Handling Evolution in Digital Libraries ⋆

Andrea Baruzzo, Paolo Casoto, Antonina Dattolo and Carlo Tasso

Department of Mathematics and Computer Science - University of Udine, Italy
{andrea.baruzzo, paolo.casoto, antonina.dattolo, carlo.tasso}@dimi.uniud.it

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, and so on). Moreover, starting a digital library initiative requires to handle several evolution issues (the need of new roles, workflows, and types of contents, the availability of new applications to integrate on the top of digital archives, etc.). To catch all these aspects, we outline a conceptual model based on three complementary domains: informational, technological, and social. This model tackles the typical issues affecting a digital library, especially concerning the evolution (i.e. change) of content, infrastructure (software tools), user roles, and related workflows in content creation, publication, and exploitation. These issues are addressed in the model by means of three elements: a suitable XML data metamodel, a service-oriented architecture, and a multi-agent infrastructure.
The XML metamodel abstracts the physical representation from the logical definition of data, making easier future changes. The multi-agent infrastructure helps to preserve the consistency of the stored archives when their schemata need to be changed. Finally, the service-oriented architecture simplifies the integration of new applications at the top of the digital library. As part of this architecture, we describe in particular a specific component: the PIRATES framework. This module introduces in the digital library a set of semantic services aimed at assisting final users to select from the archives the most appropriate content. Integrating semantic aspects helps to handle the evolution of both contents and user needs (i.e. interests). 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.

1 Introduction

Data preservation in digital libraries is often addressed mainly by means of reliable storage mechanisms, and long-term accessibility of digital supports, both aimed at ensuring that library's contents will remain sustainable, authentic, accessible, and understandable over time. *Data evolution*, at the same time, is addressed by means of scalability of the *physical* system, concerning the modification of the stored information either in data formats or in space needed to archive them. Following this vein, many works in literature consider preservation and evolution issues mainly as technological factors

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

(e.g. [3]). In this paper, we take a wider approach, recognizing also the role of other important aspects outside the technological domain. In particular, we explicitly address the *semantic aspects* of a (textual) content stored in a digital archive. We agree with Ross in claiming that digital preservation is more than keeping the streams of 1s and 0s used to represent information [20]. Preserving information is about maintaining the *semantic* meaning of both the digital object and its content, maintaining its provenance and authenticity, retaining its interrelatedness, and securing information concerning the context of its creation and use [21].

Moreover, a digital library is more than the sole data stored in its archives: it serves as an infrastructure to publish, retrieve, and access information fulfilling the final user needs. The user community can be huge and heterogeneous, thus *social aspects* can also play a critical role in the evolution of a digital library. For example, content production and exploitation processes involve a wide range of different users. The way users exploit their respective tasks may change over time and with experience, as well as with the upcoming of new information needs. The digital library infrastructure should then be aware of the evolution related to social aspects in order to provide users with the ability to easily update the way information is generated, accessed, aggregated, classified, and delivered. The effort needed to align this infrastructure with the evolution of user requirements is not trivial. New requirements often demand for new functional services which must be integrated into the legacy architecture in order to be effective. Hence, handling evolution aspects at the requirement level has also an important but often not recognized impact of the architectural level.

In this work we introduce a characterization of a digital library addressing specifically the evolution of both digital content and related services, taking into account the aforementioned considerations. This characterization integrates three complementary domains: *social*, *technological*, and *informational*. These domains can be incorporated into a conceptual model, as discussed in Section 4. We believe that providing such model facilitates the undertaking of the mentioned aspects. Hence, this formal model takes into account both *physical* and *semantic* evolution of the archived content, and the way such information is exploited by final users. In particular, semantic concerns are handled in the model at the architectural level by a specific component, the PIRATES framework, as described in Section 6.1. Such "semantic layer" is a first step in the direction of fully supporting the semantic digital library vision [16].

This paper is based on the results of an experimentation within the EU-India E-Dvara project[1]. This project is concerned with the development of a digital platform devoted to e-content management in Indian heritage and sciences. In previous works, we presented the overall project goals [11], the technical details concerning data representation metamodels, and the general software architecture [4–6]. Here, we extend previous works, focusing especially on the characterization of a digital library according to its evolution aspects. In particular, after a brief survey of related work, we extend the ideas proposed in [6], presenting a conceptual model to handle the evolution of digital archives along multiple dimensions (Section 4 and Section 5). Then, we discuss a set of semantic, adaptive, and personalized services which can be introduced on the

[1] http://edvara.uniud.it/india

top of a digital library, supporting final users in retrieving, annotating, classifying and organizing the information archived in the library (Section 6).

2 Related Work

In the last few years several research projects have been proposed in order to cope with data preservation and organization [7, 18, 8]. Storage of XML-based documents has been proposed in Greenstone [2, 26], 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 a standard schema (e.g. Dublin Core[2]), either imported or manually provided by librarians. However, Greenstone does not provide any role for managing the content submission process. Moreover, it does not provide functionalities concerning the evolution management of both contents and collection templates.

