mat. Cornelia-Flavia VEJA

PhD THESIS

(abstract)

COLLABORATIVE DIGITAL REPOSITORY MANAGEMENT THROUGH SEMANTIC MEDIAWIKI

Supervisor:
Prof.dr.eng. Mircea-Florin VAIDA

2011
List of publications in the context of the thesis

BDI Journals:


Conference Proceedings:


7. Christoph Schindler, **Cornelia Veja**, Marc Rittberger, Denny Vrandecic, **2011**. How to Teach Digital Library Data to Swim into Research. In *Proceedings of i-Semantics conference, September 7 - 9, 2011, Graz, Austria*, pp. 142-149, ACM 978-1-4503-0621-8.


1. Introduction

Digital repositories and digital libraries are today among the most common environments and tools for managing and disseminating digital object collections of cultural, educational, scientific results and other kinds of content over the Web.

Repository technology that was originally created to manage metadata can be used to data and tool integration. The digital repositories have different data models, different schema and different underlying software and there are major issues of data integration and management.

The Web 2.0 paradigm makes the traditional producer–consumer model obsolete. It promotes new technologies that change the content granularity (becomes finer) and content typology (becomes wider). Moreover, a new phenomenon is the appearance of ad-hoc shared repositories inside of Social Software created communities, aka wikis. Rather than implementing a classic model, they are using simple protocols and services for simulating the digital repository functionality, but not the repository management. The repository management is likely to be a software feature or, mostly, a local administrative software mediated decision.

Objective: This thesis presents, starting from the current repository management models and following the Mass Collaboration Systems paradigm, a new model for repository management, named Collaborative Digital Repository Management (CDRM) Model. The novelty of this idea lies in involving the Social Network community formed around the shared or classic repository into the entire process of repository management. To overcome the trade-off between needs for advanced functionalities, low level of technological expertise and low cost management, the infrastructure needs to support protocols, tools, automatic processes and semantic technologies.

The specific goal of the thesis is to focus on the collaborative metadata repository management, inside of a small or medium community characterized by a low level of technological expertise.

This thesis explores how the new era of Semantic Web technologies can help librarians and researchers to better fulfill data provider’s and end-user’s requirements. Specifically supported by Social Semantic technologies, the whole process of repository resource management will be shifted from specialist to everyday users as much as possible. Users should be supported – but not forced to – in collaboratively structuring, organizing, formalizing and validating their metadata and other resources, in a customizable environment, for better retrieve and re-use.

The main questions the thesis research addresses are:

- Which factors influence requirements and benefits in metadata repository management?
- What is a suitable model to represent and use artifacts in a uniform fashion with different degrees of formalization?
- Which methods are suitable to involve user’s participation in the entire process of digital content management?
- How to build a Web-based integration layer for existing tools and application in the digital content management area? How to deal with the low level of technological expertise on the providers and user side and sophisticated functional and non-functional requirements that the final framework should fulfill?
- How to employ Semantic Web technologies like ontologies to unify different metadata standards and how to make this process transparent for the end-users?
- How to ensure an automatic workflow process of metadata exchanges between data providers, Digital Repositories and the integration layer for end-users?
How to validate metadata quality and how to deal with different interpretations of the metadata delivered by data providers by adopting an un-obtrusive method?

How to build a conceptual model for a collaborative virtual research environment when the metadata providers and researchers who will annotate the digital resources are technologically agnostic?

At the end of the thesis we will present an answer for every research question, as result of our research.

2. Definitions and Foundations

The foundations of this thesis are laid in Chapter 2. Longer sections highlight relevant concepts and references from models in a general sense, and from documents. Both have been major conceptual influences for this thesis.

On a more technical side, digital repository systems as software artifacts, hypertext, software engineering, semantic web, wikis, workflow management, and semantic annotation are described.

To establish communication among multiple systems, tools and data, metadata need to be integrated. The rapid growth of Internet resources and digital collections has been accompanied by a proliferation of metadata schemas, each of which has been designed based on the requirements of particular user communities, intended users, types of materials, subject domains, project needs, etc.

Repository technology that was originally created to manage metadata can be used to data and tool integration [17].

The digital repositories have different data models, different schema and different underlying software and there are major issues of data integration and management. In addition, the publishing of datasets as part of the digital object lifecycle process is becoming more evident and there are issues of provenance related to the dynamic nature of the datasets.

Repository management must be considered within the context of the wider integrated information environment. It must be viewed as a component of an organization’s content management systems (CMS) which will want to consider digital resource management alongside managing other assets such as web pages, staff and other various digital records, course administration, documentations, etc. This broader environment in which digital repositories co-exist in and exchange information with contains structured and unstructured data. It represents the organization’s valuable explicit knowledge [9].

