Toward Semantic Digital Libraries: Exploiting Semantic Web in Cultural Heritage

Andrea Baruzzo, Paolo Casoto, Prasad Challapalli, Antonina Dattolo, Nirmala Pudota, Carlo Tasso
Department of Mathematics and Computer Science - University of Udine, Italy

Abstract

Developing and maintaining a digital library requires substantial investments which are not simply a matter of technological decisions, but include also organizational aspects (which user roles are involved in content production, which workflows are needed, etc.). Moreover, starting a digital library initiative requires to tackle several issues such as the introduction of new user roles, workflows, and types of contents. These issues are often handled by approaches based on a physical perspective which treats the stored information either in terms of data formats or physical space needed to archive them. All these perspectives ignore quite completely the semantic aspects of the digital contents. In this paper we address such semantic perspective. More specifically, we propose a service oriented architecture which explicitly includes a semantic layer that provides primitive services to the applications built on the top of the digital library. As part of this layer, a specific component is described: the PIRATES framework. This module assists final users to complete several tasks concerning the retrieval of the most relevant content with respect to a description of their information needs (a search query, a user profile, etc.). Techniques of user modeling, adaptive personalization, and knowledge representation are exploited to build the PIRATES services in order to fill the gap existing between traditional and semantic digital libraries.

I. INTRODUCTION

The improvements in digitalization lead, in the last decades, to a huge evolution in the way cultural digital libraries and archives are conceived, designed, and used. Both the transition of library materials from traditional to digital formats, and the large (and continuously growing) availability of digital content, pose new challenges. More sophisticated software tools are needed to face the expectation of users, which are often exacerbated by the classical information overload problem. *Searching* everything everywhere is becoming an habit also in digital libraries, but *finding* exactly what it is needed remains a very hard job [Celino et al., 2006]. Data interoperability and sharing is another issue which must be faced when developing tools concerning digitalized cultural heritage: contents and archives should often be shared across different platforms and applications, usually by means of a Web-based infrastructure. Moreover, according to the growth of the Web 2.0 philosophy, new ways to access such contents should be provided to users: they could add their own contributions to the collections, share such contributions with other users and improve the effectiveness of information access.

These trends in design and development of digital libraries have not been fulfilled as a whole; *preservation* of multimedia contents and data is still addressed mainly by means of technological factors, e.g. reliable storage mechanisms able to guarantee long-term accessibility of digital supports [Barkstrom et al., 2002]. *Evolution*, at the same time, is typically considered a matter of scalability of the *physical* system, concerning stored information either in terms of data formats or physical space needed to archive them. These perspectives, as claimed by Ross in [Ross, 2006], ignore quite completely the *semantic aspects* of the preserved objects.

In this paper we embrace the "semantic" perspective of digital libraries, introducing a general approach to digitalized cultural heritage exploitation (preservation, evolution, classification, and access) not restricted only to the physical preservation of the contents available in a digital library. Our approach is focused on delivering a platform specialized in management of cultural heritage collections which:

1) promote the (logical) independence of data from their physical representations;
2) promote data interoperability at both archive and platform level;
3) integrate a set of primitives aimed at providing support to evolution at different levels of abstraction (archived contents, user requirements, technological infrastructures, user roles and workflows);
4) integrate a set of tools aimed at supporting users during information access, classification, and retrieval;
5) integrate a set of services aimed at supporting the semantic digital library vision [Kruk and McDaniel, 2008].

In particular, we aim at developing a platform which provides to its users an environment capable of dealing with information retrieval tasks where it is not important the presence of the "exact word", but of the *intended meaning*. Our proposal involves the creation of a system that will be able to provide accurate search results exploiting several tools coming from automatic categorization algorithms, information filtering and retrieval techniques, personalization, adaptation, and Web 2.0/Semantic Web features. More specifically, we are interested, in designing and developing a digital platform able to maintain the *semantic*

The authors acknowledge the financial support of the Italian Ministry of Education, University and Research (MIUR) within the FIRB project number RBIN04M8S8.

meaning of each digital object and its content, maintain its provenance and authenticity, and retain its interrelatedness, as suggested by [Ross, 2007]. In this way, we aim at providing the "semantic layer" on the top of the digital archives, giving to digital libraries a semantic connotation.

This work is based on a three-years experimentation with the EU-India E-Dvara project[1]: a digital platform devoted to Indian and Italian cultural heritage, as described in Section III. E-Dvara represents our current development and experimentation in the area of digital libraries. In previous work, we presented the overall project goals [Challapalli et al., 2006], a conceptual model to handle evolution issues in digital libraries [Baruzzo et al., 2009a], and the technical details concerning the E-Dvara software architecture [Baruzzo and Casoto, 2008], [Baruzzo et al., 2008]. The main and new contributions of this work exploit both automatic and manual tagging as a way of implementing the exploitation process and improve the effectiveness of information access. Such features are provided by integrating in E-Dvara the PIRATES framework [Baruzzo et al., 2009b], an automatic tagging environment. Other approaches to information access and visualization could be plugged in onto the proposed architecture, like a Virtual 3D Museum, without affecting the way contents are stored.

