Building an Ontology Based on Folksonomy: An attempt to represent knowledge embedded in filmed materials

Reyad Binzabiah, Steve Wade
University of Huddersfield
Huddersfield, UK

Abstract

Ontology, usually, build upon Taxonomy, which is considered as the backbone of it. The problem is that Taxonomy entirely depends on controlled vocabulary which has many drawbacks, particularly in the environments that depend on social networking, which shaped and formed generally from contribution of nun-specialist communities. On the other side, Folksonomy found as a way to deal with free tagging systems. Although it has a many weaknesses, it could be very Useful in dealing with the requirements of the beneficiaries within those social networking environments. This article discusses part of recent work on developing an ontology that can be used to represent the knowledge inherent in filmed materials. The ontology is intended to be used as the semantic basis for a retrieval system. The focus of the paper is on the method used to develop the ontology. The method is influenced by success that has been achieved in developing Folksonomies.

Keywords: Ontology, Folksonomy, Knowledge representation, Taxonomy, Information retrieval

1. Introduction

Information retrieval systems and knowledge representation approaches follow one of two methods during the design process. The first method uses free words, while the other uses controlled vocabulary. A tagging system is a good example of the first method and has proved popular in social network applications such as Facebook. This approach can lead to the development of a "Folksonomy". A folksonomy has the following features: "social, flexible, dynamic, lightweight, user-dependant content creation and classification as in ''collaborative tagging'' in a variety of prominent Web based services (e.g. del.icio.us:http://del.icio.us/, CiteULike: www.citeulike.org/, Flickr: www.flickr.com/, etc.)". On the other hand, the controlled vocabulary method uses more strict and formal tools such as thesauri, subject headings, or classification schemes, etc. These kind of tools are characterized by several features of formality, solidarity and immutability [13].

2. What is Folksonomy?

Folksonomy (also known as social classification, social indexing, and social tagging) is the collective tagging practice and method of collaboratively creating and managing a set of keywords, the so-called "tags", to annotate and categorize content [14]. It is a type of distributed classification system [5] gathered usually socially by means of a social network, this happens when users add tags to their contributions online, whether it is a text, picture or video. Folksonomy is a term coined by Van der Val [17], to signify what he called a "user-generated classification, emerging through bottom-up consensus" [6]. Folksonomies evolve as users create keywords (tags) which enable them to organize and retrieve information stored in the network [7]. Perhaps, the most famous sites which use the tagging systems are Flickr, del.icio.us, LibraryThing, youtube, CiteULike, IMDB.

Although Folksonomy achieves a degree of success in social tagging systems used in many social networks in web 2.0 services, it "lacks of organization and precision… each folksonomy's tag is unconnected with each other." [18]. Tagging systems may inherit the recognized drawbacks of free text indexing; these include the ambiguity in the meaning (polysemy), Tag variation (synonymy) or the flat organization of the tags.

3. What is Taxonomy?

If folksonomy is weak classification where the purpose is indexing rather than structure, taxonomy is a classification that is organized along a structural hierarchy [9]. It can be seen as a science of classification, organizing information in a ranked hierarchical structure consisting of controlled vocabularies defined by experts [18]. A good example of a taxonomy is the Dewey Decimal System (DDS) widely used in libraries to classify books and help determining its places into shelves according to a fixed categorization scheme [8].

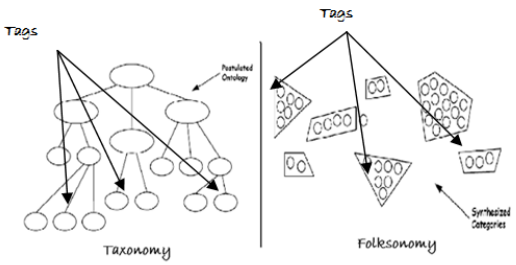

Figure 1: Flat organization of folksonomy in the opposite of hierarchy organization of taxonomy

Table 1 shows the main differences between two categories: Taxonomy, Ontology and Controlled vocabulary in one side, and in the other side Folksonomy and Free tags