The Digital Library (DL) provides wide access to a large, managed collection of information over an organized repository wealth of knowledge. Digital libraries provide high quality and well-organized information and they usually focus on categorizing and cataloguing resources. This system of information storage and dissemination has its limitations, though: they suffer from the ambiguity of natural language, neglecting the importance of metadata and not engaging users in the process of sharing knowledge.

As a special class of CMS, repositories are currently being used to manage metadata for tool and data integration. Repositories are ‘collections of digital objects’. The main characteristics of a digital repository, as Joint Information Systems Committee (JISC) emphasizes [9], are:

- the repository architecture should be flexible enough to manage content as well as metadata;
- the content is accessible through a basic set of services (i.e., put, get, search, etc.), and
- the repository must be sustainable and trusted, well-supported and well-managed.

Repository functions include document acquisition, safe storage and preservation, version control, finding known documents, and document presentation [11].
Using the repository specific terminology, **repository management** is a set of activities related to collecting, aggregating, ingesting, versioning and finding of digital objects. Related activities are curating, reporting and also longtime preserving.

These more technical topics are analyzed in depth in Chapter 3.

A fusion of classic digital repository and wiki works was the starting point of this work. The decoupling from classical digital repository is described in the last works. Later these ideas led the author to start deep research on semantic wikis and digital resource management, create the first article on Semantic Wiki related with digital resource management, and then to this thesis.

A relation to associated approaches can be found in the fields of Annotations, Social Semantic Web, Workflows and Interoperability, which are also introduced.

### 3. Literature Survey - Repositories, Digital Libraries and Content Management Models

**Chapter 3** looks into DRM use-cases, goes over to existing models and introduces a new collaborative digital repository management following the information life-cycle model. Literature analysis reveals further requirements. The existing DRM models of popular systems from the fields introduced in Chapter 2 are analyzed with the intention to find commonalities among them and gather structural features that should be present in future DRM systems. A similar analysis is performed for commonly used relations in systems that are often used for doing collaborative tasks, aka wikis.

Besides commonalities in interest, repository management is a highly customizable set of processes and practices, depending on the **repository system model**, the **repository content and destination** and the **community** who wants to use and derive benefit from it.

**Main models** for repositories and Digital Libraries are broadly divided in the following categories:

A. **formal model** – 5S (Streams, Spaces, Structures, Scenarios, Society) is the best-known formal repository and Digital library model [5].

B. **reference model** – Reference Model for an Open Archival Information System (OAIS) [22]. Derivative models from OAIS are:
   a. Three Tier Framework Reference Model, also known as Delos Model [1]; and
   b. CIDOC Conceptual Reference Model [2].

C. **semantic digital library model**. This model is based on OAIS model but allows more granularity in structure and end-uses level services provided [13].

These models are balancing two extremes: either the too much formalization of the 5S model, but very good description at the structure level, or the Delos Model, not very easy to be adopted by a small and medium community with low level of technical expertise.

These models are usually only guides for implementers, but cannot impose standards for mandatory services and protocols. Implementers can choose one of the aforementioned models, but usually they use a hybrid model. Recently, the most common hybrid model chosen by implementers is OAIS enhanced with semantic digital library model features, like Cornell University adopted for Fedora Commons.

The main drawback of these hybrid implementations is the repository management process. Frequently, this is an administrative decision and not a software feature. Sometimes, a third party optional implementation is adopted, that only partially addresses the problem. Mostly, these plugins are not available for the entire community.

A possible solution is to use a semantic repository system as backend.
Semantic repositories systems are still at their advent, even though there are several implementations already available. Pure semantic repositories systems are out of the box solutions. Such solutions require experts to configure and maintain. It fits better for the case of institutional repository development (academic, governmental).

Semantic repositories added some metadata management overhead, since they in turn rely on the evolution of ontologies [128]. Choosing a full-fledged semantic repository model is beyond the expertise of small and medium communities with low-level of technological expertise, they cannot understand ontologies very well. Thereby, such a community trying to create a DL for eLearning or Virtual Research Environment (VRE) can initially employ several experts to help on repository system configuration and setup, but there after needs an automatic protocol for repository management and maintenance and a transparent framework for Semantic Web technologies.

This current chapter presented an up-to-date study of Semantic wikis with emphasis on Semantic MediaWiki. We chose SMW because it presents a balanced solution for enabling semantics in wikis. We also presented a hybrid method for structuring wiki information, which combines semantic templates along with automatic metadata repository workflow management.