This paper is organized as follows: Section II describes the state of the art in the field of semantic digital libraries; Section III introduces the E-Dvara project; Section IV summarizes the E-Dvara service-oriented architecture. Then, in Section V we illustrate our approach to semantic digital libraries, whereas in Section VI we discuss our strategy of exploiting Semantic Web tools (e.g. ontologies and tagging) to annotate, classify, retrieve, and recommend contents in cultural heritage digital libraries. Finally, Section VII outlines both conclusions and future work.

II. RELATED WORK

In the last few years several research projects have been proposed for effective cultural heritage content organization, preservation, and integration [Bekaert et al., 2005], [Lutzenkirchen, 2002], [Candela L. and Pagano, 2007]. Storage of XML-based documents has been proposed in Greenstone [Bainbridge et al., 2001], [Witten et al., 2000], a digital library designed to provide librarians with the ability to create and publish heterogeneous collections of digital contents on the Web like text, images, videos and e-books. Each content in Greenstone can be described using *metadata* compliant with the Dublin Core[2] standard.

D-Space [Tansley et al., 2003] is a digital library aimed at providing long-term preservation of heterogeneous contents, by improving some of the limitations affecting Greenstone. Authors usually submit their documents to the system, and define metadata for them; for such reasons D-Space is also referred as an *author oriented* digital library. D-Space introduces also a multi-roles approach to content publishing, identifying the following actors: *authors* and *organizations*, which provide the contents, *librarians*, which perform content validation, and *users*, which are interested in content retrieval. Content-based workflows can be customized in order to cope with the needs of specific organizations and to delegate different tasks to different stakeholders.

In order to provide a flexible and reusable solution to data preservation and organization, the Fedora [Lagoze et al., 2005] project explored a service-oriented approach to data interoperability in digital libraries, by designing and developing a distributed architecture for contents publishing, aggregation, and retrieval. Composite information is obtained by aggregating physical contents, viewed as bit-streams, located worldwide into the Fedora repositories. Fedora allows content editors and archivists to define semantic connections between archived contents, treated as set of physical contents.

Other works related with content preservation in digital libraries are described in [Bekaert et al., 2005], [Lutzenkirchen, 2002]; the aDORe project, in particular, adopts the MPEG-21 DID content representation model in order to provide preservation and retrieval of heterogeneous multimedia contents.

The above mentioned systems are centred on contents, defined as *binary resources* enriched by metadata devoted to preservation, storage and retrieval purposes, but not intended for data structuring. Preservation and evolution of a data model in those approaches is implemented as a low-level mechanism, where data is processed as bit-streams instead of as instances of well-defined structures (i.e. XML Schema).

Several research projects, on the other hand, have been focused on improving the effectiveness of digital libraries in cultural heritage by moving towards a deeper semantic representation of the stored data, integrating ontologies and tools devoted to content annotation [Woroniecki et al., 2007].

CultureSampo[3] is a platform aimed at combining and accessing heterogeneous archives of cultural heritage related contents. Each metadata schema used to represent data has been mapped onto a shared ontology, the ONKI ontology, in order to provide semantic interoperability between contents. This semantic enrichment leads to new approaches to information access: CultureSampo introduces a perspective-based access to contents, where each perspective is represented by a subset of semantic features of the stored contents, such as temporal of geographical information.

CultureSampo provides a set of functionalities required to content publication, annotation and retrieval; content retrieval is exploited by means of both relation-based and semantic features-based approaches. Collaborative content generation, according

[1] http://edvara.uniud.it/india
[2] See for more details: http://dublincore.org/
[3] http://www.kulttuurisampo.fi/

to the Web 2.0 flavour, has been introduced into the CultureSampo infrastructure, in order to improve the amount of semantic information added to the contents; such tasks have also been partly automatized by introducing domain independent annotation agents, based on common thesauri and ontologies.

Interoperability of cultural heritage datasets and schemas between different platforms available on the Web has also been exploited by the *AMA: Archive Mapper for Archeology* project [Eide et al., 2008], as a part of *EPOCH*[4]. The tools developed during the *AMA* project are aimed at providing semi-automated mapping of cultural heritage custom data to the CIDOC-CRM, a formal ontology devoted to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information.

CCHO: Cantabria's Cultural Heritage Ontology [Hernandez et al., 2008], is also aimed at effectively integrating cultural heritage data in the region of Cantabria. Contents have been properly annotated by using the CCHO and can be browsed by a semantic-based search engine according to several perspectives, as in CultureSampo, like geographical maps, historic event timelines, and semantic relations between items.

