Schindler, Christoph; Veja, Cornelia; Rittberger, Marc; Vrandecic, Denny

How to teach digital library data to swim into research

Chiara, G. [Hrsg.]; Ngonga Ngomo, A.-C. [Hrsg.]; Lindstaedt, S. [Hrsg.]; Pellegrini.T. [Hrsg.]: Proceedings of I-SEMANTICS 2011: 7th International Conference on Semantic Systems, Sept. 7-9, 2011, Graz, Austria. New York : ACM International Conference Proceedings Series 2011, S. 142-149

Quellenangabe/ Reference:
Schindler, Christoph; Veja, Cornelia; Rittberger, Marc; Vrandecic, Denny: How to teach digital library data to swim into research - In: Chiara, G. [Hrsg.]; Ngonga Ngomo, A.-C. [Hrsg.]; Lindstaedt, S. [Hrsg.]; Pellegrini.T. [Hrsg.]: Proceedings of I-SEMANTICS 2011: 7th International Conference on Semantic Systems, Sept. 7-9, 2011, Graz, Austria. New York : ACM International Conference Proceedings Series 2011, S. 142-149 - URN: urn:nbn:de:0111-dipfdocs-66297 - DOI: 10.25657/02:6629

https://nbn-resolving.org/urn:nbn:de:0111-dipfdocs-66297
https://doi.org/10.25657/02:6629

How to Teach Digital Library Data to Swim into Research

Christoph Schindler
German Institute for International Educational Research (DIPF)
D-60487 Frankfurt am Main, Germany
schindler@dipf.de

Cornelia Veja
Telecommunications Department, Technical University of Cluj-Napoca
Cluj 400027, Romania
cornelia.veja@com.utcluj.ro

Marc Rittberger
German Institute for International Educational Research (DIPF)
D-60487 Frankfurt am Main, Germany
rittberger@dipf.de

Denny Vrandečić
Karlsruhe Institute of Technology (KIT)
D-76128 Karlsruhe, Germany
denny.vrandecic@kit.edu

ABSTRACT

Virtual research environments (VREs) aim to enhance research practice and have been identified as drivers for changes in libraries. This paper argues that VREs in combination with Semantic Web technologies offer a range of possibilities to align research with library practices. This main claim of the article is exemplified by a metadata integration process of bibliographic data from libraries to a VRE which is based on Semantic MediaWiki. The integration process rests on three pillars: MediaWiki as a web-based repository, Semantic MediaWiki annotation mechanisms, and semi-automatic workflow management for the integration of digital resources. Thereby, needs of scholarly research practices and capacities for interactions are taken into account. The integration process is part of the design of Semantic MediaWiki for Collaborative Corpora Analysis (SMW-CorA) which uses a concrete research project in the history of education as a reference point for an infrastructural distribution. Semantic MediaWiki thus provides a light-weight environment offering a framework for re-using heterogeneous resources and a flexible collaborative way of conducting research.

Categories and Subject Descriptors

D.2.12 [Interoperability]: Data mapping; D.2.10 [Design]: Methodologies; E.2 [Data Storage Representations]: Linked representations; H.1.2 [User/Machine Systems]: Human factors; H.2.5 [Heterogeneous Databases]: Data translation; H.2.8 [Database Applications]: Scientific databases; H.3.7 [Digital Libraries]; H.4.1 [Office Automation]: Workflow management; H.5.3 [Group and Organization Interfaces]: Computer-supported cooperative work; J.5 [Arts and Humanities]

General Terms

Management, Design, Human Factors, Standardization

Keywords

Virtual research environment, Semantic Web, data integration, e-humanities, Semantic MediaWiki

I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria

1. INTRODUCTION

Virtual research environments (VREs) are connected to the aims of e-science, e-humanities, cyberinfrastructure, or e-research to enhance research practice using the possibilities of networked technologies, distributed resources and computational power.[1] Lately, libraries and archives have digitalized a range of their objects that can be addressed from VREs by encompassing different tools to conduct research. Nevertheless, the VRE community identifies interoperability as one of its main challenges and demands for light-weight tools [2][5]. This paper argues that VREs combined with Semantic Web technologies could provide the missing link between libraries and research by taking into account their heterogeneous practices and styles of using data and schemata. This argument is exemplified by the integration process of bibliographic data and metadata of a digital library to a VRE while offering capacities for interactions of the different stakeholders.

The integration process is part of the VRE project "Semantic MediaWiki for Collaborative Corpora Analysis" (SMW-CorA)[2] which aims to adjust Semantic MediaWiki (SMW) for research: enabling researchers to create a research corpus, enrich, annotate and analyze digital objects. The SMW-.CorA project takes insights from previous failures into account by balancing the fallacies of false abstractions and concreteness. The project adjusts its VRE in different ways: a) by following concrete

[1] The term virtual research environment was established in the UK's JISC funding program and has since spread towards a wide range of European countries and beyond. For a more detailed introduction to VREs see [1][2][3][4].

