reports intended for publication. Although grant agencies were contacted in 2010, it has not been possible to establish cooperation. There were various reasons for this: some agencies highlighted a lack of manpower, the excessive difficulty of establishing cooperation, or refused without supplying a reason. The Technology Agency of the CR, established in 2009, was contacted with a proposal for cooperation in 2012. The agendas for the provision of support for research, which in previous years were maintained by the individual ministries, have been collected under this agency. The Technology Agency did not accept the proposal for cooperation either. Its reason was in particular the sensitivity of the data from research, which could be misused if published. There was also a legal problem – the Technology Agency does not have the right to transfer reports it receives to third parties – in this case the NRGL. The Technology Agency's contracts with support beneficiaries would have to be amended and the latter would have to agree with such amendments. The Technology Agency would be willing to commence cooperation if the publication of these research reports was imposed by law.

The production of grey literature in the ministerial environment can be split into two groups. The first group is represented by reports from projects and grants from public funds, where the ministries play the role of support providers. The support beneficiaries are usually public research institutions, universities, private societies or companies. This area is influenced by Act No 130/2002, on the Support of Research and Development from Public Funds (the Act on the Support of Research and Development) and Act No 121/2000 (the Copyright Act). Pursuant to Section 16 (1) and (2) of Act No 130/2002 ministries only have ownership rights to the fruits of a public contract, which are the minority. The support beneficiaries have ownership rights to the other results. Here it is necessary to negotiate directly with the individual support beneficiaries. Since 2012, in addition, some ministries have transferred the agenda for the provision of support to the Technology Agency mentioned above.

The second group is made up of grey literature produced by ministries and which is intended for publication. These are annual reports, yearbooks, bulletins, overviews, studies, statistics, analyses or reports on the state of the specific area that the ministry focuses on (e.g. the Report on the State of Human Rights in the Czech Republic in 2009 issued by the Ministry of Foreign Affairs of the CR, and so on). Even presentations and conference speeches by ministry employees are no exception to this. Usually these documents have already been published on the websites of the ministries. These materials have the character of employee work and the ministries have ownership rights to them. From the legal perspective, nothing prevents them from storing and publishing these materials intended for the public in the NRGL. The problem lies in the fact that grey literature is not centrally collected at ministries and the ministries do not usually have the staffing capacity to be able to designate an employee who would collect grey literature from throughout the ministry and place it in the NRGL. So far cooperation has only been commenced by the Ministry of the Environment, which only publishes records of research reports through the NRGL, by the Ministry of Justice, where the job of collecting grey literature has been taken over by the Press Department, and by the Ministry of Defence.

Public research institutions are legal entities whose main subject of activity is research supported, in particular, through public funds defined by Act No 130/2002, on the Support of Research and Development from Public Funds.[2] None of these institutions maintains its own database of grey literature as yet. If, therefore, public research institutions commence cooperation, they will use the NRGL to directly store their grey literature – in particular research and grant reports, certified methodologies, annual reports, analyses, studies and conference materials. So far eight public research institutions are cooperating with the NRGL, while discussions are underway with others. In the event of a refusal to cooperate, the reason most frequently given is the small quantity of grey literature, the appearance of sensitive information in grey literature, or lack of capacity.

Museums and galleries are specific producers of grey literature. They focus, among other things, on research activity and produce grey literature falling within the field of research, development and education. This results in conference materials, project reports, research reports, and certified methodologies. They also produce printed matter for displays and exhibitions (exhibition catalogues, exhibition guides etc.) On the other hand the main activity of museums is to create collections, which includes special digitisation of two-dimensional and three-dimensional objects. Galleries and arts-focused schools (e.g. The Academy of Fine Arts, Prague) emphasise the archiving and publication of works of art in digital form (e.g. pictures, sculptures, music, dance etc.) For these works a separate repository of collectors' objects and works of art should be created, because their handling and accessibility (description, formats etc.) significantly differ from grey literature. In spite of this the NTK is establishing cooperation with museums and galleries, which only includes grey literature falling within the current typology of NRGL documents, and is collecting data for the potential design of a project to expand the NRGL to include collectors' objects and works of art. To date only the Museum of Western Bohemia has commenced cooperation, contributing its annual reports to the NRGL. The main discussions with museums and galleries are only now gathering pace.

Grey literature in Czech libraries is found in particular in those that are part of the organisations mentioned above (Academy of Sciences, universities etc.) If we focus on independently established libraries, these have a relatively small quantity of grey literature, which they produce themselves or which they collect from other institutions. The exception is the National Medical Library (NML), which manages a collection of grant reports of the Ministry of Health of the CR. The NML has expressed its agreement with the collection of grey literature by signing a licencing contract with the NTK, and data is already being collected for the NRGL. The Moravian Library in Brno has recently commenced cooperation with the NRGL. It will publish through the NRGL both grey literature that it produces itself, and also records of grey literature that it has collected in its fund. Another potential partner for the NRGL is the National Library of the Czech Republic, which has already become a research organisation and is running many projects through which it produces grey literature. The National Library of the Czech Republic is planning to create an institutional repository, in which it would also collect grey literature that it produces, the records and documents of which it would subsequently also transfer to the NRGL repository.

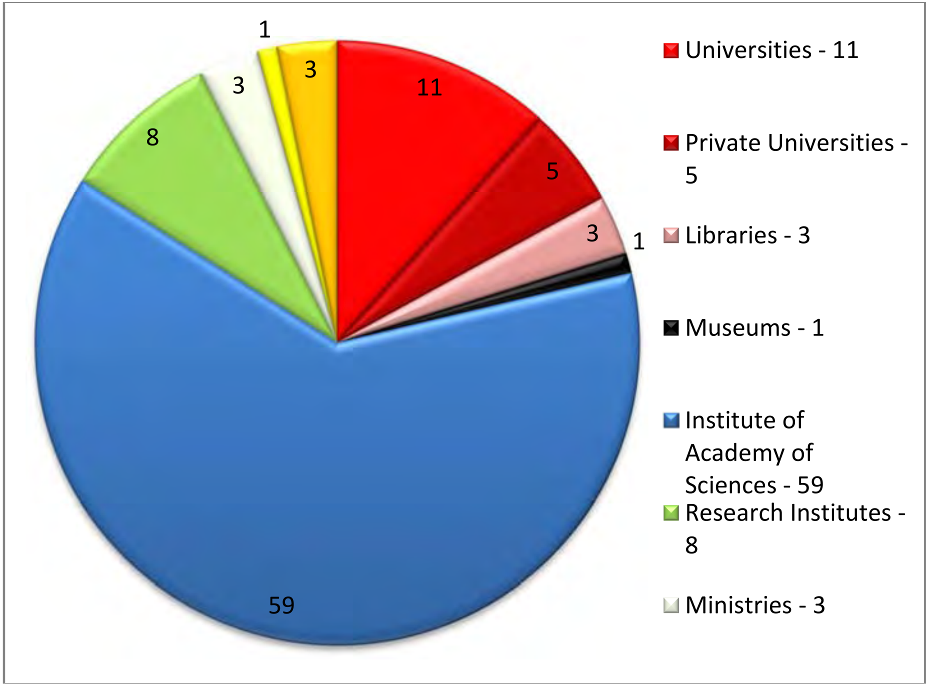


Figure 1: Types of partners

91 institutions were participating in the NRGL partnership network at the end of November 2012. The largest part of the partnership network is formed by institutes of the Academy of Sciences. These are followed by the already mentioned state-run universities, public research institutions, private universities, ministries, libraries and one museum. The Czech National Bank has not yet been mentioned. So far it is the only institution cooperating from this area; other institutions of this character have not yet been contacted. Within the framework of the NRGL there is also the possibility for scientists and researchers to establish their own personal archives of unpublished grey literature as private persons. Three persons have taken advantage of this possibility to date.

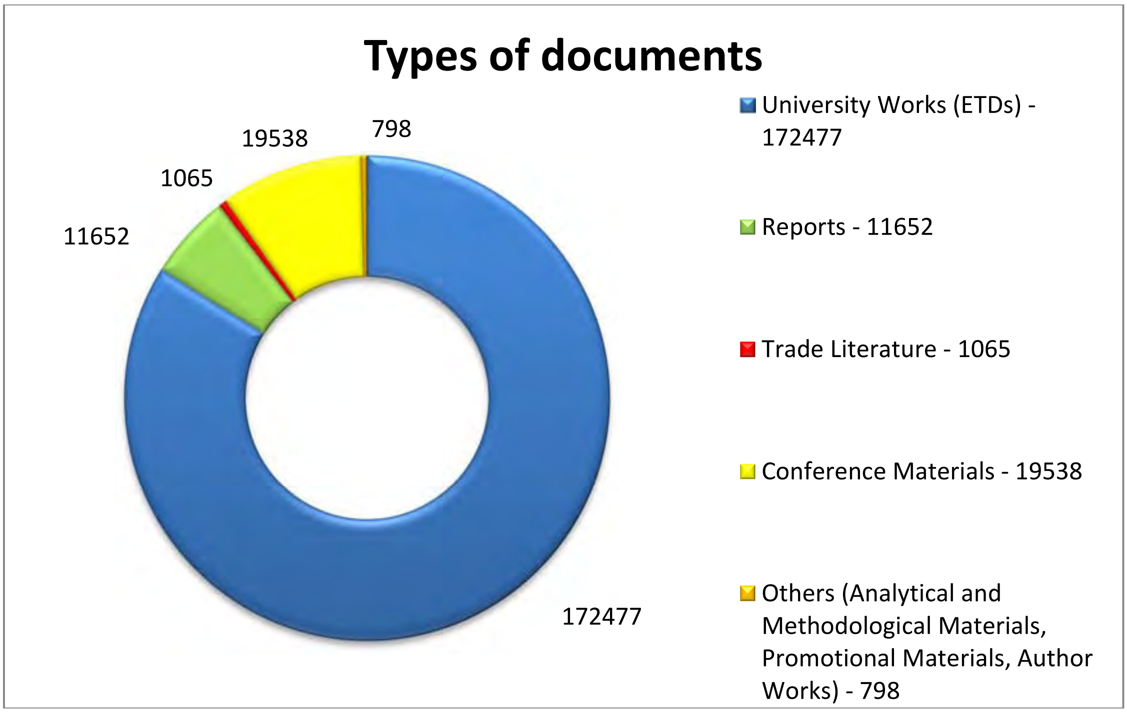


Figure 2: Types of documents

The graph shows that the most common type of document in the NRGL is university works. This is due to the fact that universities place records of university works from their own repositories into the NRGL, and this is always a large number of documents. These are followed by conference materials, of which there is also a large quantity collected in the NRGL. Various types of reports, trade literature, analytical and methodological materials, promotional materials and author works are less common.

Access to grey literature at national level is ensured through the Central Search Interface. At the end of November 2012 access to over 205,000 records was ensured through the NRGL. In view of the diversity of the access provided to documents, there is information about the availability of the primary document for each record. A document is either available directly in the Digital Repository of the NRGL, or a link is provided to the source repository of the partner institution, where access methods also differ. Only some of these repositories provide online access to digital documents. In the NRGL it is also possible to search for records of printed grey literature, and in such a case contact details are provided for the service that supplies copies of the document or that can lend it. Since the autumn of 2012 the Central Search Interface has also been adapted for mobile devices.