In contrast to the previously described projects, based on a wide and formally defined ontology such as the CIDOC-CRM, the *OCHRE: Online Cultural Heritage Research Environment* project[5] adopts an approach based on a lightweight, extendable and general ontology called the *Core Ontology*. This ontology covers the domain of cultural heritage by means of a small set of highly general concepts and relationships, in order to grant an higher level of abstraction. The OCHRE's ontology can be extended and refined for each different project, according to the amount of specialized semantic information required to characterize a given collection.

Digital libraries specialized on cultural heritage management have been further improved by integrating social practices, like social and collaborative tagging, arising from the Web 2.0 experience. Using tags and annotations, provided both manually by the users or in a semi-automatic way, contents could be semantically enriched, in order to improve the effectiveness of both navigation and retrieval tasks.

In [van der Sluijs and Houben, 2008] an example of the effectiveness of integration between the Web 2.0 approaches and a cultural heritage devoted DL is exploited: the *CHI* system, designed and developed by the *RHCe* (Regional Historic Centre Eindhoven). CHI is devoted to storage and access of photo and video archives; to each of these archives a specific set of metadata has been assigned. Users can search, browse and visualize the collections hosted by the RHCe by accessing them according to different dimensions, each one identified by a specific set of metadata, as in CultureSampo. However metadata could refer both to a specific domain ontology (e.g. OWL time ontology [6] used to represent the temporal dimension) or to a user defined set of keywords (tags) assigned to a specific resource by the users, in a collaborative way.

Annotations regarding a specific content could also be harvested and collected from the network, looking at metadata used by different platforms and users to describe the same contents (e.g. the metadata assigned to the same painting into two different collections, hosted by different digital libraries devoted to cultural heritage).

The *HarvANA: Harvesting and Aggregating Networked Annotations* [Hunter et al., 2008] system uses a RDF model to represent tags/annotations and OAI-PMH[7] to harvest the annotations/tags of a specific content (e.g. a book characterized by a specific ISBN code) from a network of heterogeneous digital libraries.

III. THE E-DVARA PROJECT

E-Dvara is a project focused on the development of a new platform for storage of digital contents [Challapalli et al., 2006]. Since its inception, it has been explicitly designed to overcome some limitations that characterize the process of building a digital content. In particular, E-Dvara was initially meant to:

1) reduce the effort required by the archivist to define the data structure used to represent data into the archives;
2) provide to archivists with no expertise in data management a set of wizards devoted to data schemata creation in a completely automatic and transparent way (with respect to the physical database);
3) allow content providers to easily share their archives on the Web by means of a build-in Web interface or with several other applications, allowing archivists and system administrators to define the way data should be rendered to final users.
4) allow archivists to provide for each archive of digital contents a specific visualization template and a set of search forms.

More recently, we started the introducing in the E-Dvara platform of a set of semantic services aimed at assisting final users to select from the archives the most appropriate content with respect to their current information needs [Baruzzo et al., 2009a].

In order to cope with all these requirements, the E-Dvara platform has been designed to adopt a layered modular architecture, as illustrated in Figure 1. At the bottom of the architecture, digital archives are stored in XML databases by a core layer devoted to data storage and persistence (*Data layer*). The access to these archives can be provided directly by the Data layer services, or by a *Semantic layer* which exploits a Knowledge Base in order to find (suggest) the most relevant content fulfilling the user needs (e.g. a user query submitted from a search engine). The *Integration layer* forms the "architectural glue" that

[4]EPOCH is a network of about a hundred European institutions collaboratively producing applications involving digital versions of Cultural Heritage.

[5]http://ochre.lib.uchicago.edu/index.htm

[6]http://www.w3.org/TR/owl-time/

[7]OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting provides an application-independent interoperability framework based on metadata harvesting.

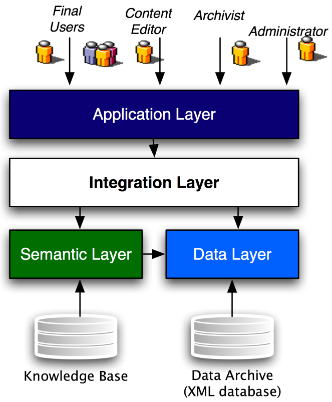

Figure 1: The high-level E-Dvara software architecture.

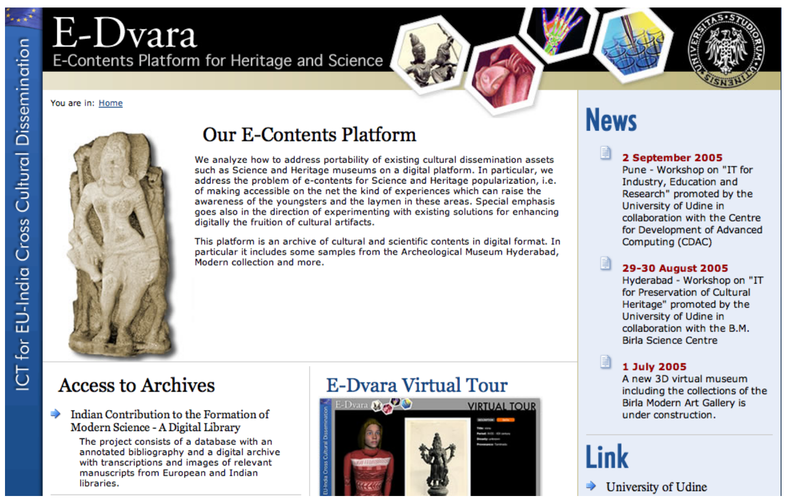

Figure 2: The home page of the first prototype of E-Dvara platform.