Table 1: Comparison between Ontology, taxonomy and Folksonomy

Formal Taxonomies or Ontologies	Folksonomies and free tagging
categorization or model is seen as something static that can be created in advance	something that is created and updated as a part of an ongoing activity
ontologies often are based on hierarchical structures	folksonomy creates an entirely flat namespace
hierarchical structures provide much more expressiveness and support for reasoning of various kinds	Less expressiveness and support for reasoning of various kinds
hierarchical structures they are also more sensitive to changes	Less sensitivity for changes
The namespace in a Ontology is normally entirely Closed. Users are free to choose whatever tags they want to describe an entity	The namespace in a folksonomy is normally entirely open. Users are free to choose whatever tags they want to describe an entity
provide a framework to handle structured information and to extract conclusions from such structured information	Does not provide such a framework
ontologies are difficult to maintain	Easy to maintain
On the spectrum of knowledge representation systems, the most expensive in creation and maintenance is an ontology	easier to create, edit, use and reuse
requires consensual agreement on its contents from community members	Does not require such consensual agreement
metadata is generated only by experts	metadata is generated not only by experts but also by creators and consumers of the content
Usually, controlled vocabulary are used	Usually, freely chosen keywords are used instead of a controlled vocabulary
	Folksonomic tagging is intended to make information increasingly easy to search, discover, and navigate over time
	The number of websites that support tagging has rapidly increased since 2004
	multidimensional: users can assign a large number of tags to express a concept and can combine them.
	Uncontrolled tagging can result in a mixture of types of things, names of things, genres and formats.

4. How can Ontology play a compromising role between Taxonomy and Folksonomy?

The meaning of sharing the information and information resources which carried out by web 2.0 environment required popularization of using and describing the knowledge resources circulated around the internet or within one website, or even one system. This led to arising of many problems in usage of social tagging as a result of poorly chosen and applied tags. The following problems have been identified [6], [11], [19]:

- The probability of using two levels of specificity by different users (animal, dog).
- **Tags variation**, the possibility to different expressions for the same concept (cat, feline) and **proliferation of synonyms** (beauty, prettiness, handsomeness)
- Usage of Special terms, meanings, languages. (*viewfrommywindow)(monamour)*
- **Tags ambiguity**, One word could lead to different meaning. (*Play "theatre*) (*play "verb*) (*Ford, the car*) (*Ford, the industrial*)
- Singular versus Plural usage. (*tooth, teeth*) (*Plants, Plant*)
- Using of hyphens, symbols, foreign characters.
- Spilling issues. (*centre, center*)
- Usage of multiple styles of the same meaning (*blog, weblog, blogs, blogging*).

Obviously, these problems emerge from the fact that tags do not linked to each other, and have no means to show relations between terms. Ontology could be used to provide many features such as determining the meaning, the level of narrowing or broading of terms and the relation between terms. Simply, ontology could confer simple, spontaneous and flat tags more deep dimension of meaning. The researcher argue that ontology could play this role. "The term ontology is used in information systems and in knowledge representation systems to denote a knowledge model, which represents a particular domain of interest. A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that is held among them." [13]. According to many previous experiences, ontology can

solve many of above mentioned drawbacks resulted from free tagging and could add additional benfets as [15] discussed:

> "Hierarchical ontology should be used to classify and visualize keywords, topics, and other metadata that users and applications generate. A well-defined annotation dictionary (such as MPEG-7) is desired as it allows the standardization of various multimedia contents descriptions. For search formulation, ontology-based classification can help users in redesigning their query if it is too specific. For example, instead of looking for ''aloe vera'', users can be suggested to search on ''green plants''. Moreover, a unified indexing on keywords and semantic summaries will enable search engines to support users in finding related topics."

To achieve these goals, some conditions should be taken on consideration, in the process of building the ontology or in use it as integral part of a retrieval system.

5. a building process for the proposed ontology

There are many methods used to build ontology. Methontology consider as the most important methodology in this area. However, it is important to know that the results of the same methodology in the same area not always analogical. Results depend on the details of the building process and the materials used. Therefore, although the building process will follow the Methontology methodology, it will take in consideration some details and methodize some techniques applied in other areas like library science. These techniques include Literary Warrant and Faceted Analysis Approach. Figure 2 illustrates the main stages of the process and shows the distinct steps with some details.