D-Space [24] 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. 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, to structure content and to delegate proper activities to different stakeholders.

In order to provide a flexible and reusable solution to data preservation and organization, the Fedora Project [17] 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 to content preservation in digital libraries are described in [7, 18]; the aDORe project, in particular, adopts the MPEG-21 DID content representation model to provide preservation and retrieval of heterogeneous multimedia contents.

The above mentioned systems are centered 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).

3 The E-Dvara Project

In this section we introduce the E-Dvara digital library. We examine its initial requirements and the results of an experimentation activity performed in the last three years. Building from these requirements, and overcoming some of the criticalities emerged

[2] See for more details: http://dublincore.org/

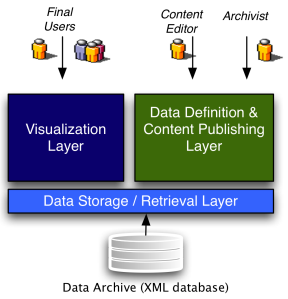

Figure 1: The high-level E-Dvara software architecture

during experimentation, we formulate the proposal for a new version of the digital library which is discussed in the rest of the article.

E-Dvara is a research project which represents our current development and experimentation in the area of digital libraries. E-Dvara is focused on the development of a new platform for the access and storage of digital contents [11]. Since its inception, it has been explicitly designed to overcome several limitations that characterize the process of building 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 platform;
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 built-in Web interfaces or with several dedicated applications, allowing archivists and system administrators to define the way data should be displayed to final users.
4. allow archivists to provide for each archive of digital contents a specific visualization template and a set of search forms.

In order to cope with these requirements, the E-Dvara platform has been designed adopting a three level modular architecture, illustrated in Figure 1. This architecture is constituted by a core layer devoted to data storage and persistence, a set of tools for data definition and content publishing, and a visualization layer devoted to content delivery and rendering.

Digital contents are represented and stored as XML documents, in order to fulfill both the requirements related to data interoperability and separation of concerns between data representation and data visualization. Data visualization is achieved by means of an XSLT engine, which can transform each XML document into HTML, PDF and WAP files, according to a specific style sheet defined by the archivist.

Data definition and content publishing are provided by a set of Web applications devoted to archivists and content publishers; archivists can easily create new data schemata, defining the required simple or complex fields, their multiplicity and data types. Con-

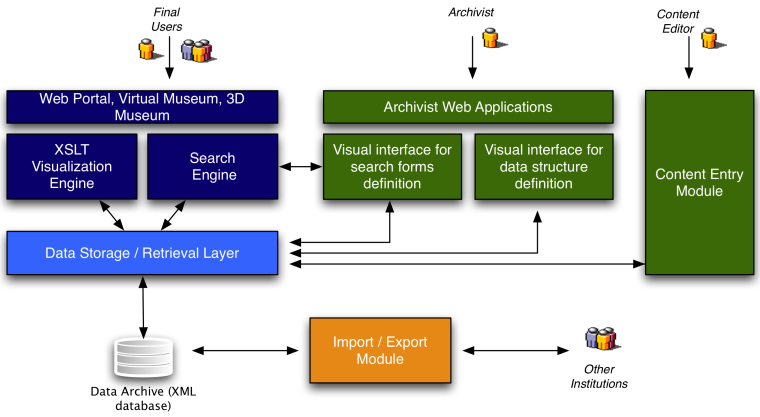

Figure 2: The architecture of the first prototype of E-Dvara platform

tent providers, according with data schemata created by archivists, can publish their data using Web forms automatically generated from the selected data structure.

The third layer concerns the visualization and delivery of the available contents to end users which can access data using either a Web browser or a WAP browser on a mobile phone. Data visualization is exploited automatically according to the device employed, the user preferences, and the style sheets provided by the archivists. Visualization layer is also responsible for providing to final users the search forms defined, for each archive, by the archivists. Search forms are represents as XML documents too. In Figure 2 an overall representation of the original architecture of E-Dvara is provided; the XML archive is the core component of the proposed architecture, on which the three described layers are based. A prototype of the E-Dvara platform was developed in 2005.

In the last three years, E-Dvara has been largely tested by expert users involved into professional content publishing for cultural heritage. From this experimentation, we have identified several problems with the first prototype. In particular, the evolution issues, the technical weaknesses, and the mistakes done led us to formulate the conceptual model and a new software architecture, both of which are discussed in the next sections of this paper.

4 Modeling evolution

To better handle the evolution aspects of a digital library, we propose an explicit conceptual model. We discuss here its main characteristics and the evolution issues which have inspired it.

