

JeromeDL – a Semantic Digital Library

Sebastian Ryszard Kruk, Tomasz Woroniecki, Adam Gzella, and Maciej
Dąbrowski

Digital Enterprise Research Institute, NUI Galway, Ireland*
<firstname.lastname@deri.org

Abstract. JeromeDL is a Semantic Digital Library engine. It uses Semantic Web and Social Networking technologies to improve browsing and searching for resources. With JeromeDL's social and semantic services every library user can bookmark interesting books, articles, or other materials in semantically annotated directories. Users can allow others to see their bookmarks and annotations and share their knowledge within a social network. JeromeDL can also treat a single library resource as a blog post. Users can comment the content of the resource and reply to others' comments and this way create new knowledge.

All data stored in the library is available in RDF format for querying and processing by other applications. Also the result of every search or browse action is immediately available as a link to RDF. Innovative MBB (MultiBeeBrowse) component offers multifaceted navigation and provides SOA for integration with other applications. TagsTreeMaps allows to easily filter out resources using clustered tags and presented with treemaps layout. Users can also navigate through the presented search results using SIMILE Exhibit component.

The home page of the project is <http://www.jeromedl.org> and a demo can be found at <http://bleedingedge.jeromedl.org>.

1 Introduction

Typical digital libraries usually focus on categorizing and cataloguing resources. Information retrieval in such libraries relies primarily on text search engines and free browsing. This approach proved to be useful, however it suffers from ambiguity of natural language, neglecting the importance of metadata; it also does not engage users in the process of sharing knowledge. Simple searching still returns too many results which have to be filtered somehow. Pageranking algorithm helps with websites but cannot be easily applied to books or e-learning objects. On the other hand, having a look on a friend's bookshelf can give us much clearer view on what is worth reading in a particular domain than digging through a thousand books or websites published this month. The semantic digital library is an attempt to restore the collaborative approach to sharing knowledge.

* This material is based upon works supported by Enterprise Ireland under Grant No. ILP/05/203. Authors thank Stefan Decker and Bill McDaniel for all their help, and all members of the Corrib community for fruitful discussions on this project.

The semantic services help to interconnect systems and exchange data, while social services let people benefit from expertise of others. Together, they improve knowledge sharing in a digital library.

2 Generic Architecture of a Semantic Digital Library

We present a three-layered architecture of metadata management on top of a digital library system (see Fig. 1). Each layer enriches basic information gathered in a library with semantic annotations and provides additional capabilities to browsing and searching.

The bottom layer handles typical tasks required from a digital objects repository, that is, keeps track of physical representation of resources, their structure and provenance (see Sec. 3). With an extensive use of a structure ontology the bottom layer provides a service for a flexible and extendable electronic representation of objects; it is especially significant in expressing relations to other resources.

The middle layer lifts up legacy bibliographic descriptions to a semantic level; it uses an extensible ontology capable of representing information originally provided in any of popular existing formats like Dublin Core, MARC21 or BibTEX (see Sec. 4). Services provided by the middle layer concentrate around storing, delivering and managing documents' metadata.

Furthermore, the middle layer offers information retrieval and identity management services. All the services are supported by semantic technologies; for example, the natural language queries take advantage of a social network specified using FOAF ontology [9].

The top layer in the semantic digital library stack utilizes benefits from engaging community of users into annotating and filtering resources (see Sec. 5). In today's Internet the influence of user communities cannot be overestimated; collaborative efforts in information sharing and management proved to be the right way to go and led to the success of many of the Web 2.0 sites.

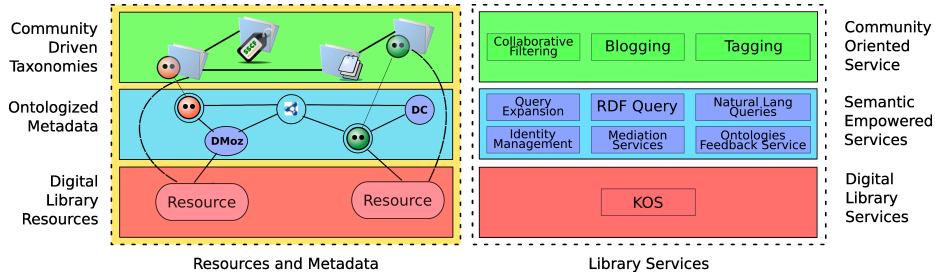


Fig. 1. Metadata and services architecture of a semantic digital library

3 Digital Library Services in JeromeDL

Knowledge Organization Systems There might be many definitions of a digital library system; the most accurate, however, will always adhere to knowledge organization systems (KOS). JeromeDL allows librarians to maintain and use the following controlled vocabularies: *authority files* - with a list of authors, editors and publishers; *classification taxonomies*, such as DMoz or DDC, for annotating resources with topics; *WordNet dictionary*, for specifying keywords. An AJAX interface delivered by the JOnto¹ project was used to support librarians in managing KOS and describing resources using controlled vocabularies.

