Online Journal of Applied Knowledge Management
A Publication of the International Institute for Applied Knowledge Management

Volume 2, Issue 2, 2014

The concept of semantic platform for addressing the information needs of university researchers and educators

Ilona Paweloszek, Politechnika Częstochowska, Poland, ipaweloszek@zim.pcz.pl

Abstract

University libraries contain varieties of documents from newspaper articles to academic journals and even multimedia collections. These resources are usually described using metadata for easier access, storage and retrieval. Using metadata alone, however, is not enough to describe the semantics of documents, and therefore enhanced search is usually not possible. This paper presents an on-going project in developing a semantic digital library for an academic institution. The concept of a semantic platform called SemLib grew out of many academic discussions and it reflects the information needs of researchers and educators from Technical University of Czestochowa. The proposed method of needs assessment is especially designed to create the starting point to build semantic structure of the digital library. The author proposes an approach for managing, organizing and populating knowledge by semanticizing existing library information resources that exists in digital and traditional paper form. The prototyped solution is based on the Semantic MediaWiki software, that allows for cooperative resource building, maintaining, and offers enhanced querying capabilities. Experimental results demonstrate the potentials and effectiveness of the proposed system as well as some obstacles reported by users that are the subject to improvement by further development of the SemLib platform.

Keywords: digital library, information needs, information retrieval, Semantic MediaWiki, semantic technology.

Introduction

In a world teeming with information technology and an overwhelming amount of data, the intelligent content management tools are a crucial factor for every large organization. Contemporary content management tools must effectively deal with the breadth, depth and complexity of information in terms of publishing, searching and presenting data in a user-friendly manner.

University faculties often have hundreds of workers in teaching and scientific positions, whose tasks are "knowledge intensive" and require the access to the up-to-date resources including books, journals, white papers, educational multimedia, etc. A lot of these resources are created by the academics themselves and published in printed or electronic form. Surprisingly, both electronic and printed forms appear to be poorly usable in terms of searching for particular information. Having hundreds of volumes such as manuals, conference proceedings, journals close at hand doesn't mean the information they contain is easily accessible. It is always easier to use a search engine than to leaf through a book searching for a definition, reference or a person

who is recognized as an authority on the given subject. The Author' s own experience as well as the multiple discussions with coworkers lends oneself to acknowledge that even having the files in the electronic form (i.e.: pdf, doc or html files) is not very convenient for searching for particular information and it is one of the most time-consuming tasks in the work of academics. Most of the local and global search engines serve a keyword-based method of finding information and are not able to respect the context of the query or to process the detailed query attributes. Finding information is a part of the everyday work of every educator or scientist.

There are many reasons arising from the policies of educational institutions as well as global circumstances that compel the authors to promote their work to be cited by other experts in their domain. On the other hand, the authors are supposed to contribute into the domain knowledge and take into account "the state of the art" in the subject they are working on. Therefore the scientiometric evaluation is *important* for all of the academics.

In all types of scholarly and research it is necessary to attribute the author and the source of information that underpin particular concepts, positions and arguments with citations. This good practice is important due to a number of reasons:

- The citations help readers to identify and retrieve the source work to verify the information, or to learn more about issues and topics addressed by the work,
- Citations provide the evidence that the subject is important and grounded in prior research,
- By citing, one gives credit to the author of an original concept or theory presented.

For these reasons, it is important to build the systems that facilitate the bibliographic description of information cited in an organized and thorough manner. At the same time, there is a growing need for creating effective search engines oriented towards supporting the retrieval of scientific information.

Having these objectives in mind, this paper presents the concept and the practical challenges facing the creation of the semantic platform for sharing professional knowledge by university employees.

The Prototype of the proposed solution called SemLib has been developed and is tested in the Department of Economics Computer Science at the Faculty of Management of The Czestochowa University of Technology.

Related Research

The movement of publications from paper-based systems to electronic and online systems has proceeded rapidly over the last decade. Electronic publishing has become a common way of distributing high-ranked scientific journals and conference proceedings. Digital library technologies are by now well-established and understood throughout the higher education community and the creation of digital collections, either in the form of 'born-digital' materials or the conversion of standard library materials into digital form, is now a well-established part of the activities of most higher education institutions (Gartner, 2014).