4.1 The conceptual model

Our conceptual model is inspired by that provided by Yates [27], and incorporates the vocabulary suggested by Rowlands-Bawden [22], which is more suited to describe mod-

ern digital environments. Moreover, we introduce the concept of *evolution dimension* [6], which best addresses the evolutionary nature of a digital library. This new model highlights three domains:

- The *Informational domain*, which describes the knowledge organization and description (e.g. metadata) of digital archives.
- The *Technological domain*, which describes the knowledge organization and discovery provided by an appropriate technological infrastructure (e.g. software agents), the technical impacts on information transfer chains, and technology factors such as human-computer interactions.
- The *Social domain*, which describes human and organizational factors, information laws and policies, social impacts on the information transfer chain, and library management concerns.

On the basis of these domains, we define three specific *evolution dimensions* by means of a pairwise combination of domains. In fact, we recognize a mutual influence among domains when we consider evolving aspects in a digital library. More specifically, the open issues faced during our experimentation with E-Dvara may be classified along three dimensions:

1. *Informational-Technological* dimension, which identifies all *data evolution* problems due to changes in the underlying data model (e.g. the invalidation of entire archives of documents that conform to the old schema version).
2. *Technological-Social* dimension, which identifies problems concerning the need to adapt the technical infrastructure of a digital library in order to fulfill new user requirements (e.g. the integration of heterogeneous services to support the interaction with new user roles).
3. *Social-Informational* dimension, which concerns the diverse workflows needed to support the activities of such different community of users, and their impact on documents (e.g. a virtual museum curator has to describe the items of a document taking into account constraints imposed by user interfaces in order to effectively show tool-tips when a visitor moves the mouse over a particular exhibit in the scene). New roles can have a different view of documents, so the digital library should provide them the information required with formats suitable to their needs.

For each evolution dimension, we propose a conceptual model element which links together two model domain, as illustrated in Figure 3. We discuss in depth each element in Section 5.

4.2 Three classes of evolution problems in digital libraries

In this section we briefly summarize three classes of open issues concerning the evolution of digital libraries that we have identified during our experimentation with the first prototype of E-Dvara. In this context we also introduce the solution to each issue we are proposing in the new prototype which is under construction. Such solutions are detailed in Section 5. A more exhaustive presentation of the problems with representative examples is provided in [6].

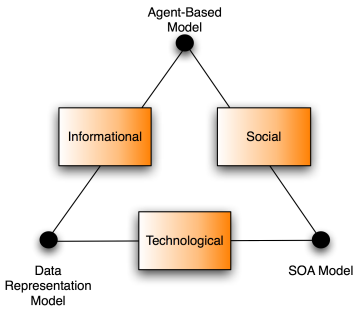

Figure 3: The evolution conceptual model

The data evolution problem The first prototype of E-Dvara provides users a flexible way to define and update the metadata associated to each digital archive. In particular, users can define a set of *schemata* which, associated with a specific archive, supplies the structure adopted for storing documents. However, metadata definition can take place during the whole life-cycle of the digital collection, leading to the problem of correctly handle the evolution of data. Such an approach requires the introduction of methodologies devoted to perform *data validation* accordingly to the schema evolution. In fact, each schema update should be properly spread to the previously validated archives, in order to automatically adapt the existing content to the new schema without introducing potential inconsistencies between data. Clearly, in order to both handle these issues and grant the preservation of yet existing contents, a set of dedicated tools should be provided. Hence, in the new prototype we are integrating the concept of *mutable templates*. To understand what is a mutable template, consider Figure 4 which describes the reference data model included in our conceptual model (a complete description of such model follows in Section 5.1). The main idea is to introduce flexibility by separating the data definition (in level M1) with the description of the corresponding types (in level M2). Furthermore, we would allow more than a single possible map between data definitions and types in order to cope with changes in both data formats and structure (e.g. consider the same set of bibliographical items which can be showed to different users using different formats). This is exactly the purpose of mutable templates: to allow type variations in data mappings. The evaluation we performed in E-Dvara has suggested us that the flexibility of *mutable templates* (i.e. evolving schemata) must be considered as an essential feature for our platform.

Technical infrastructure adaptation Other issues we have faced is the need for integrating new heterogeneous modules at the top of the digital library (e.g. virtual museums, meta-search engines, or applications for mobile devices). These requests posed unforeseen challenges on the software infrastructure. For each new application, we needed to rewrite a lot of ad-hoc business logic, without mentioning the fact that we had to duplicate some system services in order to adapt them to a new programming interface.