brings the digital library beyond the scope of a single application, unifying the interfaces of different subsystems into the same interoperable environment. The standard set of Web service technologies (XML, SOAP, WSDL) provides the means to describe, locate, and invoke a Web Service, simplifying the integration of new applications in the E-Dvara platform. The Integration layer uses a component called Enterprise Service Bus (see Figure 3) which implements the orchestration of different application-specific services in order to integrate these applications in the digital library. Finally, at the top of the architecture, the *Application layer* hosts the programs used by the digital library users (archivists, content editors, final users, etc.) in order to edit, publish, and accessing the contents stored in the archives.

Figure 2 illustrates the homepage of the E-Dvara project, developed in 2005. During the last three years, a first prototype has been largely tested by expert users involved into professional content publishing for cultural heritage. From this experimentation, we have gathered several evolution issues, weaknesses, and mistakes which led us to rethink the entire project. Currently, a second prototype is under development which introduces several heterogeneous services more oriented toward realizing the semantic digital library vision introduced in Section I.

IV. A SERVICE-ORIENTED ARCHITECTURE FOR INTELLIGENT INFORMATION ACCESS IN DIGITAL LIBRARIES

The second prototype of E-Dvara is based on a service-oriented architecture presented in [Baruzzo and Casoto, 2008], [Baruzzo et al., 2008]. Here we concentrate on describing the Semantic layer which is the new contribution of this paper. As illustrated in Figure 3, this layer exposes its services to the user applications through the Enterprise Service Bus located in the Integration layer. Two main components characterizes the Semantic layer:

- *PIRATES Framework*, which communicates with a Knowledge Base in order to retrieve or suggest potentially relevant information from the archives. This framework provides primitive services for automatically classify, annotate, recommend a specific content using techniques based on natural language processing. PIRATES is composed of three components, a *Cognitive Filtering Tools* module, an *Automatic Tagger*, and a *Knowledge Base Builder*, which are described in Section V-B.

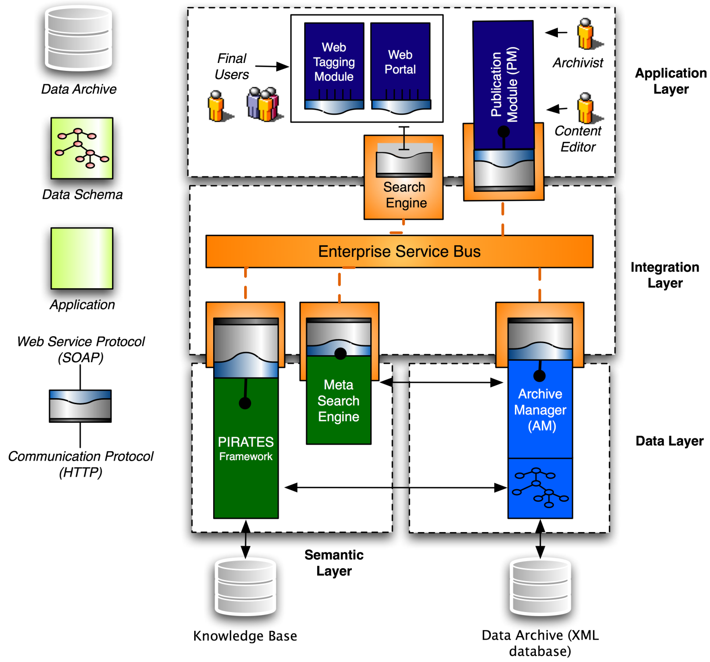

Figure 3: E-Dvara 2 Service Oriented Architecture.

- *Meta Search Engine*, which exploits the document annotations provided by PIRATES in order to recommend similar contents with respect to those retrieved by a traditional search engine fulfilling user queries. This module can also be used for refining a user query which has not provided enough results (query reformulation).

The presence of the Semantic layer is aimed at improving the information access mechanism of the E-Dvara digital library, empowering its environment by semantic services.

V. ACCESSING DIGITAL LIBRARY ARCHIVES BY MEANS OF "SEMANTIC SERVICES"

Semantically enabled technologies are expected to bring a number of benefits to the users of digital libraries such as helping people find relevant information more efficiently, giving better access to that information, and aiding the sharing of knowledge within the user community. Starting from these motivations, in this section we outline an approach of adding semantics to archives using tags suggested by an automatic system (the PIRATES framework) which is based on information extraction techniques. Before to introduce PIRATES, we describe the ways an archivist can exploit tagging services to annotate a digital content. We discuss also some notable limitations inherent to the use of manual tags which lead us to propose an automatic approach to tagging. Finally, in Section VI we present an example illustrating different usage scenarios of PIRATES in the E-Dvara digital library.

