МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2015
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2015
Proceedings of the Forty Fourth Spring Conference
of the Union of Bulgarian Mathematicians
SOK "Kamchia", April 2–6, 2015

APPLICATION OF SEMANTIC TECHNOLOGIES IN BUILDING DIGITAL LIBRARY SYSTEMS*

Maria Nisheva-Pavlova

The paper discusses some current trends of research and development in the field of digital libraries. The presentation is focused on the concept of semantic technologies and the role of their particular components in providing semantic interoperability between digital library systems as well as in building new generations of digital libraries – the so-called semantic digital libraries and social semantic digital libraries. Special attention has been paid to the development of a new search paradigm for heterogeneous digital libraries – intelligent (also known as semantic or ontology-based) search. As an illustration of the suggested ideas a particular academic digital library has been considered.

Introduction. The existence of a great amount of Web available documents of various types troubles the access to them and therefore enforces their organization in digital libraries. Generally, the term "digital library" means a library that maintains collections, preserved in digital (electronic) formats and accessible with a computer. The resources of such kind of library may be stored locally or remotely, within a system of computer networks.

Academic digital libraries are faced to the additional challenge of providing digital preservation of valuable collections of scholarly information and giving innovative methods for adequate access to their contents. From this point of view, the design and development of academic digital libraries is a serious research task.

Current Results and Trends of Research and Development in Digital Libraries. The significant achievements in the area of Digital Libraries during the last decade played a fundamental role in the development of the so-called Digital Library Reference Model [1]. This model is a result of the cooperation of several European research groups as a part of the DELOS Network of Excellence on Digital Libraries (`http://www.delos.info/`). It has the aim to come to an agreement between experts with respect to the main concepts, structures and activities in digital libraries.

The reference model points to six main aspects that should be considered in each project of a digital library: (1) content; (2) architecture; (3) user; (4) functionality; (5) policy; (6) quality.

*2010 Mathematics Subject Classification: 68P20, 68T35.
Key words: digital library, semantic technology, ontology, semantic interoperability, search engine, information retrieval, sentiment analysis.

Figure 1 illustrates the three-tier reference architecture for digital libraries [1, 2] which corresponds to the three different levels of their conceptualization. As it can be seen, three types of systems play a central and distinct role in the digital library referent architecture [1]:

- Digital Library – "an organization, which might be virtual, that comprehensively collects, manages and preserves for the long term *rich digital content*, and offers to its *user* communities specialized *functionality* on that content, of measurable *quality* and according to codified *policies*";
- Digital Library System – "a software system that is based on a defined (possibly distributed) *architecture* and provides all functionality required by a particular digital library". Users communicate with a digital library through the corresponding digital library system;
- Digital Library Management System – "a software system that provides the necessary software infrastructure both (i) to produce and administer a digital library system incorporating the suite of functionality considered fundamental for digital libraries and (ii) to integrate additional software offering more refined, specialized or advanced functionality". Digital library designers, application developers and system administrators interact with this kind of software system.

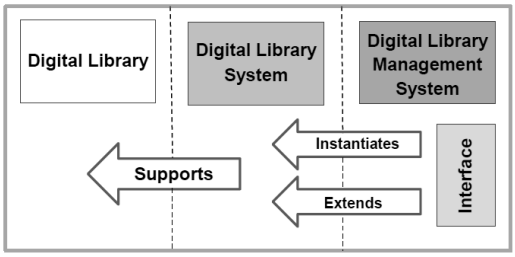

Fig. 1. Reference architecture for digital libraries

The discussed reference architecture is in fact an architectural design pattern that suggests an abstract solution which implements the concepts and relationships identified in the reference model. The design process of a digital library should include decision making not only about the content, functional facilities, main types of users and their roles, quality criteria and policies concerning the interaction with users, but also about the concrete system architecture.

Lately the research and development activities in digital libraries have been concentrated in several main directions including: strategies for preservation and management of rich digital content; interoperability between digital library systems; economic, legal and social issues.

Considerable progress has been achieved in building, management and integration of digital collections. The application of Cloud computing technology is expected to drive to better understanding and more efficient implementation of large part of digital library services [3].

At the same time, the developers of digital libraries and tools for access to their content are still faced with a number of challenges. One of these challenges is the provision

of sufficiently precise and rich in content answers of the user queries. A next significant challenge is the necessity of development of search methods and techniques which will be appropriate for repositories containing documents of different types and multiform content, available in various digital formats. There are still some open questions concerning the personalization of information resources access [4, 5]. A number of issues in the field of integration of software for writing, reading and visualizing rich media documents in networked environment [6] with tools for creation of academic digital libraries have been under consideration as well.