5.1. Methontology Methodology:

METHONTOLOGY is a method in building ontologies. It is based on the experience gained in developing ontology in the domain of chemicals.

Originally, the method suggested seven stages as following:

- **Specification:** it is the juncture where targets and purposes are set in general by normal languages.
- **Knowledge Acquisition:** It is the phase that which is the stage of collecting information and data relating to ontology by using deferent means that used usually in research collecting data and information.

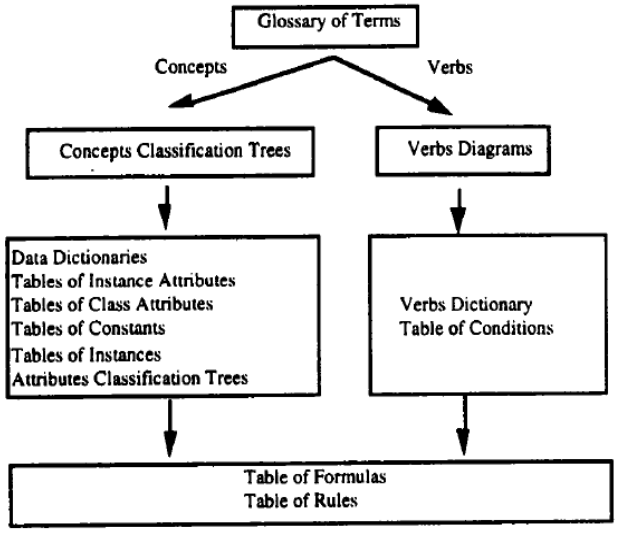

Figure 1: Conceptualization phase.

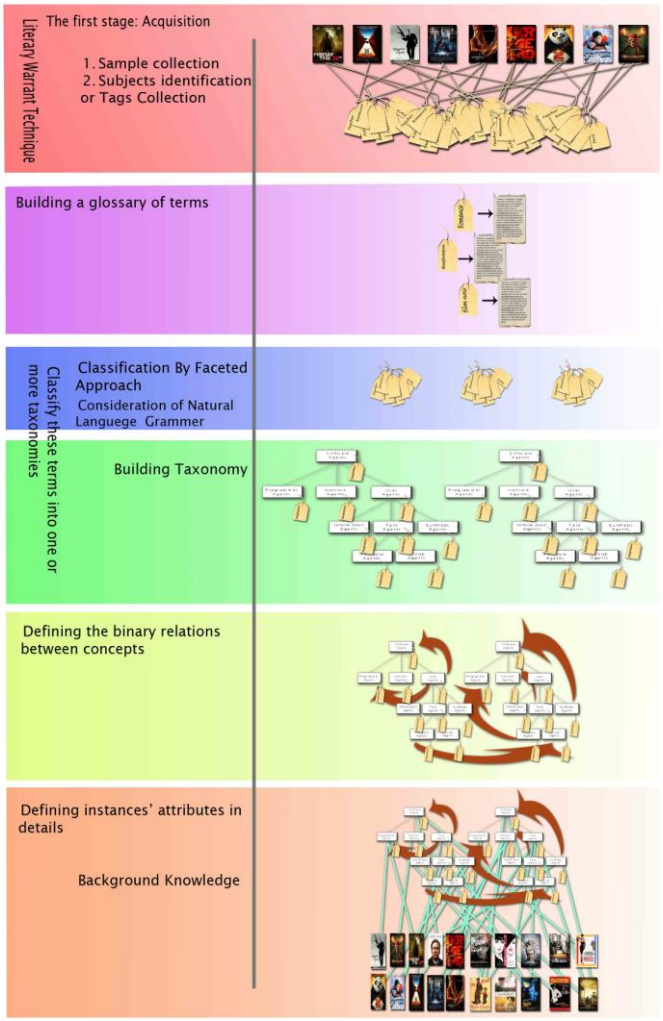

Figure 2: Building process of a proposed ontology of folksonomy for filmed materials