A. Adding Semantics to Digital Archives: the Tagging Approach

Tagging is a textual annotation technique based on *meta-data information* (i.e. tags). A *tag* is a keyword users use to annotate a content, in order to organize knowledge, describe it, correlate it with other contents, or simply to easily retrieve it in future searches. The tagging activity may be *manual* if is provided by a human user, or *automatic* if is it generated by a dedicated software. Archivists can employ tags differently because they can be guided by different tasks. Typically, tagging is used with the explicit intent of:

1) *classifying a content* by means of a corpus of concepts which are familiar to the archivist (e.g. taxonomies, thesauri, or any bag of keywords representing meaningful categories for him/her);
2) *summarizing a resource content* by means of a short list of keywords representing the user-generated content description;
3) *express a polarity judgment* about a content by means of proper adjectives provided as tags (e.g. *"sad"*, *"wonderful"*);
4) *correlating tagged resources with people and their skills* such as the level of expertise, the reputation, or the importance of a person mentioned in the resource content (e.g. *"guru"*, *"geek"*, *"vip"*, *"bill-gates"*, etc.);
5) *create dichotomic classification criteria* in order to describe resources as belonging or not to a particular category (e.g. *"clinical"/"not-clinical"*, *"statistical"/"not-statistical'*, *"accepted"/"rejected"*, and so on);
6) *providing a temporal information* to a resource, for example annotating the date of an event correlated to that resource.

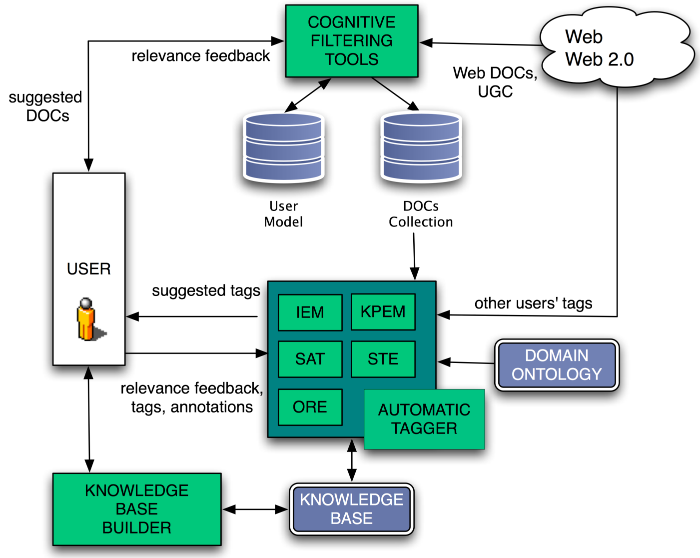

Figure 4: Pirates main modules.

To some extent, all these forms of tagging express a classification intent targeted to establish effective schemata for organizing the knowledge and facilitating content retrieval.

Tagging allows users to freely determine suitable labels for their resources without relying on any predetermined vocabulary or hierarchy [Mathes, 2004]. Moreover, tags can be very effective for serendipitous browsing a digital archive of documents (or bookmarks) in order to find relevant information. Hence people tag the content with their own vocabulary and ultimately their mental models in order to facilitate the process of recall. Besides with these potential benefits, manual tags suffer with some of notable limitations [Dattolo et al., 2009]:

- *Ambiguity*: with an uncontrolled vocabulary, many tags can be ambiguous. Indeed in tags we can find the same ambiguity that we find in natural language (e.g., homonymy, polysemy, synonymy, spelling mistakes, disambiguation, words which have more than one common spelling or morphology etc.).
- *Undistinguished concerns*: social tagging systems does not enforce, or even propose, a schema for distinguishing the purpose of a meta-data value. Tags might be, indifferently, subject descriptors, genres, self-reminders; tangential remarks (such as colors or years, especially for non-textual information such as pictures); or proper names.
- *Independence of terms*: social tagging does not provide relations to connect and relate different terms: each tag is independent of the others, and no inference is possible. In other words, the structure of a tag system is “flat”.
- *Effort*: systematically (and consistently) tagging Web resources is tedious, error prone, and rather wearying.

In order to alleviate these limitations, we propose an automated approach which assists the user when (s)he tags a Web resource. A software system analyzes the textual document and provides new tag suggestions/recommendations by exploiting information extraction tools [Cunningham, 2002] and ontologies. Using this approach, we try to achieve two different goals:

- *use a controlled, ontology-based vocabulary*, not necessarily present in the original Web resource, in order to classify it as result of the automatic tagging process; our vocabulary is a structured form of knowledge representation (the ontology) and provides entities (classes), instances, and relations (is-a links between entities).
- *reduce the manual effort* required to tag a Web resource.