Semantic Web and Semantic Technologies. The term "Semantic Web" was suggested and popularized by Tim Berners-Lee in [7]. The vision for the Semantic Web consisted of two main parts: (1) to help people to share information and services and make it easier to aggregate heterogeneous data from different sources and (2) to create a Web that would be understandable and processable by computer systems.

The variety of tools, technologies, and specifications that form the foundation for the Semantic Web can be organized into four major layers (see Figure 2): data and metadata, semantics, enabling technology, and environment.

Environment Layer	Security Privacy Trust	Cryptography Integration Standardization	Peer-to-Peer Semantic Grid Social Network
Enabling Technology Layer	Agents Search Web Services	Composition Visualization	Personalization Repository Management Natural Language Processing
Semantic Layer	Ontologies (OWL) Rules (RIF/RuleML/SWRL) Queries (SPARQL)		Logic (FOL, DL) Reasoning Trust
Data and Metadata Layer	RDF and RDF Schema XML Unicode and URI		

Fig. 2. Semantic Web layers [8]

"Semantic technologies" is a general term for any software that involves some kind and level of understanding the meaning of the information it deals with. In a narrow sense is has been used as a synonym of technologies for the Semantic Web. Semantic technologies play a significant role in the contemporary progress in the area of digital libraries.

Semantic Interoperability of Digital Library Systems. The provision of *interoperability* between different digital library systems is one of the main aspects of research in the area of digital libraries ever since they came into view. Interoperability in general is concerned with the capability of different information systems to communicate. If two or more systems are capable of communicating and exchanging data, they are displaying *syntactic interoperability*. *Semantic interoperability* is the ability to automatically and accurately interpret the meaning of the exchanged information in order to produce useful results as expected by the end users of both systems. It implies appropriate representation of the information in a way, corresponding to its implied meaning, regardless of its

source. The processing of the shared information is expected to be consistent with its intended meaning.

The main aspects in which semantic interoperability is important for digital library systems include [9]:

- improving the precision and recall of search and document retrieval;
- enabling advanced types of search;
- facilitating reasoning over document collections and knowledge bases;
- integration of heterogeneous information resources.

Standardization as well as *building and use of proper ontologies* are indicated as the most effective instruments for providing and maintaining semantic interoperability between digital library systems.

In its role of an instrument for achieving semantic interoperability, standardization may comprise the form and meaning of metadata and content description schemata. It may also have an influence upon the use of names and construction of identifiers for concepts and real world items.

Standardization may direct to better convenience for developers and in some sense for users of digital libraries. The available information resources can be kept in a single form. Similarly, information can be communicated (transferred, integrated, etc.) without intermediate transformation or alteration.

Standardization issues are relevant to different degree in most aspects of the development of digital library systems. As most important from research point of view one may mention the following types of standards: standards for creation of digital repositories; archiving and preservation standards; metadata standards; standards for electronic publishing of books, periodicals and other media.

The utilization of proper ontologies is one of the well-accepted mechanisms for achieving semantic interoperability of digital libraries.

Ontologies are considered as metadata vocabularies providing semantic context in determining the relevance of resources [10]. Ontologies are usually developed in order to define the meaning of concepts and terms used in a specific domain. So the use and sharing of proper ontologies helps to achieve a consistent understanding of the content of the preserved resources amongst the corresponding software systems of different digital libraries.

According to [11], semantic interoperability depends mainly on the existence and use of well-formed and accepted upper and core ontologies, in which the basic concepts and relationships are defined. Then, the concepts defined in the upper and core ontologies, should be extended and substantiated by appropriate domain ontologies.

Semantic Digital Libraries. Semantic digital libraries [12, 13] are digital library systems that apply semantic technologies to achieve their specific goals. They may be characterized as a particular class of knowledge based systems in which the inference engine has a limited functionality and plays mainly a support role for the corresponding intelligent search engine.

Semantic digital libraries provide new search paradigms for the information space – intelligent (semantic or ontology-based) search and community-enabled browsing. They also provide interoperability on the data level – integrate metadata from various heterogeneous sources and in this way support interconnecting different digital library systems.

Ontologies play a major role in semantic digital libraries to cope with the variety of problems caused by the structural differences of existing systems, the semantic differences of metadata standards, the maintenance of various types of intelligent search, etc.

As defined in [14], an ontology is an explicit specification of a conceptualization. In this sense ontologies are the main and almost the only one form of description of conceptual knowledge used in semantic digital libraries.

