JeromeDL - Adding Semantic Web Technologies to Digital Libraries

Sebastian Ryszard Kruk[1], Stefan Decker[1], and Lech Zieborak[2]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway, Ireland
<firstname.lastname>@deri.org
[2] Main Library, Gdansk University of Technology, Gdansk, Poland
zieborak@pg.gda.pl

Abstract. In recent years more and more information has been made available on the Web. High quality information is often stored in dedicated databases of digital libraries, which are on their way to become expanding islands of well organized information. However, managing this information still poses challenges. The Semantic Web provides technologies that are about help to meet these challenges.
In this article we present JeromeDL, a full fledged open-source digital library system. We exemplify how digital library content management can benefit from the Semantic Web. We define and evaluate browsing and searching features. We describe how the semantic descriptions of resources and users profiles improve the usability of a digital library. We present how digital libraries can be interconnected into one heterogeneous database with use of semantic technologies.

1 Introduction

Increasing investment in research and development as well as the trend to produce more and more written information are challenges for Digital Libraries. Metadata - one of the main cornerstones of the Semantic Web - has been used in libraries for centuries. E.g., tree-structured classifications schemes used to coordinate materials on the same and related subjects lead to the hierarchical organization of catalogs. The catalog approach to manage resources has been successfully adapted in on-line directories like Yahoo! or the Open Directory Project. Since more and more information becomes available also a number of search engines have emerged. Even the search engines utilizing algorithms like PageRank[1] still seem to not always find the high quality information we desire.

The Semantic Web effort continues to explore the usage of metadata on the Web, and comes up with additional ways how to manage resources. Semantically annotated content of the digital library's database can be retrieved by new properties. These properties interlink different resources and provides new meaningfull information to the existing fulltext and bibliographic descriptions.

In this paper we present JeromeDL, an open-source digital library, and exemplify which parts of JeromeDL benefit from the Semantic Web.

2 Architecture of Digital Library and the Semantic Web

The JeromeDL[3] architecture[5] is an instantiation of the architecture presented in [2] with special focus on the exploitation of the Semantic Web based metadata (RDF, FOAF, and ontologies). The main components of the JeromeDL system consists of:

- **resource management**: Each resource is described by the semantic descriptions according to the JeromeDL core ontology. Additionally a fulltext index of the resource's content and MARC21, and BibTEX bibliographic descriptions are provided. Each user is able to add resources via a web interface. To satisfy the quality of delivered content, each resource uploaded through the web interface has to be approved for publication. The administrative interface for librarians (JeromeAdmin) allows to manage resources and associated metadata (MARC21, BibTEX and semantic annotations) as well as to approve user submissions.
- **retrieval features**: JeromeDL provides searching and browsing features (see section 4.1) based on Semantic Web data.
- **user profile management**: In order to provide additional semantical description of resources[4], scalable user management based on FOAF (see section 3.2) is utilized.
- **communication link**: Communication with an outside world is enabled by searching in a network of digital libraries. The content of the JeromeDL database can be searched not only through the web pages of the digital library but also from the other digital libraries and other web applications. A special web services interface based on the Extensible Library Protocol (ELP)[8] (see secion 4.2) has been developed.

3 Semantic Description of Resources in Digital Libraries

There are several approaches to constructing the resource description knowledge base for digital libraries. Conventional catalogs and fulltext indexes are just the most popular examples. In addition one can use bibliographic descriptions like MARC21 or BibTEX. MARC21 consists of few keywords and free text values, without a controlled vocabulary. Therefore machines are not able to utilize much of a MARC21 description.

Text values are not enough to support machine based reasoning.

To perform more intelligent interactions with readers, the knowledge base must be equipped with semantics. The concept of ontology introduced by the Semantic Web is a promising path to extend Digital Library formalisms with the meaningfull annotations. Not exploiting existing standards in Digital Libraries would be a waste of resources. Therefore it is important to introduce ontologies to the digital libraries domain. The ontologies have to be compatible with already existing bibliographic description formalisms.

[3] JeromeDL - e-Library with Semantics: http://www.jeromedl.org/

3.1 JeromeDL Core Ontology

The main purpose of the bibliographic ontology is to *annotate* resources. There is no need to completely capture and formalize the content of the resources. The JeromeDL Core ontology (see Fig. 1) requires only a high level core, which is used to capture the essence of bibliographic annotation. This corresponds to the notion of an upper level ontology, as e.g., discussed in [7] for Digital Libraries. The upper level ontology for JeromeDL aims at compatibility with existing standards. So a good starting point for building an ontology for bibliographic description is DublinCore Metadata.