From the underlying technology point of view, wikis are applications profiles for a common model, Semantic MediaWiki model.

The added markup information is used by the semantic wiki engine to facilitate semantic search. To conduct searches, a special search language is used.

The impressive searching capability provided by semantic wiki engine can hardly be matched by any traditional wiki site, and all is the result of simple markup that users are willing to add.

At the end of this chapter, several use-cases and applications using SMW have been mentioned. The following chapter will present a now model for collaborative repository management, architectures and methods summarized here as application profiles:

- An online authoring and publishing platform [7]
- Usage of MediaWiki templates as white-space invariant for involving non-technical users in the entire process of digital content management [1]
- Usage of the Semantic MediaWiki model for metadata management inside wikis [24].


In Chapter 4 the model solution is presented. The chapter starts with in depth analyze of social factors in DRM from which first requirements are derived. The DRM representation models are used to derive repository requirements. The chapter presents later a concise requirements table. The solution consists of a simple collaborative management model for digital repository management. This model represents content of varying granularity as well as semantic links among them and is accompanied by a workflow ontology that contains the commonly used operations for metadata integration tasks. It derives and explains a hierarchy of relationship types, suitable to describe digital repository management models.

A bidirectional reference model developed for collaborative repository management. The new created model reconsiders parts of the main models from literature and enhances their flexibility by additional features inspired from Mass Collaborative systems principles. This model encapsulates different aspects of reference model logic: the application ontology model (AO), the workflow ontology model (WO), and the repository ontology model (RO). The reference model is decoupled from the repository content model, covering mostly the problem of metadata integration regardless of the specific storage environment.
The application ontology defines precise semantics for repository management activities, while workflows are mainly designed to orchestrate them.

The idea started from the general model using workflow ontology for this thesis approach.

This model refines the levels of abstractions and in which workflows can be defined, [173] introduces a middle layer of mapping, having social software model as support. This is the level of abstraction that support for workflow management, monitoring and control. Rather than a typical data-driven workflow, the model is adopting a simple hybrid data-flow language with limited transactional control constructs. This approach will stand the best chance of being interoperable with most tools and frameworks for fulfill our purposes.

This idea consists in taking elements of the repository ontology, as they are defined in 5Ss model. These elements are transferred to “Process” class as building blocks (i.e. tasks) of workflows, and using the repository vocabulary as parameters of the processes (i.e. digital object, collections). Processes are executed using on bespoke toolset applications (ranging from wiki editor to custom built applications) to execute the whole protocol.

We indirectly use a standard available in the workflow community: BPMN notation and its XPDL serialization trough Semantic MediaWiki process editor extension [4], which in turn implements a plugins to Oryx web-based editor[

This extension builds on the capability to query for semantic properties which is provided by Semantic MediaWiki, and displays these query results as a process graph. It adds a special query printer to SMW which transforms the result of a special process-property query into the GraphViz DOT language and returns the corresponding process graph created with DOT. A semantic approach for workflow formalization, complementary with this is using the concept of Community Processes modelling notation and it is derived from Petri Nets.

In this approach we take the formalized processes in RDF, transform them into simple Petri Nets and use them in a process execution engine. Users can modify either Wiki pages containing templates or the process model.

For the workflow description, we rely on the SMW feature to form a light-weight ontology given by user created semantic wiki pages. We combine this feature with semantic wiki templates to represent the workflow description. Every activity can be represented by a semantic wiki page. In this way, the entire ontology can be split into components.

Then, is presented a reference architecture based on the Mediator pattern, that reflects the design and cover the requirements of reference model and also capture metadata information gathered over time.

5. Reference Architecture and Implementations

Chapter 5 presents concrete architectures and different tools based on the CDRM model. The Collaborative Ingestion Service for Fedora Common Repository is a complex service using a wiki for repository management, on which several other tools have been built by third parties. The Collaborative Integration Tool is used further for heterogeneous data integration into a semantic wiki. These specific bespoke wiki-based tools are also presented. This chapter specifies use-cases and implementation constraints for each of which concrete architecture.

The first approach assumes the existence of external multimedia sources on the provider side, which induces heterogeneity in data sources and intended usage. Tools, data and heterogeneity of data sources require an integrative and collaborative environment.

The second approach assumes the existence of media in the wiki (born & bred). A SMW approach fits better in this case and allows reducing the number of layers by using the built-in

1 Oryx web-based Business Process Editor http://code.google.com/p/oryx-editor/
SMW technologies. Moreover, from the user side, the identification keys created in the SMW can be used independently from the wiki where they are stored [24].