Making effective use of the information resources is dependent on the creation of good quality metadata, along with powerful and intelligent search engines. Without those two elements, the information resources cannot be effectively retrieved by users, nor developed by administrators in a controlled and reasonable way.

Another important issue is the standardization of metadata so that it would be possible to link together the repositories belonging to many institutions, and sharing the information resources for mutual benefit of the stakeholders.

In most of the current interfaces to digital libraries, users pose keyword-based queries to perform document retrieval. However, these keywords do not have the ability to express the semantics of the user's information needs. This pitfall has been addressed by many authors proposing ontology-based retrieval systems (Bloehdorn , Cimiano, Duke, Haase, Heizmann, Thurlow, Völker, 2007), some of them use general ontologies and other are domain–specific solutions (Noah, Alias, Osman, Abdullah, Omar, Yahya, Yusof, 2010).

The study by E. Reitrer (2013) presents the concept of result ontologies that describe a search results in a comprehensive and coherent way. The work of T.M.M. Swe (20014) presents the domain ontology-based model of intelligent information retrieval, which consists of identifying domain concepts in user's query and applying expansion to them. The work of G. Fu , C.B. Jones and A.I. Abdelmoty (2005), describes how ontologies, both domain and geographical, developed in the EU Semantic Web project SPIRIT were used to support retrieval of documents that were considered to be spatially relevant to users' queries. However the aforementioned solutions are based on the predefined ontology and there is a limited end user engagement in their development.

A number of papers report successful use of MediaWikis with Semantic Mediawiki extensions for research purposes. The SMW is often used in the domains such as: medical sciences (Jiang, Solbrig, Iberson-Hurst, Kush, Chute, 2010), genetics (Kumar, Schiffer, Blaxter, 2012), information technology knowledge management (Alquier, McCormick, Jaeger, 2009).

The above-mentioned research efforts and publications concentrate many aspects of wiki usage and semantic technology achievements. In this study, we propose the collaborative approach to building semantic resources by assessment of users' information needs.

Needs Assessment

Creating the effective system to support the unique information needs of academics is a complex project that requires end-users participation in defining goals for the system and examining alternatives for achieving these goals. A first step to building the system is the needs assessment. Because it is a typical qualitative research aiming to gain the users' insight, we decided to conduct a study using an open-ended questionnaire.

Open-ended questions are helpful in finding out the diverse expectations of the end-user community and specifying objectives for the new system. A first part of the questionnaire contained the questions aiming to find out the users' opinion about the library system currently in use. The second part was to find out what are some other sources of knowledge used by educators and researchers during their daily work. The third part of the questionnaire was

concerned with the specific information needs, namely, what information the users search for and what is the desirable structure of this information.

The conclusions from the survey that served as a base to design and build the prototype of the knowledge sharing platform SemLib are as follows:

1. The appraisal of an existing library catalog:

    - Missing features: the tables of contents of publications, abstracts, full-text base,
    - The existing tagging system of library catalog is not very useful and the categories are too general
    - Lack of integration between library databases – one has to figure out which source will be potentially useful and check them all out consecutively whether they contain the desired information
    - Lack of one, intuitive search engine for all the library resources.
    - The information about the authors is very limited

2. Other sources of knowledge:

    - External libraries such as ibuk.pl - lack of full text search engine.
    - External bases of journals like: EBSCO, ELSEVIER, EMERALD – restricted to one journal series
    - Web search engines – resources of different quality and reliability
    - Private collections of printed or electronic publications – limited access, difficulties in cataloging and searching.

3. Specific information needs are presented in table 1. To learn about the information needs of the surveyed persons, the respondents were asked to formulate queries reflecting their information needs in a natural language, for example:

    - Find a definition of concept A published after the year 2010,
    - Find a publication about subject B authored by a person who works at the Technical university C,
    - Find all the articles published by authors from University C on the subject B,
    - Find persons interested in subject B from University C.
    - Find a publication containing a figure showing concept X,
    - Find a publication containing a table presenting data from research Y.