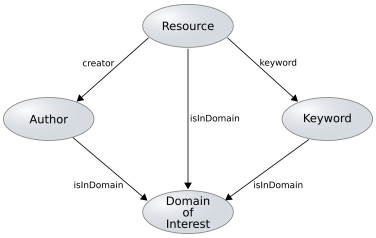

Fig. 1. JeromeDL core ontology

WordNet[4] was a potential candidate as a part of the JeromeDL core ontology. However, some properties defined for the keyword concept in the JeromeDL core ontology are not accessible within the WordNet ontology.

Resources may also be described in BibTEX and MARC21 format. The MarcOnt Initiative[5] is aiming to extend the MARC21 with ontological information. The annotation information is available in RDF format, and can be easily combined with other information. In JeromeDL this feature of RDF is exploited to connect resource information with social networking information and user profiles.

3.2 Semantic Social Collaborative Filtering for Digital Libraries

A classic library requires its users to identify themselves in order to be able to assign a resource (e.g., a book) to a reader. Digital resources in the Internet are often easily duplicable and often a reader does not have to identify himself before viewing specific content, with the exception of restricted resources. However, a reader of a Digital Library can benefit in many ways from the identification.

Registered readers are able to annotate, evaluate and classify resources stored in the JeromeDL database. Electronic bookmarks are popular on the WWW. Everyone can organize already examined resources the way he perceives the world. To identify categories a reader is interested in information on previously read books, electronic bookmarks, annotated resources and highly evaluated resources, are automatically collected. Each resource is described by some categories. Based on the collected categories JeromeDL identifies the categories a reader is interested in (see Fig. 2).

On-line communities introduced the idea of online social networking [9] as a way to interconnect members of a community at give community members a way to explore the community.

[4] http://www.cogsci.princeton.edu/~wn/
[5] MarcOnt Initiative: http://www.marcont.org/

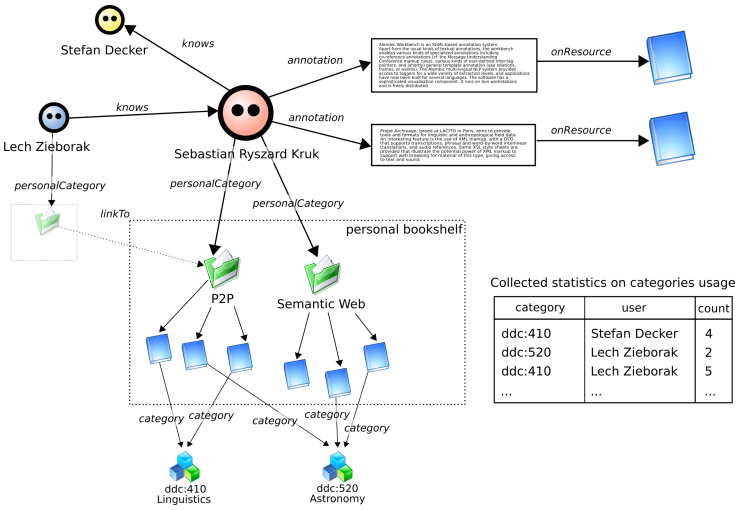

Fig. 2. The user's profile in JeromeDL

To manage the users' profiles in the JeromeDL system the FOAFRealm[6][3] library is used. It provides a semantic social collaborative filtering[4] features to resource retrieval. The FOAF based representation of user-profiles enables one to interlink FOAF-profiles with arbitrary other metadata, such that user profiles and social network information can be exploited automatically in searches.

The users of the JeromeDL are able to browse bookmarks of their friends and link some folders (categories) into their own structure. Readers can also state how much their interests are similar to their friends. Later on each of categories created by the reader have a unique ACL (access control list) that defines which friends are able to see or use the content of this category. The ACL entries are based on the distance and similarity level between issuer of the category and the user that is willing to read the content of this category.

3.3 Interconnecting Heterogeneous Ontologies

The multiplicity of different types of descriptions used in a digital library system can cause many problems when trying to interconnect them. Legacy bibliographic formats, such as MARC21, Dublin Core or BibTeX may take a form of binary file, text file with specific formatting or (if we are lucky) XML or RDF file. To take advantage of information which they contain, an framework must be created to manage and translate between different metadata.

Description of resource's structure, different bibliographic metadata (MARC21, BibTEX, DublinCore), user profiles and access control lists are hard to manage. The semantic description of resources provides a unified way to make the different types or even distributed metadata interoperable.