The following table shows the growth in records of documents over 3 years since 2010, when work began on building the NRGL partnership network.

Year	Amount of records
2010	34 290
2011	107 266
2012	205 530

Figure 3: Number of records in the Central Search Interface

Searches are performed in particular through navigation by document type, author, key words, document format, and connected databases. A timeline is also offered. The Central Search Interface is intended for end users and is available at http://www.nusl.cz.

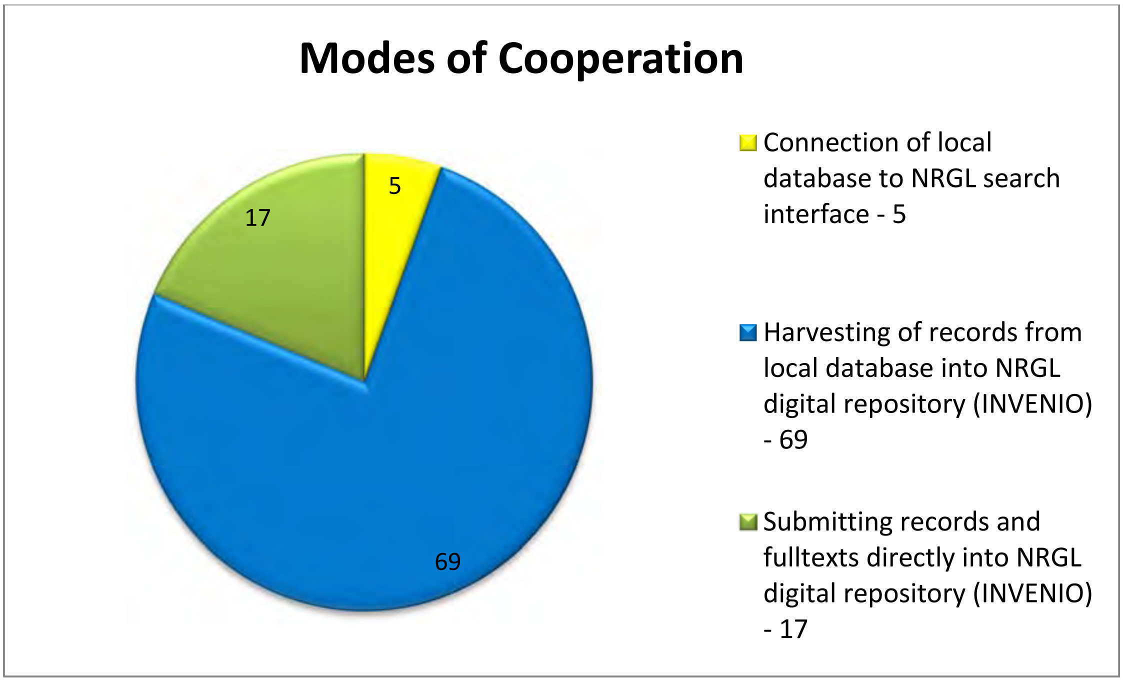


Figure 4: Modes of cooperation

The NRGL offers partner institutions several ways to cooperate. If an institution has its own repository (e.g. in the Digitool or DSpace system) for the long-term storage of grey literature documents, the repository can be connected to the Central Search Interface as a source (yellow colour – 5 institutions). Data from the repository of the institution are harvested using the OAI-PMH protocol, but are not long-term archived in the NRGL. Records from the Central Search Interface of the NRGL then link to the repository of the institution, which ensures long-term access to the digital documents.

If an institution has its own repository, but is also interested in the long-term archiving of metadata and full texts in the Digital Repository of the NRGL at the NTK, it can transfer the metadata as well as digital documents using the OAI-PMH protocol directly to the Digital Repository of the NRGL for long-term storage (blue colour – 69 institutions).

If an institution does not have its own repository and does not have the option of installing and subsequently managing its own institutional repository, it can enter metadata and digital documents directly into the Digital Repository of the NRGL (green colour – 17 institutions). For such an institution its own collection is created in the Digital Repository of the NRGL on the landing page of the Digital Repository of the NRGL. The institution is allocated its own user name and password for the collection in question, while an authorised employee is trained in the use of the Invenio system, and directly creates records and attaches digital documents to them in the repository. All the metadata appears in the Central Search Interface of the NRGL.

Through various means the NRGL team is working both to provide support to institutions from the NRGL partnership network, and also to use these means to help expand awareness about grey literature, related issues and so on. An important ancillary means is the website of the NRGL service http://nrgl.techlib.cz/. There is information for both potential and existing NRGL partners, news relating to the National Repository of Grey Literature as well as related activities, and basic information about grey literature. There are also all the necessary standards, manuals and recommendations for work with grey literature and cooperation with the NRGL, informative documents and analyses of Czech legislation that applies to grey literature. The website also includes a section dedicated to the annual seminar and the proceedings from this action http://nrgl.techlib.cz/index.php/Workshop. Information is also uploaded to Facebook in addition to the website: http://www.facebook.com/nusl.cz.

Important information is regularly sent to partner institutions within the framework of a moderated email conference. Support for cooperation with the NRGL includes the provision of local installation of the Invenio system, which interested institutions can install on their own hardware and create and manage their own digital repository.

Another step towards successful cooperation was the preparation of an analysis of legal relations that will arise between the NTK, producers of grey literature and database operators. For the needs of the project, JUDr. Radim Polčák of the Faculty of Law at Masaryk University in Brno prepared an opinion on the legal issue entitled “Digital Processing of Grey Literature for the National Repository of Grey

Literature", which is available from the project website. This analysis also contains six types of recommended formulations for licencing contracts, which are used depending on the mode of cooperation and the type of institution.

The legislative framework for processing, archiving and enabling access to digital grey literature documents is made up of several different legal norms. In Czech legislation we do not find any specific legal regulation that focuses purely on grey literature, and therefore it is necessary to comply with and verify the validity of several legal norms at the same time. This does not mean only the Copyright Act and the Act on Higher Education Institutions, but also the Act on the Support of Research and Development. Producers of grey literature are also the target of the text "Submitting Grey Literature to the National Repository of Grey Literature", which serves as a guide on how to proceed in individual cases of the publication of various types of grey literature, and which has been placed on the website of the NRGL service.

The Creative Commons licence supports easier publication, in particular electronically through the NRGL. The Czech Creative Commons licence was presented in 2009. In the NRGL every work is protected pursuant to the Copyright Act No 121/2000, and it is also possible to add to this the Creative Commons licence, which precisely defines the possibilities for the use of a work.

In order to improve the accessibility of documents the Open Access initiative is globally supported. This is an alternative to the traditional method of scientific communication. The objective is to reduce the access, financial and technical problems connected with scientific publication and this in particular through the use of the possibilities offered by the internet. In the Czech Republic the Working Group on Open Access was established in 2010, of which the NTK is a member.

Thanks to cooperation between the NRGL and international repositories and portals, access is enabled to Czech grey literature through the European grey literature database OpenGrey (formerly OpenSIGLE), DRIVER (Digital Repository Infrastructure Vision for European Research), ROAR (Registry of Open Access Repositories), OpenDOAR (The Directory of Open Access Repositories), Base, and the Ranking Web of Repositories. The NRGL is also indexed by the Google search engine, which significantly contributes towards its visitor rates.

[1] The National Repository of Grey Literature was established within the framework of the project "The Digital Library for Grey Literature – Functional Model and Pilot Implementation" running from 2008 to 2011.

[2] The source of information for public research institutions is the Register of Public Research Institutions. http://rvvi.msmt.cz/

Tools and Resources Supporting the Cultural Tourism

Eva Sassolini, Alessandra Cinini, Stefano Sbrulli, and Eugenio Picchi
Istituto di Linguistica Computazionale, ILC, Italy

Introduction

The diffusion of internet and the information technologies are creating continuous information flows. There is a widespread awareness of the added value and of the role that the Web has in dissemination, exploitation and promotion of the cultural tourism, especially in a country like Italy, where the cultural heritage is very important. Moreover, an open philosophy causes problems of authoritativeness in the production of contents because it is characterized by a strong interaction among users thus creating a distance between knowledge and communication. The spread of internet has brought the significant changes of communication paradigm. Nowadays the competition decreases among contents, even among from sources published in potential competition with them. In network logic, all nodes are interdependent and represent a single large hypertext. The proliferation of paths boosts a free circulation of ideas and can bring out most interesting contents[1].

The Projects

Methodologies, strategies, resources and tools created in our research group allowed us to take part in some national initiatives:

- "On-line dissemination of the historical artistic and landscaped, regional heritage" *(WeBasCH)*: The project was born within the framework of collaboration between the ILC-CNR Pisa and the APT Basilicata (i.e. Agenzia di Promozione Territoriale della regione Basilicata) to experiment and implement strategies for promotion and dissemination of regional heritage.
- *"SmartCity*: new solutions for content engineering and ambient intelligence as support of cultural tourism" is a Tuscany region project ("POR Creo" session), funded by the European Community (FESR). The specific project purpose is to allow important steps forward in terms of productivity, versatility and adaptability of digital content, both textual and multimedia, through knowledge engineering techniques. Particularly we have developed and tested tools for a better preservation and enhancement of cultural heritage, identifying methodologies and solutions to meet new demand of cultural spaces in particular for tourist purposes. Partners of this project were ILC-CNR and a consortium of companies: Space, Rigel Engineering and Meta.

The specific goal of both projects is experimenting and implementing strategies for promotion and dissemination of regional heritage through an effective communication style. By exploiting potentialities of the Web 2.0, we offered a set of chosen documents and cultural routes that increase knowledge of historic and landscaped resources. The aim is to join maintenance and preservation activities and enhancement and promotion of the cultural heritage initiatives with an increasing opening to the general user[2].

Text material for cultural tourism

The cultural tourism in Italy aims at identifying town of artistic interest or cultural tours in wider areas linked by historical events or traditions as the "Via Francigena", " Parco Archeologico Storico Naturale delle Chiese Rupestri ", "the wine roads", etc.. In this context, information available is in continuous evolution and so various that the database should be continually updated. The updating process of textual resources is an open question. The manually generation of resources is expensive and it requires the review of human experts. The manual approach is prone to errors of omission. Moreover, an approach based on an automatic updating of resources is not efficient because these are constantly evolving, this reason advises against an exclusive use of automatic applications. Especially when it comes to tourism, for which the Web offers constantly new ideas: a new park, an archaeological site discovered, etc..

There are many European initiatives within Cultural Heritage that aim to develop knowledge and enhancement of digital cultural heritage that have been undertaken. Among these, "Minerva" and "Michael" have been coordinated by the Italian Ministry for Cultural Heritage. Minerva has developed a platform of guidelines and recommendations, which are shared by European member states, for the digitization of cultural heritage and its network access. Similar information can be considered the reliable source of knowledge, especially because they were created with specific objectives:

- ✓ accessibility and visibility improvement of European digital cultural resources;
- ✓ support development of the European Digital Library for a better access to the cultural resources;
- ✓ contribution to increasing the interoperability among existing networks;
- ✓ promoting the use of digital cultural resources by business and citizens.

However the creation of this kind of resources is also referred to written text material such as brochures and web advertisement, that often is difficult to find, and retrieve. Such data regard an important source of information but since it tends to be original, recent and ephemeral can be considered a typical material of Grey Literature.