The third approach represents work in progress. The VRE might fit well in the use-cases of visual arts research, history and biology. This model facilitates collaboration amongst researchers and research teams providing them with more effective means of collaboratively collecting, manipulating and managing data, as well as collaborative knowledge creation.

The comparison among three architectures showed that all three are in place and fulfill their requirements, depending on the use-cases and the problem’s context.

Additional tool for semantic image annotation is presented also in this chapter.

6. Framework Evaluation

Next, different kinds of evaluation are presented in Chapter 6. A part of the evaluation is the rational analysis of fulfilled criteria taking in consideration classical models for digital repository management. Furthermore, an original method for evaluation, using a semantic wiki as support and ingestion user logs, is presented.

We followed a data-driven method for evaluating the framework result.

Firstly, we transformed every wiki page revision into a separate semantic wiki page, to make the data available for analysis and querying.

The tool also search for the corresponding collection ingests from the first survey to compare the results.

The results have been assessed and compared in various ways with a classic approach for metadata repository management.

We have used the results from two surveys, analyzing the error rate and warning rate, globally and per collections. The conclusion is that the second survey method fits a collaborative environment better. The feedback is useful only if the collaboration is still available. Further enhancements might consider also a feedback to automatic wiki tools (bots).

For all considered aspects, the evaluation demonstrates significant improvements that are achieved by an application of Semantic MediaWiki for digital repository management.

In other words, we prove that user-centric approaches are benefic with the respect of the following points:

(i) saving of task processing time,
(ii) facilitating and improving the learning curve with collaborative facilities,
(iii) improving usability,
(iv) providing an overview of the process including status changes at run-time,
(v) increase the degree of user’s collaboration,
(vi) increasing metadata quality, and
(vii) increasing user satisfaction.

7. Conclusions and Outlook

In Chapter 7, the conclusions along with a number of future research themes are identified and presented. Finally, the main contributions of this thesis are summarized.

The main goal of the thesis is improving the quality of metadata management on the Web in general and thus gaining advantages from better tools, better collaborative environments and increasing the availability of semantic web technologies, e.g., ontologies by providing usable methods for non-expert users. Instead of aiming for the best methods that tell us if a model for
repository management is good, we settled for the best approach for a context given. We have shown that online collaborative environments help to lower the maintenance costs for repository management by providing methods that point out possible errors well in advance. The results of these evaluations are presented, indicating the usefulness of the framework. After the evaluation, it should be clear that these methods, either alone or in combination, can guarantee the best result only along with community involvement.

Achievements
The result of this thesis is a reference model and three comprehensive frameworks for collaborative repository management. The frameworks organize methods and tools in three dimensions: technologic (engineering methods, technologies and technical agreements), social and organizational (collaborative, contextual, adaptability, conciseness, consistency, and organizational fitness) and semantic aspects (vocabulary, syntax, structure and representation). We added a number of new techniques to the community toolbox aiming to build a digital repository, to manage it and to integrate data from heterogeneous sources.

We have proved that with little human and material resources, the frameworks can support the infrastructure of a digital repository and VREs.

The frameworks in this thesis are also a novelty in as far as they put some emphasis on the evaluation of the collaborative environments from the user perspective.

To sum up, we will review the initial several research questions and requirements.

The main challenge for the repository management is to build repository content and deliver the benefits of repositories without burdening content creators and end-users with additional process. Our approaches fulfill this by using semi-automatic or automatic workflow management.

<table>
<thead>
<tr>
<th>Research question</th>
<th>Thesis answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which factors influence requirements and benefits in metadata repository management?</td>
<td>User’s requirements and requirements for a semantic content model repository.</td>
</tr>
<tr>
<td>What is a suitable model to represent and use artifacts in different degrees of formalization in a uniform fashion?</td>
<td>Using a semantic wiki over the whole aspect of knowledge acquisition and shared repository management. Extend wiki’s functionalities to fulfill user’s requests.</td>
</tr>
<tr>
<td>Which methods are suitable to encourage user’s participation in the whole process of digital content management?</td>
<td>Collaborative methods. Give a consistent feedback and rely on crowdsourcing. Choose the supporting technologies to achieve this, adopt noninvasive methods, help user to “snap to grid”. Believe in consensus choice and agreements. Relieve user of burdening tasks.</td>
</tr>
<tr>
<td>How to build a Web-based integration layer for existing tools and application in the digital content management area? How to deal with low level of technological expertise of providers and end-users and high levels of functional and non-functional requirements that the final framework should fulfill?</td>
<td>Use a wiki as infrastructure layer. Shift the power from the developer to the user. Adopt automatic methods for repository management. Adopt semantic wiki and all these environments have the best: agreements. Bear in mind that technologies are changing, agreements last forever.</td>
</tr>
</tbody>
</table>
This will not only improve the availability and usefulness of digital resources on the Web, but also point out a path to reconcile the two research streams on linked data and digital resource management, highlighting the mutual benefit they can gain from each other.