[6] FOAFRealm project: http://www.foafrealm.org/

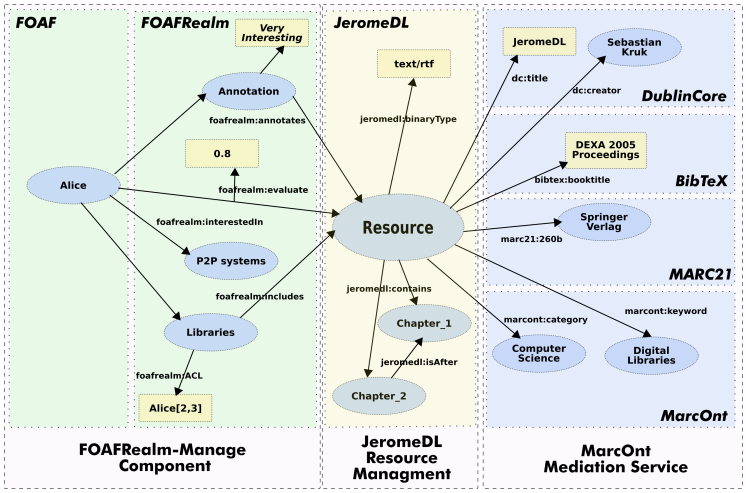

Fig. 3. Components and ontologies in JeromeDL

JeromeDL uses additional libraries to cope with the multiplicity of metadata being managed. Figure 3 represents the overview of ontologies and components used in JeromeDL to annotate resources. FOAFRealm-Manage component is responsible for use management based on FOAF and semantic social collaborative filtering. MarcOnt Mediation Service enables to describe resources with legacy bibliographic descriptions like MARC21 or BibTeX translated into MarcOnt ontology. It is possible to generate the legacy metadata description of a resource from existing semantic one. JeromeDL resource management handles descriptions of structure (e.g. chapters) and content of resources in RDF. And RDF-based object consolidation makes it possible to provide a semantically enhanced retrieval features (see section 4).

Transforming description of resource in legacy format such as MARC21 to semantic description requires few steps. An example flow of transforming MARC21 description to MarcOnt semantic description would be: (1) Parse a binary MARC21 file and create a MARC-XML file. (2) Transform a MARC-XML file to a MARC-RDF file using XSLT. (3) Transform a MARC-RDF graph to a MarcOnt semantic description.

The third step represents the most difficult task - translating one RDF graph into another one, using a different vocabulary. In other words - it requires specifying a set of rules, where a single rule identifies existence of one set of triples as a requirement of creating another set of triples. Translating MarcOnt semantic descriptions back into specified legacy format requires going in other direction on the same way. To perform translations between RDF graphs RDF Translator[7] has been used.

[7] RDF Translator: http://rdft.marcont.org/

4 Semantically Enhanced Resources Retrieval in Digital Library

4.1 Query Processing in JeromeDL

To initially find resources a search over the complete content of JeromeDL is possible. The search algorithm of JeromeDL consists of three major steps. Each step requires different metadata sources that describe resources in specific ways:

step A – the first step is the fulltext index search on the resources' contents and users' annotations on resources,
step B – the next step is the bibliographic description search consisting of MARC21 and BibTEX formats,
step C – the last step finally is a user-oriented search with semantics, based on the semantic description of the resources and information about most interested categories (regarding the user that issued the query).

Query object When issuing a query to JeromeDL a reader has to submit a query object (see example Fig. 4), usually using an HTML-form.

Each query contains several entries which state what information the search algorithm should look for in the resource description. The reader can choose from Dublin Core Metadata, MARC21-based and BibTeX-based properties. A special property that indicates the content of the resource is additionally provided. Each property contains the list of possible values, that the reader expects to find in the description of the resource. The user can specify which values are required and which are prohibited. Additionally each value may have a ranking value specified, so results containing a desired value are higher ranked. It is possible to define a maximum distance between words consiting the phrase value.

Result object A result object (see example Fig. 5) contains the information about the resources that have been found and additional information on the query process that has been executed. Each of the resources is described with: the URI of the resource, title and authors, categorizations and keywords, summary – digest, information on the type of the resource (like XSL:FO, SWF, an antique book's page scans), and the ranking of the resource in this result set. Additionally some debugging information and history of the query processing is included in the result object.

Semantically Enhanced Search Algorithm The search algorithm with semantics implemented in the JeromeDL system[5] processes the query object according to the flow described earlier and returns a set of result objects.