Specific strategies for texts acquisition

We have dealt the creation of the text corpus with different strategies in the two projects. Different needs of each project have guided the choices of the best acquisition strategy and all text material has collected in two steps:

1. We built a starting text corpus (hereafter C0) which was exploited to generate new linguistic resources and enriching the ones available from TextPower[1][3] project.
2. On the basis of these resources, we enlarged C0 to create the reference corpus.

CO in *SmartCity*

The first acquisition strategy in *SmartCity* project concerned the creation of a domain corpus “Empoli e dintorni", composed of historical, artistic and tourist documents related to the specific geographical area. Among activities of the project it has been possible collaborate with the library of Empoli and Cerreto Guidi for the retrieval of textual material.

On a first group of documents provided by the library of Empoli, we attempted to select material of interest in order to deal all the most important themes of Empoli and surroundings. On the basis of these texts we generated the Training Corpus whose size is almost 110,00 occurrences of words.

CO in *WeBasCH*

A different approach has been followed about *WeBasCH* project where all text material has retrieved on-line. The APT Basilicata provided us with a list of institutional websites. The items of the list more related to historical, artistic and landscaped regional heritage were selected and browsed by using automatic spiders and parsers, for the creation of a text corpus. Much of documentation available in internet can be assimilated to the “grey literature”, since it has been produced by the authors and institutions outside publishing, particularly the websites of regional entities.

Text corpus to linguistic resources

C0 was exploited to generate new linguistic resources and integrating the ones already available.

All textual material obtained has indexed and, after a tagging phase with the PiTagger[2] tool, we could identify all lemmas and relative POS in each document[4].

PiTagger associates each word to the related lemma by using the morphological component of the Italian language PiMorfo[3]. Then it solves the ambiguities by following a statistical approach and with the help of a training corpus.

Later on, all *multiword* expression (MWE) were extracted from C0 by exploiting pattern matching techniques. Typically for the Italian syntactic construction, the most productive linguistic patterns are N-preposition-N and ADj-N/N-ADj. Statistical algorithms analyze the distributions frequency of each pattern identified. On the basis of results we extracted a set of semantically relevant terms and concepts for the cultural heritage domain. The analysis of the collected texts by means of linguistic tools (morphological engine and tagger) is fundamental for productive application of the statistical functions of extraction[5].

We exploited enriched text material, that composes C0, to build weighed domain lexicons besides. Starting from a small set of relevant pivot terms a lexicon is obtained by means of *mutual information* criteria. Statistical algorithms analyze and weigh the frequency of the co-occurrency of each word with the pivot terms. The domain lexicons can be used to evaluate the relevance of a document for that domain, in this case it is most important to establish a minimal threshold.

Reference (text) Corpus

Since textual documentation collected didn’t cover exhaustively any events and cultural resources, we tried to increase the amount of textual material available, in order to find a maximum coverage of information and knowledge. In both projects we used a new search strategy of text material retrieval, namely, we used extracted knowledge and specialized crawlers that work on a bulk of text material available on-line. In a next phase we have developed other tools that, by using specific semantic filters,

[1] Project made for building of terminological resources and enrichment and annotation of textual material

[2] PiTagger is an important component for text lemmatization and tagging and constitutes a software module of PiSystem: integrated system for processing of textual and lexical materials.

[3] PiMorfo: system for morphological analysis of the Italian language.

make a ranking the documents and evaluate their relevance with the specific domain. All extracted documents were joined in C0 corpus to build the reference (text) corpus.

The (text) reference Corpus created in SMARTCITY project consists of 2219 text units:

- 1634 units come from the material provided by the library of Empoli and Cerreto Guidi, about 1.000.000 (997299) words;
- 585 units are related to texts retrieved on the Web, little more than 850.00 words (857355).

At the end of the project the specific Corpus has been reorganized in 650 documents in XML format.
In the *WeBasCH* project, the (text) reference Corpus instead is constituted of almost 2 million and half words. As it described, the creation required two phases, in which we retrieved documents from the Web and built the reference (text) corpus.

DBT-Faccette

The intelligent browsing system of text, named "DBT Faccette", is a customization of the categorization system used in librarianship. Its uniqueness lies in possibility of exploiting the semantic relevant elements (or "micro semantics") identified in the text for suggesting a further search. This feature makes most important the availability of an annotated corpus.
The text corpus is originally a set of text files. Inside these texts specific tools have identified annotations of various kinds: information related to words, phrases and piece of text. However it is necessary to organize the annotated texts in a coherent way and orderly, for a better storage, management and expansibility over time.
The corpus becomes a collection of digital sources containing accurate semantic annotations and necessary information to managing of the textual sources.
A useful way to imagine a good browsing system is to assume the final usage scenarios. Many developers often tend to adopt "reference systems" implicit, made of the individual experiences. Since the "reference system" of a user is not necessarily identical to ours, it's easy to fall into misunderstandings harmful to the design of a complex product.

The browsing system is able to exploit knowledge extracted from textual material in several ways:

- expanding the search by exploiting the search for MWE and then offering suggestions to research incomplete or vague. For example, by proposing the query "Pontormo" are returned as results also the occurrences of "Jacopo da Prontormo", "Jacopo Carucci" and "Pontormo";
- improving the correlation process among documents identified, by using specific linguistic resources;
- improving the ranking of results in the classification of answers;
- better organizing the display of the results. The system shows in reply at the queries a variety of contexts. In these results are highlighted all relevant entities or "micro-semantics".

An example of browsing in WeBasCH shows that some material constitutes examples of grey literature. For example, the search of "premio" proposes the site of Aliano, particularly the web-page where planned activities for the year 2008 are displayed. This material would be lost over time, even though it provides important information about the natural and cultural heritage of a region.
In our approach, the creation of linguistic resources is designed to the development of navigation and information retrieval systems, that are able to exploit them. These tools capture, organize, classify and distribute the information in according to the desired objectives. In an open domain, as the Cultural Heritage, information can rarely be classified just with hierarchical criteria. An approach based on principles of "semantic similarity" is more efficient. This approach allows to link information crosswise, seemingly belonging to different categories, but that match to the same informative need.
The experience in the treatment of large amount of data has not only allowed the refinement of extraction tools for semantically relevant information, but also the creation of terminological resources toolkit.
The more a text is "enriched" with annotations, the better it can be processed by tools for analysis, categorization, browsing and Information Retrieval. The system is not limited to the identification and classification of entities; it also identifies the particular relations between the entities involved.

Conclusion

Our work proposes to overcome the traditional categorization systems and their rigidity, by means of a set of open and adaptive terms classes, which can guide the end user in refining of his/her search.
Such knowledge systems are valuable support on the one hand to networks of e-participation and e-government, on the other hand they offer more information and better performance.

References

1. Spadoni F., Tariffi F., Sassolini E., (2011). SMARTCITY: Innovative Technologies for customized and dynamic multimedia content production for Tourism applications. In: EVA 2011 Florence Electronic Imaging and the Visual Arts. (Firenze, 4-5-6 may 2011). Proceedings, Cappellini Vito (ed.). Pitagora Editrice Bologna, 130 - 135.
2. Granieri, G., Et Al., (2009) Linguaggi digitali per il turismo. Edizioni Apogeo, November 2006.
3. Picchi, E. Et Al. (2009). “Text Power”: tools for Cultural Heritage. In Proceedings of in 4th International Congress on “Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin”, IMC (CNR) Rome, Italy, pp. 277--278.
4. Picchi, E. (1994). Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian. In Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Proceedings of Euralex ‘94, Amsterdam, The Netherlands.
5. Picchi E., Et Al., (2004). Linguistic Miner. An Italian Linguistic Knowledge System. In: LREC Fourth International Conference on Language Resources and Evaluation (Lisboa-Portugal, 26-27-28 May 2004). Proceedings, M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silvia (eds.), 1811 – 1814.

Semantic networks for improved access to biomedical databases

Eva Sassolini, Sebastiana Cucurullo, and Eugenio Picchi,
Istituto di Linguistica Computazionale, ILC, Italy

Introduction

The development of strategies and tools to access and analyze large amounts of data, so to discover correlations between seemingly unrelated data, capture associations and draw conclusions, is a research area of recent development in the acquisition of knowledge.

It is a fact that the study of innovative technologies has enabled an exponential growth of knowledge in the biomedical field, but the large amount of information available and the heterogeneity of information sources are a severe constraint to the full exploitation of such knowledge. The availability of systems for collecting and aggregating data as well as of analysis systems has therefore become a priority, mainly in fields which public health.

State of the art

Some research groups are developing tools and methods for summarizing of medical documents; others use specific Information Extraction (IE) techniques for building medical and biomedical ontologies, for example the infrastructure AMBIT (Acquiring Medical and Biological Information from Text), developed within the CLEF[1] and myGrid[2] projects. AMBIT aims to provide a intelligent access to large and unstructured biomedical data resources [1].

Recently, many efforts have been directed to the creation of large-scale terminological resources that merge information contained in various smaller resources: large thesauri based on a normalized nomenclature[2], extensible lexical and terminological databases like TERMINO[3] and the specialized Lexicon (e. g. BioLexicon[4], its peculiarity is to combine features of both terminologies and lexicons, within the project BootStrep[3]).

SUBITO project

SUBITO (Unique Social Network for Innovation in Biomedical Tuscany) is a "POR Creo" project, promoted by the Tuscany region and funded by the European Community (FESR). The project goal is creating an archive and a website to collect all specific domain information, regarding institutional or private players in the field of life science. The project inherits and improves some previous experiences carried out in Tuscany, like ORBIT, THRAIN, Net-TLS. These projects have already identified some synergies existing in the territory, regarding technical knowledge, activities, skills and potential scientific-technological.

The project involved the Institute of Computational Linguistics “Antonio Zampolli” (hereafter ILC), the Institute of Clinical Physiology (IFC) and a consortium of private companies. Particularly, ILC has developed tools and resources for the extraction and classification of textual data in order to enable a more efficient browsing.

Textual database

We created the knowledge base with the retrieval of abstracts and other information from three main websites: PubMed.gov, Espacenet.com, ClinicalTrials.gov.

PubMed is known as the most reliable and used repository for the publishing of biomedical articles, since it is a service of the U.S. National Library of Medicine and the National Institutes of Health that comprises more than 22 million citations for biomedical papers from MEDLINE and journals with content related to life sciences. PubMed has become currently the standard of reference for the scientific papers in biomedical domain.

This type of text is available in the Web but its consultation is difficult, especially if we want a selection of documents related to a specific sub-domain: typical problems that a information retrieval system has when a user wants to retrieve information from any knowledge base. It is important to improve and innovate the quality of services offered to users. The above-mentioned text material can be identified as "grey literature" because internet has transformed the electronic publishing. The Web offers new tools and channels for producing, disseminating and assessing scientific literature. Author/producer and reader/consumer changed their roles. The transformation of the research environment and the birth of

[1] The Clinical e-Science Framework (CLEF) project provides a repository of structured and well-organized clinical information which can be queried and summarized for biomedical research and clinical care.

[2] The project myGrid presents research biologists with a single unified workbench through which component bioinformatics services can be accessed using a workflow model.

[3] BootStrep (Bootstrapping Of Ontologies and Terminologies STrategic Project) is a STREP project of the FP6 IST (call 4), that involves six partners from four European countries (Germany, U.K.,Italy, France) and one Asian partner from Singapore.

new channels of scientific communication show clear that grey literature needs a new conceptual framework[1].

