Search engine in a class of academic digital libraries

Maria Nisheva-Pavlova, Pavel Pavlov

Faculty of Mathematics and Informatics, Sofia University
5 James Bourchier Blvd., Sofia 1164, Bulgaria
{marian, pavlovp}@fmi.uni-sofia.bg

Abstract

The paper discusses some aspects of an ongoing project aimed at the development of a methodology and proper software tools for building and usage of academic digital libraries. A particular functional model of academic digital library has been proposed and analyzed. The emphasis falls on some solutions of the large set of problems concerning the development of adequate mechanisms for semantics-oriented search in multilingual digital libraries. An ontology-based approach is suggested in order to standardize the semantic annotation of the library resources and to facilitate the implementation of the functionality of the search engine. The main features of a prototype of knowledge-based search engine for a multilingual academic digital library with research and learning materials are discussed. This search engine uses proper ontologies describing the conceptual knowledge considerable for the chosen domains and in this way it is capable of retrieving and filtering documents by their semantic properties.

Keywords: Digital Library; Metadata; Semantic Annotation; Ontology; Search Engine

1. Introduction

Research and practical activities in the field of Digital Libraries during the last two decades lead to significant results in the development and management of digital collections, in the innovation in scholarly publishing and the long-term preservation of digital information. Many institutions are actively involved in building suitable repositories of the institution's books, papers, theses, and other works which can be digitized or were “born digital”. In

particular, universities and other academic institutions participate successfully in lots of projects directed to the development of different types of *academic digital libraries*. Academic digital libraries are committed to maintaining valuable collections of scholarly information. To this end, essential information resources should remain available and accessible into the future – a real challenge in the cases of digital resources that are increasingly transient and at risk.

The paper is aimed at the presentation of an ongoing project which is directed to the development of a methodology and corresponding software tools for building academic digital libraries. A special attention has been paid to the elaboration of means for semantics-oriented search in multilingual digital libraries. The study and the practical experiments are oriented to the development of DigLib-CI – a digital library with research and learning materials (articles, dissertations, monographs, lecture notes, textbooks, presentations, example program sources, data sets, quizzes, manuals etc.) created at the Department of Computer Informatics of the Faculty of Mathematics and Informatics (FMI), Sofia University, or especially selected from among the scholarly materials freely available on the Web.

2. Related Work

Digital Libraries can mainly be characterized as a converging point where disparate communities have been meeting to address common issues related with the creation, management and usage of digital information [1]. The goal of a digital library and especially of an academic library is to provide access to selected intellectual works. Moreover, academic digital libraries are usually aimed at some specific challenges like digital preservation of valuable scientific heritage collections and investigation of innovative methods for automatic indexing, metadata extraction, document search and retrieval etc. In this sense, academic digital libraries are the front-rankers in the discussed area.

The digital libraries of Cornell University [2], the University of Michigan [3] and Carnegie Mellon University [4] are considered as leaders in the field of academic digital library creation and management.

The Cornell University Library is the eleventh largest academic library in the United States, ranked by number of volumes held. In 2005 it held 7.5 million printed volumes in open stacks, 8.2 million microfilms and microfiches, and a total of 440,000 maps, motion pictures, DVDs, sound

recordings, and computer files in its collections, in addition to extensive digital resources and the University Archives.

The Cornell Library Digital Collections Project integrates online collections of historical documents. Featured collections include the Database of African-American Poetry, the Historic Math Book Collection, the Samuel May Anti-Slavery Collection, the Witchcraft Collection, and the Donovan Nuremberg Trials Collection.

The University of Michigan Digital Library Project (UMDL) is based on the traditional values of service, organization, and access that have made libraries powerful intellectual institutions in combination with open, evolving, decentralized advantages of the web. The content of UMDL will emphasize a diverse collection, focused on earth and space sciences, which can satisfy the needs of many different types of users. The content will be supplied by publishers, although the project will eventually allow all users to publish their work.