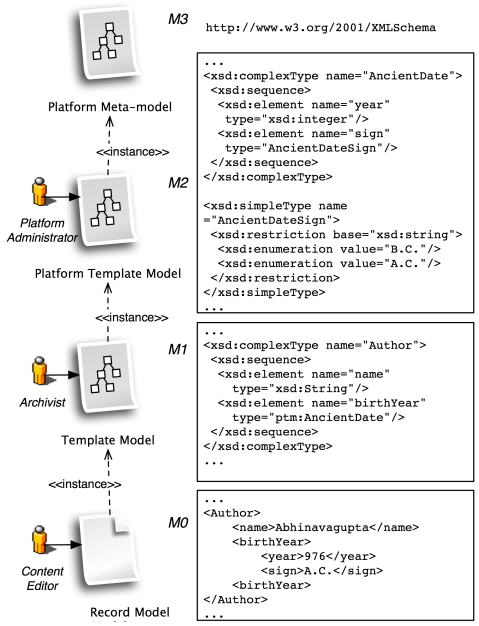

Figure 4: XML Representation of Data Model

Another tricky aspect was sharing similar behaviors and data located in different components which were based on incompatible communication protocols. These issues clearly demanded for a *reusable integration layer* that we lacked as part of our initial software architecture. Moreover, we have learned also that integrating several applications in a common environment requires a substantial investment in understanding and implementing their *orchestration*, in order to handle incompatibilities between different business logics in a standard and transparent way. In conclusion, in our first prototype we failed to recognize the importance of these issues, which are the building blocks to describe the interaction between different applications. If we had realized that since the beginning, composing together heterogeneous functionalities would be simpler to achieve. Furthermore, also the resulting deployment would be expedited, especially by moving the orchestration logic from the inside of a component to an external configuration file (i.e. by means of a XML description file associated to each component), enabling a flexible and dynamic setup.

Coordination issues of different user roles and workflows The continuous integration of different applications on the top of the existing software infrastructure were typically a manifestation of new requirements involving user roles and information access policies. An example of this situation is an external service which, based on its own data management policy, defines *when* a particular workflow is required to organize the archived contents. In E-Dvara, a *workflow* expresses a set of roles, related activities, and constraints that define together the structure of the information manage-

ment process. As a typical workflow, consider the curator of a digital museum which has to arrange a new gallery, composed by paintings, ancient books, and movies hosted in three archives, respectively, and owned also by three different archivists. When the curator wants to incorporate in this gallery a set of features to search, organize, and enrich the existing records, he may want to add new fields describing the position that each item should have in the virtual museum scene. Moreover, final users may also improve the exhibition quality, by creating new relations between existing contents (e.g. opinions and links to a specific related content in a typical Web 2.0 style). All these scenarios pose many issues that must be faced to provide flexibility in data management. Such issues concern *intellectual property* (Who is allowed to use/modify a content?), *coupling between archives* (How do the schemata of archive B evolve according to the evolution of both schemata and data in archive A, if any dependency between A and B exists?) and *coupling between workflows* (Do the activities in the workflow A overlap those in workflow B?).

5 Handling evolution of content production workflows

This section describes three conceptual model elements (the XML metamodel, the service-oriented architecture, and the multi-agent infrastructure), each one associated to a specific evolution dimension introduced in Section 4.1. Moreover, we describe also how these model elements are characterized by mutable templates, integration layers, and workflows management which are discussed in previous section.

5.1 The Informational - Technological dimension

In order to handle the evolution problems concerning the changes in data format and schemata described in Section 4.2, our conceptual model is based on a four-layer data representation model (Figure 4). At the bottom of the hierarchy, we place the *records* (level M0, Record Model), aimed at representing the archived data (documents). A record is an instance of a document stored in the digital platform. Every document must also conform to a *document template* (level M1, Template Model), which provides structural definitions (e.g. the document contains the `Title`, `Author`, and `Date` fields) and constraints (e.g. the `Data` field must conform to the `mm/dd/yy` format or the `Title` field is mandatory). Document templates are themselves conformed to *platform template* (level M2, Platform Template Model) devoted to define both business rules and data types the archivists can use to build document templates (e.g. each record in every archive must contain the `Creation Date` and `Owner` fields). Finally, platform templates are instances of a more general layer, the *platform meta-model* (level M3, Platform Meta-Model), which defines a set of common low-level structures (e.g. primitive data types as `xsd:String`) and operations (e.g. data sequencing) available in order to define more complex data structures. This level corrisponds to that of the OMG XML Schema specifications[3].

The overall data model involves the interaction with three different actors:

[3] http://www.w3.org/XML/Schema

- *Content editor*, devoted to data entry, with respect to a specific document template; however, he is not allowed to perform any template change.
- *Archivist*, devoted to document templates definition.
- *Platform administrator*, devoted to the management of platform templates (e.g. the templates provided by archivists should be validated against the platform template model each time they are created/modified or when the platform template model itself is updated).

This hierarchical data model provides *automatic data validation policies* which play a central role in our vision. Indeed, validation is applied both to the templates and (recursively) to all the records stored in the platform archives. Templates which do not respect the business rules defined in the platform template model should be manually updated by either archivists or content providers in order to become consistent. This type of validation is propagated then to the platform meta-model (level M3) which acts as a template for the platform template model (level M2). In order to develop the proposed model, we present an implementation approach based on the XML technology and standards, focusing our attention on the features provided by XML Schema.