Contributions

The main contributions of this thesis are:

A. Comprehensive analysis of Digital Repository Management (DRM) and Semantic MediaWiki. Firstly, we describe a model-oriented digital repository and digital library implementation on existing open source software, which identifies the main underlying reference models and the inherent boundaries and limitations. Second, we study the Semantic Wikis model and implementations, which identifies features, functionalities and approaches offered by recent systems. Thirdly, we describe the feature-oriented studies on (Semantic) MediaWiki we performed. A comparative study with other Content Management Systems has been conducted also.

B. A reference model which has been introduced further. The reference model is constructed in an abstract manner, covering requirements from different use cases and implied workflows. Following the digital object life-cycle, the model described (i) phases, sub-processes, and tasks, (ii) connecting transitions, (iii) approaches to specific tasks, and (vi) forms of data generated and processed by proposed approaches.

C. Introduction of three separate frameworks for digital repository management. The first architecture uses MediaWiki features for heterogeneous metadata harvesting, collection and aggregation in an external repository based on Fedora Commons content model. The second architecture uses Semantic MediaWiki features for to semantically enable the metadata generated by automatic tool used to create identification keys inside wikis. The third architecture uses Semantic MediaWiki features to allow researchers to collaboratively annotate the corpora. These annotations are designed to create instances of domain ontologies and their related properties.

D. Introduction of a high-level ontology for collaborative repository management

E. Implementation of the theoretical results in practical systems making the results of this thesis accessible to the user. The frameworks consider the context in which they should work. Therefore, the first implementation assumes the existence of media on the data provider side. A MW approach fits better in this case and it allows for using an external metadata repository and several Java-based tools for repository management. The second implementation assumes the existence of media on MediaWiki. A SMW approach fits better in this case and allows using the shared wiki metadata repository. A jQuery extension and Java-based set tools complete the integration. The third implementation represents a virtual research environment and is used in history of education use-case. An annotation tool for image annotation, SMW based, extends SMW with new features.

F. A separate tool that deals with Intellectual Property Rights issues. MediaWiki requires careful approach for media content. The license and license statements are explicitly required in “File” namespace. Usually, a further media processing take place in MediaWiki. These derivatives works should be protected at ingest time (upload). This separate tool is uploading media in MediaWiki shared repository, attaches IPR metadata and a backlink to the original resource, on data provider server.

G. Evaluation of the framework and the implementation using Semantic MediaWiki against the SoSI model of interoperability has proved framework usability. User-centric approaches proved to be benefic with the respect of the following points: processing time saving, making an easier and improved learning curve with collaborative facilities, improving usability, providing an overview of the process, including status changes at run-time, increasing the degree of user’s collaboration and mostly increasing metadata quality.

The thesis claims to address the following main challenges in digital repository management: to build up repository content and deliver the benefits of repositories without burdening
content creators and end-users with additional processes. Empowering everyday user with all the prerogatives of the metadata administrator only restores a natural relationship between digital content and those who created it.

**Technological Achievements**

Technological solutions can be subdivided into the following sub-ordinate targets: Metadata Repository Management Platform Definition, Metadata Repository Management Tools, Automatic Workflow Execution, Workflow Visualization, Explicit User Feedback, and Process Definition.

The chapter continues by presenting a number of approaches related to the different parts of the solution.

**Further questions**

Some newer semantic digital library system and repositories take into account the fundamental role of discourse and agreement between different stakeholders. They identify the “shared-ness” of the resources (digital resources but also ontologies) as an explicit value and provide processes to guarantee that this shared-ness is achieved. Even though we do not regard shared-ness as an aim of its own, it is an integral part of our conceptualization. Shared-ness is an aspect of semantics and touches on a number of criteria: accuracy, clarity, and organizational fitness.

Future research would aim to a deeper compounding of semantic technologies with social approaches. We propose two directions of professional development:

- Contributing to improved trust and shared-ness in metadata management by recommendation systems and metadata provenance tracking applied to linked data also;
- Developing methods and tools for bootstrapping linked data in order to discover new, free and sharable digital resources.

Inside the same group of users, using informal or semi-formalized agreements and maintaining the content themselves, the emphasis is shifted on **trust** and **mutual caring**.
References


