CallimachusDL: Using Semantics to Enhance Search and Retrieval in a Digital Library

Juan Miguel Gómez-Berbís* , Ricardo Colomo-Palacios, and Ángel García-Crespo

Universidad Carlos III de Madrid, Computer Science Department
Av. Universidad 30, Leganés, 28911, Madrid, Spain
{juanmiguel.gomez,ricardo.colomo,angelgarcia}@uc3m.es

Abstract. Among the challenges of classifying, locating and accessing knowledge in Digital Libraries tackling with the huge amount of resources the Web provides, improving Digital Libraries by means of different strategies, particularly, using semantics remains a promising and interesting approach. In this paper, we present CallimachusDL, a semantics-based Digital Library which provides faceted search, enhanced access possibilities and a proof-of-concept implementation.

Keywords: Digital Library, Semantic Web, Ontologies, Faceted Search.

1 Introduction

Digital Libraries represent a new breed of software applications whose aim encompasses categorizing, classifying, archiving and providing access to the vast constellation of Web resources. Currently, Digital Libraries (DL for short) are facing a new paradigm shift coping with various challenges which include overcoming traditional browsing or keyword-based strategies. Fundamentally, DL infrastructure improvement attempts have been trying to increase the quality of information retrieval, from query expansion to the collaborative filtering or multi-faceted browsing [1]. However, current approaches are still not fulfilling expectations, leading the user in many cases to frustration.

On the other hand, semantic technologies are evolving to a more mature state in which ontologies [2], its backbone technology, provide a formal representation of a domain. The use of semantics in DL can outperform the current endeavors that require finding data spread out across the DL structure and dynamically drawing inferences, something continually hampered by their reliance on ad-hoc, task specific frameworks in present DL technologies.

In this paper, we present CallimachusDL, a semantics-based DL which uses semantic information gathering and browsing to enhance search and retrieval.

The remainder of the paper is organized as follows. In section 2, we present the related work on DL and the state of the art. In section 3, we discuss a number of

requirements and the benefits of tackling them with our semantically-enhanced approach. Section 4, we describe CallimachusDL in detail, its architecture and proof-of-concept implementation. Finally, section 5 concludes the paper providing a number of conclusions and section 6 summarizes our future work.

2 Digital Libraries: Been There, Done That

Digital libraries provide high quality and well-organized information. Many of the powerful characteristics of Digital Libraries rely on Metadata. Librarians describe the resources of catalogues and other collections trough metadata in order to facilitate efficiently the delivery of information. The use of metadata in its formats and functionalities has been an object of study in the past in the field of Digital Libraries: ej. XML [3] and RDF [4], [5]. The use of ontologies in the context of Digital Libraries could be interesting in order to incorporate new functionalities by describing the relationships between elements. The concept of ontology introduced by the Semantic Web is a promising path to extend Digital Library formalisms with the meaningful annotations [6]. Several authors have proposed ontologies for describing the relationships between all the elements which take part in a digital library scenario [7], [8] that goes beyond different standards of digital libraries description formats like MARC21, Dublin Core and BibTeX.

The new and promising digital libraries content management tool generation comes from the joint of the Semantic Web and the new social aspects of the so called Social Web. Here we find several initiatives such as the ambitious JeromeDL project [1] or DLibra [9]. Jerome DL uses MarcOnt Ontology [8] mediates with several legacy metadata standards (MARC21, BibTeX & Dublin Core) and offers a number of search and retrieval services based on Semantic technology.

Fundamentally, our approach is radically different to the ones detailed before since we propose semantic navigation, faceted search and browsing, metadata representation format and usability as the main building principles and cornerstones of the whole approach. Those features are detailed in the next section.

3 Using Semantic Information Gathering and Browsing to Enhance Search and Retrieval

Since its inception, Digital Libraries on the Web had to strive for classifying, locating and accessing resources. However, the advantage of the simplicity in Digital Libraries leads to its great drawback, the increasing number of information being stored without a clear structure. Actually, most current DL cannot be used as fully-fledged environments to create and search knowledge in an efficient way, since the information collected through these systems lays unused by computers, mainly due to the human language in which the resources are written. As further processing is needed, new formal approaches are used to make computers "understand" the Web content [10] or, more precisely, applying semantics.

The Semantic Technologies paradigm is based on this statement, where the traditional Web is enhanced with formal knowledge placed below the current information.