Table 1. Information needs and query context parameters

Needed information	Query context parameters
definitions of concepts	Author, year of publication, Author's affiliation
research reports and results	Author, institution, year of publication, geographical region, branch
citations of one's own publications	Years of the referencing work
figures, tables, charts	Subject, year of publication, relevant definition
multimedia	Creator, subject, year of publication, file type
bibliographic references	Year and place of publication, isbn or issn, author/authors
Information about potential partners in project realization or coauthoring.	People, interests, authorship, participation in projects
Information about conferences	Subject, place, publication type, price

The transcription and analysis of the data from questionnaires revealed that the database of publications exposed by the university library is hardly usable. It is difficult to find the desired information and the search process is iterative and rather daunting. One of the unexpected problems revealed during the study was the lack of up-to-date information about the publications of colleagues from the same faculty in the light of searching for partners to realize projects and cooperate in authoring publications. The surveyed academics experienced problems in sharing knowledge about their work and communicating their information needs to coworkers from the same faculty.

Digital libraries contain varieties of documents from newspaper articles to academic journals and even audio and video collections. These collections of documents are usually described using metadata for easier access, storage and retrieval. Using metadata alone, however, is not enough to describe the semantic of documents And therefore enhanced search is usually not possible. An adaptation of the Dublin core metadata can be used to represent the general knowledge about the resources such as author, title, language etc, but it is not able to convey the content of the books, journal articles, location of which the particular research were carried out, and the connection with the other information resources.

It is undeniable that every source or knowledge should be dynamic in terms of meeting the user's changing information needs. The same is true of digital libraries. The content of university library changes every year, it is extended by new publications, journals, conference proceedings etc.

On one hand, semantic technology is the remedy that seems to offer solutions to the aforementioned challenges in digital libraries. On the other hand, the user participation in

creating the resource metadata is needed. Therefore, we had to find a consensus between using rather complex yet powerful semantic technologies and at the same time making the whole system easy to use and maintain.

To overcome the bridging of the gap between complexity and user-friendliness the author analyzed Web 2.0 publishing tools and the possibilities to enrich them with semantics. Therefore, for building the proposed solution the author decided to use Semantic Mediawiki, which is a collaborative knowledge sharing environment extended by semantic features that fulfill the assumptions of the Web 3.0 paradigm.

The Specifics of Semantic MediaWiki as a Collaborative Knowledge Sharing Environment

The Use of wikis is popular in higher education since it supports the creation of large and well-structured websites for various purposes, including individual note taking, managing documentation of cases, bibliographies, analyzing data and many more (Schneider D., DaCosta, 2014).

Collaborative authoring is the effective way of creating knowledge proved by many examples, of which the most spectacular is Wikipedia. The concept of "wiki" was coined in 1994 by Cunningham (Cunningham, 2006). He sketched out the numerous essential features for a platform facilitating cooperative knowledge sharing. According to Cunningham, "wiki is created by community" and it is designed in a way that allows everybody to create as well as edit whatever and whenever one wants if they find the legacy content incorrect or incomplete (Richardson, 2006).

Currently, when the Web 2.0 applications are flourishing and the Web 3.0 initiatives are emerging, the original concept of wiki proposed by Cunningham can be supplemented with some features such as sharing knowledge, interaction, social networking and intelligent technologies. The concept of cocreation by community refers not only to the content of the wiki, but also to the Mediawiki software which is distributed under the Open source license.

There is no single definition of "open source" The basic idea behind it is that the programmers can read, redistribute, and modify the source code for a piece of software free of charge. The software is then improved, adapted and corrected much faster than would otherwise be the case (Epstein, Politano, 2002).

Therefor there are hundreds of so called extensions that power MediaWiki software and make it customizable for many purposes. One of these extensions is Semantic Mediawiki package.

The idea behind the semantic Web technology is providing meaning to the words and sentences published on the HTML pages so their content can be understood and processed by machines (for example search engines).