[2] This project is funded by the German Research Foundation (DFG) entitled: "Entwicklung einer Virtuellen Forschungsumgebung für die Historische Bildungsforschung mit Semantischer Wiki-Technologie - Semantic MediaWiki for Collaborative Corpora Analysis" in the domain of "Scientific Library Services and Information Systems" (LIS). It is realized in a co-operation between the German Institute for International Educational Research (DIPF), the Karlsruhe Institute of Technology (KIT), the Library for Research on Educational History (BBF) and researchers of educational history.

research practices, b) by building a community, c) by offering an infrastructural distribution to other scholarly communities, and d) by taking into account capacities of interactions between libraries, research and the Semantic Web. Considering this, the stabilization of the VREs and their interconnected infrastructures challenges the interlocking stakeholders and their communities to align their heterogeneous practices, tools and objects for possible interactions.

The starting point of the design of SMW-CorA is a concrete research project in the field of the history of education. The research project aims to analyze educational lexica spanning nearly 200 years, focusing on networked relationships in the domain of education. Therein, bibliographic metadata are used for bibliometric analysis; they are enriched and annotated by the educational researchers in a collaborative wayfor qualitative analysis.

This article summarizes the approaches, technologies and engineering issues that were addressed for the data and metadata integration to the collaborative VRE. The Section 2 discusses recent developments and affordances of VREs and digital libraries related to research practices in the humanities. The objectives of the SMW-CorA project are described, the features of MediaWiki (MW) and SMW are elucidated, followed by a discussion of different approaches for the integration of metadata into SMW. Section 3 describes design approaches and methods of the SMW-CorA project. Section 4 explains the concrete realization of the data integration process with heterogeneous requirements from different stakeholders and a schemata mapping. The article closes with a conclusion and outlook (see Section 5).

2. SETTING THE OBJECTIVES

2.1 Linking VREs, Digital Libraries, and Scholarly Research Practices

A wide range of papers identify the possibilities of VREs in the humanities [3][2][8] or particularly for historical research [9]. Nevertheless, interoperability is identified as one main challenge [2]. VRE tools should be able to interact with other tools used in research as well as archives and libraries and their objects.

It should be pointed out that scholarly research practices in the humanities have adjusted the e-science agenda in one main aspect: While the so called “data deluge” [7], the amount of new ways of gathering research data, was seen as a major driving force in scientific research, complex relationships contrasted to this the so called “complexity deluge” [10] in the humanities. These two agendas outline the scope of challenges and possibilities of VREs and Semantic Web technologies focused in the SMW-CorA project, i.e. how to offer researchers digital objects for their study while giving them the capacities to re-arrange these and create new relations, properties and objects. By discussing several problems and objectives of the involved communities, their possible alignment is outlined in the following.

Recently, the library and infrastructure communities have become aware of the potential of VREs for their assignments [14] and they regard research data as a new major field of their activities. Especially in the humanities, the need is seen to offer cultural artifacts as research objects in a digital way whether these are texts, images, or audio recordings. In this context, data management and curation to facilitate data sharing and re-use are seen as a major new aspect [16]. Furthermore, new possibilities for re-using the scholarly value chain are identified while calling for a main challenge in e-research: building relations between research data and further artifacts [6].

The digital library community has started to use Semantic Web technologies and to align their data and technologies[3] Cultural artifacts are increasingly offered in digital formats, corresponding to Semantic Web standards. Hence, a body of resources is available for research, but adequate tools for conducting such research are still lacking.

Some VREs[4] have started to use Semantic Web technologies to reach a higher level of interoperability. However, one of the main challenges of e-research - linking research data with further artifacts and offering them for re-use - is only captured in a limited way. The VRE and information infrastructure communities try to catch the complexities of research while seeing research lifecycles and scholarly primitives [15] as heuristics to articulate requirements from research practice. Nevertheless, in the humanities the fuzziness, incompleteness, overlapping and distribution of data and research practices are highlighted [10] and described as nontransparent or *invisible infrastructures* [11]. This pushes the problem of interoperability beyond using standardized protocols and vocabularies.

The digital humanities and e-humanities communities have recently addressed the interoperability problem by discussing the design of digital editions. Their aim to represent research objects like books comprehensively has lost ground against a more pragmatic orientation: In addition to a digital representation of an object, the capabilities of doing things interactively are focused. In this respect the challenge for designers of editions has been articulated in a dramatized way: “We had better start swimming, or we’ll sink like a stone”[5] [17].

The SMW-CorA project follows this approach while focusing the capacities for interaction between research, libraries and the Semantic Web. This involves a translation from a library to a research object which sets new requirements beyond schema matching: a researcher’s capability to create new entities, relationships, re-arrange their own classifications on the fly and request these entities. Additionally, the flow of enriched digital objects back to libraries has to be taken into account.

2.2 Semantic MediaWiki for Collaborative Corpora Analysis

Instead of designing a VRE in an isolated way, SMW-CorA uses a participatory design approach which aims to empower relevant research communities to engage actively in the design process. Therefore the SMW-CorA project is aligned to a concrete research project and its research community.[6]

[3] A list of library related initiatives is available at http://ckan.net/tag/library

[4] Examples of VREs which use Semantic Web technologies are www.ourspaces.net/, www.myexperiment.org, and Wikinger [13].

[5] This expressive illustration of the situation of libraries in a networked world inspired the authors to use this metaphor as a leitmotif and use it in the article’s title.