Terminological resources

In a more general context if is have systems for extraction, management and browsing of semantic relevant information, it is also possible to experiment new approaches for the automatic selection of those terms that are able to identify a specific domain of interest, even if these not are ontological nodes known. For example if we want to identify all the articles dealing with "rare diseases" or if we want to study what the "emergent diseases" in Tuscany, we need to identify a different domain.

Our research team works in the ILC and builds upon experiences in NLP techniques (text mining, text analysis). As recent research pointed out, the knowledge, intended as relationships and dependencies among the various relevant information contained in a text, can be extracted by means of text mining techniques and particular linguistic-statistical algorithms.

Typical text mining tasks do include text categorization, text clustering, concept/entity extraction; for our purpose, however, this is not sufficient. As a matter of fact, analyzing the collected texts material through linguistic tools (morphology and tagger) and resources (terminological dictionaries, lists of proper names, last names, geographical places, etc.) is fundamental for a productive application of the statistical functions of extraction, that would not by themselves offer guarantees to ensure the validity of the extracted data. Our aim is to develop not only tools for the analysis and synthesis of linguistic evidences, but also terminological data bases and specialized linguistic resources for textual analysis and named entities recognition.

Correct identification of terms and of all the semantic information in a text is essential for building a system of textual analysis, but also for classification and browsing of the text. Such a system is able to create relationships among semantically relevant information and also suggest synergies among private companies and public institutions, such it is required in SUBITO. The goal is to build a network of "knowledge" useful for an intelligent browsing, which is the real richness of the web services we offer to the project.

For this reason we developed a specific browsing system, text classification tools and semantic knowledge extraction systems.

The extracted features are mostly proper names, names of institutions, names of places and other relevant terms that characterize the specific domain.

Reference (Text) Corpus

In a first phase, we applied our attention to the creation of a reference (text) corpus of the biomedical domain. This training corpus was made up of a set of documents (in particular abstract of scientific articles) extracted from the PubMed website, where all texts are in English. All the resources and tools that constitute our background are in Italian, so it was necessary to adapt them in order to work in English. The same strategy will be adopted for the three other types of text documents considered: descriptions of projects, patents (EP, US e WO categories) and clinical trials (extracted by clinicalTrials.gov), in case the text size him will permit.

The creation of a specific reference corpus is a really important task and constitutes the basis of the whole process of creation of the specialized resources; the adaptation of the procedures to the project requirements, as well as the final editorial phase, are quite important since they can suggest new adaptations and improvements to the whole process.

1. Multi-word term extraction

After the creation of a specific reference corpus we extracted the relevant terms that will constitute the semantic information of the specific domain.

The creation of a biomedical ontology remains a valuable starting point for the extraction of knowledge and semantic associations by means of our statistical and linguistic tools. In order to meet the project requirements, we started the documents categorization using the tree MeSh[2] as knowledge base, like in PubMed. On the basis of MeSh tree we have then enriched the terminology by using our classification tools.

The extracted terms are not only those existing in online thesauri and dictionaries and belonging to different categories such as genes, proteins, drugs and molecules, etc. but are also those retrieved by semantic analysis procedures.

[1] D. J. Farace & J. Schöpfel (eds.) (2010). Grey Literature in Library and Information Studies. De Gruyter Saur

[2] The Medical Subject Headings (MeSH) is a huge vocabulary created by the National Library of Medicine (NLM) of the United States, with the goal of indexing scientific literature in the biomedical field. The 2008 version of MeSH contains a total of 24,767 subject headings, also known as descriptors. Because of these synonym lists, MeSH can also be viewed as a thesaurus.

For example, through the automatic extraction of ontological trees:

- ✓ Acid (.. acid, etc.);
- ✓ Agent (.. immunosuppressant, etc.);
- ✓ Alcohol (methyl .. etc..);
- ✓ Rare diseases.

The extraction of terms (simple and compound words) linked to a domain terminology can be another example:

- ✓ immuno-suppressant agent, chromosome

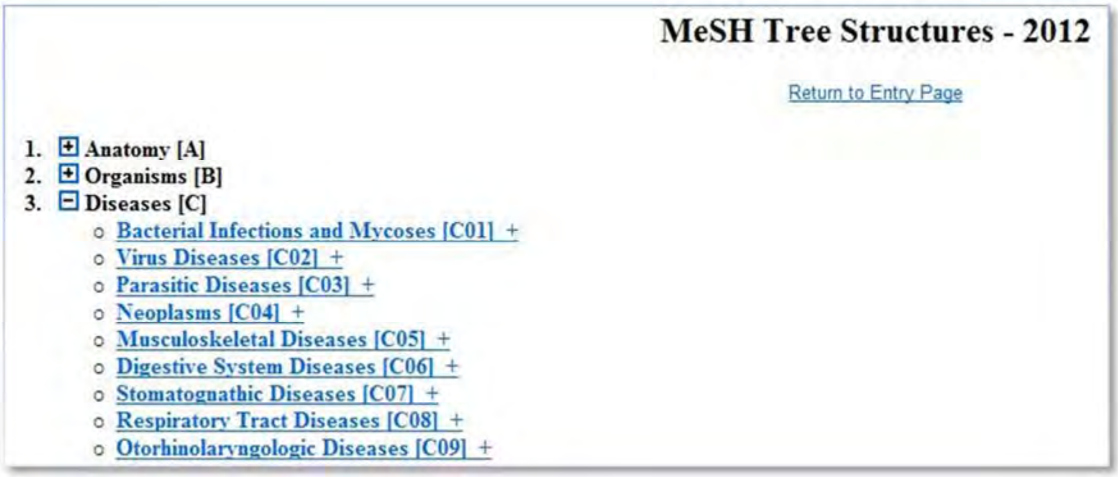
It is also possible to extract the events:

- ✓ tumor growth, low blood sugar, cardiovascular collapse.

Once the reference corpus is built, the elaborate terminology can be extracted and used for the creation of a knowledge network.

Semantic filtering

From the same set of features, we extracted the terms for the creation of domain dictionaries, which, in our case, coincided with the main MeSh sub-tree, for example, starting from the category “Diseases [C]”, vocabularies have been created for the 23 subcategories: “Bacterial Infections and Mycoses [C01]”, “Virus Diseases [C02]”, etc.



MeSH Tree Structures - 2012

Return to Entry Page

1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊟ Diseases [C]
   - ○ Bacterial Infections and Mycoses [C01] +
   - ○ Virus Diseases [C02] +
   - ○ Parasitic Diseases [C03] +
   - ○ Neoplasms [C04] +
   - ○ Musculoskeletal Diseases [C05] +
   - ○ Digestive System Diseases [C06] +
   - ○ Stomatognathic Diseases [C07] +
   - ○ Respiratory Tract Diseases [C08] +
   - ○ Otorhinolaryngologic Diseases [C09] +

Figure 1: Example of MeSh tree structures

Each terminological lexicon was created with statistical procedures that measure the relevance of a term to the domain, in order to create the semantic filters or “topics”.

The term "topic" identifies an area of interest chosen according to the project requirements. In fact the whole MeSh tree contains sub-trees that, after assessment, were deemed inadequate for the construction of a specific domain lexicon.

In general, the creation of a domain lexicon begins with the selection of pivot terms that have a high semantic value for the same domain, in this case specifically are the MeSh nodes. The lexicon also includes those terms having a higher co-occurrence value with the pivot terms, but that are not necessarily MeSh nodes.

Hence, the decision to acquire all nodes as basic terminology, but to use only some of these to create the semantic filter. In the light of the above considerations, we made a targeted decrease of nodes and sub-nodes, aimed at selecting the categories of greatest relevance to the areas of new technologies and research in the biomedical field.

Text browsing system

The browsing system “DBT-Faccette” provides primitives to be integrated in the project website using the terminological basis identified and allows the automatic re-organization of content, based on the salient concepts.

This approach allows the user to dynamically discover the concepts semantically relevant for the domain, and to carry out search refinements through the interrelated concepts.

An alternative access to content is the search for topics, which is as important as a traditional browsing of textual content. This research modality provides the user with a selection of crucial documents, ordered by their relevance to the topic. In this way it is allowed to measure the ranking of an document with respect to the topic.

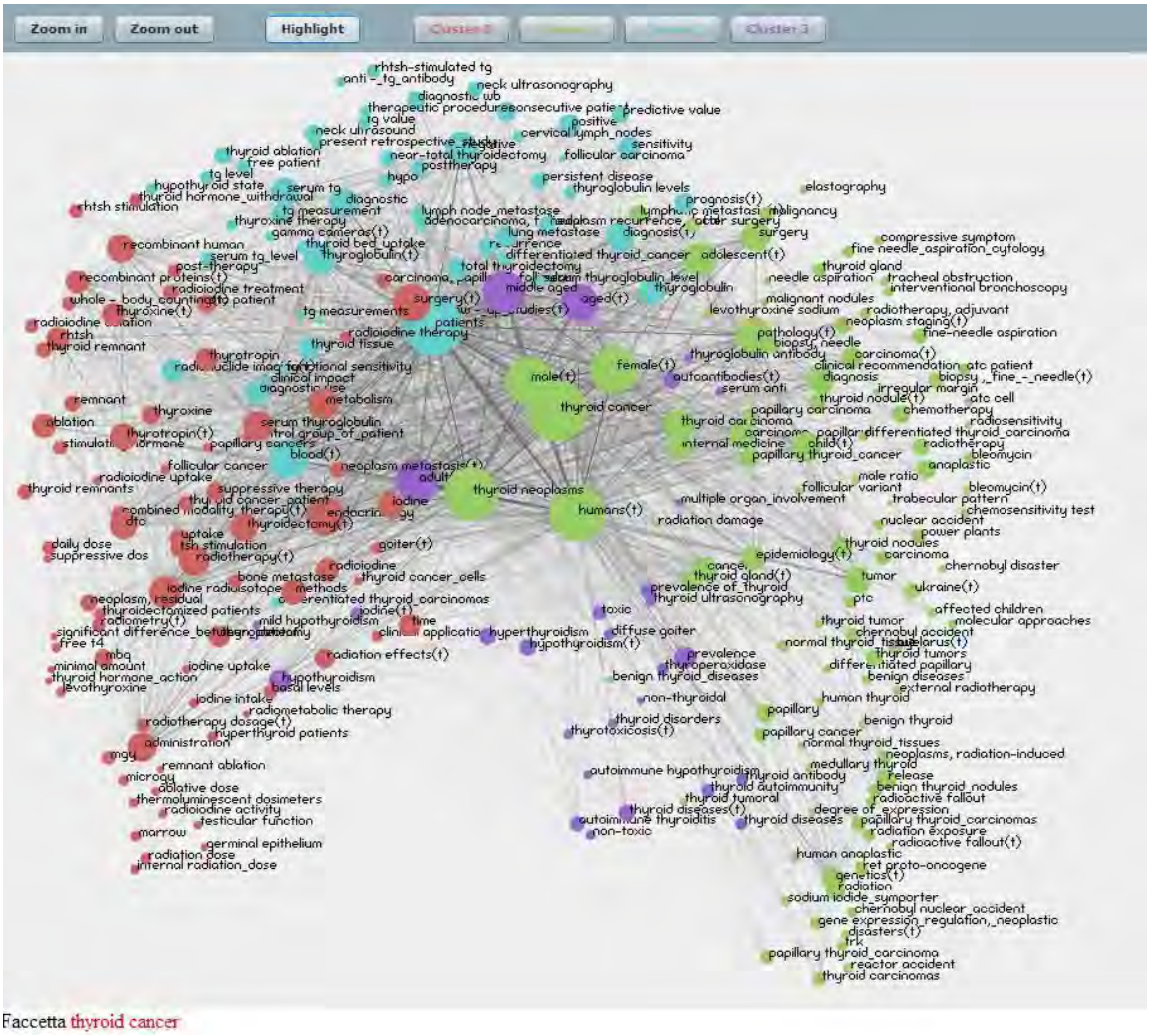


Figure 2: graph of "thyroid cancer" query

2. Conclusion

Through the semantic browsing tools, SUBITO can allow researchers of the Tuscany region to benefit from a set of information that will facilitate the development of new synergies with consequent positive effects on employment, economic growth and citizens' welfare.