JeromeDL's search algorithm was designed based on the following requirements:

– The query should return resources where descriptions do not directly contains the required values.

IsSemanticQuery	`true`
IsConjunction	`false`
property	name="keywords" value=P2P(*mustExists*) value="Semantic Web" (*ranking*=10)
property	name="category" value=AI(*mustNotExists*)
...	
fulltext	value="semantic routing" (*proximity*=4)

Fig. 4. The search query object

resource	**uri**=http://jeromedl.org/show?id=... **title**="EDUTELLA: A P2P ... **author**="Wolfgang Nejdl, ... **categories**=[distributed systems, ...] **keywords**=[P2P, RDF] **summary**="Metadata for the WWW ... **bookType**=pdf **hits**=3
...	
info	"... *semantic web* is to general ...
...	

Fig. 5. The search result object

– The meaning of values provided in the query should be resolved in the context of users' interests.

These goals can be achieved by combining fulltext search as well as searching the bibliographic description and semantic descriptions of resources. The semantically enabled search phase includes query expansion based on the user's interests.

Bibliographic descriptions To provide support for legacy bibliographic description formats like MARC21 or BibTEX a digital library system needs to utilize the information provided by these descriptions.

RDF query templates In the last phase (phase C) of the search process the RDF query is performed. Since reader has no knowledge on the actually representation of the semantical description of resources. The literals provided in the query object are translated into the paths queries, based on the predefined ontology.

Semantically enabled query expansion If the size of the result set is outside the predefined range <`MIN`, `MAX`>, the query expansion procedure is called[5].

The information about the readers' interests and semantic descriptions of the resources is exploited to tailor the result set. Unnecessary resources are removed. Previously omitted resources are added to result object. All entries in result object are ranked according to user's interests.

The query expansion is performed iteratively. The decision which property to choose for the query expansion in an iteration depends on the number of the results received from the previous iteration.

Extrapolated profile When a reader has just registered to the JeromeDL system the profile information is incomplete. JeromeDL is *extrapolating user profiles* by asking for friends, whose profile is used to extrapolate the readers profile. During the search process the search engine is able to exploit categories defined by the reader's friends.

4.2 Searching in a distributed digital libraries network

A recent trend in digital libraries is to connect multiple digital libraries to federations where each digital library to able to search in other digital libraries systems. JeromeDL supports federated digital libraries by providing a communication infrastructure for a distributed network of independent digital libraries (L2L) similar to communication in a P2P network. Utilizing XML encoded query and result objects enabled building a SOAP based protocol prototype - Extensible Library Protocol (ELP)[8]. The use of Web Services for building the P2P network of digital libraries will enable connecting JeromeDL in the future to the ongoing projects like OCKHAM[8].

The idea of the ELP is to allow communication in the heterogeneous environment of digital libraries. Each library has to know about at least one other digital library, so it could connect to the L2L network. Each query is processed across the minimal spanning tree of the L2L network.

The minimal requirement imposed on the digital library is to support at least the DublinCore Metadata. If two digital libraries describe the resources with semantics, like JeromeDL system, the communication between them is automatically upgraded to the semantic description level. It allows to use the search algorithm with semantics in the L2L communication.

5 Evaluation of the search algorithm with semantics

The aim of the search algorithm presented in the previous section is to reflect the readers' expectations and to reduces the time required to find the specified resources in JeromeDL. An evaluation of the search algorithm needs to cover the computable effectiveness measures and users' satisfactory level. The quality of retrieval features in JeromeDL depends on user oriented resource description (FOAFRealm-manage component) and bibliographic description (JeromeDL resource management, MarcOnt mediation service).

The semantic social collaborative filtering supported by FOAFRealm has been evaluated in [4]. It has been assumed that high quality information is collected by experts. The results of experiments revealed that each user can find an expert on particular subject within 7 hops in social network graph.

In order to measure the improvement of effectiveness of the semantic enabled search algorithm[5], the database of the prototype system has been filled with 100 resources. After a little time of browsing 50 queries have been processed with and without the semantic query expansion phase. To evaluate the gain in effectiveness produced by the semantic phase of the semantic searching process, tree metrics have been calculated: precision, recall and waste [11].

The results have shown that the semantic query expansion phase in the search algorithm improves the results by 60% compared to the search process without the semantic (user-oriented tailoring) phase.

[8] OCKHAM: http://www.ockham.org/

6 Future work

The evaluation of the JeromeDL search algorithm revealed that the results depend strongly on the semantic parts of resources' descriptions. That leads to the conclusion that better quality of the semantic description will result in higher effectiveness of the searching process.