- **Conceptualization:** this is meaning that structuring the acquired knowledge in a conceptual model in a way that describes the problem and the solutions of this problem. This can be done by identifying the domain vocabulary
- **Integration:** this stage include process of re use other ontologies and include them in yours.
- **Implementation**: this phase will result in an ontology codified in a formal language. This mean the whole process related to codifying the ontology in one of known languages such as OWL.
- **Evaluation**: This stage includes two terms Verification and Validation. Verification refers to the technical process to ensure the absence errors in the consistency of the ontology, while Validation is to ensure that the ontology matching the aims and the purposes which the ontology is formed fore.
- **Documentation**: is the final stage, which implicates the process of gathering, collecting and archiving all the documents related to the ontology in all stages for the purpose of documentation, which can be useful in circulating the ontology cycle life.

In short, by focusing only on the most important process in this methodology we can defined the stages which can be recounted as following [3], [4], [10]:

The first stage: Acquisition

1. Sample collection: In this step the researcher will try to collect data of balanced sample of the filmed materials represent the community to the maximum.
2. Subjects identification: Identify the subjects contained in these materials

The second stage: Conceptualization

1. Building a glossary of terms
2. Classify these terms into one or more taxonomies.
3. Defining the binary relations between concepts.
4. Building the dictionary of concepts.
5. Defining binary relations in detail.
6. Defining instances' attributes in details.
7. Defining classes attributes in details.
8. Defining the constancies in details and construct a constant table.
9. Describing the formal axioms.
10. Defining the rules.
11. Introducing the instances details.

The third stage: Evaluation

This stage involves the following three aspects according to the consistency, completeness and conciseness criteria:

1. Ontology verification, in terms of the ontology being free of errors.
2. Ontology validation, in terms of whether the ontology will be represents the real world.
3. Ontology assessment, by the judgments from the end users point of view

Justifications for the use of this method are many, include: Firstly, this approach is the most detailed. Secondly, it is the most commonly used; therefore it is the mostly experimental and confident. Finally, Data collection phase in the Knowledge Acquisition stage, cited previously, fit perfectly with Literary Warrant technique, which come talk about it later.

5.2. Faceted Approach:

This technique which is followed by many libraries around the world in creating their classification schemes, could be useful in building the taxonomy of this ontology.

Using faceted approach, subjects can be separated according to their key components so that it can access to those topics through one part or more of those parts according to the need of the beneficiary. This method is the best way to combine between browsing and searching online.

Faceted classification could overcome hierarchical classification restrictions by classifying of documents to multiple categories organized from the bottom-up in multidimensional taxonomy. The categories resulting from faceted classification are determined by analyzing the domain knowledge and hierarchy is made by constructing the metadata in the way that expected that users will prefer it, which would require some human efforts (Uddin and Paul, 2007). Faceted classification can achieve to some extent the following [2]:

- The capacity to express through synthesis the complexity of subject content that is typical of digital documents
- A system syntax that ensures this is managed in a regular and consistent manner
- A rigorously logical structure that is compatible with machine manipulation at whatever level
- A structure that is compatible with a graphical interface for end-user navigation and query formulation;
- The facility through variation or rotation of the citation order to allow approaches from a number of angles (i.e. cross domain searching);
- A structure and methodology that permits conversion to other index language formats (i.e. subject heading lists and thesauri)
- And features of these integrated tools that allow modifiable keyword searching through mapping vocabularies and vocabulary control via the thesaurus, and provide tools for browsing and display via the subject heading list.

5.3. Natural Language facets

It can be suggested that this analysis could be follow the natural language grammar in classifying the tags, since the ontology will deal with free uncontrolled tags, which closer to the natural language than controlled vocabulary. It can be suggested that the facets could be equivalent to linguistic divisions. For example: (verbs-adjectives-adverbs..etc.) This could facilitate the queries when come as a sentence not just one word, in addition to the original feature which is dealing with free tags. Furthermore, it could be create facets inspired by the lexicon divisions to give a further dimension for nouns. For example: (Professions, Cities, Animals…etc.) or (Situation, Jobs, Position…etc.)