The implementation of the current prototype of UMDL requires the integration of numerous agent technologies for knowledge exchange, commerce, learning, and modelling. Recently, the efforts have been concentrated on developing technologies that, for example, manipulate ontological descriptions of the elements of a digital library to help agents find services and auctions for exchanging goods and services under various conditions. These technologies allow flexibility in the UMDL configuration policies, extensibility and scalability by using demand as incentive for replicating services.

Carnegie Mellon University Libraries became very popular with the Million Book (or the Universal Library) project which was aimed to digitize a million books by 2007. The activities within the project include scanning books in many languages, using OCR to enable full text searching, and providing free-to-read access to the books on the web. As of 2007, they have completed the scanning of the planned number of books and have made accessible the corresponding database.

The research within the Million Book project includes developments in machine translation, automatic summarization, image processing, large-scale database management, user interface design, and strategies for acquiring copyright permission at an affordable cost.

Compared to these well-known large scale initiatives, our project is of a significantly smaller scale, but in contrast to all of them, it investigates the use of a set of subject ontologies to provide flexible, semantics-oriented access to the library resources for users with different profiles and language skills.

3. Architecture of DigLib-CI

DigLib-CI is designed as a typical academic digital library. It has been under development at FMI in order to provide open access to various kinds of scholarly and instructional content, mainly in a wide range of subfields of Computer Science and Information Systems. The functional structure of DigLib-CI is shown in Figure 1.

The content repositories include research and learning materials of different types (books, dissertations, periodicals and single articles, manuals, lecture notes, presentations, source code of computer programs, data sets, tests, quizzes etc.) in the areas of Computer Science and Information Systems. These library resources are available in various digital formats: pdf, html, plain text, doc, ppt, jpeg etc. Most of them are developed by faculty members, the others are especially selected from among the scholarly materials freely available on the Web. The content repositories are stored in a small number of locations. The materials in them are written in Bulgarian or in English language.

The metadata catalogues are destined to facilitate the identification of the needed research or learning materials by the search engine. They contain descriptive metadata stored in XML format and support the reusability of all library resources and facilitate their interoperability.

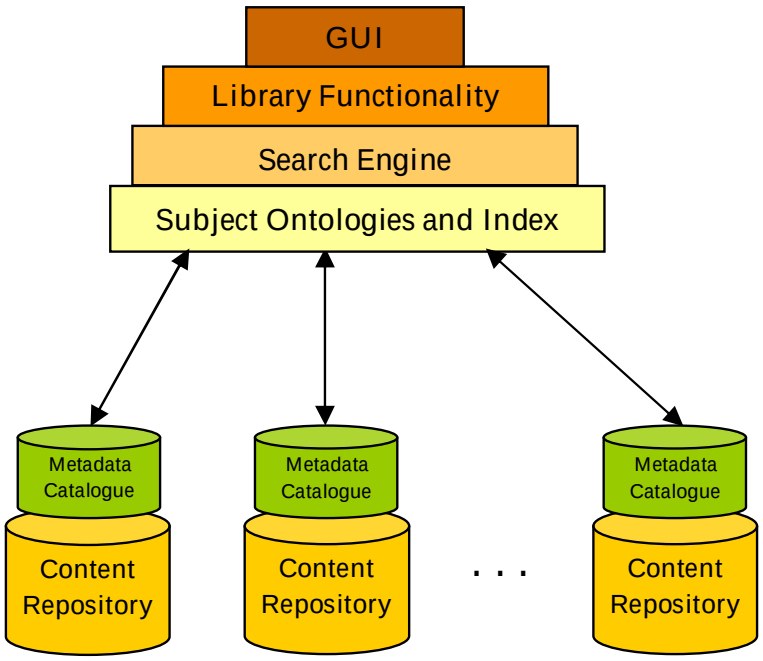

Figure 1: Functional Model of DigLib-CI

The subject ontologies include large sets of concepts of the areas of Computer Science and Information Systems, with description of their properties and different kinds of relationships among them. They play a significant role in the implementation of the full functionality of the search engine.

The purpose of the search engine is to provide adequate access to the complete palette of resources stored in DigLib-CI.

The library functionality and the user interface of DigLib-CI are designed in accordance with the expected needs and requirements of the basic types of users of the library. The interface module provides adequate online access to the corresponding library resources and supporting software tools.