Furthermore, these initiatives will allow to an audience of European researchers to use such information through the pages of the portal CORDIS, that provides to the regions a space for the dissemination of research activities on its territory.

References

1. Harkema H., et al.: Information Extraction from Clinical Records. In S.J. Cox (ed.), Proceedings of the 4th UK e-Science All Hands Meeting, Nottingham. UK (2005)
2. Kors, J.A., et al.: Combination of Genetic Databases for Improving Identification of Genes and Proteins in Text. In: Proceedings of the BioLINK 2005. ACL (2005)
3. Harkema, H., et al.: A Large Scale Terminology Resource for Biomedical Text Processing. In: Proceedings of the BioLINK 2004, pp. 53–60. ACL (2001)
4. Quochi V., et al.: A Standard Lexical-Terminological Resource for the Bio Domain. In: Lecture Notes in Artificial Intelligence, vol. 5603 pp. 325 - 335. Human Language Technology - Challenges of the Information Society. Z. Vetulani and H. Uszkoreit (eds.). Springer Berlin / Heidelberg. (2009)
5. Picchi E., et al.: The "Micro semantics" for intelligent browsing. In: CHC 2011 - 4-th Intl. Congr. Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin (Istanbul, 22-25-11 2011). In: Proceedings of Congress, pp. 286. Valmar, Roma (2011)

List of Participating Organizations

Australian Council for Educational Research, ACER	Australia
Bangalore University	India
Battelle Memorial Institute	United States
Benaki Phytopathological Institute	Greece
Biblioteca Centrale "G. Marconi"; CNR	Italy
BMS College of Engineering	India
British Library, BL	United Kingdom
Centre National de Recherché Scientifique, CNRS	France
Dalhousie University, DAL	Canada
Data Archiving and Networked Services, DANS	Netherlands
EBSCO Publishing	United States
Elsevier	Netherlands
European Commission, EC	Belgium
European Food Safety Authority, EFSA	Italy
Federal Library Information Network, FEDLINK	United States
Food and Agriculture Organization of the United Nations, FAO	Italy
Government P.U. College	India
Grey Literature Network Service, GreyNet	Netherlands
Health Information Network Calgary, HINC	Canada
Information International Associates, IIa	United States
Institut de l'Information Scientifique et Technique, INIST	France
Institute of Information Science and Technologies, ISTI-CNR	Italy
Institute of Marine Sciences, ISMAR-CNR	Italy
Institute of Research on Population and Social Policies, IRPPS	Italy
Irvine Valley College Library	United States
Istituto di Linguistica Computazionale, ILC	Italy
Istituto Superiore di Sanità, ISS	Italy
Japan Atomic Energy Agency, JAEA	Japan
Japan Science and Technology Agency, JST	Japan
Korea Institute of Science & Technology Information, KISTI	Korea
Library of Congress, LC	United States
Mysore University Library	India
National Diet Library, NDL	Japan
National Documentation Center	Greece
National Research Council, CNR	Italy
National Technical Library, NTK	Czech Republic
New York Academy of Medicine, NYAM	United States
Open University	United Kingdom
PricewaterhouseCoopers, PwC	Netherlands
Research Centre on Scientific and Technical Information, CERIST	Algeria
Science & Technology Facilities Council, STFC	United Kingdom
Senate Library	Italy
Slovak Centre of Scientific and Technical Information, CVTISR	Slovak Republic
Springer Verlag GmbH	Germany
Swinburne University of Technology, SWIN	Australia
Technische Informationsbibliothek, TIB	Germany
TextRelease	Netherlands
Thomson Reuters	Italy
Université Charles de Gaulle Lille 3	France
Université Claude Bernard Lyon 1	France
University of Bergen, UiB	Norway
University of Calgary	Canada
University of California, Irvine Libraries, UCI	United States

Author Information

Massimiliano Assante holds a master degree (M.Sc.) on Information Technologies received from the University of Pisa, where he is undertaking a Ph.D. on Information Engineering. He is research staff at the Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" (ISTI), an institute of the Italian National Research Council (CNR). He joined ISTI in 2007 and is currently member of the iMarine EU Project and EUBrazilOpenBio Project. In the past he has been member of D4Science II, D4Science, DILIGENT and DRIVER European Projects. His research interests include Data Infrastructures, Next Generation Digital Libraries, Information Systems and NoSQL Data Stores. Email: massimiliano.assante@isti.cnr.it

Anne Asserson holds a master from the University of Bergen, UiB. She has been working with Research Documentation, and has participated in substantial parts of CRIS developmental work, locally and nationally since 1992. Anne Asserson has been part of the establishing and implementing of several CRIS both at the UiB and nationally. For several years she was the chairwoman of the Steering Group of the national CRIS system and project secretary of a National system for academic administration. Anne Asserson is presently representing UiB in the national project group of CRIStin. She has also participated in The CORDIS funded European-wide project on " Best Practice" 1996 and was a member of the working group set up 1997 that produced the report CERIF2000 Guidelines (1999) www.cordis.lu/cerif, coordinated by the DGXIII-D4. euroCRIS is now the custodian of the CERIF model www.eurocris.org. Anne Asserson is a member of the euroCRIS board with the responsibility Member Strategy and External Relations. anne.asserson@fa.uib.no

Alessia Bardi received her MSc in Information Technologies in the year 2009 at the University of Pisa, Italy. She is a PhD student in Information Engineering at the Engineering Ph.D. School "Leonardo da Vinci" of the University of Pisa and works as graduate fellow at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. Alessia is currently involved in EC funded projects for the aggregation, curation and export of library, archival and museum digital objects and metadata records. Her research interests include Digital Library Management Systems, data interoperability, compound object management, and service-oriented infrastructures.
Email: alessia.bardi@isti.cnr.it

Stefania Biagioni graduated in Italian Language and Literature at the University of Pisa and specialized in data processing. She is currently a member of the research staff at the Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" (ISTI), an institute of the Italian National Research Council (CNR) located in Pisa. She is head librarian and member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She has been the responsible of ERCIM Technical Reference Digital Library (ETRDL) and currently of the PUMA (PUblication MAnagement) & MetaPub, a service oriented and user focused infrastructure for institutional and thematic Open Access repositories looking at the DRIVER vision, http://puma.isti.cnr.it. She has coauthored a number of publications dealing with digital libraries. Her activities include integration of grey literature into library collections and web access to the library's digital resources, including electronic journals and databases. She is a member of GreyNet since 2005.
Email: stefania.biagioni@isti.cnr.it

Leonardo Candela is a researcher at Networked Multimedia Information Systems (NeMIS) Laboratory of the Italian National Research Council - Institute of Information Science and Technologies (CNR - ISTI). He graduated in Computer Science in 2001 at University of Pisa and completed a PhD in Information Engineering in 2006 at University of Pisa. He joined the NeMIS Laboratory in 2001 and was involved in various EU-funded projects including CYCLADES, Open Archives Forum, DILIGENT, DRIVER, DELOS, D4Science, D4Science-II and DL.org. He was a member of the DELOS Reference Model Technical Committee and of the OAI-ORE Liaison Group. He is currently involved in the iMarine and EUBrazilOpenBio projects. His research interests include Digital Library [Management] Systems and Architectures, Digital Libraries Models, Data Infrastructures. leonardo.candela@isti.cnr.it

Joseph Candlish is a Technical Assistant and GIS Analyst at Information International Associates, Inc. (IIa). He received his M.S. in Biosystems Engineering Technology from the University of Tennessee, Knoxville and his B.S. in Environmental Studies: Natural Resources from Sewanee: The University of the South. At IIa he provides secretariat services to CENDI agencies, the federal scientific information manager's group and also fulfills a mapping role for the USGS National Biological Information Infrastructure's Southern Appalachian Information Node (SAIN). jcandlish@iiaweb.com

Lydia Chalabi is a researcher at Research and Development in Information Sciences Division in Algerian Research Center on Scientific and Technical Information (CERIST). She studied Library Sciences at university of algeirs2 where she obtained her first graduate diploma by studying the on-line Algerians scientific journals. And a Master's Degree related to the provision of on-line bibliographic services: a case study of Algerian books. Since 2010, she is preparing a PhD in information sciences about the impact of open access on the Algerians researcher's science. Since 2009, she teaches the documentation networks course in the CERIST e-learning platform. For 2006 to now, she held a various research activities and projects in the field of information sciences and currently, she is a research project leader concerning the impact of open access on an Algerian scientific literature. Her research special fields of interests are on Open Access to information sciences, open archives and open repositories, e-publishing, impact factor, open access citation impact and electronic resources. Email: net_lydia@yahoo.fr

Narayanappa Chowdappa obtained a Master's Degree in Geology and also Master's Degree in Library and Information Science from Bangalore University, Bangalore with a distinction. Obtained Doctorate degree for his Thesis on "Organization and Use Patterns of Grey Literature in Engineering Research Institutions" from the University of Mysore, Mysore in Library and Information Science. Serving as Chief Librarian at BMS College of Engineering, Bangalore for the last 28 years. Special interest includes promoting the use of grey literature among faculty and researchers in engineering discipline. Organized 24 professional and extension programs for Teachers and Librarians in Science and Technology. Served as resource person in library and information science, and Academic Counselor for Indira Gandhi National Open University, New Delhi. Presently Dr. Chowdappa is holding the position of President, AKELPA - All Karnataka Engineering College Library Professionals Association, Bangalore. Areas of professional interest are: Scholarly Communications, Facilitating Research Programs and Reference Service Email: ncbmsce@yahoo.co.in

June Crowe is the Senior Researcher and Group Manager, Open Source Research Division at Information International Associates, Inc. (IIa). She received her AMLS from the University of Michigan, Ann Arbor and her M.Ed. in geographic education from the University of Georgia, Athens. She has extensive experience in the management and operations of library services across government, public, academic, and special libraries. At IIa she manages the open source research division which focuses on medical, socio-cultural, science and technology and business research. Her primary interests are open source information in Grey Literature, digital repositories, and open source intelligence tools. Email: jcrowe@iiaweb.com

Elizabeth M. De Santo is an Assistant Professor in the Marine Affairs Program at Dalhousie University, Halifax, Canada. Dr. De Santo holds a BA (Honors, Zoology) from Connecticut College, a Master of Environmental Management from Duke University, a Master of Science (International Relations) from the London School of Economics and Political Science, and a MPhil & PhD (Geography & Law) from University College London. Prior to her appointment at Dalhousie University, Dr. De Santo was the Marine Protected Areas Coordinator with the International Union for Conservation of Nature, based in Washington, DC; Program Manager with the World Environment Center, New York; and a Researcher with the American Museum of Natural History, New York. Email: elizabeth.de.santo@dal.ca