[6] Therefore the project has associated co-operators in the research community of history of education as well as the Research Library for the History of Education (BBF) www.bbf.dipf.de/.

A group of researchers of history of education distributed across Germany articulated their concrete need for a collaborative environment for conducting their research. The group intends to analyze the disciplinary landscape of educational science by following networked relations of the lexica covering a period of nearly 200 years. To do this in a collaborative way, the educational researchers are interested in a research environment which provides bibliographic data of the library (editors, authors, date of publication …). These formalized data serve as a basis for further enrichment activities like adding new entities and annotating the content in order to pursue the research interests.

The collection of lexica that is of interest here is mainly available at the digital library Scripta Paedagogica Online (SPO),[7] hosted by the Library for the History of Education (BBF) at the DIPF which has indexed the lexica and rendered them accessible online as image files. The corpus contains a total amount of nearly 22,000 articles and more than 20 lexica. Each lexicon is bibliographically described as a collected edition in the library database allegro-C.[8] Therein, three levels of entities, their properties and relations are formalized (lexicon, volume, article). Scripta Paedagogica Online accesses this database and creates via Perl scripts a front-end which connects the bibliographic data with the images of the lexica pages. Since the Z39.50 protocol contains restrictions for data exchange, currently Allegro-XML files are used for the integration process. An OAI-interface will be used in the near future.

While following the educational research project in practice, it is possible to articulate the requisite capacities for interaction. Nevertheless, to avoid designing an isolated application for a particular research group, the alignment to this concrete research practice is seen as a starting point for an infrastructural distribution to other research contexts. Thereby, heterogeneous local practices and actors in research and libraries are taken into account, which is seen as a main requirement for a sustainable stabilization of standardization processes and infrastructures [21][22][23][24]. Furthermore, this approach offers the possibility to identify re-usable parts of the scholarly value chains like data, classifications or instruments of SMW-CorA.

Currently, the following needs are articulated in the SMW-CorA project: data integration, adding new entities, annotating and re-arranging of data and metadata, monitoring and quality checking, analyzing and visualizing results, writing articles, exporting research data and its metadata. These issues address the continuous challenge of the SMW-CorA project designing the VRE in alignment with research practice. The data integration process exemplifies the first step of the project while it has to address further implications of the research process in the future. MediaWiki and SMW already offer a range of possibilities to enhance these activities but need to be aligned with research requirements.

2.3 MediaWiki and Semantic MediaWiki

Several VREs were analyzed on the basis of heuristics[9] and SMW was chosen as a basic environment, owing to its collaborative aspect inherited from MediaWiki and the wide range of characteristics of SMW and further extensions:

- Flexible, mature software product, open source and wide acceptance in the humanities as a software supporting Wikipedia.
- Allows for collaborations on the level of Semantic Web technologies [20] permitting data interlinking and schemata re-arrangement.
- Possible re-use of data and schemata while using ontologies and controlled vocabularies.
- Offers extensions for import, export (CSV, RDF, BibTeX) and analysis (visualizing, requests)
- Interoperability with libraries (OAI[10], JSON), with other Semantic Web standards (RDF) and research tools (XML[11]).
- Customizable user interfaces (forms, templates, widgets) and re-usable parts of the research environment.

In order to fulfill researchers' needs, SMW has to facilitate additional aspects. Besides usability problems of MediaWiki, the main technological drawback concerns the restricted possibilities for annotating at the level of texts or media objects. Nevertheless MediaWiki and SMW offer a high degree of flexibility and many possibilities to enhance research. As these capacities are invisible to researchers, they have to be aligned to their needs. Therefore unified interfaces for specific research tasks like analyzing with inline queries against the wiki semantic properties has to be put into practice. To fulfill this, SMW has to adopt the appropriate design for becoming a research tool and offer re-usable parts.

MediaWiki provides an object oriented document storage system for the data integration described in this article wherein wiki pages are the elementary entities. This allows storing, retrieving and describing digital objects as wiki pages in a collaborative way. Additionally, the namespace mechanism allows for allocating wiki pages in the storage system and offers to address a range of functionalities. Examples for this are digital media objects (binary plus metadata) stored in the "File" namespace, programming and rich text parts in the "Template" namespace and a hierarchical lightweight schema development in the "Category" namespace.

SMW expands the MediaWiki functionalities by adding semantic technologies and by providing a higher granularity of the storage system through properties of pages and their relatedness. SMW enables use of a wiki page as an instance of a class or the subject of an RDF triple. Additionally, SMW reuses the "Category" namespace to define ontology classes and allows the declaration of sub-class and sub-property relations. From the functional point of view, SMW can be used to extend the MediaWiki template mechanism by offering to process semantic properties as parameters of the template. On this basis, semantic forms[12] facilitate user input by hiding the complex syntax of the semantic templates.

[7] The lexica are available at http://bbf.dipf.de/digitale-bbf/scripta-paedagogica-online/digitalisierte-nachschlagewerke.

[8] See http://www.allegro-c.de/

[9] The following heuristics were assessed while planning the project: range of available tools, participation and collaboration, interoperability, sustainability and openness, possible organizational embeddedness.