This is possible thanks to the extensibility of the Web with metadata and metadata processing, which allows computational reasoning and intelligent capabilities. In the following, we analyze the problems raised in the way to reach a semantically-enhanced DL environment, including technical and social factors:

- Metadata representation format: Metadata support for the actual information on DL resources must be explicitly declared. In some of the current social tools such as the emerging Web 2.0 applications like Flickr (http://www.flickr.com) or del.icio.us (http://del.icio.us/) apply the so called "folksonomies" to add meta-information in form of tags chosen by the user [11]. In this case, tags are different among different users, because they are chosen freely, so they cannot be fully exploited in a community. Besides, its storage has nothing to do with Semantic Web.
- Navigation. Ordinary DL base the relationship between pages in explicit hyperlinks. This links relate one page to another basically by user considerations. If the relation between DL resources were represented by means of semantics, the application would be able to provide mechanisms to semantically navigate between related resources with real meaning.
- Search. Given a set of resources, the basic type of querying in current DL is the keyword-based search. Structured requests for more advanced information retrieval are needed to make a DL a really useful knowledge repository. In addition to simple full-text searches, users would recover information by querying or selecting the semantic knowledge.
- Usability. Communities need a critical mass of users. Not only the number of users is crucial, but also their participation in the communities. Without his mass, the systems underlying communities will be abandoned by the users [12]. Semantic Web community has to grasp this principle and make it their own. For that purpose, applications enhanced with semantic functionalities have to be designed with maximum usability and minimum cognitive load for every user, including both Semantic Web experts and Internet users with no knowledge about semantics.

The gist of this work is providing an answer to these requirements. In CallimachusDL, we focus on the aforementioned requirements, solve them and propose an integrated solution that uses semantic information gathering and browsing to enhance search and retrieval. In the next section, our approach for CallimachusDL will be discussed.

4 CallimachusDL: Bringing the Library Mess into Order

4.1 The CallimachusDL Description

Given the aforementioned problems that traditional DL cope with, we explain here our approach, based on several design principles to avoid these drawbacks, and built as the kernel to develop a fully-fledged semantic working environment for the final users. These design principles are as follows:

- Metadata Representation Format: Bearing in mind that metadata processing requires a controlled and well-defined vocabulary, the Semantic Web saw in the ontologies the best mechanism to represent, share and reuse the knowledge behind. One of the most well-known definitions of an ontology is the one stated by Borst [13], extending Gruber's one [14], define it as a formal specification of a shared conceptualization. Since semantic knowledge must be represent in form of well-designed ontologies, it is time then to choose the models and languages in which this representation will be brought to life. For that, we recommend the selection of the different World Wide Web Consortium (W3C) proposed standards. Resource Description Framework (RDF, http://www.w3.org/TR/rdf-primer/) and RDF Schema (RDFS, http://www.w3.org/TR/rdf-schema/) can be perfectly suitable for defining the semantic information needed. Other languages such as the Web Ontology language (OWL) can also be suitable, but its further inference mechanism can be too much for the real necessities of the application.
- Multi-ontology approach for defining DL resources: Once ontologies representation has been defined, the scope of the used ontologies must be explicitly declared. Since DL resources are basically resources in the Web, they should be described this way first. For this, Dublin Core initiative (DC, http://dublincore.org/) fits perfectly as the main ontology for describing the whole wiki pages. Once identified, the DL resources must be described as far as its content is concerned. Therefore, a second ontology or several ontologies must be used for formalizing the real domain of the DL resources content.
- Semantic Navigation: As ordinary hyperlinks are not enough to show the related information in a DL, another approach is needed to offer the user all the information semantically similar to the one they are viewing. This is mainly due to the fact that the user interface must enable navigation to semantically related items [15]. For that, we propose semantic links, semalinks, which are ordinary hyperlinks in appearance but built upon semantic information. This semantic information, consisting both on the ontology concept which a certain part of the content is referring and its value, will lead to the user to pages with a semantic similar content as the semalink indicates. That is, if a set of words have been used to form a semalink, with a property x and a value, when mouse over this link, the nodes appearing will make reference to other pages with same property x and value, and as many more references as properties directly related with property x exists in the repositories, with same value.
- Usability: Authoring a semantic wiki must be made just as easy as authoring a traditional wiki. For that purpose, editing the semantic links must be done at the same time and in the same view as editing the rest of the page. Semantic annotations are the answer to fill this gap. Annotate a document means adding semantic data to these documents [16]. Users will be provided with semantic information to add; therefore, while editing a page, they will be able to annotate a word or a set of words with semantic data, just as easy as marking the selected words and associate them to a

property or vocabulary concept from the ontology domain. Usability is also reflected in the functionalities of browsing and searching seen in the previous subsections.