Rosa Di Cesare is responsible for the library at the Institute for research on populations and social policies of the National Research Council (CNR). She worked previously at the Central library of CNR where she became involved in research activities in the field of Grey literature (GL) as member of the Technical Committee for the SIGLE database. Her studies have focused on the use of GL in scientific publications and recently on the emerging models of scholarly communication (OA and IR). Email: r.dicesare@irpps.cnr.it

Marta Dušková studied library and information science at Comenius University in Bratislava (Slovakia). Since July 2010 she works in the Slovak Centre of Scientific and Technical Information in Bratislava (Slovakia) in Publication Evaluation Department. She deals with grey literature and The Central Registry of Publication Activity. She coordinates activities associated with obtaining and making grey literature available and cooperates with the processing and verification data publications included in The Central Registry of Publication Activity. She also participates as a team member at three national projects: National Information System Promoting Research and Development in Slovakia - Access to Electronic Information Resources, Infrastructure for R&D - Data Centre for Research and Development, and

Author Information

National Infrastructure for Technology Transfer in Slovakia. From 2012 she is studying PhD study at Comenius University in Bratislava (Slovakia) with the theme: Knowledge communication with support of grey literature. Email: marta.duskova@cvtisr.sk

Helen Galatis is a Research Fellow with the Australian Council for Educational Research (ACER) responsible for research, information retrieval, and online aggregation and publishing for the Digital Education Research Network. Formerly, Galatis has managed all aspects of the online component of the Education Network Australia service for the Australian Government. Galatis specialises in online communities, web services, metadata and archiving. Email: helen.galatis@acer.edu.au

Julia Gelfand is the Applied Sciences and Engineering Librarian at the University of California, Irvine and has written and presented extensively for over two decades on different aspects of grey literature and its challenges in collection development for libraries and the emerging technologies that support grey literature. Email: jgelfand@uci.edu

Silvia Giannini graduated and specialized in library sciences. Since 1987 she has been working in Pisa at the Institute for the Science and Technologies of Information "A. Faedo" of the Italian National Council of Research (ISTI-CNR) as a librarian. She is a member of the ISTI Networked Multimedia Information Systems Laboratory (NMIS). She is responsible of the library automation software "Libero" in use at the CNR Research Area in Pisa and coordinates the bibliographic and managing activities of the ISTI library team. She cooperates in the design and development of the PUMA (PUblication MAnagement) & MetaPub, an infrastructure software for institutional and thematic Open Access repositories of published and grey literature produced by CNR. Email: silvia.giannini@isti.cnr.it

Mayuki Gonda works as librarian at the Central Library of JAEA (Japan Atomic Energy Agency). He joined JAERI (former JAEA) in 2005, and had been working for management and dissemination of JAEA research results information. Since 2009, he is in charge of selection, classification and indexing for INIS (International Nuclear Information System) Database. He is also a member of editorial committee of the Journal of Information Science and Technology Association (INFOSTA) since 2008. He holds a degree in information science (M.A.) from the Graduate School of Library, Information and Media Studies, University of Tsukuba. gonda.mayuki@jaea.go.jp

Armand Gribling is working in the Fisheries & Aquaculture Branch Library in FAO and is involved in the contents development of the Aquatic Commons repository, and member of the AC Board. He is co-author of a paper presented at the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC) Conference in 2010: Visibility and access through the Aquatic Commons. Email: armand.gribling@fao.org

Despina Hardouveli is a senior Information Professional in Greece having an experience of nearly 30 years in the field. She graduated in Biology and holds a master degree. Since 1983 she works for National Documentation Centre (NDC/NHRF) in Greece. From 1993 to 2006 was Head of the S+T Information Services Department of NDC/NHRF, responsible for the provision of scientific & technological information to the Greek research and academic community. Since 1997, she has been a member of the project group for National Information System for Research and Technology (NIRST), a project elaborated within the NSRF (National Strategic Reference Framework) and the European Operational Program, responsible for several tasks, among them the settlement of international databases into the NDC's information systems, the support to the creation of the Electronic Reading Room, the transition from traditional retrieval services to digital content services, the conceptual design and content management of special bibliographic collections, the digital library integration. In charge of the creation of the NHRF's OA Repository "Helios", she is responsible for the content maintenance and upgrading, providing supportive services to the stakeholders/users. She was instrumental in the development of a funder repository of mainly grey literature material of diverse types, produced under the funding programmes of the Hellenic Ministry of Education (co-financed by the EU). She has served on numerous project committees and was demonstrator & trainer of the digital content services of NDC. From 2004 to 2010 was a member of GRNET S.A (Greek Research & Technology Network) Administrative Board. Email: dxardo@ekt.gr

Misa Hayakawa works as librarian at the Central Library of Japan Atomic Energy Agency(JAEA). She Joined JAEA in 2010, and has been working for managing metadata of both papers and oral presentations published by JAEA researchers. In addition, she disseminates such information on the Internet via "JAEA Originated Papers Searching System"(JOPSS). She got the master's degree in Library and Information Science at the Graduate School of Library Information and Media Studies, University of Tsukuba (in Japan). hayakawa.misa@jaea.go.jp

Nikos Houssos works at the National Documentation Centre/NHRF in Athens, Greece as Head of the Software Development Unit. He is the software architect of the Greek "National Information System on Research and Technology" (EPSET) which comprises a variety of scholarly communications systems such as CRIS, repositories, e-publishing platforms and bibliographic systems. He has designed a number of grey literature systems including the Hellenic National Archive of Doctoral Dissertations. He participates in various EU FP7 projects like OpenAIRE/OpenAIREPlus, Arrow Plus, ENGAGE and PAERIP. He is a member of the euroCRIS Board and a contributor to the development of the CERIF data model. He has participated in EU FP5 and FP6 IST projects related to mobile networking and services (1999-2004), and lectured at the Technical University of Crete (2004-2007). He holds a Ph.D. in Computer Science from the University of Athens and has co-authored more than 30 peer-reviewed publications in international journals and conferences. nhoussos@ekt.gr

Keith Jeffery is currently Director International Relations at STFC (Science and Technology Facilities Council) based at Rutherford Appleton Laboratory. Keith previously had strategic and operational responsibility for ICT with 360,000 users, 1100 servers and 140 staff. Keith holds 3 honorary visiting professorships, is a Fellow of the Geological Society of London and the British Computer Society, is a Chartered Engineer and Chartered IT Professional and an Honorary Fellow of the Irish Computer Society. Keith is currently President of ERCIM and President of euroCRIS, and serves on international expert groups, conference boards and assessment panels. He had advised government on security and green computing. He chaired the EC Expert Groups on GRIDs and on CLOUD Computing. keith.jeffery@stfc.ac.uk

Hye-Sun Kim is the manager of dept. of NDSL Service at the Korea Institute of Science and Technology Information (KISTI). She has Ph.D in Library and Information Science from Ewha Women's University (2012) and a Master of Art degree in Library and Information Science from Ewha Women's University (1994). Her doctoral dissertation title is 'study on the factors influencing foreign journal subscription in university and college libraries'. Her research interests include: information services, collection development for journals, Open Access, institutional repositories. hskim@kisti.re.kr

Maria V. Kitsiou is born in Ioannina, Epirus, in north-western Greece. She holds a BSc in Archive and Library Science and a MSc in Library Management from Ionian University (Corfu, Greece). She has worked as a Librarian in a range of Libraries (Public Library of Lefkas, Special Library of Technical Camper of Greece, Academic Library of Technological Educational Institute of Ionian Islands). Since 2009 she has been working in the Library of Benaki Phytopathological Institute acting as a Head Librarian. Her interests include library assessment, quality systems implementation, information literacy-user education, grey literature management systems, e-repositories, open access, etc. Also, she has a publication in Library Management (vol. 29, issue 6/7, 2008) titled "Issues and perceptions for ISO 9000 implementation in Greek Academic Libraries". Email: m.kitsiou@bpi.gr

Július Kravjar graduated in Mathematics at the Comenius University of Bratislava and later in Informatics. He is currently responsible for the „Central Repository of Theses and Dissertations" and „Plagiarism Detection System for Slovak Academic and Research Institutions" projects at the Slovak Centre of Scientific and Technical Information (SCSTI). Both nationwide systems are in the real operation from April 2010. He also participates as a team member at three other national projects: National Information System Promoting Research and Development in Slovakia - Access to Electronic Information Resources, Infrastructure for R&D - Data Centre for Research and Development, and National Infrastructure for Technology Transfer in Slovakia. Prior to joining the SCSTI, he held several positions in software development and software and ICT services marketing in the private sector. julius.kravjar@cvtisr.sk

Sandro La Bruzzo received his MSc in Information Technologies in the year 2010 at the University of Pisa, Italy. He is working as graduate fellow at the Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. His research interests include Digital Library Management Systems, compound object management, and Service-oriented Data Infrastructures. Email: sandro.labruzzo@isti.cnr.it

Amanda Lawrence is a researcher with Swinburne University's Institute for Social Research. Ms Lawrence is managing editor of Australian Policy Online and a guest lecturer in information architecture at the Royal Melbourne

Author Information

Institute of Technology (RMIT). She holds a Graduate Diploma in Library and Information management (RMIT) and BA (Hons) Arts (Melbourne). She is qualified for membership of the Australian Library and Information Association (ALIA). Email: alawrence@swin.edu.au

Seon-Hee Lee is Senior Researcher at the Korea Institute of Science and Technology Information (KISTI). She has a master's degree in Library and Information Studies (MLIS) from the University of California, Los Angeles (1996) and a Master of Art degree in Philosophy from Ewha Women's University (1988). Her research interests include: collection development, grey literature, e-journals, information services, and collaborative digital reference services (CDRS). Email: wisdom@kisti.re.kr

Anthony Lin formerly a Business Librarian, is currently the Head of Reference and Technical Services at the Irvine Valley College Library, Irvine, CA, USA and this is the 4th major presentation that these authors have collaborated on. alin@ivc.edu

Yongtao Lin has been working as a health information network librarian at the Tom Baker Cancer Knowledge Centre in Calgary Alberta since 2008. She provides library services to support health care professionals in their evidence-based practice. The library is part of a provincial patient-centered education strategy supporting cancer patients and their families. She was a hospital librarian in rural Nova Scotia for a few years before moving to the University of Calgary. Her prior experience as an instructor has led her to integrate education into various aspects of library programs. Yongtao is interested in the impact of grey literature in health care and a strong believer in evidence-based practice. Yongtao was the awarded the Canadian Hospital Librarian of the Year in 2011. yolin@ucalgary.ca

Daniela Luzi is researcher of the National Research Council at the Institute of research on populations and social politics. Her interest in Grey Literature started at the Italian national reference centre for SIGLE at the beginning of her career and continued carrying out research on GL databases, electronic information and open archives. She has always attended the International GL conferences and in 2000 she obtained an award for outstanding achievement in the field of grey literature by the Literati Club. Email: d.luzi@irpps.cnr.it