5.2 The Technological - Social dimension

Sometimes content production workflows change because new actors emerge, playing roles that was not foreseen in advance. Developers of digital libraries are then forced to consider new requirements when it is more expensive to integrate them in an existent technical infrastructure. Hence, to handle evolution issues concerning the adaptation to such new requirements (e.g. the integration of heterogeneous services described in Section 4.2), we conceive the second prototype of E-Dvara according to a Service-Oriented Architecture (SOA) model (Figure 5), characterized by:

- The introduction of an explicit *Integration layer*, which forms the "architectural glue" that brings the digital library beyond the scope of a single application, unifying the interfaces of different subsystems into the same interoperable environment.
- The migration toward *services*.
- The adoption of a *peer-to-peer, message-based communication protocol* supported by the *Enterprise Service Bus* (ESB)

The standard set of Web service technologies (XML, SOAP[4], WSDL[5]) provides the means to describe, locate, and invoke a Web Service. However, it is often necessary to compose different services with a specific business logic in order to complete a task. This is where orchestration plays a crucial role, deploying sophisticated and complex Web services as a single, whole functionality. Thus, the orchestration engine (the ESB component) acts as a centralized authority to coordinate interaction between services and applications. At the top of our SOA architecture we have applications such as administration interfaces to manage users and archives, publication interfaces to produce new content in the digital library, or virtual museums to exhibit a document

[4] http://www.w3.org/TR/soap
[5] http://www.w3.org/TR/wsdl

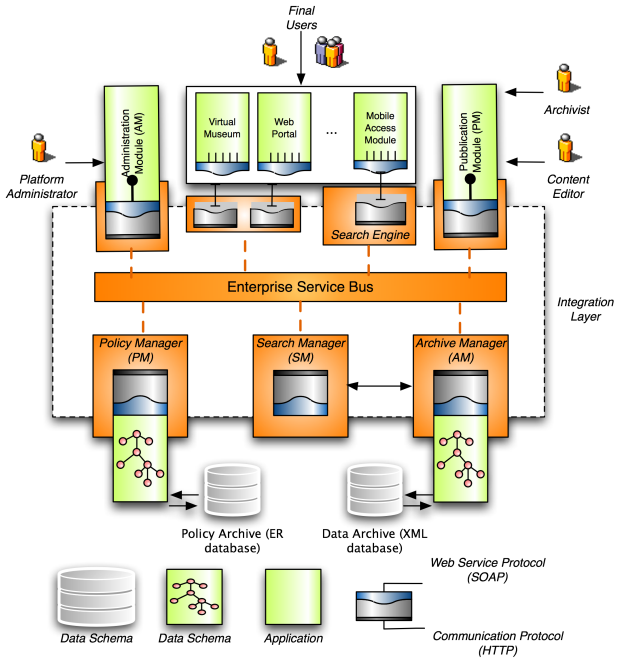

Figure 5: The architecture of the second prototype of E-Dvara

archive in a "museum-like" setting. All these heterogeneous modules can exploit any reusable service available in the Integration layer, (e.g. to perform searches in the platform archives). Finally, the archives are placed at the bottom of the architecture, which are managed by two custom applications: the *Archive Manager* which stores and retrieves documents, and the *Policy Manager* which manages users, accessing policies, and archives. The Archive Manager isolates the business logic needed to realize the data-model described in Section 5.1, whereas the Policy Manager implements the data validation rules, decoupling them from other architectural components.

5.3 The Social - Informational dimension

The introduction of mutable templates in content representation provides the ability to update a schema during the whole life-cycle of a data collection, but leads also to several challenges such as the evolution and re-validation of existing archives. In this section, we introduce a *multi-agent approach* to tackle the problem, aimed (when possible) to automatically resolve evolution issues.

The levels from M1 to M3, proposed in Figure 4, can be affected by updates during the digital library life-cycle. In particular, such updates can involve XML Schema definitions (level M3, with a low frequency), Platform Template Models (level M2, with a low-medium frequency) and Template Models (level M1, with a rather high frequency). Each schema is connected by a dependency link with the schemata on its top for val-

idation purposes. However, in a collection one level can be related to another also by means of relations between different data types (e.g. an instance of the template `Book` in M1 can be related with one or more instances of the template `Author` in M0). At the same time, we can also have a relation connecting templates in different collections (e.g. instances of the template `GalleryRoom` in a virtual museum application can be related with instances of `Book` and `Painting` templates taken from different collections). Hence, such dependencies requires evolution mechanisms that must be propagated both in a specific level and across multiple levels. This propagation mechanism is achieved by means of a multi-agent system. Each agent is assigned to a specific schema, monitoring its evolution; an agent can interact with other agents assigned to depending schemata, send them messages and apply evolution to the instances of its schema.