4. Catalogue Metadata

The library catalogues contain metadata which support the identification of the requested resources by the search engine. These metadata are stored in XML format and comply with the IEEE Standard for Learning Object Metadata [5].

Typical examples of relevant attributes of most kinds of research and learning materials are: type of the material; author; title of the material; language(s) (human and/or programming one(s)); digital format; location; version; date of creation; completion status; restrictions on use; semantic annotation – list of concepts from a proper subject ontology describing the Computer Science or Information Systems subfields and/or concepts covered or treated by the material. Learning materials have been characterized also by their educational level and the principal types of users for which the corresponding material was designed; officially published research materials and textbooks are supplied with the corresponding bibliographic metadata.

Each catalogue entry (i.e., each resource description) consists of two equivalent parts in which the element values are texts in Bulgarian or English language, respectively. The search engine examines the corresponding parts of the descriptions according to the language of the user query.

The elements <ontologyRefs> and <keywords> of the resource descriptions play the role of semantic annotations of the corresponding library materials. The values of the child elements of <ontologyRefs> are concepts of the suitable subject ontologies (names of classes in these subject ontologies) which present most precisely the content of the corresponding document.

The concepts of the subject ontologies are too general from the point of view of the expectations of the typical users of DigLib-CI. For that reason one

can include in the resource descriptions additional lists of keywords which describe the content of the corresponding documents at the necessary level of abstraction. These keywords are set as values of the child elements of the <keywords> resource description elements.

The names of the respective subject areas and names of the files containing the suitable subject ontologies have been assigned as values of the child elements of the catalogue description elements <subjects> and <ontologies> respectively.

5. Subject Ontologies

The subject ontologies include a large set of concepts in the fields of Computer Science and Information Systems, with description of their properties and the different kinds of relationships among them. Two subject ontologies are included in the current version of DigLib-CI. The Computer Science ontology is based on the Computer Science Curriculum 2008 of ACM and IEEE/CS [6]. Using the curriculum as a guideline, this ontology defines the atomic knowledge units for the University courses and available research materials in the field of Computer Science and makes them sharable and reusable. Its current version includes approximately 300 concepts with their relationships.

The Information Systems ontology has been under development using the Model Curriculum and Guidelines for Undergraduate Degree Programs in Information Systems of ACM, AIS and AITP [7].

The subject ontologies are designed in order to play the role of information sources describing the hierarchy and the other relationships between the main concepts in the discussed domains. A dictionary of synonyms has also been under development with the purpose of providing the search engine with other viewpoints to the conceptual structure of the areas of Computer Science and Information Systems.

The body of knowledge in the areas of Computer Science and Information Systems is formulated in the terms of a considerable number of common concepts, therefore the two subject ontologies discussed above contain many common classes (with equal or similar names and intersecting properties and restrictions on them). Because of that our further plans include the development of an approach to the integration of domain ontologies relevant to the contents of multilingual academic digital libraries which will be based on some of our former results [8].

6. User Interface

The library functionality and the user interface of DigLib-CI are designed in accordance with the expected requirements of the basic types of users of the library. The interface module provides adequate online access to the corresponding library resources and supporting software tools.

The current version of the user interface allows one to formulate queries in Bulgarian or English language. It is intended for four types of users:

- FMI students – they may read/download textbooks, open lecture notes and presentations from all public sections of the library as well as all manner of other kinds of materials (monographs, dissertations, articles, periodicals, degree theses, lecture notes, presentations, exercises, programs, data sets, quizzes, tests etc.) from fixed public library sections;
- FMI lecturers and researchers – in addition to the students' access rights, they may upload materials to fixed public sections as well as create and update private sections and use materials in some of them;
- librarians (library administrators) – they have full access to all public resources of the library (may download and upload materials destined for all public sections of the library);
- general citizen – they may read and download public materials of fixed types (e.g., dissertations, textbooks, open lecture notes and presentations).