Analysis of tags contained in the IMDB website indicates that they belong to one of the following groups: Noun-Adjectives-actions-processes. Nouns can be divided into: places-countries-geographical areas- animals – organizations-characters-names-music-dance-occupation-plants-events-relations-objects-situations.

5.4. Literary Warrant technique:

Generally, Literary warrant in classification context can be consider as a determination mean that according to it the decision can be made about the classes or concepts should be taken into account , what order, and how they divided [1].

This technique, which suggested by Brian Campbell Vickery, is a method for deriving facets from selected sample of a certain library resources for the purpose of constructing a scheme eligible for classifying the whole library. This can be achieved by extracting some terms form the sample and compilation of similar under on group which called in ontology building conceptual clustering. This technique is what might call in the ontology building bottom-up building. Diaz think that a combination between two methods, bottom-up and top-down would be better, so that "a high level ontology is postulated, then it is revised and validated based on a bottom-up analysis of existing domain specific documents" [12].Thus, when the higher level shaped by specialists based on foundations stemmed from the domain itself, the bottom-up process "keywords and phrases are extracted from domain documents using standard text analysis tools". This is mean that this method by using Literary Warrant technique in building ontologies could permit developing it collaboratively by using folksonomy or social tagging too. Thereafter,

> "The Literary Warrant technique is then used to build a domain specific faceted classification scheme. The resulting scheme is used to group phrases into categories thus creating clusters that represent concepts in the domain."[12]

Thus the ontology will be in develop steadily can develop steadily with the addition of new tags which will find their place in the scheme easily.

5.5. Background Knowledge

This ontology intended to be as a catalogue by itself. Thus, this ontology could be part of a retrieval system and can play the pivotal role in this system, in terms of permitting retrieve bibliographical information about filmed materials. The contribution of this work resides in providing a new paradigm in analyzing contents of the selected collection. Unlike traditional catalogues, which could answer the queries regards a specific topic by proposing a specific document, the new paradigm could answer not by just a specific documents but might be a person, a place, an event or any other kind of object (which called background knowledge) .For this to be achieved, it is required a level of subject analysis of documents content This level of analysis comprise structural components such as (fonts, paragraphs, line breaks) and basic bibliographical information such as (title, author, date) and the body of the documents contents such as (names, dates, references to other objects, events…etc.). These information shape the ontology structure which consists of five main components: (documents, objects which include people-places-companies-organizations, subjects, document modalities, events which include conferences-wars-battles-meetings) as illustrated in figure3.

This division enables many kinds of relations without any links with subjects, so persons could be an author to a book, or an organization could be a sponsor of an event. Thus, through this structure, it would be possible to cover all faces of what called background knowledge not just subjects.

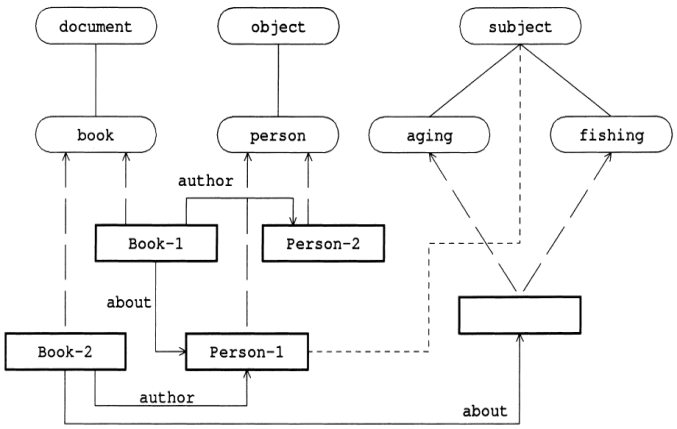

Figure 3: The structure of the ontology to meet the requirements of Background knowledge (Welty and Jenkins, 1999)

This project revolves around modeling subjects, and how to treat the instances in the ontology. How to treat the subject, with or without a relation with background knowledge, as

there are many obstacles arise when adopting each method (e.g. a book-1 as an instance under agricultural policy concept, book-1 as an instance under agricultural policy books concept, or book-1 an instance of agricultural policy as an individual under agricultural policy as a concept). Finally, the proposed method was as shown in figure3 dealt with this issue by creating a place in this ontology as an individual represents a subject or many subjects, as it composed by a combination of subjects. This will keep the taxonomy of subject maintained and will ensure that the users could narrow their search as they wish based on subjects.