Bertrum MacDonald is a Professor of Information Management in the School of Information Management at Dalhousie University, Halifax, Canada. With a background in science (BSc, Biology), history of science (MA), and information science (MLS, PhD), he pursues research that investigates the dissemination and use of scientific information in historical and contemporary contexts. He pursues interdisciplinary research, particularly within the Environmental Information: Use and Influence initiative (www.eiui.ca), since this work tackles large questions from the point of view of several relevant disciplines. He has been Director of the School of Information Management and Associate Dean (Research) in the Faculty of Management at Dalhousie University. He can be seen speaking about research projects at local, national, and international levels, and he holds executive positions with national and international associations. In 2004, he won the International GreyNet Award with his research colleagues, Ruth Cordes and Peter Wells. He is the recipient of the Marie Tremaine Medal, the highest award of the Bibliographical Society of Canada, and he was awarded a Dibner Research Fellowship at the Smithsonian Institution in 2001. bertrum.macdonald@dal.ca

Paolo Manghi received his PhD in the year 2002 from the Dipartimento di Informatica of the University of Pisa, Italy. He is presently working as a researcher at the Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Consiglio Nazionale delle Ricerche (CNR) of Pisa, Italy. His research interest include Data Models for Digital Library Management Systems, Types for Compound Objects, data curation in Digital Libraries, service-oriented ICT infrastructures with special focus on data ICT infrastructures. Paolo.Manghi@ISTI.CNR.IT

Claudia Marzi is a research fellow at the Institute for Computational Linguistics "A. Zampolli" (ILC), National Research Council (CNR), Pisa. PhD in Acquisitional and Computational Linguistics, University of Pavia, in progress. Laurea degree in Modern Languages (English) at Pisa University in 1998, with the dissertation "The power of words: language creativity in Edgar Allan Poe's narrative". Programme Coordinator of the European Science Foundation Research Networking Programme "The European Network on Word Structure. Cross-disciplinary approaches to understanding word structure in the languages of Europe". Member of board at Institute for Computational Linguistics. Main areas of interest: Computer modelling of the Mental Lexicon; Second language acquisition; Child language; Document and knowledge management. Email: claudia.marzi@ilc.cnr.it

Chrysostomos Nanakos has extensive experience in Open Source technologies. He has successfully executed several projects with different complexities and sizes in the Public and Private sector based on Open Source architectures. Since 2011 he cooperates with the National Documentation Center as a Senior Systems Administrator and Open Source Developer. He received his diploma in Electrical and Computer Engineering from the National Technical University of Athens (NTUA) where he is currently working towards the Ph.D. degree in Computational and Applied Electromagnetics. Email: cnanakos@ekt.gr

Marianna Nobile graduated in Linguistics, Faculty of Letters and Philosophy at the University of Rome La Sapienza. She had an internship in the library of the Senate of the Republic (Giovanni Spadolini Library), where she worked with the Books Acquisition. Currently she is collaborating with the library of the Institute for Research on Population and Socials Science of the Italian National Research Council and she is involved in the activities of indexing electronic resources as well as the development of an e-publishing service. marianna.nobile@irpps.cnr.it

Pasquale Pagano is a Senior Researcher at the Networked Multimedia Information Systems Laboratory of the "Istituto di Scienza e Tecnologie della Informazione A. Faedo" (ISTI) of the Italian National Research Council (CNR). He received my M.Sc. in Information Systems Technologies from the Department of Computer Science of the University of Pisa (1998), and the Ph.D. degree in Information Engineering from the Department of Information Engineering: Electronics, Information Theory, Telecommunications of the same university (2006). The aim of his research is the study and experimentation of models, methodologies and techniques for the design and development of distributed virtual research environments (VREs) which require the handling of heterogeneous resources provided by Grid and Cloud based e-Infrastructures. Pasquale has a strong background on distributed architectures. He participated to the design of the most relevant distributed systems and e-Infrastructure enabling middleware developed by ISTI - CNR. Pasquale is currently the Technical Director of the Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources (iMarine) and member of VENUS-C European project. He is also involved in the GRDI2020 expert working group a he serves EUBrazilOpenBio initiative as consultant. In the past, he has been involved in the D4Science-II, D4Science, Diligent, DRIVER, DRIVER II, BELIEF, BELIEF II, Scholnet, Cyclades, and ARCA European projects. Email: pasquale.pagano@isti.cnr.it

Gabriella Pardelli was born at Pisa, graduated in Arts in 1980 at the Pisa University, submitting a thesis on the History of Science. Since 1984, researcher at the National Research Council, Institute of Computational Linguistics "Antonio Zampolli" ILC, in Pisa. Head of the Library of the ILC Institute since 1990, responsible for the Archives of the ILC Institute since 2005. Her interests and activity range from studies in grey literature and terminology, with particular regard to the Computational Linguistics and its related disciplines, to the creation of documentary resources for digital libraries in the Humanities. She has participated in many national and international projects including the recent projects:- BIBLOS: Historical, Philosophical and Philological Digital Library of the Italian National Research Council, (funded by CNR); - For digital edition of manuscripts of Ferdinand de Saussure (Research Programs of Relevant National Interest, PRIN - funded by the Ministry of Education, University and Research, MIUR). Email: gabriella.pardelli@ilc.cnr.it

Fabrizio Pecoraro. Degree in Computer Engineering in Rome and Philosophy Doctorate in Bioengineering at the University of Bologna. During his doctorate studentship period he also held the position of assistant researcher at the University of Strathclyde, Glasgow, Scotland. Since 2007 he works as a researcher at the National Research Council – Institute of Research on Population and Social Studies, Rome Italy – where his research activities mostly focus on the following aspects: business process analysis, development of conceptual models based on standard of clinical data such as HL7 and CDISC, design and development of information systems and definition of relational databases. fabrizio.pecoraro@irpps.cnr.it

Petra Pejšová studied information science and librarianship at Charles University. She works as an information specialist in the State technical Library, Czech Republic. Actually she is leading a project Digital Library for Grey Literature – Functional model and pilot. Email: petra.pejsova@techlib.cz

Eugenio Picchi graduated in Computer Science, at Pisa University, is Research Director at the Institute of Computational Linguistics (ILC) of the National Research Council. He is currently responsible for the research line "Computational models and tools for research in humanities, with a special

Author Information

focus on linguistic and literary disciplines and on lexicography". From 1972 to 1983 he was responsible of the Systems and Programming Division of the Linguistics Section of CNUCE (Pisa). Since 1983 he has been responsible of the Division "Methodologies and Tools for Lexicology and Computational Linguistics" of ILC. In the last few years he has been scientific director of national and international projects, among which: "International Network of Linguistics Data-Bases and Workstations" and "New Technologies for Language Engineering", within the Project "Natural Languages Processing"; ILC Unit of Research of Esprit Basic Research Actions "Aquilex - Acquisition of Lexical Knowledge for Natural Language Processing Systems", Action n. 3030; ILC Unit of Research of European Project MULTEXT "Multilingual Text Tools and Corpora". He has also been Technical Manager of CNR in the European Project "EUROSEARCH" for an European Federation of multilingual WEB browsers and Scientific Director of the Project "Italian-Arabic bilingual Corpora and Tools" within the CLUSTER "Computational Linguistics: monolingual and multilingual Researches", funding by Italian law 488/98. He has authored/co-authored a number of publications dealing with Computational Linguistics. eugenio.picchi@ilc.cnr.it

Hélène Prost is responsible for studies at the Institute of Scientific and Technical Information (INIST-CNRS). The different studies concern the evaluation of collections, document delivery, usage analysis, grey literature and open access to information. Expertise in statistical tools and knowledge in library information science allowed her to participate in various research projects and writing of several publications. Email: helene.prost@inist.fr

Kevin Quigley is Associate Professor and Director of the School of Public Administration at Dalhousie University, Halifax, Canada. Dr. Quigley holds a BA (English) from Queen's University (Kingston, Canada), a MSc (Public Administration and Public Policy) from the London School of Economics and Political Science, and a PhD in Public Policy Studies from Queen's University Belfast. Prior to his appointment at Dalhousie University he was a Post-Doctoral Fellow at the University of Edinburgh and a Visiting Scholar at the American Political Science Association in Washington DC. He has also held public sector appointments with the Government of Ontario, Canada. His research focusses on critical infrastructure protection, security, risk regulation, and public policy. His research has been funded by the Social Sciences and Humanities Research Council of Canada, the Canada School of Public Service, Defence Research Development Canada, Public Safety Canada and the UK's Economic and Social Research Council. Email: Kevin.Quigley@dal.ca

Chowbiny Ramasesh obtained his Master's Degree in Philosophy from the University of Mysore in 1976 with specialization in Vedanta Philosophy. He also obtained a Master's Degree (1978) and Doctorate Degree (1989) in Library and Information Science from the University of Mysore. Ramasesh is the recipient of Dr. S.R. Ranganathan Memorial Gold Medal for securing first rank in the Master's Degree. He served as professional librarian for three decades and is responsible for the organization of several extension programmes. Worked as the Deputy Director of Centre for Information Science and Technology (CIST), Mysore and coordinated for the implementation of Quality Procedures under ISO 9001 Standards of Quality Management System. Compiled Quality Manual, delivered special lectures and served as Quality Auditor/ Performance Auditor at CIST. Presently serving as University Librarian of the University of Mysore and supervising research programs in the field of 1) Grey Literature 2) Institutional Repositories of Heritage Collection and 3) Use Pattern of Online Journals. Email: cpramasesh@gmail.com

Alexandra Roubani holds a Bachelor of Science (BSc), Technological Educational Institute of Athens, Department of Librarianship and Information Systems, with a Master's Degree in Library Sciences (MLIS), Ionian University, Department of Archives and Library Science. She also studied Cultural Administration in the Department of Communication, Media and Culture, Panteion University of Social and Political Sciences (BSc). She is a cataloguing and metadata expert at Institutional Repositories of the National Documentation Centre (EKT)/National Hellenic Research Foundation since 1997, and her main occupation is focused on the gradual alteration/transformation of the National Union Catalogue of Serials. Her professional interests also include Authority Files, Digital Preservation Metadata Standards, Interlibrary Loan and Digital Curation. Email: arouba@ekt.gr

Roberta Ruggieri is librarian at the Senate of the Republic where she is responsible for the supervision of a digitalization project on Senate parliamentary print documents for the 'I to X Legislature'. Her activity in managing digitalization project also includes document addition and classification in the electronic Senate catalogue. From 2004 she has been collaborating with the Institute for research on populations and social policies of the National Research Council (CNR) in research activities related to the field of Grey literature and Institutional repositories. Email: biblio.irpps@irpps.cnr.it

Ioanna Sarantopoulou is the Head of the Digital Library Department of the National Documentation Centre (EKT) in the National Hellenic Research Foundation (NHRF). She is involved in organizing and maintaining Greek digital content in Library's collection development and also in providing intermediated library information services. She participated in several workgroups within the EKT for the implementation of the National Information System for Research and Technology. She formerly worked for many years at the EKT, using online systems for documentation and information retrieval, to facilitate information seeking for scientists on Natural Sciences and Engineering. She holds a Bachelor's degree from National Technical University of Athens in Chemical Engineering. Email: isaran@ekt.gr

Manuela Sassi graduated in Foreign Languages and Literature at Pisa University, 110/110 cum laude. Since 1974 she has been working in Pisa at the Institute for Computational Linguistics of the National Research Council. Her interests and experiences range from linguistic to textual data processing and in providing linguistic resources on-line. She has been responsible for many national projects and has participated in numerous international projects. Email: manuela.sassi@ilc.cnr.it