[10] See the MediaWiki extension OAIRepository at http://www.mediawiki.org/wiki/Extension:OAIRepository

[11] MediaWiki offers an XML export and further extensions http://meta.wikimedia.org/wiki/Help:Export

[12] See http://www.mediawiki.org/wiki/Extension:Semantic_Forms

By using Semantic MediaWiki (SMW), semantically defined template schemata can be annotated using standard ontologies (Dublin Core, FOAF, SIOC). This allows direct semantic metadata search and inference on as well as exposure in the OWL/RDF format. Semantic queries can be embedded, creating dynamic content in wiki pages. Additionally, the Semantic Web technology offers an interoperable platform for sharing and re-using research resources.

2.4 Metadata Integration and SMW

While a comparison of data integration into MediaWiki and SMW as a framework for VRE[13] lacks concrete comparable realizations, Virtual learning environments (VLEs) [2] deal with similar issues and drawbacks in the field of data integration. This offers the opportunity to take their exchange practices into account while addressing the specific needs of research environments.

A VLE project using SMW as support for semantic technologies is UNESCO's OceanTeacher.[14] This is a digital library of knowledge related to oceanographic data and information management. It includes texts, images, objects (PDF and Word documents, software installation packages, audio and video files) and links to web pages and objects on other web sites. The project builds an e-learning platform in the domain. The marine metadata are automatically collected from sensors and translated into SMW properties. Afterwards, the metadata are manually verified and validated. While the project imports different documents, it does not integrate heterogeneous metadata which are of interest in the SMW-CorA project.

The biodiversity project KeyToNature[15] provides a collaborative environment which uses a new method for metadata collection, web-based repository interface, metadata repository management and search tools. The project is based on MediaWiki and SMW mechanisms for addressing the issues related to interoperability in a collaborative space. In KeyToNature, the metadata are collected from data providers using MediaWiki templates and submitted to the wiki by a push mechanism. The automatic integration process triggered by data providers targets a Fedora Commons based central metadata repository [28]. Because this project deals with similar problems regarding the integration of metadata, the tools and workflows are re-used and adjusted in the SMW-CorA project.

3. APPROACH AND METHODS

The design approaches described in Section 2.2 were taken into account while creating the integration process. Additionally these are combined with iterative, agile computing[16] aiming to get fast feedback on the technical and organizational level of implementation and framework re-use.

The matching mechanism of the integration process between the library, SMW, and MediaWiki was established iteratively. Three main interlacing steps are carried out:

1. Identifying the relevant entities of the digital library (collections, digital objects, metadata catalogs, and metadata structures) and its capacities for doing research within SMW-CorA.
2. Engineering of light-weight ontology for the library objects matching with MediaWiki and SMW and offering a user interface for research.
3. Configuring a workflow manager for the integration process that addresses elements, creates templates, and generates instances.

To align and evaluate the integration process with research needs and practices of researchers, different methods of gathering data were used: talks, site visits, group discussions, document analysis, and rapid prototyping. Besides the usage of SMW as a prototype, drafts of the user interface were created to articulate future aspects of the VRE. These approaches allowed an early start of the explorative part of the educational project and an iterative feedback of the educational researchers while designing the VRE. The experiences of the heterogeneous stakeholders are presented, discussed and documented on a wiki-based platform.

4. FROM VISION TO PRACTICE: SWIMMING INTO RESEARCH

4.1 Data and Metadata Integration into SMW-CorA

The data integration process presented here is mainly formed by the alignment between the digital library, research practices and Semantic Web technologies. While using MediaWiki and SMW, options as well as limits to the interactions between the stakeholders are focused in this article. For this purpose the data integration and the transfer back to the library and to the Semantic Web were taken into account as far as it is possible at this stage of the project (see Figure 1). Further aspects of the research life cycle and of the educational researchers' needs like annotating the content have to be articulated by realizing the next project phases of SMW-CorA and of the educational research project.

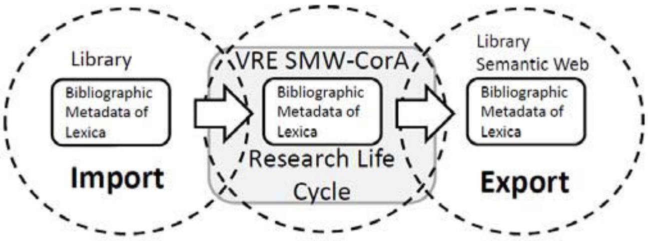
Library
Bibliographic Metadata of Lexica
Import
VRE SMW-CorA
Bibliographic Metadata of Lexica
Research Life Cycle
Library Semantic Web
Bibliographic Metadata of Lexica
Export

Figure 1. Metadata Integration Process.

The life cycle of digital content can be split up into 6 primary phases [26]: create, update, publish, translate, archive and retrieve. In this respect, SMW could be regarded as a repository which collects and aggregates heterogeneous metadata and media resources. This integration of metadata into SMW partially covers the first three steps of the data life cycle of digital content.

The used semi-automatic workflow for the metadata integration which is described below is built on two mechanisms, i.e.

[13] The authors have researched for comparable SMW-VREs several times without a concrete result.