A *coordinator agent* is assigned to each instance of the platform, in order to monitor the updates of the Platform Template Model and to activate the agents connected to each schema when required. The coordinator agent is also devoted to the creation of a new agent every time a new schema is defined (even if it does does not act directly on data because such task is delegated to agents located at level M1).

A *schema agent* is devoted to the evolution of contents related to a specific template at level M1. They can perform a set of actions on the existing data, accordingly to the updates affecting related schemata. Agents perform several evolutionary operation on data, in order to preserve data validity and, at the same time, to prevent archivists and content editors to spend a lot of time re-entering the whole set of existing contents. In [14, 15] a complete taxonomy of updates, which can affect a generic XML schema, is described; actually only a subset of the listed operations has been implemented in E-Dvara, covering the set of updates which can be performed by archivists. For example, we provide the utilities to rename or add elements and attributes of the Template Model (level M1). In order to cope with the complexity of the evolution tasks and the amount of data yet available in E-Dvara, our attention is focused on simple updates which commonly occur during the life-cycle of a collection. A typical evolution task is represented by the extraction of a vocabulary (a closed list of predefined strings) from the set of values assigned to a free-text `String` element. In our experience such an update is rather frequent, specifically when we are not able to know a priori *all* the values assignable to a specific element. In this case, when an archivist decides to change the type of the element `Name` from `String` to `Vocabulary`, the agent assigned to that schema should access each instance of the template and perform a `change_item_type`, verifying if the old values assigned to `Name` are validated with respect to the values admitted by the new element type. When this task is completed, the agent should notify the schema updates to the related agents (according to the dependency chain between schemata), in order to grant the consistency of any inter-dependent data.

6 Toward semantic digital libraries

In Section 5 our attention was focused on handling evolution issues from the content providers point of view, more focused on data schemata and publication services. Here we take into account aspects more related to the semantic nature of the information stored in a digital archive. Our perspective, now, is that of a typical final user which uses

the digital library services to fulfill a specific information need (e.g. submitting textual queries to the digital library search engines). In this context, we recognize that such need has a dynamic nature. The interactions between final users and the library may evolve, reflecting changes in information needs. (e.g. which is the semantics of a user query? Which content archived in the library best fits the query?) Moreover, according to the growth of Web 2.0 philosophy, new ways to access such information should be provided to users: they could add their own contributions to documents (e.g. tags, comments, etc.), share them with other users improving in such a way the effectiveness of information access.

In conclusion, a digital platform should provide to its users an environment capable of dealing with information retrieval tasks where it is not important the presence of the "exact word" (string matching approach), but of the *intended meaning* underlying the information need. 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. To achieve this goal, we have designed the PIRATES framework. Its integration on the top of the E-Dvara platform is aimed at providing a "semantic layer" into the digital library.

6.1 The Pirates framework

PIRATES (Personalized Intelligent Recommender and Annotator TEStbed) is a general framework for text-based content retrieval and categorization and exploits social tagging, user modeling, information filtering, and information extraction techniques. The main feature of PIRATES concerns a novel approach that automates in a personalized way some typical manual tasks (e.g. content annotation and tagging). The framework operates on the contents archived in the Information Base (IB) repository (e.g. the digital library archives) and suggests for these some personalized tags recommendations and other forms of textual annotations (e.g. key-phrases) in order to classify them. The original contents are then annotated with these tags, forming enriched archives that we store in a Knowledge Base (KB) repository. These two different types of archives denote our particular approach to access and manage information provided by digital libraries. In our view, annotating a specific content with semantic information such as that potentially conveyed by tags or key-phrases, we shift from the perspective of *data* to the perspective of *knowledge*.

Personalization is achieved exploiting user profiles (which represent the user interests), personal ontologies, personal tags, etc. Furthermore, PIRATES provides several mechanisms of user feedback that help to adapt the content retrieved and showed by the digital library services to the needs of the final user, even where these needs change.

6.2 The PIRATES modules

The PIRATES architecture is illustrated in Figure 6. On the left-hand side, all the possible input sources are shown: single textual documents, specific IB repositories which can be contained within an e-learning or knowledge management environment, and the Web, with specific (but not exclusive) focus on Web 2.0 portals, social networks, etc.

The right-hand side shows the annotations automatically suggested to the user and the resulting KB repository. The main modules of our PIRATES framework can be classified in two categories: modules that discover new content (*Web agents and filters*), and modules that extract information from these new contents, in the form of textual annotations (*content annotators*). Typically, the modules devoted to retrieve new contents are started first, initializing repositories and the "semantic environment".