Structure Ontology Modern digital library systems not only store bibliographic metadata; they also manage an electronic representation of the content itself. The structure of the content might, however, depend on the type of the resource, e.g.: a book can be decomposed into chapters, a multimedia presentation into media parts. We proposed [7] an ontology for defining the structure of resources in RDF; it defines concepts related only to the structure of the resource; this ontology provides a universal layer for metadata and content retrieval; it allows us to extend the structure description with new concepts, without violating the integrity of existing data.

4 Semantic Services in JeromeDL

Support for Legacy Information The key feature of every digital library system is making bibliographic resources accessible. A semantic description does not only provide bibliographic information; it also involves domain (topic) categorization of a resource from the WordNet dictionary. The uploading process in JeromeDL adopts research on folksonomies [12]; librarians who manage resources in JeromeDL are free to use any properties they find appropriate to annotate publications.

Bibliographic ontology There are many approaches to the development of library resource description formats. Heterogeneity becomes one of the most important problems in the digital library domain. There are many, more or less popular, legacy formats being used, e.g., MARC21, BibTeX or Dublin Core; different resources can be described by different metadata. We need a mediation standard such as MarcOnt Ontology [10] which defines concepts used in JeromeDL for bibliographic description.

User feedback - MarcOntX concept As mentioned in the previous paragraph - a librarian can use any properties they find appropriate to annotate a resource during the uploading process. Any new properties are automatically retrieved from known instances of JeromeDL with a MarcOntX agent. They are suggested as proposals for the community involved in the MarcOnt ontology development

¹ JOnto: <http://jonto.sf.net/>

process; the community decides whether it is necessary to incorporate these concepts into a new version of the MarcOnt ontology.

MarcOnt Mediation Services MarcOnt ontology also serves as a mediation standard for MarcOnt Mediation Services [4]. It allows users to translate the bibliographic description of the resource in JeromeDL between any of other formats aligned with the MarcOnt ontology.

Community Profile Management FOAFRealm [5], a distributed identity management system based on FOAF allows JeromeDL users to control their profile information; it also manages an authority list (authors, editors, publishers). During registration user constructs a FOAF profile or uploads an existing one. FOAFRealm also allows the system to manage user preferences. In distributed FOAFRealm, called D-FOAF [8], instances of the system are connected in a P2P network; this allows users to register only in one instance and sign-in to any other across the network.

FOAFRealm extends the FOAF vocabulary with the notion of friendship level properties; together with `foaf:knows` they are used to deliver access control based on social networks; for example only the good friends of the system administrator could add a new content to the library.

JeromeDL utilizes the social network's information (a FOAF digraph) as a base for social services (see Sec. 5).

Search and Browsing Services JeromeDL offers a number of search and retrieval services, which help to improve users' experience in digital library navigation.

Direct RDF Query Service Searching for resources in the network of digital libraries became a commodity; protocols like Z39.50 or OAI-PMH are used for the communication purposes. Semantic web service technologies are still on their way to becoming industrial standards; in the meantime, many services are being mashed-up together [1]; standards like SPARQL define means for these and other solutions based on semantic-enabled services [11]. JeromeDL delivers a direct RDF query service to be able to act as one of the mash-up services; it supports various query languages, such as RDQL, SeRQL and SRQL; the query results in RDF can be serialized to XML, N3, N-Triples, TURTLE, and JSON format, or presented in HTML. Each query is evaluated by the Sesame engine on a secure snapshot repository, which contains all information about resources and only public information from community profile management components: FOAFRealm (see Sec. 4) and SSCF (see Sec. 5).

Natural Language Query Templates The direct RDF query interface is very expressive; it is also fairly unusable for an average library user. Therefore we have investigated applying solutions that would allow users to query the semantic database of the library using natural language [3]. We decided to adopt a simple,

yet quite powerful, solution based on natural language query templates [9]. The templates were created from favorite questions that users of a certain JeromeDL instance were likely to ask; each template consist of a set of regular expressions matching various grammar and language forms for one RDF query template.