All types of users of DigLib-CI may use the standard input interface which provides convenient means for entering, editing and submitting queries for various kinds of document search and retrieval. FMI lecturers and researchers as well as the library administrators may play the role of authors of library resources and have an access to the author's part of the user interface. This part of the user interface places at the authorized persons' disposal appropriate forms enabling one to enter and edit catalogue descriptions of all types of library resources (Figure 2). More precisely, the user may enter the values of some of the elements or pick out the values of others from previously drawn lists. In particular, the available subject ontologies can be properly visualized and the necessary concepts in them can be picked out as values of the child elements of the element <ontologyRefs>.

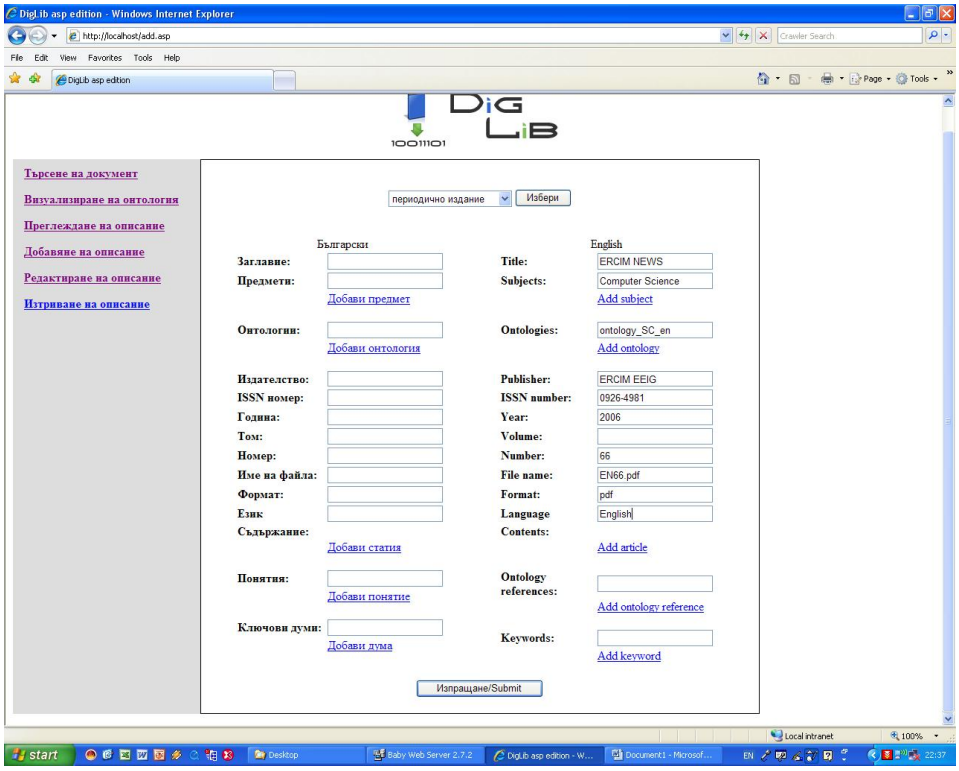

Figure 2: User interface of DigLib-CI (author's view – form for entering catalogue metadata of periodicals)

7. Working Principles of the Search Engine

The purpose of the search engine is to provide adequate access to the complete palette of resources stored in DigLib-CI.

The search engine maintains several types of search and document retrieval within DigLib-CI. The user queries define restrictions on the values of certain metadata attributes of the required research or learning materials. Generally, the search mechanism may be formulated as follows: the document descriptions included in all permissible user sections of the library are examined one by one and these descriptions which have a specific element (determined by the type of the user query) with a value matching the user query, are marked in order to form the search result. The matching process is successful if the value of the element or the value of one of its child elements is equal to the user query. The documents pointed by the marked descriptions are retrieved and the user is given an access to these documents and their catalogue descriptions.

The current implementation of the search engine supports four types of search and document retrieval:

- full search – search and retrieval of all available library resources, ordered by title, by author, by category, by date of creation or by date of inserting in the library;
- author search (search and retrieval of the documents created by a given author) – the search is performed in the value of the element <authors>;
- ontological search – the search is performed in the value of the element <ontologyRefs>;
- keyword search – the search is performed in the value of the element <keywords>.