The understanding of data enables the machines to combine, compare and analyze them. These are very useful in times of unlimited access to human readable information residing on the Web.

Semantic Mediawiki (SMW) is the package of extensions to MediaWiki platform, that allow to realize in practice the assumptions of the Semantic Web.

The Basic advantages of using the SMW platform are:

- Automatic lists generations,
- Improved data structure, thanks to semantic annotations,
- Improved search capabilities,
- Greater coherence of different language versions,
- Possibility of exporting data in formats like: CSV, JSON I RDF,
- Possibility of exploiting Wikipedia and other wiki services as sources of knowledge.

Before presenting the essence of SMW, it is necessary to emphasize the fact that wiki software is created to serve as a platform for an online encyclopedia, so each wiki page should describe an object of a particular category, a category, or an attribute of a category. The first impression on semantic mediawiki pages is that they do not differ at all from a standard wiki pages; they are readable for humans. However, the difference lies in the code of the page which contains annotations defined in the special semantic wiki syntax. The annotations inform the wiki search engine about the meaning of the elements such as text, numbers, pictures, etc. The annotations assign the page to a given category and define attributes for that category. Annotations added in the page text allow the SMW platform to function as an object database. The users can easily create their own categories with attributes. Adding the new attribute or category does not influence the existing resources. Adding annotations to a text is a facile task based-on syntax similar to MediaWiki hyperlinks (Semantic Mediawiki, 2014)

[[Attribute name:: attribute value]]

In the above syntax, only a text after two colons is displayed on the wiki page and it looks like a hyperlink. That text is understood as a value of an attribute; clicking on it moves the user to the page where the specified attribute value is defined.

One of the advantages of using SMW in comparison with other semantic solutions is the fact that to create semantic resources in SMW, it is not necessary to use ontology editing tools nor to have the skills of semantic Web languages like OWL or RDF. The definition of categories and attributes is accomplished by creating wiki pages describing with the syntax shown above. It is also very easy to add new attributes and categories not influencing the already existing structure.

The aforementioned features make the SMW an ideal platform for collaborative creation of semantic resources by a community of non-technical users (PhD students, teachers, researchers).

Having in mind the information needs of the users along with the specific of the MediaWiki software, we decided to build a solution that would serve as a semantic middle layer facilitating the access to heterogeneous and dispersed information resources. The simplified idea of the proposed solution is portrayed in Figure 1.