Definition of evaluation experiment for the search algorithm. The JeromeDL search algorithm utilizes tree types of information: (1) implicit descriptions, including semantic description; (2) descriptions provided by readers: annotations, personal bookshelves, history of usage; (3) information about relations between readers.

To evaluate the whole search subsystem of JeromeDL, we propose a staged experiment, that would cover all aspects of usability. In each experiment performed the efficiency measures: precision, recall and waste[11] are computed.

The database of JeromeDL system is filled with a mass of resources and MARC21 and BibTEX descriptions translated to MarcOnt ontology Readers perform some browsing in the categories that are interesting to them. **Experiment 1**: Readers are querying the system two times: with and without the query expansion with semantics. With the knowledge on the database content of the digital library, learned during the browsing part, they calculate the metrics: precision, recall and waste of each query result. **Experiment 2**: Readers register to the JeromeDL system and continue browsing its content, annotating some resources and creating personal bookshelves. Later on, readers performs the queries once again, computes the metrics and compares them to the metrics obtained from Experiment 1. **Experiment 3**: Each reader indicates his friends registered in the JeromeDL system. Readers provides ACLs to the categories in their personal bookshelves and links categories created by their friends into their own personal bookshelves. Readers performs the queries for the last time and compares the results with the previous experiments.

Building network of federated libraries. The work started by the Library of Congress on MARC21-based web services allows to expect that these technology will also enable communication with ELP-based digital libraries. To simplify the use of the distributed environment of digital libraries the current work initiates connects the L2L network to e-Learning environments[6], on-line communities and P2P networks.

To overcome the problems that can arise in the P2P network of digital libraries (called L2L networks), semantic routing algorithms can be applied. Possibilities include HyperCuP[10] and categorization based multicasting. That would also improve scalability of the L2L network by limiting the required bandwidth in the nodes.

7 Conclusions

In this paper we presented JeromeDL, a digital library that deploys Semantic Web technology for user management and search. The FOAF vocabulary is used to gather information about user profile management, and semantic descriptions are utilized in the search procedure. JeromeDL is actively deployed in several installations and is continually enhanced with semantic features. JeromeDL is implemented in Java and available under an open-source license. Parties interested in setting up JeromeDL are invited to join our library P2P network.

References

1. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
2. J. Frew, M. Freeston, N. Freitas, L. L. Hill, G. Janee, K. Lovette, R. Nideffer, T. R. Smith, and Q. Zheng. The alexandria digital library architecture. In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, pages 61–73. Springer-Verlag, 1998.
3. S. R. Kruk. Foaf-realm - control your friends' access to the resource. In *FOAF Workshop proceedings*, http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/foaf_realm/, 2004.
4. S. R. Kruk and S. Decker. Semantic social collaborative filtering with foafrealm. In *submitted to ISWC*, 2005.
5. S. R. Kruk, S. Decker, and L. Zieborak. Jeromedl - a digital library on the semantic webgi-. In *submitted to ODBASE*, 2005.
6. S. R. Kruk, A. Kwoska, and L. Kwoska. Metadito - multimodal messanging platform for e-learning. In *International Workshop on Intelligent Media Technology for Communicative Intelligence*, pages 84–87. Polish-Japanese Institute of Information Technology, PJIIT - Publishing House, 2004.
7. C. Lagoze and J. Hunter. The abc ontology and model. *Journal of Digital Information*, 2(2), 11 2001.
8. M. Okraszewski and H. Krawczyk. Semantic web services in l2l. In T. Klopotek, Wierzchon, editor, *Intelligent Information Processing and Web Mining*, pages 349–357. Polish Academy of Science, Springer, May 2004. Proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004.
9. I. O'Murchu, J. G. Breslin, and S. Decker. Online social and business networking communities. In *Proceedings of the Workshop on the Application of Semantic Web Technologies to Web Communities*, Valencia, Spain, August 2004. 16th European Conference on Artificial Intelligence 2004 (ECAI 2004).
10. M. Schlosser, M. Sintek, S. Decker, and W. Nejdl. Ontology-based search and broadcast in hypercup. In *International Semantic Web Conference, Sardinia*, http://www-db.stanford.edu/ schloss/docs/HyperCuP-PosterAbstract-ISWC2002.pdf, 2002.
11. P. C. Weinstein and W. P. Birmingham. Creating ontological metadata for digital library content and services. *International Journal on Digital Libraries*, 2(1):20–37, October 1998. ISSN: 1432-5012 (Paper) 1432-1300 (Online).