During the ontological search the user query is augmented with regard to the concepts searched out in the semantic annotations of the required research or learning materials. The more specific concepts from each of the subject ontologies indicated by the user are added to the original one in the resulting query. Then the search engine retrieves all documents in the library containing in their descriptions at least one component of the augmented query as the value of a child element of <ontologyRefs>. In this way the ontological search enables one to find documents described by ontology concepts which are semantically related to the concept defining the user query.

Till now, we have no disposal of an accomplished proper dictionary of synonyms of the concepts in the areas of Computer Science and Information Systems neither in Bulgarian, nor in English, but our idea is to provide a possibility for two-stage augmentation of the user query. At the first stage the request for ontological search will be extended with the more specific concepts (its successors) from the indicated subject ontologies. At the second stage the synonyms found in the dictionary will be added to the main (given by the user) concept and its successors.

We allow in the current version of the implementation of the search engine only "atomic" user queries that do not contain conjunctions or disjunctions of words or phrases. The next step will be to elaborate a sophisticated version of the search engine which will be capable to analyze and execute queries in the form of conjunctions or disjunctions of phrases of interest for the user. Some of our former ideas suggested in [9] will be used for the purpose.

The discussed working principles of the search engine of DigLib-CI are designed in order to support flexibility, interoperability and reusability. These principles could be applied in the implementation of the search engines of a whole class of academic digital libraries that provide semantics oriented access to their resources.

8. An Example of Ontological Search

Let us suppose for example that the user defines a request (a query) for ontological search concerning the concept “fundamental constructs”. First an extension of this request will be generated. It will include all ontological concepts which are special cases of the concept given by the user (with respect to the ontologies indicated by the user). For this purpose, breadth-first search in the graphs that represent the ontologies will be performed, starting in each one from the concept chosen by the user.

Assume that the Computer Science ontology is chosen by the user. In this case the extended request (the augmented query) will include the concepts “fundamental constructs”, “basic syntax and semantics”, “binding and scope”, “conditional structures”, “declarations”, “expressions”, “functions and procedures”, … , “variables”, “bindings”, “blocks”, … , “simple variables”.

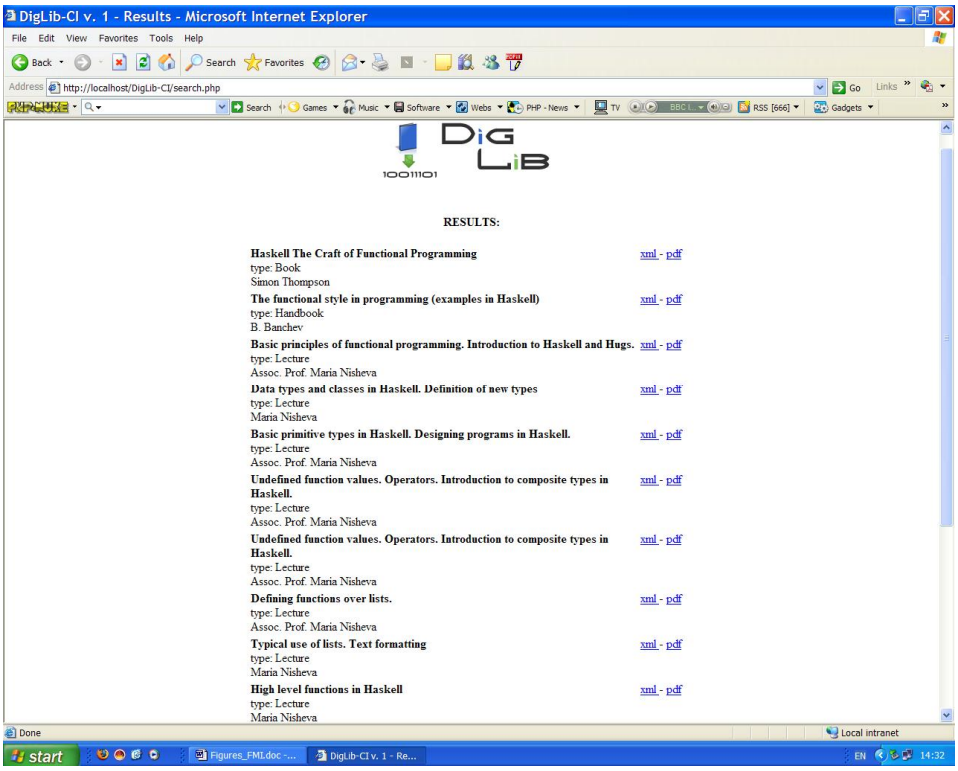

Figure 3: Some search results for the query “fundamental constructs”