Web agents and filters

- `IFT Web Agents`, which continuously monitor the Web (and the blogosphere) looking for new information, cooperates with IFT to filter contents according to the user model, loads and updates the IB repository as soon as new relevant information is available.
- `IFT` (*Information Filtering Tool*), which evaluates the relevance (in the sense of topicality) of a document according to a specific model of user interests represented with semantic (co-occurrence) networks [1]. IFT and its Web agents form together the Cognitive Filtering module discussed in [9].

As soon as new content is available, one or more annotators can be executed. The number of annotators and the exact order of execution is user defined.

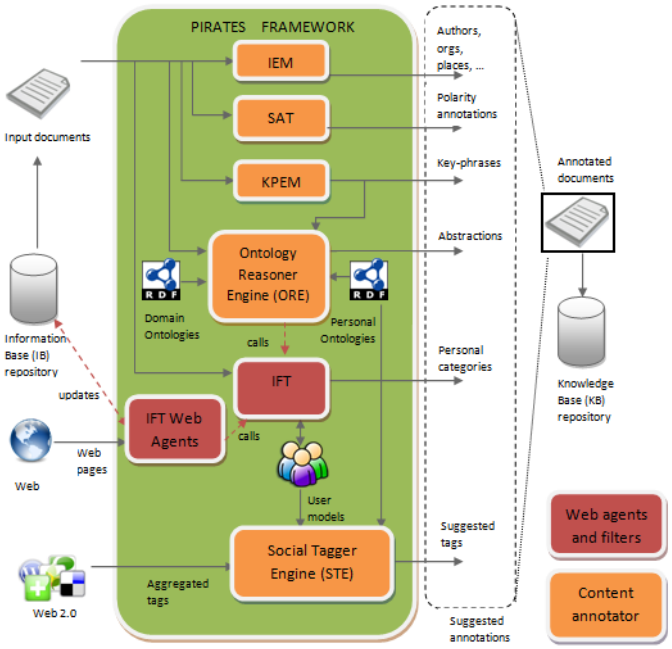

Figure 6: Overall architecture of PIRATES.

Content annotators

- `IEM` (*Information Extraction Module*), which is currently based on the GATE platform [12] to extract named entities, adjectives, proper names, etc. from input documents, contained in the IB.
- `SAT` (*Sentiment Analysis Tool*), which is a specific plug-in for personalized sentiment analysis (typically to be activated for marketing intelligence applications), that is capable of mining consumer opinions in the blogosphere and classify them according to their polarity (positive, negative, or neutral)[10].
- `KPEM` (*Key-Phrases Extraction Module*), which implements a variation of the KEA algorithm [13] for key-phrase extraction. KPEM identifies n-gram key-phrases (typically with n between 1 and 4) that summarize each input document. This information is provided to the user, but is also given in input to possibly subsequent modules.
- `ORE` (*Ontology Reasoner Engine*), which suggests new *abstract concepts* by navigating through ontologies, classification scheme, thesauri, lexicon (such as WordNet), etc. An abstract concept is identified by looking for a match between the annotations found by the other modules (IEM, KPEM, IFT, and STE) and the concepts stored in ontologies. When a match is found, ORE navigates through the ontology, looking for the common parent node which represents the more abstract term suggested as annotation. ORE also assists users in creating personal ontologies with techniques similar to those described in [23].
- `STE` (*Social Tagger Engine*), which suggests new annotations for a document relying on *aggregated tags*, i.e. the user's personal tags (tags previously exploited) and the more popular tags used by the community of people that classify the same document in social bookmarking sites such as Del.icio.us[6], Faviki[7] or Bibsonomy[8]. This social information is integrated with content-based analysis techniques as discussed in [25].

Our main goal in building the PIRATES framework is to empower the services provided by digital libraries, allowing users to exploit social bookmarking tools in order to easily add new contents in the library archives and categorizing such content by means of keywords (tags) in a personalized and adaptive way. This work is a first step toward the generation and sharing of personal information spaces described in [9]. We have designed PIRATES keeping in mind several applications where it can provide innovative adaptive tools enhancing user capabilities:

- in e-learning for supporting the tutor and teacher activities for monitoring (in a personalized fashion) student performance, behavior, and participation;
- in knowledge management contexts (including for example scholarly publication repositories [19]) for supporting document filtering and classification and for alerting users in a personalized way about new posts relevant to individual interests;
- in online marketing for monitoring and analyzing the blogosphere where word-of-mouth and viral marketing are nowadays more and more expanding.