- Faceted Search: As keyword-based searches or other different syntactical queries are not an efficient retrieval mechanism, and providing that semantic information is underlying our system, a more advanced search is required. A facets-based search is the solution. With faceted metadata [17], the information space is partitioned using orthogonal conceptual dimensions of the data. These dimensions are called facets, and represent the characteristics of the information elements. These facets are used then to select or filter the relevant elements in a certain information space, leading users to the exact information needed. These facets are the properties defined in the domain ontologies.

Once we have described CallimachusDL and its main features, we will describe the CallimachusDL architecture in the next section.

4.2 The CallimachusDL Description

The CallimachusDL architecture is heavily based on the SWAN architecture [18]. Having into account these apparently different levels of knowledge (ontologies, resources and semantic information), we explicitly divide this knowledge into three layers:

- Resource layer: This layer stores the DL resources and all the objects related to those resources.
- Domain layer: This layer deals with the ontologies used for formalized the semantic information for both the DL pages (DC vocabulary) and contents (RDFS vocabulary or vocabularies).
- Application layer: This layer is supported on top of the previous one and will be built with the domain ontologies the CallimachusDL system requires, and applied to the resources in the first layer.

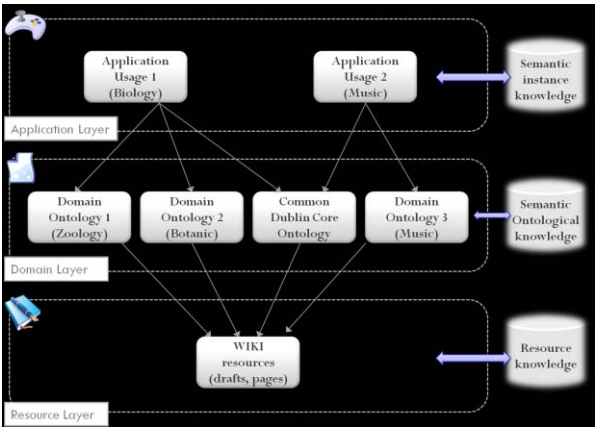

Fig. 1. The SWAN architecture as a basis for the CallimachusDL architecture

Keeping these knowledge layers conceptually separated, implementation will guarantee the flexibility and reusability of the CallimachusDL application for every sort of domain. [Fig. 1] shows the framework of this approach, along with named examples for better understanding.

Domain layer keeps the different domain ontologies that can be used. The Application layer will use one or more domain ontologies depending on the sort of topics the DL application is going to deal with. The Dublin Core Ontology will be always used to represent the basic metadata concepts of every resource.

4.3 Using CallimachusDL

The CallimachusDL implementation is based on the SWAN architecture successfully SWAN architecture has been successfully deployed on CoolWikiNews, a Semantically-enhanced Wiki devoted to online news publishing [18]. CallimachusDL implements the MVC pattern by means of Ruby on Rails (RoR, http://www.rubyonrails.org) [19], a MVC-based framework which eases the task of building this architectural pattern. The common ontology used for describing the resources is Dublin Core. Its terms allow defining the metadata related to the whole page. The MarcOnt ontology is used for the annotation of more complex data. Both ontologies are developed with RDF Schema, and serialized in N-Triple syntax (http://www.w3.org/TR/rdf-testcases/#ntriples). The DL pages are presented to the user in XHTML 1.0 syntax (http://www.w3.org/TR/xhtml1/), and visual graphics for navigation are made with JavaScript libraries such as CoolTip (http://www.acooltip.com). Persistence repositories are MySQL server for resources information and SQLite- based RDFLite for semantic information. Finally, CoolWikNews uses ActiveRDF [20], a library for abstracting the queries for RDFLite within the implementation in RoR. Finally, we will show how CallimachusDL is used with a motivating scenario. Recently, a new breed of user generated content aware technologies which have been encompassed by the “Web 2.0” buzzword umbrella have turned up to provide a huge amount of metadata and information about the user as a particular entity. Web 2.0 technologies as outlined in [11] are exemplified by blogs, namely easy to update websites about a particular subject where entries are written in chronological order, picture-sharing environments such as Flickr or Photobucket, social bookmarking sites such as Del.icio.us, video-sharing such as YouTube or music preferences such as Last FM. A number of common features of Web 2.0 applications have been identified in [21]. This Web 2.0 user generated content is a perfect battlefield for our example.