Ontologies have been a key factor in the evolution of digital libraries. They can be used in semantic digital libraries to [15]:

- organize bibliographic descriptions;
- represent and expose document contents;
- share knowledge amongst users.

They have been used in building *semantic annotations* of the information resources available on the Semantic Web.

The support of functionalities for *intelligent search* [16, 17], also known as *semantic search*, is one of the main features of semantic digital libraries. Ontologies play a key role in this kind of search. Three types of ontologies have been identified as a support for semantic search: *bibliographic ontologies*, *subject ontologies*, and *community-aware ontologies*. Bibliographic ontologies describe metadata standards. Subject ontologies are useful as knowledge sources which define the meaning of most domain concepts, their hierarchy, properties and relationships. Community-aware ontologies are oriented to the description of the different types of users, their requirements and interactions.

As most popular examples of successful semantic digital libraries one can mention JeromeDL (`http://www.deri.ie/content/jerome-dl`) and BRICKS (`http://www.brickscommunity.org/`).

The careful study of the experience in development and use of digital libraries shows that current (semantic) digital libraries are not enough because [18]:

- digital libraries should not be for librarians only but for average people;
- they concentrate on delivering content/information, not on knowledge sharing within a community of users;
- digital libraries have lost the human part of their predecessors.

The so-called *social semantic digital libraries* are suggested as a solution of these problems (see Figure 3). As formulated in [18], they have the aim to make users/readers

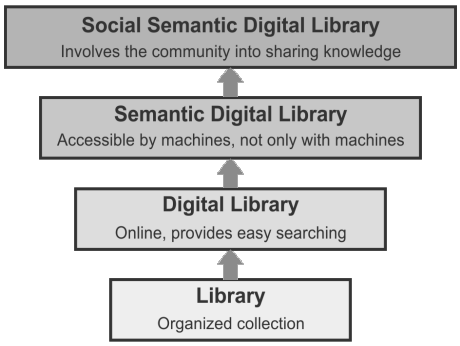

Fig. 3. Evolution of libraries

involved in the content annotation process, allow users/readers to share their knowledge within a community, and provide better communication between users in and across communities.

Main Characteristics of the Semantic Digital Library DjDL. As an example illustrating some specific aspects of the application of semantic technologies in building academic digital libraries we discuss here DjDL – a semantic digital library with Bulgarian folk songs.

DjDL preserves over 1000 Bulgarian folk songs presented with their notes, text and music and provides adequate access to the available digital content. The development of its prototype was supported by the Bulgarian National Science Fund within a project titled "Information technologies for presentation of Bulgarian folk songs with music, notes and text in a digital library" [19, 20]. The collection treasured in DjDL constitutes a considerable part of the digitized archive of Prof. Todor Dzhidzhev published in [21].

The functional structure of DjDL includes six main components: a metadata catalogue, a repository, a subject ontology, a search engine, a module implementing the library functionality, and an interface module.

The metadata catalogue consists of descriptions in RDF format of the particular folk songs included in the repository. These descriptions contain various types of metadata, for example: the title of the song, the song genre in accordance with different classification schemes, the folklorist who gathered the song, the singer(s), the date and place of record, etc. More accurately, each catalogue entry contains the text (i.e., the lyrics) of a particular song accompanied with the corresponding metadata.

The repository of DjDL contains heterogeneous resources of three types: lyrics of songs, notations of songs, musical (MP3) files with the authentic performances of the songs.

The subject ontology includes concepts of different areas related to the contents of the lyrics of songs, with description of their properties and various kinds of relationships between them. It plays a significant role in the implementation of the functionalities of the search engine. The purpose of the search engine is to provide adequate access to the resources stored in DjDL.

The library functionality and the user interface of DjDL are designed in accordance with the examined requirements of the typical users of the library. The interface module provides flexible online access to the library resources and some additional software tools.

The subject ontology describes conceptual knowledge in several domains, relevant to the content of Bulgarian folk songs. It contains definitions of the main domain concepts, their properties/relationships and representative instances. This subject ontology consists of a set of interrelated subontologies supporting the search engine of DjDL [20]:

- ontology of folk songs which includes various genre classifications of folk songs (e.g. by their thematic focus, by their cultural functions, etc.);
- ontology of manner of life and family (professions, instruments, clothing, ties of relationship, feasts, traditions and rites, etc.);
- ontology of impressive events and natural phenomena;
- ontology of social phenomena and relationships (exchanges, elections, unrest, etc.);
- ontology of historic events;
- ontology of administrative division – combines the current administrative division

of Bulgaria with the one from the beginning of 20th century.

The properties "synonym" and "form" provide the search engine with suitable synonyms and grammatical or colloquial forms of the terms used as names of ontology concepts.