TagsTreeMaps (TTM) TagsTreeMaps² allows to represent flat set of tags as a tree of clustered categories, using treemaps algorithm. It allows to filter out tags based on name of the tags (and sub-tags) and frequency of using tags. Users can select tags and create summary or conjunctive filter on current set of results. TTM implements zooming paradigm, and users can easily move vertically on the tree of clustered tags.

TTM solves one of the fundamental problems of tagging space; when trying to browse through a large space of tagged information, it is very hard to get a hold on a large number of tags used. Currently used solutions like tag clouds present static approach to present and differentiate presented tags. Eventually, user has problem when trying to find appropriate tags to browse the information. Very often the tags cannot be filtered by any means and they are always represented as a flat structure.

TagsTreeMaps allows to perform zoom in/out actions (multi-level) to a selected group of tags, clusters, bundles. User can also choose between *union mode* – at least one of the tags in the selected view should be used in the information filtered, or *conjunctive mode* – all tags must be used, as opposed to current solutions where only conjunctive mode is delivered.

JeromeDL builds the list of tags from keywords and topics assigned to resources by librarians, and by the community of users (see Sec. 5). The list of providers can be easily extended to other systems.

MultiBeeBrowse (MBB) MultiBeeBrowse allows to browse unstructured meta-data represented as an RDF graph. System is build according to SOA (Service Oriented Architecture) paradigm, coupled with AJAX-based user interface. All the services build for MBB are RESTful³. An argument for REST, in the context of the MultiBeeBrowse service, is that GET action defines an idempotent request, i.e., subsequent calls of the same URL should return the same results. This can ensure that user will get the same results, each time given URL is called. In MBB it is vital for and handling history of results. Our goal was also to construct a meaningful URL representing single browse operations, as well as, whole chain of operations building up a browsing query. This would allow advanced users to quickly construct their queries, directly in the web browser address field.

Browsing services, deliver the primary functionality of the MultiBeeBrowse component. It consist of: access to resource, search services, filter service, similar service, related service, combination service (conjunction, sum, difference, binding, on two given sets of results). Meta-services allow to render results in one of RDF serializations, or in one of feed (RSS, Atom) formats.

² TagsTreeMaps: <http://sf.net/projects/tagstreemaps>

³ <http://wiki.s3b.corrib.org/MBB/SOA>

Based on these two goals: handling many paths of back and forth refinements, and access to a structured history of operations, we have identified 4 views of browsing context, which allow user to access effortlessly each of aforementioned features. (1) Basic browsing view (2) Structured history view allows users to view their current results in the context of previous and following (if any) operations. (3) HoneycombTM view presents users a comprehensive overview of their current browsing context (4) Life-long browsing history

JSON Serialization and Exhibit Navigation Service All search and browsing results in JeromeDL, are in fact an RDF graph; therefore, users can choose to serialize them in one of RDF formats. JeromeDL can also represent the resulting RDF graph in JSON format. This features has been especially delivered so that we could integrate Exhibit⁴ navigation component into JeromeDL. This component allows users to easily filter out, based on various facets, a set of presented resources. It can also render results on the timeline and google maps.

Support for JSON and Exhibit in JeromeDL brought also other benefits. Search and browsing results can be accessed by other (semantic) web services. JeromeDL can also integrate arbitrary RDF graph with resources handled by the system, a feature users by the aggregation service; hence, Exhibit component, the same way as TagsTreeMaps and MultiBeeBrowse, integrated into JeromeDL can also browse arbitrary RDF graph together with resources managed by the system.

5 Social Services in JeromeDL

A social network maintained by FOAFRealm allowed JeromeDL to develop in a new direction; focusing on community aspects and utilizing the benefits of social networking. A JeromeDL user is no longer only a reader of the information, she/he becomes a contributor of the content and a creator of the knowledge.

Social bookmarking Usually, when a user browses a digital library, some articles and materials seem to him more valuable than others. Common practice is to bookmark those resources. Recently collaborative bookmarking such as del.icio.us⁵ have become more and more popular. Users want to see the bookmarks of their friends, and use the knowledge collected by them. Such features are provided in JeromeDL with Social Semantic Collaborative Filtering (SSCF) [6].

The SSCF is based on two concepts: distributed collections and annotations of resources. Each user classifies only a small subset of knowledge, based on the level of expertise he/she has on the specific topic; this knowledge is later shared across the social network.

Users maintain their own collections and render them accessible to their friends. All resources are collected according to the user's point of view expressed

⁴ Exhibit: <http://simile.mit.edu/wiki/Exhibit>

⁵ <http://del.icio.us/>

by his/her folders categories. We can assume that some of the topics are better explored by some people. Each collection can be imported to a users personal bookmarks; users can take advantage of the knowledge and experience of their friends.