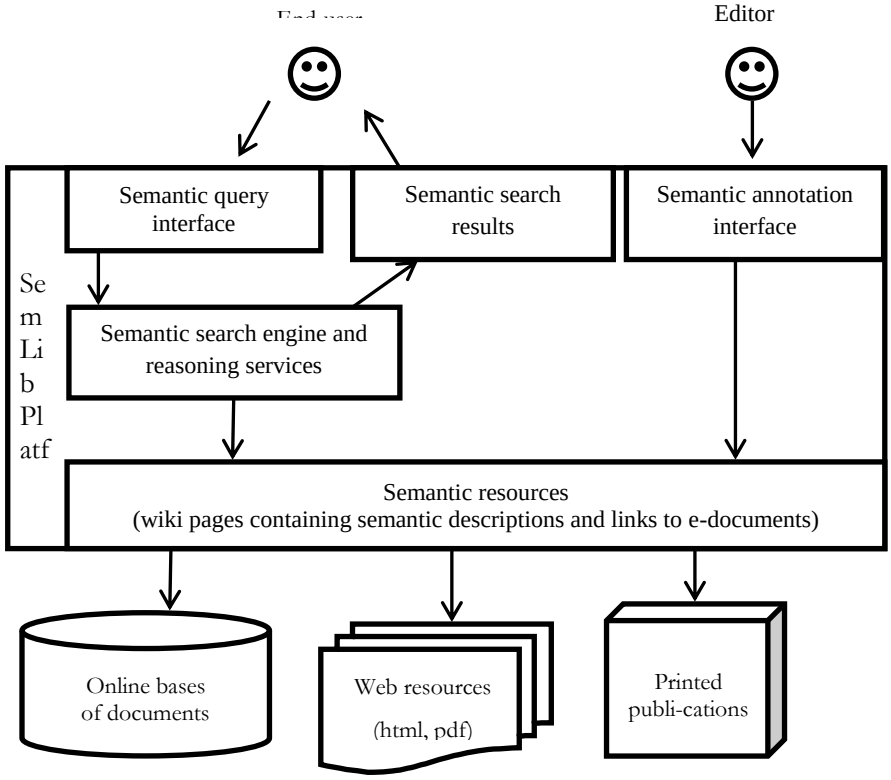

Figure 1. The concept of SemLib platform

The semantic resources are wiki pages that contain semantically enhanced metadata - descriptions of information resources with hyperlinks to documents that can be found online. The traditional printed publications that are not digitized are described in a way that facilitates the access to them. The SemLib wiki page describing a printed document can contain the information about the owner of the book, organization unit (university department or library), where the book is accessible. SemLib is especially helpful while dealing with many printed publications (conference proceedings, journals, monographs) physically accessible in our department. Actually the first idea of creating the semantic platform assumed it would be a local catalog of printed conference proceedings an journals to make them more useful.

The Semantic Structure of the SemLib Platform

The semantic structure defines the categories, subcategories and attributes of the entities described in the digital library. The attributes are also categories. The semantic structure was designed according to the users' information needs unveiled by the survey. The structure and the frequency of the information needs were analyzed and so the categories and attributes were established. Figure 2 presents the simplified class diagram with relations between categories defined and applied in the SemLib prototype system. Some of the relationships were omitted on a diagram to make it more legible. For example "publication" is a general class which subclasses

are "book", "journal", and "article". In this context, an article can be understood as a part of the journal or as the book chapter. We decided not to define the separate class for the entities like book chapter and article because they would have exactly the same attributes. The property "is written by" can contain multiple values selected with coma, each of the values indicates one entity of the class "Author".

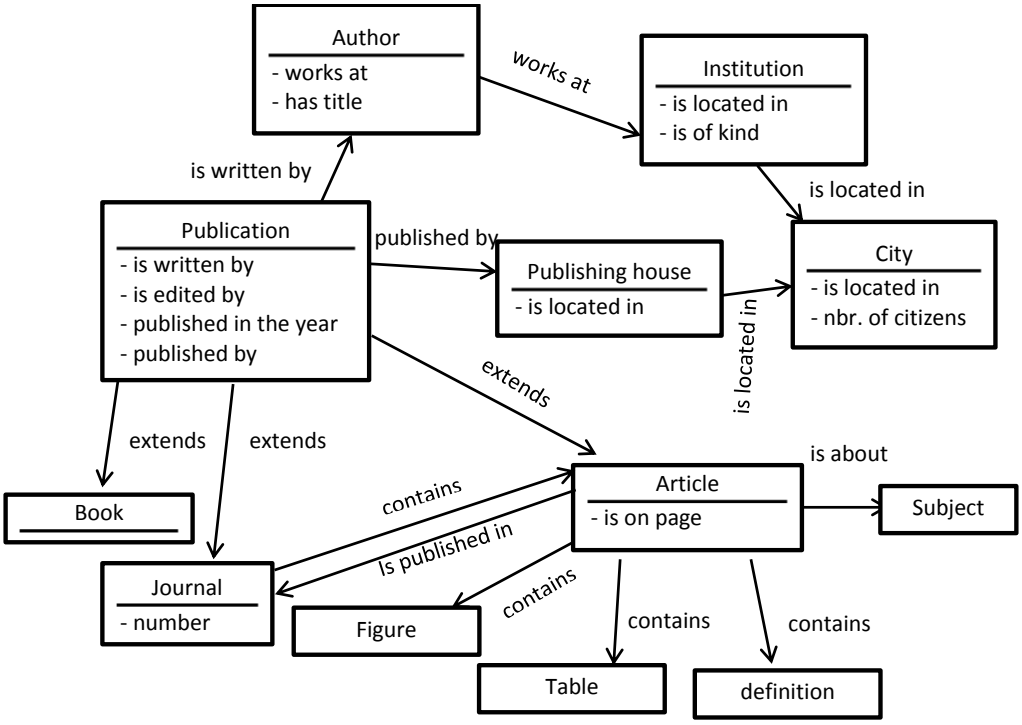

Figure 2. Classes and relationships in a prototype semantic library system

The presented structure of classes and attributes allows users to ask compound queries respecting many context parameters.