[14] See http://classroom.oceanteacher.org

[15] See www.KeyToNature.eu

[16] The agile computing approach in the context of e-research and social network sites is discussed in [19].

MediaWiki templates and an extended flexible model of a workflow ontology. This workflow is based on the theoretical work of [30] and the flexible workflow model exposed in [28]. The model developed in the framework of the KeyToNature project is extended and adapted in this project to make use of semantic templates.

The main component fulfilling the data integration process is a Java-based tool which consists of other tools for transforming the XML and images and uploading the digital resources into the wiki [26]. These are independent reusable components and open-source technologies [27]. Triggered by a general XML configuration file stored on a wiki page, these tools can run either in combination or as stand-alone systems. The data transfer between the tools is based on XML format. The tools' orchestration depends on the workflow template which in turn is mapped on the specified wiki metadata template for harvesting

This mapping is specified by means of a general configuration wiki page in the "MediaWiki" special namespace. This wiki page contains parameters and comments about the user interface and the functionality of the wiki. The integration tool harvests only wiki pages, having a lexicon file in XML format as attachment. For example, the following configuration parameter specifies the name of semantic properties provided to the XML file attachment:

- *attachParam = Wikitext Metadata Attachment*
  (Configuration parameter as semantic property referring to the lexicon in XML format)

This example specifies that wiki pages containing the semantic property *Wikitext Metadata Attachment* carry the last XML files with metadata for harvesting. The configuration wiki page can be different on another wiki. The only constraint required is that both have enabled a SMW extension.

The main processes in digital resources management imply collecting, aggregating and storing metadata. The workflow templates concerning this metadata management and the tools developed to achieve them are: *IntegrationLexiconTemplate*, *AchiveImageResouce* and *UploadImageToWiki*. The latter two templates are constrained as follows: Any *UploadImageToWiki* instance should follow only after the corresponding *AchiveImageResouce* has succeeded.

Any workflow instance is actually mapped on a lexicon entity. A workflow instance can perform the following actions:

a. The metadata providers (i.e. Digital Libraries) submit the metadata of their media in a specified form (metadata harvesting).
b. Data in XML format are queried by an automatic Java-based tool.
c. XML and XSLT transformations assisted by the metadata exchange agreement validate metadata and create the digital object in wiki pages.
d. Digital resources (images) are downloaded from a digital library and uploaded in the wiki media repository, having automatically created metadata specifying the license, copyright statement and copyright holder.
e. The metadata and/or media resources are stored in a wiki repository (Repository Object Architecture created as wiki pages).

The contribution of XML to implementing uniform data output format facilitates an automatic integration for similar library resources. Instead of creating various data transformation tools, XSLT transformation was applied. For an additional metadata provider only the adjustment of the XSLT is necessary. For heterogeneous syntactic and structural data integration, a hybrid strategy is considered: Global and Local as View (GLAV) [32] that can create a sight over sources by generating a view over global schema described by source descriptions. Because this strategy is not flexible enough, it is usually necessary to adapt the crosswalk method [33] for addressing interoperability. In order to enable flexible, dynamic mapping between complex metadata descriptions which mix elements from multiple domains, an application profile is created having the versatile support of SMW (Equivalent Properties, controlled vocabulary import[17], RDF export). Natural language for tools orchestration [31] is a possible option for future developments to hand the configuration of the workflow management over to the end users.

4.2 Identifying relevant Entities, Needs and Interactions

4.2.1 Identifying relevant Entities, Needs and Actions in Research

The educational research project and its usage of bibliographic data offer an adequate setting for identifying relevant entities and needed actions within SMW-CorA. This is seen as a starting point for articulating concrete needs of research practice. At this stage of the project, the educational researchers enunciated following the entities and actions for the research corpus:

- collections of lexica (BBF), their entities (lexicon, volume, article, image object) as well as metadata elements
- research interests in further properties, new entities (persons, institutions, etc.) and relations
- developing and re-arranging a classification schema on the fly, categorizing articles and coding the texts of articles

Besides the question of relevant entities and metadata elements, the researchers addressed specific kinds of requirements which have to be realized in the user interface. So far, the following research needs have been articulated:

- the possibility to browse in the book and the structure of the paper-based lexicon (table of content, previous and next page, list of authors)
- processing additional metadata to the entities (related with authors, articles)
- using understandable descriptions which enable computational processing
- interoperable formats for further tools (qualitative and quantitative analysis, bibliographic tools …)

4.2.2 Identifying relevant Needs and Actions of the Library

While establishing the integration process, the digital library articulated the need to gain feedback about outcomes of scholarly work like quality improvements of data. This offers the possibility to establish a workflow between the library and the educational research group. The collaborative mechanism is so far addressed and will be implemented with the OAI interface. Possible enhancements of the digital library services are:

- identification and correction of inaccurate data
- enrichment of entities with properties (persons, institutions)

[17] See http://semantic-mediawiki.org/wiki/Help:Ontology_import

- linking of own resources to further identifiers (for example GND[18], DBpedia[19]) and other resources

4.2.3 Identifying Relevant Entities for Interoperability and Re-Use

For a while, the research community in the history of education has studied disciplinary publication practices and published objects like scholarly articles and books.[20] This research interest correspond with developments in other research communities such as scholarly or science studies, bibliometrics, or network studies. All these research communities have in common that they use bibliographic data concerning authors, titles, references, affiliations and their relations from digital libraries or repositories as main research objects. The bibliographic data and tools for analyzing these data disregard main interoperability issues, which can be seen as a barrier for re-use. To offer these bibliographic data in a re-usable format like RDF and to create tools for analyzing these are capabilities for re-usable parts of the scholarly value chains in these communities.