B. The Pirates Framework: Merging Cognitive Filtering and Semantic Web

This section presents the PIRATES framework: a Personalized Intelligent Recommender and Annotator TEStbed for text-based content retrieval and classification. Using an integrated set of tools, this framework lets the users experiment, customize, and personalize the way they retrieve, filter, and organize the large amount of information available on the Web. Furthermore, the PIRATES framework undertakes a novel approach that automatizes typical manual tasks such as content annotation and tagging, by means of personalized tags recommendations and other forms of textual annotations (e.g. key-phrases). The PIRATES architecture, shown in Figure 4, is formed by three major components:

- The *Cognitive Filtering Tools* module implements the IFT (Information Filtering Tool) module. The IFT algorithm [Tasso and Asnicar, 1997] is used to build representations of user interests (*IFT user models*), to provide mechanisms of relevance feedback, and to classify the textual content of a document belonging to an incoming stream of documents.

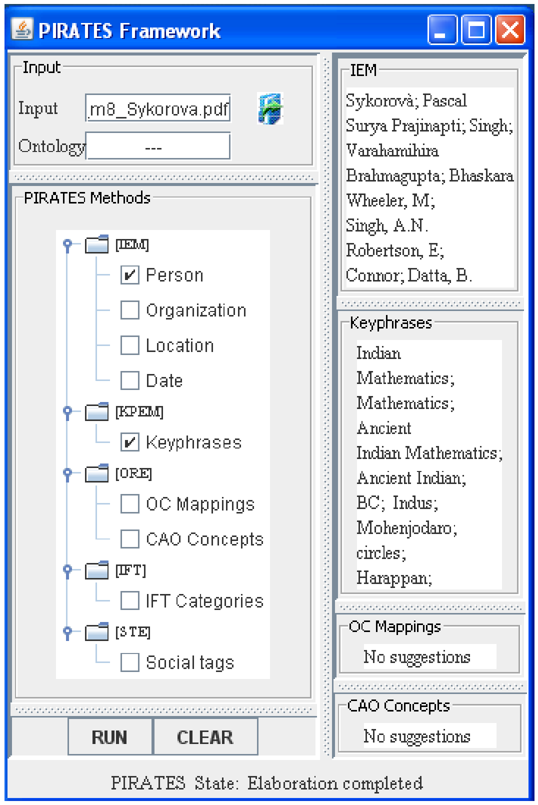

Figure 5: Tags suggested by PIRATES for a document.

The classification process produces evaluations of the relevance (in the sense of topicality) of a document according to a specific model of user interests represented by semantic networks.

- The *Automatic Tagger* module implements several modules devoted to automatically annotate an incoming stream of text (the content of a document) by means of tag recommendations: the submodule IEM (Information Extraction Module) suggests entity, names, and dates, KPEM (Key-Phrases Extraction Module) key-phrases, SAT (Sentiment Analysis Tool) polarity judgments, STE (Social Tagger Engine) tags used by a community of Web 2.0 users, while ORE (Ontology Reasoner Engine) tags extracted from an ontology. The user can choose the combination of annotators to exploit in order to obtain the tag suggestions.
- The *Knowledge Base Builder* module organizes documents in a Knowledge Base repository, producing annotated documents.

The PIRATES framework operates on a set of input documents stored in the Information Base (IB) repository and suggests for these some personalized tags and other forms of textual annotations (e.g. key-phrases) in order to classify them. The original documents are then annotated with these tags, forming the Knowledge Base (KB) repository.

Our main goal in integrating PIRATES in the E-Dvara platform is to empower information access, allowing users to easily find new relevant contents, and automatically supporting them when categorizing documents by means of keywords (tags) in a personalized and adaptive way. We have designed PIRATES keeping in mind several applications where it can provide innovative adaptive tools enhancing user capabilities: in e-learning portals for supporting the tutor and teacher activities in monitoring student performance, behavior, and participation; in knowledge management contexts (including scholarly publication repositories and digital libraries[Omero et al., 2007]) for supporting document filtering and classification and for alerting users in a personalized way about new posts or document uploads relevant to their individual interests.

VI. Improving Information Access in Cultural Heritage Archives using Semantic Web

In order to show a usage scenario of PIRATES, consider an archivist which is going to update an E-Dvara archive concerning ancient Indian science, adding a scientific paper by Sykorova entitled “Ancient Indian Mathematics” [8]. Before to store the document, the archivist prepares a card summarizing the paper contents. As part of this card, he can provide one or more keywords which will be exploited by the E-Dvara applications (typically, search engines) to retrieve the document efficiently. The PIRATES framework assists the archivist in this task, performing both statistical and semantical analysis of the document

[8] http://www.mff.cuni.cz/veda/konference/wds/contents/pdf06/WDS06_101_m8_Sykorova.pdf

content and suggesting a set of tags, as illustrated in Figure 5. The archivist can select each annotator tool individually; in this case, he was interested in extracting names and key-phrases. Eventually, he will select from the list those tags that better represent the document and annotate it consequently.