The "subject" category that is used to define attributes of publications, has many instances of the hierarchical structure. The example fragment of marketing subject category is:

- Marketing
  - Customer relationship management
  - Marketing mix
    - Promotion
    - Advertising
  - Web marketing
  - Customer service

The structure of the subjects facilitates annotation tasks. For example, if the described article is annotated with the tag "advertising", the search engine "knows" that the article is from the

marketing domain, although it may not be explicitly written in the page describing the given article.

The example query syntax that can display the list of all the articles from marketing domain written by authors from Technical University of Częstochowa, will look like that:

[[Category:Article]][[Is about::marketing]][[Written by.Works at::Technical University of Częstochowa]]

The list displayed by the above query will also contain articles about promotion and advertising, although they are not annotated with the word "marketing".

While designing the semantic library portal based on SMW there is no need to define the whole category structures in advance. For example the categories of subjects are created or developed when there is a need to describe the particular publication from a given domain.

Some of the information needs specified by the user community pointed to conferences, projects and persons. To address the aforementioned issue the site category structure can be extended with entities such as conferences and projects. The new categories can be linked to existing resources, so it would be easy to list, for example "all the publications from conference Y authored by person X" or "all the conferences in which the people from University C took part".

There are a lot of combinations of attributes and categories. Such lists can be settled according to date, place, institution, person and practically every attribute defined in the system. This functionality is very useful while preparing reports about the activities of the University departments or individual employees.

Conclusions

At present the prototype of the SemLib system contains the detailed information about 80 articles, and 67 authors and the database is continuously developed. There were two groups of users taking part in a described case study:

- Active users (6 persons), who participated in the system development, added the semantic descriptions of new publications and also used the semantic information resources in their daily work,
- Passive users (8 persons), who were only using the system by posing queries to the semantic base.

As the practice shows the wiki passive users (learners, teachers, researchers) encounter many difficulties, the most often reported ones were:

- Lack of knowledge about wiki functions in general,
- Difficulties in defining semantic queries.

The active users reported difficulties regarding the administrative functions of the site. The most important obstacle was the interface for making annotations – lack of well-designed forms for every category requires a lot of manual work. The wiki forms extension is far from perfect and was assessed as neither elastic nor easy to use.

Both the active and the passive users claimed the semantic base to be potentially very promising. Having all the metadata at hand makes the work faster when it comes to find particular information in the library. It is also possible to prepare multiple reports from the scientific activities much faster.

However, it would be far more useful if all the described resources (books, journals, articles) were accessible in electronic form and linked directly to the semantic wiki site where they are described. Transformation of all the library resources to electronic form is inevitable in the future, but it requires a lot of work and changes in the copyright law. The old fashioned fossil library catalogs of publications do not have a potential to reveal the full value of information resources they describe. The semantic technology along with cooperative editing tools like MediaWiki have the potential to reveal their hidden power.

There are also many potential problems that will probably appear as the platform evolves, one of them being the possibility of inaccurate annotations. One way to overcome that problem is to engage the community of end-users. If we assume the users are experts in their subject, it is undeniable that after a short training, they can do the tagging far more accurately than the system administrators. Moreover, it is in the interest of the users to describe their own publications in a way that ensures they will be frequently found, retrieved, and consequently most often cited.

Another problem that should be solved is the lack of wiki skills. Very few users seem to understand the nature of the wiki and the need for respecting some organizational guidelines, e.g. using the categories and attributes defined earlier by other users. The SemLib platform is very elastic and the directions of its evolution should be determined by analyzing the users' needs. The needs assessment method consisting in formulating example queries posed by users proved to be an appropriate and agile way to transform the users' needs to semantic categories and attributes.