4.3 Light-Weight Ontology for Mapping

4.3.1 Mapping Metadata

The iteratively articulated needs of researchers, digital libraries, and requirements of SMW and MW mechanisms have led to a lightweight mapping ontology aiming to offer a multi-layered research environment. Thereby, a mapping schema (XML/XSLT) is used which addresses the bibliographic entities and metadata of the digital library and integrates these into the main MediaWiki namespace. In so doing, the main entities of the library represent wiki pages while the page hierarchy of MW is used. The following mapping relations are used and exemplified in Figure 2:

- lexicon ***isEquivalent*** with Lexicon wiki page.
- volume ***isEquivalent*** with wiki sub-page
- article ***isEquivalent*** with a wiki sub-sub-page
- every image of an article ***isEquivalent*** with a wiki page in the "File " namespace

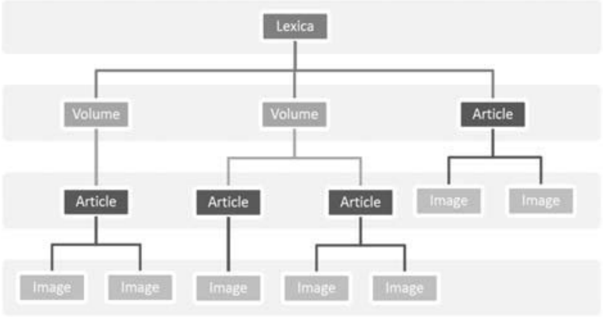

Figure 2. Library Objects integrated into the Wiki

The images of the articles are stored in the File namespace of MediaWiki, keeping the metadata of the created work and its scanned image separated. A property for the article page is created and connects the lemma with the image, offering the educational researchers direct access. Thereby, every semantic property represents a wiki page in the "property" namespace of SMW. Additionally, the mapping ontology addresses the metadata of the library objects and adds these as semantic properties to the main entities. This enables the educational researchers to create new entities and to insert formal or text-based descriptions into this new wiki page. Figure 3 visualizes main aspects of the matching mechanism:

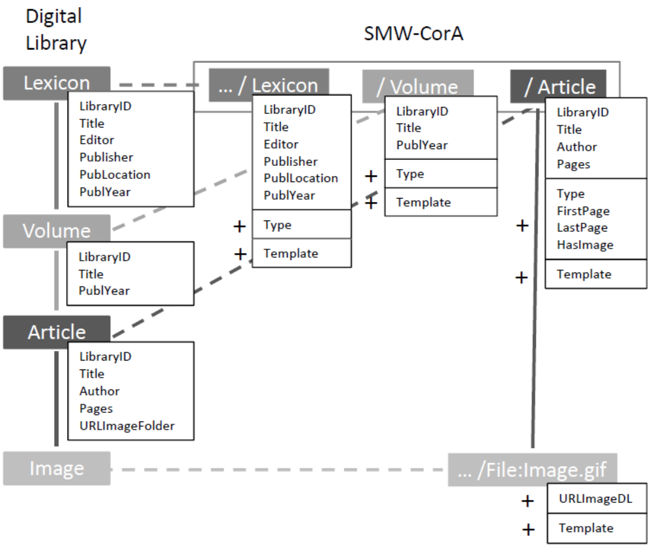

Figure 3. Mapping between DL Schema and SMW-CorA

The mapping schema addresses further articulated needs and relevant SMW features by processing new properties. For interacting with the research object and for offering detailed metadata for doing research, several additional semantic properties were created:

- firstPage, lastPage of an article, firstName and lastName
- isPartOf (volume); hasPart (article), hasImage
- Type (volume, article, reference)

To enable the digital library to address these entities and monitor the changes, their main identifier was integrated into SMW and the structure of the entities was maintained.

4.3.2 Addressing the User Interface

The researchers articulated specific needs to interact with the research object lexicon and its entities. Therefore, the matching schema addresses the MW system, the wiki syntax and wiki templates to create a user interface. As requested by the educational researchers, the possibilities for browsing articles' images and tables of contents were generated. Furthermore, the semantic forms combined with templates enable the researcher to change the value of properties like names of authors without using the wiki syntax. Additionally, the semantic forms extension offers the use of a controlled vocabulary of created classifications to reduce typing errors and regulate the annotation.

[18] The "Gemeinsame Normdatei (GND)" of the German National Library combines the existing authority files PND, SWD and GKD for persons, subject headings and corporate bodies. See http://www.d-nb.de/eng/standardisierung/normdateien/gnd.htm

[19] See http://dbpedia.org/

[20] Following the I-SEMANTICS conference, the community of researchers in educational history will hold a conference in its own right on the subject of networks and co-operations in education (http://web.fhnw.ch/plattformen/tagung-netzwerk/).