[6] http://delicious.com
[7] http://www.faviki.com/pages/welcome/
[8] http://bibsonomy.org

7 Conclusions

In this paper we have introduced three specific evolution dimensions which characterize our conceptual model for handling evolution in content production within a digital library initiative. Furthermore, we have proposed the introduction of a semantic layer on the top of the E-Dvara digital library 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 that can empower the three dimensions of our conceptual model (especially in the social domain), 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. These proposals come from the lessons learned during the experimentation with the first prototype of the E-Dvara platform. 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 new 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

1. F. A. Asnicar, M. Di Fant, and C. Tasso. User model-based information filtering. In A. Evans, S. Kent, and B. Selic, editors, *AI*IA 97: Advances in Artificial Intelligence - Proceeding of the 5th Congress of the Italian Association for Artificial Intelligence*, volume 1321 of *Lecture Notes in Artificial Intelligence*, pages 242–253, Berlin, 1997. Springer-Verlag.
2. D. Bainbridge, G. Buchanan, J. Mcpherson, S. Jones, A. Mahoui, and I.H. Witten. Greenstone: A platform for distributed digital library applications. In *ECDL '01: European Digital Library Conference*, pages 137–148, Berlin, 2001. Springer-Verlag.
3. B.R. Barkstrom, M. Finch, M. Ferebee, and C. Mackey. 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, 2002.
4. A. Baruzzo and P. Casoto. 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, 2008.
5. A. Baruzzo, P. Casoto, P. Challapalli, and A. Dattolo. 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, 2008. Springer-Verlag.
6. A. Baruzzo, P. Casoto, A. Dattolo, and C. Tasso. 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, 2009. Springer-Verlag.
7. J. Bekaert, X. Liu, and H. Van de Sompel. 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, 2005.
8. D. Candela L., Castelli and P. Pagano. A reference architecture for digital library systems: Principles and applications. In *Digital Libraries: Research and Development,* 1^{st} *International DELOS Conference*, pages 22–35, 2007.

9. P. Casoto, A. Dattolo, F. Ferrara, N. Pudota, P. Omero, and C. Tasso. Generating and sharing personal information spaces. In *Proc. of the Workshop on Adaptation for the Social Web, 5th ACM Int. Conf. on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 14–23, 2008.
10. P. Casoto, A. Dattolo, and C. Tasso. Sentiment classification for the Italian language: A case study on movie reviews. *Journal of Internet Technology*, 9(4):365–373, 2008.
11. S.R.C.P. Challapalli, M. Cignini, P. Coppola, and P. Omero. E-dvara: an xml based e-content platform. In *AICA: Associazione Italiana per l'Informatica e il Calcolo Distribuito*, 2006.
12. H. Cunningham. Gate, a general architecture for language engineering. *Computers and the Humanities*, 36:223–254, 2002.
13. E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, and C.G. Nevill-Manning. Domain-specific keyphrase extraction. In *IJCAI '99: Proc. of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 668–673. Morgan Kaufmann, 1999.
14. G. Guerrini, M. Mesiti, and R. Rossi. Impact of xml schema evolution on valid documents. In *WIDM '05: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, pages 39–44. ACM, 2005.
15. G. Guerrini, M. Mesiti, and M. A. Sorrenti. Xml schema evolution: Incremental validation and efficient document adaptation. In *Database and XML Technologies, 5^{th} International XML Database Symposium*, pages 92–106, 2007.
16. S.R. Kruk and B. McDaniel. *Semantic Digital Libraries*. Springer Verlag, 2008.
17. C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: An architecture for complex objects and their relationships, 2005.
18. F. Lutzenkirchen. MyCoRe - ein open-source-system zum aufbau digitaler bibliotheken. *Datenbank-Spektrum*, 4:23–27, 2002.
19. P. Omero, N. Polesello, and C. Tasso. 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, 2007.
20. S. Ross. Approaching digital preservation holistically. In A. Tough and M. Moss, editors, *Information Management and Preservation*, pages 115–153, Oxford, 2006. Chandos Press.
21. S. Ross. Digital preservation, archival science and methodological foundations for digital libraries. In *ECDL '07: European Digital Library Conference*, Berlin, 2007. Springer-Verlag.
22. I. Rowlands and D. Bawden. Digital libraries: A conceptual framework. *Libri*, 49:192–202, 1999.
23. M. Speretta and S. Gauch. Using text mining to enrich the vocabulary of domain ontologies. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:549–552, 2008.
24. R. Tansley, M. Bass, D. Stuve, M. Branschofsky, D. Chudnov, G. McClellan, and M. Smith. The DSpace institutional digital repository system: Current functionality. In *JCDL '03: Joint Conference on Digital Libraries*, pages 87–97. IEEE, 2003.
25. C. Tasso, P.G. Rossi, C. Virgili, and A. Morandini. Exploiting personalization techniques in e-learning tools. In *SW-EL '04: Proc. of the Workshop on Applications of Semantic Web Technologies for Adaptive Educational Hypermedia*, 2004.
26. I.H. Witten, R.J. McNab, S.J. Boddie, and D. Bainbridge. Greenstone: A comprehensive open-source digital library software system. In *ICDL '00: International Conference on Digital Libraries*. ACM, 2000.
27. J. Yates. *Control Through Communication*. The Johns Hopkins University Press, 1989.