Automatic tag recommendation may lead to several improvements in content access, one of these being *annotation-based content recommendation*. For example, if a user is accessing a document similar to the Sykorovas one, already available into the platform and yet tagged by PIRATES, the set of most similar contents may be retrieved and presented to the user. Similarity is defined as the set of common tags shared by a source content and the other contents available into the platform. In this scenario each user can provide a query to the platform search engine, browse the results and, in real-time, receive a list of suggested contents, which do not necessary contain the same keywords used in the query (tags identified by PIRATES may represent concept which are not directly referred into the resource, but obtained as result of ontology base inference).

Another scenario concerns the task of *query reformulation*. Semantic representation of available contents, provided manually or in an automatic way, may be also exploited to improve the effectiveness of traditional keyword-based retrieval. More specifically semantic knowledge can be used to augment the effectiveness of traditional keyword search, moving further toward the concept of semantic search. In order to achieve such goal, a query reformulation engine will be included into the Semantic layer of the E-Dvara platform. Using both the contents metadata and the ontologies constituting the Knowledge Base of the platform, the query reformulation engine will intercept the requests submitted by users and suggest, in addition to the retrieved contents, a list of potentially related queries.

According to the workflow and the nature of the set of modules constituting the PIRATES framework, the query reformulation task will be based two different kind of knowledge: ontology-based reformulation and annotations-based reformulation. *Ontology-based reformulation* will be used to identify concepts similar to the terms used into the query by browsing the domain-dependent ontology used by the ORE module to annotate resources. Such concepts may be included into the query or can be used to substitute existing terms. *Annotations-based reformulation*, on the other hand, is based on the tags assigned to the contents retrieved using the original query; by ranking tags and looking at the most relevant ones (relevance will be defined as a tag frequency function), reformulation engine can generate a new query. Annotations-based reformulation exploits all the different kinds of annotations provided by the PIRATES framework; such approach may be seen complementary to the one used by the ontology-based reformulation, where only knowledge occurring into the domain ontology is considered.

In the next few months we will formally define both the concepts of ontology-based and annotation-based similarity and, according to such model, we will integrate the query reformulation service into the E-Dvara Semantic layer.

VII. CONCLUSIONS

In this paper we have proposed a service-oriented architecture for the E-Dvara digital library which explicitly integrates a semantic layer. The integration of semantic services is aimed at better addressing the changes in the final users information needs and improving the effectiveness of the information access. To support this new semantic layer, we have designed a framework based on adaptive and personalized services, distinguishing the digital library from a old-fashioned DBMS/structured archive system. Give access to the semantics of contents helps to realize the vision of semantic digital library which is possibly one of the most innovative evolutions of current digital libraries.

The ideas discussed in the article come from the lessons learned during the experimentation with the first prototype of the E-Dvara platform in the last three years. We are now working to complete a second version of E-Dvara which will embody the improvements discussed in this paper. Our future plans include a validation of the overall prototype in different areas, concerning the exploitation of both information and services by means of mobile applications, virtual museums, and Web 2.0 environments.

REFERENCES

[Bainbridge et al., 2001] Bainbridge, D., Buchanan, G., Mcpherson, J., Jones, S., Mahoui, A., and Witten, I. (2001). Greenstone: A platform for distributed digital library applications. In *ECDL '01: European Digital Library Conference*, pages 137–148, Berlin. Springer-Verlag.

[Barkstrom et al., 2002] Barkstrom, B., Finch, M., Ferebee, M., and Mackey, C. (2002). Adapting digital libraries to continual evolution. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 242–243. ACM.

[Baruzzo and Casoto, 2008] Baruzzo, A. and Casoto, P. (2008). A flexible service-oriented digital platform for e-content management in cultural heritage. In *IABC '08: Intelligenza Artificiale nei Beni Culturali Workshop*, pages 38–45.

[Baruzzo et al., 2008] Baruzzo, A., Casoto, P., Challapalli, P., and Dattolo, A. (2008). An intelligent service oriented approach for improving information access in cultural heritage. In *IACH '08: Information Access in Cultural Heritage (IACH) Workshop, European Conference on Digital Libraries*, Berlin. Springer-Verlag.

[Baruzzo et al., 2009a] Baruzzo, A., Casoto, P., Dattolo, A., and Tasso, C. (2009a). A conceptual model for digital libraries evolution. In *WEBIST '09: Proceedings of 5th Informational Conference on Web Information Systems and Technologies*, pages 299–304, Berlin. Springer-Verlag.

[Baruzzo et al., 2009b] Baruzzo, A., Dattolo, A., Nirmala, P., and Tasso, C. (2009b). A general framework for personalized text classification and annotation. In *International Workshop on Adaptation and Personalization for Web 2.0 in connection with UMAP 2009, Trento, Italy, June 22-26*.