4.3.3 Interoperability, Re-Use and Semantic Web

To ensure interoperability with analyses and bibliographic external tools and inter-wiki metadata re-use, semantic technologies are used as a main platform. In the first step, low-granularity controlled vocabularies in Semantic Web standards (PRISM, SKOS, FOAF, DC) are used to simplify a RDF export to third parties. The elements of those vocabularies were integrated into the wiki using the SMW mechanism of importing vocabularies and ontologies. Thereby the added properties of the mapping between library metadata and SMW are also matched (i.e. prism:startingPage). To offer the traceability to the digital library, DC terms properties are used to address copyright aspects. The Figure 4 depicts a part of the Semantic MediaWiki RDF export:

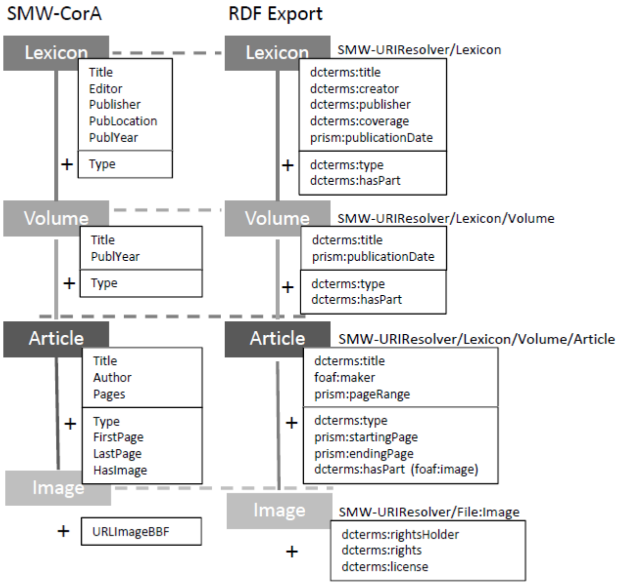

Figure 4. Controlled Vocabularies Mapping

5. CONCLUSION AND OUTLOOK

This paper argued for the capability of VREs and Semantic Web technologies to play an active role in the alignment processes between researchers and libraries. This was demonstrated by the example of a metadata integration process from a digital library to SMW-CorA to address research needs and a first mapping with Semantic Web technologies. Thereby, the respective possibilities of Social Software and Semantic Social Software to share and create knowledge in a collaborative way are used. SMW mechanisms allow for the integration of different digital resources, using the same syntax (template and semantic forms syntax) and offer the assistance of controlled vocabularies. The main achievements of the data integration process can be summarized as follows:

- Automatic metadata integration into SMW and beyond.
- Processing is extensible to a heterogeneous data integration process which requires only an attached workflow template and XSLT transformation.
- The approach is a layered candidate in achieving interoperability at the syntactic and semantic level with a low technological entry barrier.
- It allows for the integration of external tools and adaptation.
- It supports researchers in accomplishing their initial research step: annotating and categorizing articles and discussing these in a collaborative way.

This article contains the first findings of SMW-CorA and the translation processes from a library to a research object in a VRE. To stabilize the SMW-CorA as a VRE, further capacities for interactions between research data and research actions have to be realized. One main aspect that remains open is to interlink different levels of annotation and analysis of the research project and capture the adequate granularity and flexibility. One further aspect that needs to be addressed is to align SMW-CorA with further ways of doing research from other disciplines. To reach this aim, the VRE challenge will be: to teach linked data to swim into research.

6. ACKNOWLEDGMENTS

Part of this work has been carried out in the framework of the contract no. ECP-2006-EDU-410019 and co-financed from Social European Found – FSE – in the frame of POSDRU/6/1.5/S/5 project, and the other part in the framework of the German Research Foundation (DFG) project entitled: "Entwicklung einer Virtuellen Forschungsumgebung für die Historische Bildungsforschung mit Semantischer Wiki-Technologie - Semantic MediaWiki for Collaborative Corpora Analysis" (INST 367/5-1, INST 5580/1-1), in the domain of "Scientific Library Services and Information Systems" (LIS).

7. REFERENCES

[1] Fraser, M.A. 2005. Virtual research environments: overview and activity, Ariadne, 44.

[2] Voss, A. and Procter, R. 2009. Virtual research environments in scholarly work and communications. In*: Library Hi Tech*. 27, 2, 174-190. DOI = 10.1108/07378830910968146

[3] Neuroth, H., Aschenbrenner, A, Lohmeier, F. 2007. e-Humanities - eine virtuelle Forschungsumgebung für die Geistes-, Kultur- und Sozialwissenschaften. In: *Bibliothek. Forschung und Praxis*, 3, 272-279.