Our current activities are aimed at giving DjDL some characteristics of a social semantic digital library. In particular, an authoring tool has been under development in order to enable users to participate in the content annotation process, to communicate and to share their impressions and opinions.

Search Engine of DjDL. The search engine of DjDL [20] supports two main types of search: keywords-based and semantic search. The implementation of the semantic search module is mostly based on an appropriate utilization of the subject ontology.

The folklore lyrics uses lots of similes, metaphors, idioms and other sophisticated or language-dependent stylistic devices. In keeping with this essential domain feature, it is expedient to combine the use of proper subject ontologies with other Artificial Intelligence tools to provide more adequate support for the semantic search.

In this sense we defined a set of natural language-dependent patterns of typical stylistic or thematic constructs that can be matched with relatively large parts of the texts of folklore songs and have specific meaning as e.g. an expression of "unfaithfulness", "jealousy", "discontent", etc. We call them *concept search patterns*. A set of special symbols that may be used in the description of these patterns and the corresponding pattern matching rules were defined as well.

The search engine provides a set of facilities for augmentation of the user queries – automatic query reformulation according to the available subject ontology. The user may refine the resulting augmented queries. When possible, proper concept patterns are applied during the search process.

The most considerable source for augmentation of the user queries is the taxonomy (the is-a hierarchy) of concepts which serves as the basis of the subject ontology. During the augmentation of the user query, first of all an exhaustive breadth-first search in the graph representing the "is-a" concept hierarchy is performed, starting from the node which corresponds to the original user query. The names of the visited nodes that are in fact the respective more specific concepts described by the ontology, are added to the one given by the user. The resulting list of concepts is properly visualized and placed at user's disposal for further "manual" refinement.

Within the next step of query expansion, the search engine adds to the newly constructed set of queries some derivatives and synonyms of the main terms found as values of their "form" and "synonym" properties in the subject ontology. The corresponding property values from the definitions of all concepts included by that time in the expanded user query and the existing instances of these concepts are added to the query as well. Finally, the values of such properties of the newly included instances that have been explicitly specified as significant for their classes/concepts with respect to the semantic search (e.g. the property "participant" of the instances of "historic event" and the property "role" of the instances of "traditional rite"), are included in the resulting augmented query.

Thereby the user query is augmented as far as possible in terms of the subject ontology and in fact it has the form of a disjunction of all included variations of concepts and

instance names. In this form the resulting query is ready for further refinement and processing.

If there is a concept in the augmented query provided with appropriate search pattern(s), the pattern matching module performs an additional search for each of these patterns.

The search engine of DjDL supports some additional functionalities which enable the user to combine the search and retrieval of documents with a kind of sentiment analysis of their texts. For that purpose some of the subject ontology concepts are associated with proper positive or negative numbers which play the role of their sentiment estimates.

The sentiment of a song has been defined in accordance with the sum of the sentiment estimates of the particular words in the lyrics of this song. Futhermore, the specializations of ontology concepts and all their forms and synonyms inherit the sentiment estimates of the corresponding concepts. It other words, the sentiment estimates of the ontology concepts have been used as default values of their specializations and forms. Thus we use the subject ontology of DjDL as a principal form for conceptual knowledge in its software system as well as the main semantics resource for sentiment analysis of the lyrics of folklore songs. In addition, the sentiment symbolized by phrases that match existing concept search patterns, is considered first of all during the sentiment analysis process. The concept search patterns are provided with proper sentiment estimates which may override the ones of the corresponding ontology concepts. Moreover, the sentiment estimates of some particular forms of a set of distinct ontology concepts have specific values in accordance with their typical sense and cases of use.

Conclusion. Semantic Web technologies can be combined with the methodology and proper technologies of Web 2.0 in the development of a new generation of digital libraries – the so-called social semantic digital libraries. The latter integrate information resources based on various types of metadata, provide interoperability with other systems and provide user friendly and adaptive search and document retrieval tools. They enable users to participate in the content annotation process and can assist users/readers to communicate and share their knowledge and opinion. In this way social semantic libraries may serve as environments supporting collaborative learning and research.

REFERENCES

[1] L. Candela, D. Castelli et al. The DELOS Digital Library Reference Model: Foundations for Digital Libraries. ISSN 1818-8044, ISTI – CNR, 2007.

[2] L. Candela, D. Castelli, P. Pagano. A Reference Architecture for Digital Library Systems. *ERCIM News*, **66** (July 2006), 21–22.

[3] P. Teregowda, B. Urgaonkar, C. Giles. Cloud Computing: A Digital Libraries Perspective. *IEEE Cloud* 2010, 115–122.