6. The position of the proposed ontology in a retrieval system

As figure4 shown, the ontology is the centre of this system, where can be as an updatable index. It could transact as a database contain all required information in one side, and as a folksonomy contain all the information regarding the tags which are submitted previously, whether they are revised or not yet.

At the beginning, error checking, tag suggestion, synonym suggestion, widening or narrowing the term during submitting recourses or in query, all these functions can be acting through an interface interacting with the information deposited in the ontology.

New tags that do not meet the controlled vocabulary, will be stored as a new tags or concept as unconfirmed tag. These unconfirmed new tags/concept will be reviewed by experts manually to accept them and put them in the wright position, then to build there relations with other concepts.

This ontology will play the role of the library index by containing the bibliographical information about the filmed materials whither these information about the film such as: who made these films, when it made, and the relation with other films or even other materials like books, noviles, newspapers, biographies, or the contents of these films as submited by the users such as: film places, film times, and film subjects.

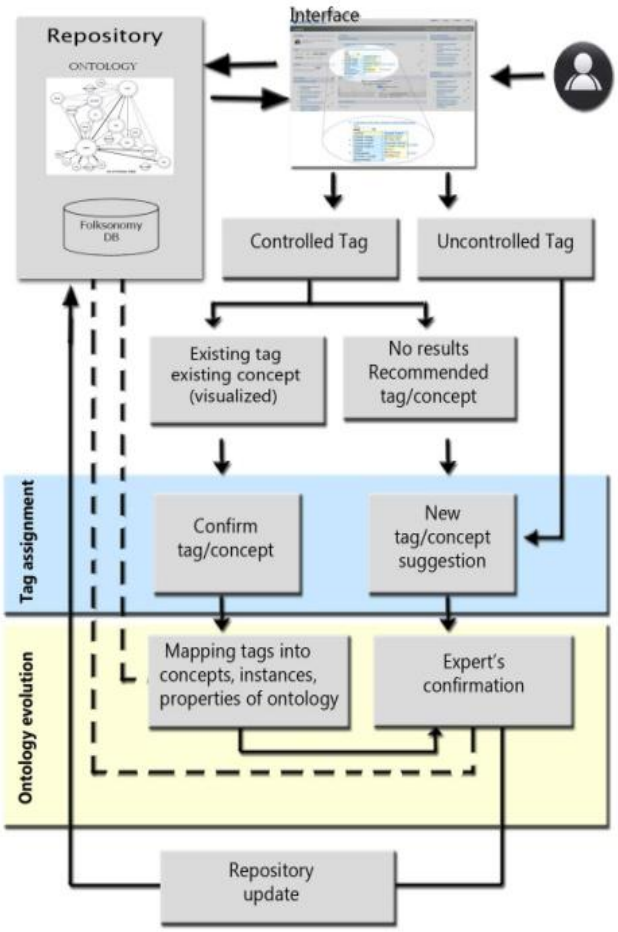

Figure 4: Model of ontology of folksonomy based retrieval system

7. Conclusions

This paper suggested the structure of an ontology that could be used as the central component of a retrieval system devoted to retrieving filmed materials. Due to the nature of filmed material, its distribution and because it is intended to be directed towards a wider public, it is apropriate to locate this work in the web 2.0 environment where information can be gathered and shared socially. Therefore, the ontology should integrate with the requirements of a free tagging system, without discarding the benefits of a controlled vocabulary goverened by a strict official taxonomy, the backpone of ontology. There is therefore a need to compromise between taxonomy and folksonomy during the construction of the ontology. In making this compromise it has been useful to consider other design techniques such as: the Methontology methodology, the Faceted Analysis approach, Natural Language facets, Literary Warr Techique and Background Knowledge.

8. Acknowledgement

I would like to thank my wife, for listening and supporting me during the work and discuss some ideas which formed the skeleton of this project. I deeply thank my Supervisor, Dr. Steve Wade, whose help, advice and supervision was invaluable.