[Bekaert et al., 2005] Bekaert, J., Liu, X., and Van de Sompel, H. (2005). aDORe: A modular and standards-based digital object repository at the Los Alamos National Laboratory. In *JCDL '05: Joint Conference on Digital Library*, pages 367–367. ACM.

[Candela L. and Pagano, 2007] Candela L., Castelli, D. and Pagano, P. (2007). A reference architecture for digital library systems: Principles and applications. In *Digital Libraries: Research and Development,* 1^{st} *International DELOS Conference*, pages 22–35.

[Celino et al., 2006] Celino, I., Turati, A., Della Valle, E., and Cerizza, D. (2006). Squiggle Med: Semantic search for medical digital library. Technical report, CEFRIEL, Politecnico di Milano.

[Challapalli et al., 2006] Challapalli, S., Cignini, M., Coppola, P., and Omero, P. (2006). E-Dvara: an XML based e-content platform. In *AICA: Associazione Italiana per l'Informatica e il Calcolo Distribuito*.

[Cunningham, 2002] Cunningham, H. (2002). Gate, a general architecture for language engineering. *Computers and the Humanities*, 36:223–254.

[Dattolo et al., 2009] Dattolo, A., Tomasi, F., and Vitali, F. (2009). Towards disambiguating social tagging systems. In Murugesan, S., editor, *Handbook of Research on Web 2.0, 3.0 and X.0: Technologies, Business and Social Applications*, volume Chapter 27. IGI-Global.

[Eide et al., 2008] Eide, O., Felicetti, A., Ore, C., DAndrea, A., and Holmen, J. (2008). Encoding cultural heritage information for the semantic web. In *Procedures for Data Integration through CIDOC-CRM Mapping, EPOCH Conference on Open Digital Cultural Heritage Systems*, pages 1–7.

[Hernandez et al., 2008] Hernandez, F., Rodrigo, L., Contreras, J., and Carbone, F. (2008). Building a cultural heritage ontology for cantabria. In *Annual Conference of CIDOC*.

[Hunter et al., 2008] Hunter, J., Khan, I., and Gerber, A. (2008). Harvana: harvesting community tags to enrich collection metadata. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 147–156, New York, NY, USA. ACM.

[Kruk and McDaniel, 2008] Kruk, S. and McDaniel, B. (2008). *Semantic Digital Libraries*. Springer Verlag.

[Lagoze et al., 2005] Lagoze, C., Payette, S., Shin, E., and Wilper, C. (2005). Fedora: An architecture for complex objects and their relationships.

[Lutzenkirchen, 2002] Lutzenkirchen, F. (2002). MyCoRe - ein open-source-system zum aufbau digitaler bibliotheken. *Datenbank-Spektrum*, 4:23–27.

[Mathes, 2004] Mathes, A. (2004). Folksonomies - cooperative classification and communication through shared metadata.

[Omero et al., 2007] Omero, P., Polesello, N., and Tasso, C. (2007). Personalized intelligent information services within an online digital library for medicine: the BIBLIOMED system. In *IRCDL '07: Proc. of the Third Italian Research Conference on Digital Library Systems*, pages 46–51.

[Ross, 2006] Ross, S. (2006). Approaching digital preservation holistically. In Tough, A. and Moss, M., editors, *Information Management and Preservation*, pages 115–153, Oxford. Chandos Press.

[Ross, 2007] Ross, S. (2007). Digital preservation, archival science and methodological foundations for digital libraries. In *ECDL '07: European Digital Library Conference*, Berlin. Springer-Verlag.

[Tansley et al., 2003] Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., and Smith, M. (2003). The DSpace institutional digital repository system: Current functionality. In *JCDL '03: Joint Conference on Digital Libraries*, pages 87–97. IEEE.

[Tasso and Asnicar, 1997] Tasso, C. and Asnicar, F. A. (1997). ifweb: a prototype of user model-based intelligent agent for document filtering and navigation in the world wide web. In *Adaptive Systems and User Modeling on the WWW, 6th UM Inter. Conf.*

[van der Sluijs and Houben, 2008] van der Sluijs, K. and Houben, G. H. (2008). Metadata-based access to cultural heritage collections: the RHCe use case. In *PATCH'2008: Proceedings of the 2nd International Workshop on Personalized Access to Cultural Heritage, workshop at the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2008)*, pages 15–25.

[Witten et al., 2000] Witten, I., McNab, R., Boddie, S., and Bainbridge, D. (2000). Greenstone: A comprehensive open-source digital library software system. In *ICDL '00: International Conference on Digital Libraries*. ACM.

[Woroniecki et al., 2007] Woroniecki, T., Gzella, A., Dobrowski, M., and Ryszard Kruk, S. (2007). JeromeDL ? a Semantic Digital Library. In *Semantic Web Challenge Co-olocated with The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference the 2nd Asian Semantic Web Conference*, Busan, Korea.