[4] P. Georgiadis, N. Spyratos, V. Christophides, C. Meghini. Personalizing Digital Library Access with Preference-Based Queries. *ERCIM News*, **66** (July 2006), 34–35.

[5] F. Seyedarabi. Personalization: An Emerging Direction for Tackling the Web Searching Barriers Faced by Teachers when Searching for Educational Resources. *Webology*, ISSN 1735-188X, **8**, No 2 (December 2011).

[6] M. Sredkov, T. Trifonov, K. Georgiev. Sophie 2.0 – A Platform for Reading, Writing and Publishing of Electronic Books in a Network Environment. Proceedings of the Fourth ISGT Conference (Sofia, May 28–29, 2010), ISBN 978-954-07-3168-1, Sofia, St. Kliment Ohridski University Press, 2010, 145–152.
[7] T. Berners-Lee et al. The Semantic Web. *Scientific American* (May 2001), 35–43.
[8] V. Sugumaran, J. Gulla. Applied Semantic Web Technologies. CRC Press, 2012.
[9] M. Patel et al. DELOS: A Network of Excellence on Digital Libraries. D5.3.1: Semantic Interoperability in Digital Library Systems. UKOLN, University of Bath, 2005.
[10] M. Doerr. The CIDOC CRM: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, **24** (2003), No 3, 75–92.
[11] N. Guarino, M. Carrara, P. Giaretta. Formalizing Ontological Committment. Proceedings of the 12th National Conference on Artificial Intelligence AAAI-94 (Seattle, Washington, July 31–August 4, 1994), Morgan-Kaufman, 2004, 560–567.
[12] S. Kruk, T. Woroniecki, A. Gzella. JeromeDL – a Semantic Digital Library. Proceedings of the Semantic Web Challenge Workshop at ISWC2007, Busan, South Korea, 2007.
[13] L. Hollink et al. Semantic Web Opportunities for Digital Libraries. 32nd Library Systems Seminar of the European Library Automation Group (ELAG 2008), Wageningen, The Netherlands, 2008.
[14] T. Gruber. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, **43** (1995), 907–928.
[15] S. Kruk et al. The Role of Ontologies in Semantic Digital Libraries. 10th European Conference on Research and Advanced Technology for Digital Libraries, Alicante, Spain, 2006.
[16] R. Guha, R. McCool, E. Miller. Semantic Search. In: Proceedings of the 12th International World Wide Web Conference (WWW 2003), Budapest, Hungary, 2003, 700–709.
[17] J. Lervik, S. Brygfjeld. Search Engine Technology Applied in Digital Libraries. *ERCIM News*, **66** (July 2006), 18–19.
[18] S. Kruk, B. McDaniel. Conclusions: The Future of Semantic Digital Libraries. In: S. Kruk, B. McDaniel (Eds.), Semantic Digital Libraries. Springer, 2009, 215–222.
[19] N. Kirov. Digitization of Bulgarian folk songs with music, notes and text. *Review of the National Center for Digitization*, ISSN 1820-0109, **18** (2011), 35–41.
[20] M. Nisheva-Pavlova, P. Pavlov. Ontology-Based Search and Document Retrieval in a Digital Library with Folk Songs. *Information Services and Use*, **31** (2011), No 3–4, 157–166.
[21] T. Dzhidzhev. Folk Songs from Thrace. L. Peycheva, G. Grigorov, N. Kirov (Eds.), Prof. Marin Drinov Academic Publishing House, Sofia, 2013.

Maria Mihaylova Nisheva-Pavlova
Faculty of Mathematics and Informatics
Sofia University
5, James Bourchier Blvd
1164 Sofia, Bulgaria
e-mail: marian@fmi.uni-sofia.bg
and
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia, Bulgaria

ПРИЛОЖЕНИЕ НА СЕМАНТИЧНИ ТЕХНОЛОГИИ ПРИ СЪЗДАВАНЕТО НА ЦИФРОВИ БИБЛИОТЕКИ

Мария Нишева-Павлова

В статията са анализирани някои съвременни тенденции на изследванията в областта на цифровите библиотеки. Изложението е фокусирано върху така наречените семантични технологии и тяхната роля за осигуряването на семантична оперативна съвместимост между софтуерните системи на цифрови библиотеки, както и при създаването на две нови поколения цифрови библиотеки – семантични цифрови библиотеки и социални семантични цифрови библиотеки. Дискутират се въпроси, свързани с разработването на методи за интелигентно (семантично) търсене. Като илюстрация на представените идеи е разгледана конкретна академична цифрова библиотека.