For example, a user called John Smith has uploaded a number of videos in YouTube about his staying in Norway. Particularly, those videos are about the Norwegian fjords so he tags them with the “fjord” and “Norway” tags. However, tags are freely chosen keywords describing a particular resource. They offer a simple way of retrieving content but they are subjective conceptualizations, being potentially aggregated to a flat bottom-up categorization or folksonomy. In [22], folksonomies have been claimed to be an interesting emergent attempt for information retrieval but serve different purposes to ontologies, the latter are attempts to more carefully define parts of the data world and to allow mappings and interactions between data held in different formats. In this scenario folskonomies had been used for creating semantic metadata

[23] or as a support to learning [24]. Hence, ontologies are defined through a careful, explicit process that attempts to remove ambiguity, whereas the definition of a tag is a loose and implicit process where ambiguity might well remain. Finally, the inferential process applied to ontologies is logic based and uses operations such as join. The inferential process used on tags is statistical in nature and employs techniques such as clustering.

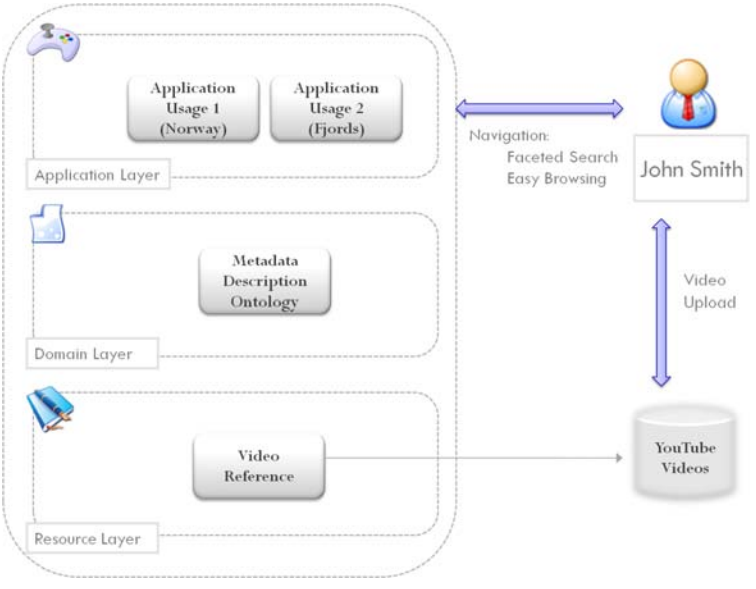

Fig. 2. John Smith videos in CallimachusDL

If John Smith chooses any traditional DL, he will face a number of problems, as we have shown in section 3. First of all, there is no metadata description, no chance of faceted browsing and problems to locate and retrieve its YouTube videos. Nevertheless, CallimachusDL offers a completely different situation. Using the three-layered architecture described in section 4.2, the Resource Layer would store references to the videos in YouTube or the videos as such. In the Domain Layer, there are metadata formally describing the videos by means of ontologies, mostly DC and the MarcOnt ontology. Finally, the Application Layer will use domain ontologies (for example, those referred to Norway and Fjords).

Finally, Faceted Search and Browsing would make very easy the life of John Smith when retrieving his videos, since he can navigate through the categories and also see related videos thanks to the semalinks, as explained in section 3.

5 Conclusions and Future Work BibTeX Entries

Callimachus (c.305-c.240 B.C.) was an ancient Greek poet, librarian, and scholar, famous representative of the Alexandrian school of poetry. Following the works of

Zenodotus of Ephesus, Alexandria Library first library director that began an inventory of the scrolls acquired by the Ptolemies, Callimachus created for the first time a subject catalog in 120,000 scrolls of the Library's holdings, called the Pinakes or Tables [25]. Following the Callimachus efforts, the man that improved subject search in Alexandria, we have presented a novel approach to improve browsing and searching in DL by adding semantics to the definition of resources. In a larger context, the problem of DL scaling may be multiplied by thousands of data structures located in hundreds of incompatible databases and message formats.