After that, a consecutive search in the catalogue descriptions follows. In this search all documents with descriptions that are juxtaposed with at least one element of the extended request are extracted. In the current implementation each document appears as many times in the result list, as many elements of the augmented query are juxtaposed with its description

(which means that the element <ontologyRefs> of the description includes a sub-element that has value, coincident with an element of the augmented query).

Figure 3 shows a screenshot displaying part of the ontological search results for the query "fundamental constructs".

If the user indicates more than one subject ontology (e.g., the Computer Science ontology and the Information Systems ontology), the procedure described above is repeated consecutively for each of these ontologies.

Our current activities are directed to the selection of a proper set of relationships between the ontology concepts that should be taken into account in the process of ontological search along with the hierarchical ones. We envisage for the near future the development of a more flexible and user-friendly mechanism for ontological search which will not expect the user to indicate explicitly the subject ontologies appropriate for every particular case.

9. Conclusions

The most considerable results of the discussed project obtained so far may be summarized as follows:

- A functional model of an academic digital library was proposed. This model provides tools for semantics oriented access to learning and research materials in various digital formats written in different languages;
- A prototype of DigLib-CI – an academic digital library with research and learning materials in the areas of Computer Science and Information Systems, was developed.

The main advantage of the suggested approach to building academic digital libraries consists in the provided facilities for flexible and adequate semantics-oriented access to the library resources for users with various professional profiles and language skills.

The complete implementation of the project will help to enhance the research activities and the exchange of teaching innovation and thus will improve the overall scholarly and teaching quality in Computer Science and Information Systems at FMI. It will also contribute to the methodology of development of innovative software systems maintaining the entire lifecycle of academic digital content.

Acknowledgements

This work has been partly funded by the Sofia University SRF within a project titled "Methods and Tools Supporting the Lifecycle of Rich Digital Content".

References

[1] BORBINHA, J. The Age of the Digital Library. In: D. Castelli, E. Fox (Eds.), Pre-proceedings of the first International Workshop on Foundations of Digital Libraries, Vancouver, Canada, 2007, pp. 31-36.

[2] Cornell University Library Digital Collections. Available at http://cdl.library.cornell.edu/ (March 2010).

[3] University of Michigan Digital Library. Available at http://www.si.umich.edu/UMDL/ (March 2010).

[4] Carnegie Mellon University Libraries: Digital Collections. Available at http://diva.library.cmu.edu/ (March 2010).

[5] IEEE Standard for Learning Object Metadata. Available at http://ltsc.ieee.org/wg12/20020612-Final-LOM-Draft.html (March 2010).

[6] Association for Computing Machinery; IEEE Computer Society. Computer Science Curriculum 2008: An Interim Revision of CS 2001. Available at http://www.acm.org/education/curricula/ (March 2010).

[7] ACM; AIS; AITP. IS 2002: Model Curriculum and Guidelines for Undergraduate Degree Programs in Information Systems. Available at http://www.acm.org/education/ (March 2010).

[8] ZLATAREVA, N; NISHEVA, M. Alignment of Heterogeneous Ontologies: A Practical Approach to Testing for Similarities and Discrepancies. In: D. Wilson, H. Chad Lane (Eds.), Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, AAAI Press, Menlo Park, CA, 2008, pp. 365-370.

[9] PAVLOV, P; NISHEVA-PAVLOVA, M. Knowledge-based Search in Collections of Digitized Manuscripts: First Results. In: Proceedings of the 10th ICCC International Conference on Electronic Publishing, FOI-Commerce, Sofia, 2006, pp. 27-35.