The SemLib platform is quite new project that is still in the development phase, but we continuously add new publications to the semantic base and at the same time we are working to expand the ontology of research topics.

We are working on the concept of the improved query interface that is based on the idea of sample queries that are presented to the user and can be easily customized to his/her information needs. The second issue is an editor's interface – we are developing and improving the set of forms to make the process of adding new publications faster and more facile.

The next research direction, that we plan to describe in our consecutive publications will be the detailed methodology of the assessment of user's information needs. The method will be especially dedicated to be applied to the processes of improvement of the semantic information systems.

References

Alquier, L., McCormick, K., & Jaeger, E., (2009). *knowIT, a semantic informatics knowledge management system*. Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym '09). ACM, New York, NY, USA, DOI=10.1145/1641309.1641340

Bloehdorn, S., Cimiano P., Duke, A., Haase P., Heizman,n J., Thurlow, I.& Völker J (2007). *Ontology-Based Question Answering for Digital Libraries*. In: Kovács L., Fuhr N. & Meghini C. (eds.) ECDL 2007. LNCS, vol. 4675, 14–25. Springer, Heidelberg.

Cunningham, W. (2006).*Design principles of Wiki: How can so little do so much?*. Retrieved 01.2014 from: http://c2.com/doc/wikisym/WikiSym2006.pdf

Epstein, M..A. & Politano F.,L. (2002). *Drafting License Agreements, Fourth Edition.* Aspen Publishers, pp.13-46.

Fu G., Jones C.B. & Abdelmoty A. I., (2005). *Ontology-based Spatial Query Expansion in Information Retrieval*. Lecture Notes in Computer Science, Volume 3761, On the Move to Meaningful Internet Systems: ODBASE: OTM Confederated International Conferences, Vol. 3761, pp: 1466-1482 Springer Berlin Heidelberg.

Gartner, R. (2008). Metadata for digital libraries: State of the art and future directions. JISC: Bristol, UK, Retrieved 03.02.2014 from http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf

Jiang G., Solbrig H. R., Iberson-Hurst D., Kush R. D. & Chute C. G..(2010). *A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki*. AMIA Summits Transl Sci Proc, pp. 11-15.

Kumar S., Schiffer P.H. & Blaxter M., (2012). *959 Nematode Genomes: a semantic wiki for coordinating sequencing projects*. In: Nucleic Acids Research, Vol. 40, Database issue D1295–D1300 doi:10.1093/nar/gkr826

Noah, S.A., Alias N.A.R., Osman N.A., Abdullah, Z., Omar N., Yahya Y. & Yusof, M.M., (2010). *Ontology-Driven Semantic Digital Library*. In: Information Retrieval Technology Lecture Notes in Computer Science Volume 6458, pp 141-150. Springer-Verlag Berlin Heidelberg.

Reiterer E., (2013). *Search Result Ontologies for Digital Libraries*. (2013). In: The Semantic Web: Semantics and Big Data, Lecture Notes in Computer Science Volume 7882, pp. 687-691. Springer Berlin Heidelberg.

Richardson W. (2006). *Blogs, Wikis, podcasts, and other powerful Web tools for classrooms*. Thousand Oaks, CA Corwin Press.

Schneider D.& DaCosta J., (2014). *Adding power to educational and research wikis with Semantic MediaWiki*. Retrieved 03.02.2014 from: http://edutechwiki.unige.ch/en/Adding_power_to_educational_and_research_wikis_with_Semantic_MediaWiki.

Semantic Mediawiki, (2014) "Properties and types", Retrieved: 01.2014 http://semantic-mediawiki.org/wiki/Help:Properties_and_types.

Swe, T.M.M. (2014). *Intelligent Information Retrieval Within Digital Library Using Domain Ontology*. Retrieved 03.02.2014 from: http://airccj.org/CSCP/vol1/csit1232.pdf