9. References

[1] BEGHTOL, C. 1986. Semantic validity: Concepts of warrant in bibliographic classification systems. *Library resources \& technical services,* 30**,** 109-125.

[2] BROUGHTON, V. 2006. The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings,* 58**,** 49-72.

[3] GARCÍA, R. 2006. *A Semantic Web Approach to Digital Rights Management (PhD thesis).* PhD PhD, Universitat Pompeu Fabra.

[4] GOMEZ-PEREZ, A., CORCHO, O. & FERNANDEZ-LOPEZ, M. 2004. *Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition (Advanced Information and Knowledge Processing)*, Springer.

[5] GUY, M. & TONKIN, E. 2006. Folksonomies. Tidying up Tags? *D-Lib Magazine,* 12.

[6] HAYMAN, S. & LOTHIAN, N. 2007. TAXONOMY DIRECTED FOLKSONOMIES: Integrating user tagging and controlled vocabularies for Australian education networks. *WORLD LIBRARY AND INFORMATION CONGRESS: 73RD IFLA GENERAL CONFERENCE AND COUNCIL*. Durban, South Africa: IFLA.

[7] JONSSON, M. Year. Using a Folksonomy Approach for Location Tagging in Community Based Presence Systems. *In:* Mobile Data Management, 2007 International Conference on, 1-1 May 2007 2007. 304-308.

[8] KNERR, T. Tagging Ontology – Towards a Common Ontology for Folksonomies. Available: tagont.googlecode.com/files/TagOntPaper.pdf [Accessed 21.10.2010].

[9] LI, J. Z., GAŠEVIĆ, D., NESBIT, J. C. & RICHARDS, G. Ontology Mappings to Enhance Interoperability of Knowledge Domain Taxonomies. *www.lornet.ca/Portals/10/presentation_i2lor.../i2lor05-07.pdf.*

[10] OSCAR, C., MARIANO, F.-L., ASUNCIÓN, G.-P. & ANGEL, L.-C. 2005. Building Legal Ontologies with METHONTOLOGY and WebODE. *Law and the Semantic Web.*

[11] PASSANT, A. 2007. Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs : Theoretical background and corporate usecase. *ICWSM*. Colorado: http://www.icwsm.org/papers/paper15.html.

[12] PRIETO-DIAZ, R. Year. A faceted approach to building ontologies. *In:* Information Reuse and Integration, 2003. IRI 2003. IEEE International Conference on, 2003. 458-465.

[13] SHARIF, A. 2009. Combining Ontology and Folksonomy:An Integrated Approach to Knowledge Representation. *Emerging trends in technology: libraries between Web 2.0, semantic web and search technology.* Italy: --.

[14] SUN-SOOK, L. & HWAN-SEUNG, Y. Year. OntoSonomy: Ontology-Based Extension of Folksonomy. *In:* Semantic Computing and Applications, 2008. IWSCA '08. IEEE International Workshop on, 10-11 July 2008 2008. 27-32.

[15] TJONDRONEGORO, D. & SPINK, A. 2008. Web search engine multimedia functionality. *Information Processing & Management,* 44**,** 340-357.

[16] UDDIN, M. N. & PAUL, J. 2007. Faceted classification in web information architecture : A framework for using semantic web tools. *Emerald,* 25**,** 219-233.

[17] VAL, V. D. 2005. Explaining and Showing Broad and Narrow Folksonomies :: Off the Top / thomas Van Der Val.

[18] WANG, Y.-H. & JHUO, P.-S. 2009. A Semantic Faceted Search with Rule-based Inference. *the International MultiConference of Engineers and Computer Scientists 2009 IMECS 2009.* Hong Kong.

[19] WEBER, J. 2006. *Folksonomy and Controlled Vocabulary in LibraryThing* [Online]. Available: http://jonathanweber.info/samples/2452-Folksonomy.pdf [Accessed 12 July 2011 2011].

[20] WELTY, C. A. & JENKINS, J. 1999. Formal ontology for subject. *Data & Knowledge Engineering,* 31**,** 155-181.