Hence, our future work will consist of evaluating our implementation and approach more carefully, validating CallimachusDL with a number of quality-aware case studies and using big-sized DL resources where pooling out of results can determine more accurately if the effectiveness of the breakthroughs of our approach detailed in section 3 take place. In a more general view, future work should further integrate social networks full potential into Digital Libraries. The unlimited potential of the Web 2.0 is an open field for technology investigators around the globe, and it is also a great opportunity for Digital Libraries researchers to put together social features and limitless content into a single package.

References

1. Kruk, S.R., Decker, S.: JeromeDL - the Semantic Digital Library. In: Proceedings Semantic Technology Conference 2007, San José, California (2007)
2. Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Berlin (2002)
3. Kim, H., Choi, C.: XML how it will be applied to digital library systems. The Electronic Library 18(3), 183–189 (2000)
4. Xing, W., Dikaiakos, M.D., Yang, H., Sphyris, A.S., Eftichidis, G.: Building a distributed digital library for natural disasters metadata with grid services and RDF. Library Management 26(4/5), 230–245 (2005)
5. Han, Y.: ARDF-based digital library system. Library Hi Tech. 24(2), 234–240 (2006)
6. Kruk, S.R., Decker, S., Zieborak, L.: JeromeDL - Adding Semantic Web Technologies to Digital Libraries. In: Proceedings of the 16 th International conference on database and expert systems applications, Copenhagen, Denmark (2005)
7. Ferrán, N., Mor, E., Minguillón, J.: Towards personalization in digital libraries through ontologies. Library Management 26(4/5), 206–217 (2005)
8. Kruk, S.R., Synak, M., Zimmermann, K.: MarcOnt - Integration Ontology for Bibliographic Description Formats. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, Madrid (2005)
9. Mazurek, C., Werla, M.: Distributed services architecture in dLibra digital library framework. In: 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems (2005)
10. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American 284(5), 34–44 (2001)
11. O'Reilly, T.: What is web 2.0? O'Reilly Network Retrieved (July 10, 2008) (2005), `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`
12. Ignacimuthu, S.: Basic bioinformatics. Alpha Science International, Harrow (2004)
13. Borst, W. N.: Construction of engineering ontologies for knowledge sharing and reuse. Doctoral Dissertation, University of Twente (1997)

14. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition 5(2), 199–220 (1992)
15. Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R.: The perfect search engine is not enough: A study of orienteering behaviour in directed search. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, Vienna, Austria, pp. 415–422 (2004)
16. McEnery, T., Wilson, A.: Corpus linguistics. Edinburgh University Press, Edinburgh (2001)
17. Ranganathan, S.R.: Elements of library classification. Asia Publishing House, New York (1962)
18. Fuentes-Lorenzo, D., Gómez-Berbís, J.M., García-Crespo, Á.: CoolWikNews: More than meets the eye in the XXI century journalism. In: Rech, J., Decker, B., Ras, E. (eds.) Emerging technologies for semantic work environments: Techniques, methods, and applications. Idea Group, Germany (2007)
19. Thomas, D., Heinemeier-Hansson, D., Breedt, L.: Agile web development with rails: A pragmatic guide. The Pragmatic Bookshelf, Raleigh, NorthCarolina (2005)
20. Oren, E., Delbru, R., Decker, S.: Extending faceted navigation for RDF data. In: Proceedings of the International Semantic Web Conference. LNCS, pp. 559–572. Springer, Berlin (2006)
21. Ankolekar, A., Krötzsch, M., Tran, T., Vrandecic, D.: The two cultures: mashing up web 2.0 and the semantic web. In: Proceedings of the 16th international Conference on World Wide Web, pp. 825–834. ACM Press, New York (2007)
22. Shadbolt, N., Hall, W., Berners-Lee, T.: The Semantic Web Revisited. IEEE Intelligent Systems 21(3), 96–101 (2006)
23. Al-Khalifa, H.S., Davis, H.C.: Exploring the value of folksonomies for creating semantic metadata. International Journal on Semantic Web and Information Systems 3(1), 13–39 (2007)
24. Lux, M., Dosinger, G.: From folksonomies to ontologies: employing wisdom of the crowds to serve learning purposes. International Journal of Knowledge and Learning 3(4/5), 515–528 (2007)
25. Bevan, E.: The House of Ptolemy. Argonaut Inc., Chicago (1968)