To facilitate knowledge sharing each folder with bookmarks can be annotated using controlled vocabularies, such as WordNet, Dmoz and DDC. Additionally, all directory has its rank value kept in SSCF, which shows which topics are especially interesting for the user.

JeromeDL as a blog JeromeDL exports users' comments, which are in the form of blog responses to a blog post (library resources), using SIOC metadata [2]; therefore, they can be easily integrated with other social semantic information sources. As a result, current readers can easily deliver new knowledge for future readers; furthermore, the knowledge created by the users along with the library resource could also be find and used outside the library world. For example, in the SSCF bookmarks interface which supports SIOC information, a user can browse the comments for each library resource.

In order to achieve the interoperability between JeromeDL and other community-based sites, our system needed to ensure compatibility between SIOC and SSCF/JeromeDL ontologies. We achieved that by creating user annotations using SIOC metadata and a delivery mediation mechanism for other SSCF/JeromeDL content.

6 Summary

With JeromeDL social and semantic services every library user can bookmark interesting books, articles or other materials in semantically annotated directories. Users can share their knowledge with others within a social network. We enriched the standard SSCF browser with an ability to bookmark and browse community based data. JeromeDL also has a feature which allows it to treat a single library resource as a blog post. With SIOC based annotations users can to comment the content of the resource and in this way create new knowledge. JeromeDL also provides various browsing, filtering and navigation solutions, such as TagsTreeMaps, MultiBeeBrowse and Exhibit.

JeromeDL has been installed in a number of locations⁶; the two most used, DERI Galway library⁷ and WBSS⁸ at Gdansk University of Technology, serve their community of users in everyday activities. DERI Galway library is used by researchers as a pre-print server to locate and share publications. WBSS maintains a set of scans of antique books and a number of books written by lecturers at GUT; the latter ones are used as learning material.

⁶ A list of JeromeDL instances: <http://wiki.jeromedl.org/Instances>

⁷ DERI Galway library: <http://library.deri.ie/>

⁸ WBSS library: <http://www.wbss.pg.gda.pl/>

For the purpose of the Semantic Web Challenge we have installed the beta of the upcoming version 2.1 (described in this article) at <http://bleedingedge.jeromedl.org/>, with the content from <http://library.derri.ie/>.

References

1. C. Bizer and T. Gauss. Rdf book mashup - serving rdf descriptions of your books, Nov. 2006.
2. J. Breslin, A. Harth, U. Bojars, and S. Decker. Towards semantically-interlinked online communities. In *Proceedings of the 2nd European Semantic Web Conference (ESWC '05), Heraklion, Greece*, volume 3532, pages 500–514, June 2005.
3. M. J. Cafarella and O. Etzioni. A search engine for natural language applications. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 442–452, New York, NY, USA, 2005. ACM Press.
4. S. Kruk, M. Synak, and K. Zimmermann. Marcont initiative - mediation services for digital libraries. In *ECDL*, 2005.
5. S. R. Kruk. FOAF-Realm - control your friends' access to resources. http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/foaf_realm/.
6. S. R. Kruk, S. Decker, A. Gzella, S. Grzonkowski, and B. McDaniel. Social semantic collaborative filtering for digital libraries. *Journal of Digital Information*, Special Issue on Personalization, 2006.
7. S. R. Kruk, S. Decker, and L. Zieborak. JeromeDL - Adding Semantic Web Technologies to Digital Libraries. In *Proceedings of DEXA'2005 Conference*, 2005.
8. S. R. Kruk, S. Grzonkowski, A. Gzella, T. Woroniecki, and H.-C. Choi. D-FOAF: Distributed Identity Management with Access Rights Delegation. In *Proceedings to ASWC'2006*, 2006.
9. S. R. Kruk, K. Samp, C. O'Nuallain, B. Davis, B. McDaniel, and S. Grzonkowski. Search interface based on natural language query templates. In *Proceedings of the poster session of IADIS International Conference WWW/Internet 2006*, 2006.
10. S. R. Kruk, M. Synak, and K. Zimmermann. MarcOnt - Integration Ontology for Bibliographic Description Formats. In *Proceedings of DC'2005*, 2005.
11. S. R. Kruk, K. Zimmermann, and B. Sapkota. Semantically enhanced search services in digital libraries. In *Telecommunications, 2006. AICT-ICIW '06. International Conference on Internet and Web Applications and Services*, 2006.
12. P. Mika. Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference*, pages 522–536, 2005.