Eva Sassolini graduated in Computer Science, at Pisa University, is CTER (Research Collaborator) at the Institute of Computational Linguistics (ILC) of the National Research Council - Pisa. She is involved currently in several national and international projects. Research Collaborator in "TextPower" (TP) project, (new technology and approach to treatment and exploitation of texts) and before in the project "Corpus Bilingue Italiano-arabo" for linguistic tools and resources for bilingual Italian/Arabic corpora realization. Junior researcher ILC in the project LINGUISTIC MINER: linguistic Knowledge system for the Italian language; working contribution in "INTERA" (Integrated European Language Data Repository Area) project, for multilanguage terms extraction. Junior researcher ILC in the project: "Progetto Iraq: navigazione nei materiali testuali del museo virtuale di Baghdad in maniera contrastava tra le tre lingue previste dal progetto (italiano, arabo e inglese)". Junior researcher ILC in the project 8: "Diffusione della cultura e valorizzazione del patrimonio letterario della lingua italiana e della lingua araba attraverso una diffusione telematica di banche di dati letterarie". Collaboration with IMSS (Istituto e Museo di Storia della Scienza) for the realisation of web applications for the query on galileian texts. Junior researcher in the project: "Corpus Bilingue italiano-arabo": in the framework of the comprehensive "Linguistica Computazionale: ricerche monolingui e multilingui". Email: eva.sassolini@ilc.cnr.it

Joachim Schöpfel is Head of the Department of Information and Library Sciences at the Charles de Gaulle University of Lille 3 and Researcher at the GERiiCO laboratory. He is interested in scientific information, academic publishing, open repositories, GL and usage statistics. He is a member of GreyNet and euroCRIS. He is also the Director of the National Digitization Centre for PhD Theses (ANRT) in Lille, France. joachim.schopfel@univ-lille3.fr

Suzuette Soomai is an interdisciplinary doctoral student in the Faculty of Graduate Studies. Her research focuses on the role of fisheries information, published extensively by national and international governmental organisations, in policy making for fisheries management. She holds a Master of Marine Management from Dalhousie University and an MPhil (Zoology/Aquatic Ecology) and BSc from the University of the West Indies, Trinidad and Tobago. Prior to her studies at Dalhousie University, Suzuette was a Fisheries Officer with the Ministry of Agriculture, Land and Marine Resources in Trinidad and Tobago and was a member of fisheries scientific working groups led by the United Nations Food and Agriculture Organisation and the Caribbean Regional Fisheries Mechanism. She has special expertise in fisheries resource and coastal zone management and has completed national and regional fish stock assessments while interacting with a diverse range of stakeholders in the Caribbean. In 2012, she was awarded a major doctoral fellowship from the Social Sciences and Humanities Research Council of Canada. Email: suzuette.soomai@dal.ca

Alexandros Soumplis studied "Engineering in Informational and Communication Systems" and holds a "MSc in Network & Data Communication" from Kingston University UK. Since 2010 is accepted by the

Author Information

Faculty of Sciences of the Hellenic Open University as a PhD student. His research interests involve informal learning environments as well as the use of innovative technologies for learning. Furthermore he has more than 12 years of active experience as a systems engineer with focus on core IT systems and in the past has worked for major computer and telecommunications companies in Greece. Since 2007 collaborates with the National Documentation Centre as a member of the "Information Systems and Networks Department" and has active involvement in several projects. Also, through his academic and professional career has submitted work and participated in several scientific and business conferences. Email: soumplis@ekt.gr

Panagiotis Stathopoulos received his diploma in Electrical and Computer Engineering and his PhD in Broadband Networks at 1999 and 2004 respectively, from the National Technical University of Athens (NTUA). From 1999 until 2006 he has been with the Computer Networks Laboratory of NTUA participating and technically coordinating research projects in the areas of broadband communications and applications. From 2006, he is the head of the Systems and Networks Unit of EKT leading the team developing a highly sophisticated IT infrastructure, for providing advanced open access applications and services. He has taught at the University of the Aegean and the TEI of Piraeus, and he has over 30 publications in peer reviewed journals and conferences. Email: pstath@ekt.gr

Ioanna-Ourania Stathopoulou received a B.Sc. in Computer Science and, later on, a Ph.D. degree from the University of Piraeus, Piraeus, Greece. Her doctoral research was sponsored by the General Secretary of Research and Technology of the Greek Ministry of Development, under the auspices of the PENED-2003 basic research program. Since May 2007, she is with the Department of Software Application Development of the National Documentation Centre (EKT) of the Hellenic Research Foundation, where she works as a software engineer participating in the design and implementation of digital repositories, e-publishing platforms and bibliographic systems. Her primary research interests are in the areas of affective computing, human-computer interaction, computer vision, and pattern recognition, and their applications in user modeling, information retrieval and intelligent software systems. She has over 30 publications in peer reviewed journals and conferences and has co-authored a monograph entitled “Visual Affect Recognition”, published by IOS Press. Email: iostath@ekt.gr

Julian Thomas has been the Director, Institute for Social Research (ISR), Swinburne University of Technology since 2005. In that role he has been responsible for a large number of publications about the influence of the Internet, government policies and social issues as well as part of the World Internet Project. Professor Thomas has initiated and led a number of online information and commentary services such as Australia Policy Online. Professor Thomas is one of the Chief Investigators of a major national Australian research project into Grey literature, policy innovation and access to knowledge about realising the value of informal publishing. Email: jthomas@swin.edu.au

Jessica (Jess) Tyndall is currently enrolled in Adelaide University's Master of Clinical Science, a postgraduate research degree that aims to train students in research methodologies and techniques specific to the needs of evidence-based healthcare, as well as the critical evaluation of evidence and research. Her research is focused on evaluating the impact of the findings of grey literature on the results of systematgic reviews on prevention of childhood obesity. Email: jessica.tyndall@flinders.edu.au

L. Usha Devi obtained her Bachelor’s Degree in Science and Master’s Degree in Library and Information science from the University of Mysore, Mysore. She has been working as an Assistant University Librarian at Bangalore University, Bangalore for the last 30 years. Smt Ushadevi also possesses M.Phil Degree in Library and information Science and Post graduate diploma in English language teaching, Postgraduate Diploma in Human Resource Management also. She has 5 years of postgraduate teaching experience in Library and information science and presented and published 14 research papers at National and International level seminars and conferences .She is a life member of professional associations: ILA, IASLIC, IATLIS, KALA and AKELPA. Her areas of research interest are: User studies, Organizations and Use of Grey literature, Citation Studies. Email: ushachowdappa@yahoo.in

Marcus Vaska is a librarian with the Health Information Network Calgary, Holy Cross Site, providing research and information support at an Alberta Cancer Care research facility. A firm supporter of embedded librarianship, Marcus engages himself in numerous activities, including instruction and research consultation, with research teams at Holy Cross. Marcus’ current interests focus on educational techniques aimed at creating greater awareness and bringing grey literature to the forefront in the medical community. Email: mmvaska@ucalgary.ca

Paul Weldon completed his PhD in sociolinguistics in 2007 and joined ACER in January 2010. For the three years prior to this he was Research Associate at Independent Schools Victoria. Dr Weldon has had a diverse career in Australia, the UK and in China, where he worked for two years as English Language Editor for the Journal of China University of Geosciences. He has worked on research projects in the fields of Social Capital, Education for Sustainability, School-Community Partnerships, Cybersafety, and on the development of school performance measures. He has published several reports and articles and he has also been a member of the Australian Cybersafety Consultative Group. Email: weldon@acer.edu.au

Peter G. Wells is an Adjunct Professor in the School for Resource and Environmental Studies and the Marine Affairs Program, Faculty of Management, and a Senior Research Fellow, International Ocean Institute, at Dalhousie University, Halifax, Canada. He holds a BSc (Biology) from McGill University, a MSc (Zoology) from the University of Toronto, and a PhD (Zoology/Aquatic Toxicology) from the University of Guelph. After over 30 years of public service, he took early retirement from Environment Canada, in 2006, to focus on his research. He concluded his work with Environment Canada as Head, Coastal and Water Science, and Senior Research Scientist, Coastal Ecosystems. He served the United Nations Joint Group of Scientific Experts on Marine Environmental Protection in various capacities for 14 years, and taught an international marine pollution course in Bermuda for 16 years. His current research includes choosing indicators for coastal ecosystem health, utilizing blue mussels for monitoring chemical contaminants, and evaluating the use and influence of marine environmental information in environmental policies and decision making. He was elected Fellow of the American Association for the Advancement of Science, and is a recipient of Dalhousie’s highest award for teaching excellence by part-time faculty. Email: oceans2@ns.sympatico.ca

Gerald (Gerry) White is a Principal Research Fellow at the Australian Council for Educational Research (ACER). He specialises in the use of digital technologies and digital media in education and currently manages the Digital Education Research Network (DERN) (http://www.dern.org) which publishes weekly research reviews about ICT in education. Formerly head of Australia’s education technology national agency for education and training, Gerry’s interests and experience are in digital diffusion, online collaboration, grey literature, teaching and learning, leadership and online communities. Email: whiteg@acer.edu.au

Index to Authors

GreyNet Members
less 20%

Tracking Innovation through Grey Literature

National Research Council, CNR Rome, Italy 29-30 November 2012

Publication Order Form 2003-

FOURTEENTH INTERNATIONAL CONFERENCE ON GREY LITERATURE

Forthcoming February 2013	*No. of Copies*	*x*	*Amount in Euros*	**Subtotal**
GL14 CONFERENCE PROCEEDINGS – Printed Edition ISBN 978-90-77484-20-3 ISSN 1386-2316 *Postage and Handling* ***excluded****)		x	99.00 =	€
GL14 CONFERENCE PROCEEDINGS - CD-Rom Edition ISBN 978-90-77484-20-3 ISSN 1386-2316 *Postage and Handling* ***included***		x	99.00 =	€
GL14 Conference Proceedings – Online Edition ISBN 978-90-77484-20-3 ISSN 2211-7199 ***Password Protected Access***		x	99.00 =	€

*POSTAGE AND HANDLING PER PRINTED COPY**)				
Holland		x	5.00	€
Europe		x	10.00	€
Other		x	20.00	€
TOTAL =				€

Customer Information

Name:	
Organisation:	
Postal Address:	
City/Code/Country:	
E-mail Address:	

Upon receipt of payment the publication(s) will be forwarded to your shipping address with an invoice marked paid.

□ Direct transfer to TextRelease, Account No. 3135.85.342, Rabobank Amsterdam
BIC: RABONL2U **IBAN:** NL70 RABO 0313 5853 42, with reference to "GL14 Publication Order"

□ MasterCard/Eurocard □ Visa card □ American Express

Card No. ______________________________ Expiration Date: ______________

Print the name that appears on the credit card, here ______________________________

Signature: ____________________ **CVC II code:** ________ *(Last 3 digits on signature side of card)*

Place: ____________________________ Date: ______________________________

NOTE: **CREDIT CARD TRANSACTIONS CAN BE AUTHORIZED BY PHONE, FAX, OR POSTAL SERVICES. EMAIL IS NOT AUTHORIZED.**

TextRelease
www.textrelease.com

GL14 Program and Conference Bureau
Javastraat 194-HS, 1095 CP Amsterdam, Netherlands
T/F +31-(0) 20-331.2420 Email: info@textrelease.com