[4] Botte, A., Rittberger, M. and Schindler, C. 2011. Virtuelle Forschungsumgebungen. Wissenschaftspolitische Erwartungen, informationswissenschaftliche Forschungsfelder und Herausforderungen. In: Griesbaum, J., Mandl, T., Womser-Hacker, C. (Eds.): Information und Wissen: global, sozial und frei? Boizenburg : Hülsbusch, Schriften zur Informationswissenschaft, 58, 422-433

[5] Eccles, K., Schroeder, R., Meyermet al. 2009. The Future of e-Research Infrastructures. In: *Proceedings of the International Conference on e-Social Science*, Köln. Retrieved November 08, 2010 from http://www.merc.ac.uk/sites/default/files/events/conference//2009/papers/Eccles.pdf

[6] Borgman, C. L. 2008. Data, Disciplines, and Scholarly Publishing. In: *Learned Publishing*. 21. 29-38.

[7] Hey, T., Trefethen, A. 2003. The Data Deluge: An e-Science Perspective. In Berman, F., Fox, G. and Hey, T. (Eds*.). Grid Computing: Making the Global Infrastructure a Reality*. John Wiley & Sons.

[8] Neuroth, H., Jannidis, F., Rapp, A. and Lohmeier, F. 2009. Virtuelle Forschungsumgebungen für e-Humanities. Maßnahmen zur optimalen Unterstützung von Forschungsprozessen in den Geisteswissenschaften. In: *Bibliothek. Forschung und Praxis*, 2.

[9] Meyer, T. 2011. Virtuelle Forschungsumgebungen in der Geschichtswissenschaft – Lösungsansätze und Perspektiven. In*: LIBREAS.Library Ideas*, 1, 18.

[10] Dunn, S. 2009. Dealing with the complexity Deluge-VREs in the arts and humanities. In: Library Hi Tech. 27, 2, 205-216 DOI = http://dx.doi.org/10.1108/07378830910968164.

[11] Borgman, C. L. 2010 The Digital Future is Now: A Call to Action for the Humanities. In: *Digital Humanities Quaterly*. 3, 3. Available at: http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html. (30.04.2011).

[12] Blanke, T. and Dunn, S. 2009. Digital Humanities Quarterly Special Cluster on Arts and Humanities e-Science. In: *Digital Humanities Quarterly*, 3,4. http://digitalhumanities.org/dhq/vol/3/4/000079/000079.html (30.04.2011).

[13] Bröcker, L. and Paal, S. 2007. WIKINGER – Wiki Next Generation Enhanced Repositories. Paper presented at the *German e-Science Conference*.

[14] Brown, S. and Swan, A. 2007. Researchers' Use of Academic Libraries and their Services. A report commissioned by the Research Information Network and the Consortium of Research Libraries. Available at: http://www.rin.ac.uk/system/files/attachments/Researchers-libraries-services-report.pdf.

[15] Unsworth, J. 2000. Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? Paper presented at symposium, *Humanities Computing: Formal Methods, Experimental Practice*, May 13, King's College London. http://www.iath.virginia.edu/~jmu2m/Kings.5-00/primitives.html. (30.04.2011)

[16] Lynch, C. A. 2008. The Institutional Challenges of Cyberinfrastructure and E-Research.In: *EDUCAUSE Review*, 43, 6. http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume43/TheInstitutionalChallengesofCy/163264 (30.04.2011)

[17] Sperberg-McQueen, C.M. 2009. How to teach your edition how to swim. In: *Literary and Linguistic Computing*. 24,1.

[18] Warwick, C., M. Terras, P. Huntington, and N. Pappa. 2008. If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data. In: *Literary and Linguist Computing* 23, 85-102.

[19] Lin, Y., Poschen, M., Procter, R., Voss, A., Goble, C., Bhagat, J., De Roure, D., Cruickshank, D., Rouncefield, M. 2008. Agile Management: Strategies for Developing a Social Net-working Site for Scientists. In*: Oxford e-Research Centre Project Management Workshop*, 10-11 April 2008.

[20] Ankolekar, A., Krötzsch, M. Thanh Tran, D., Vrandecic, D. 2008. Die zwei Kulturen. In: Social Semantic Web. X.media.press. Springer, Berlin, Heidelberg, Germany. 99-123. DOI = 10.1007/978-3-540-72216-8_6

[21] Bergman, M., J. L. King, and K. Lyytinen. 2002. Large-Scale Requirements Analysis as Heterogeneous Engineering. In *Scandinavian Journal of Information Systcrns* 14. 37-55.

[22] Star, S.L. and Ruhleder, K. 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. In: Information System Research. 7, 1

[23] Star, S.L. and Bowker, G.C. and Neumann, L.J. 2003. Transparency beyond the individual level of scale: Convergence between information artifacts and communities of practice. In. Bishop, A.P., Van House, N. A. and Buttenfield B-P. (ed.) *Digital library use: Social practice in design and evaluation*. 241-269.

[24] Schindler, C. and Rittberger, M. 2009. Herausforderungen für die Gestaltung von wissenschaftlichen Informationsinfrastrukturen durch Web 2.0. In: *Information - Wissenschaft & Praxis*, 60, 4, 215-224.

[25] Hagedorn, G., Weber, G., Plank, A. Giurgiu, M. Homodi, A., Veja, and others, 2010. An online authoring and publishing platform for field guides and identification tools. *Proc. of Conf. Tools for Identifying Biodiversity: Progress and Problems, Paris, 13-18.*

[26] Veja, C., Hagedorn, G., Weber, G., Giurgiu, M. 2010 Semantic MediaWiki Interoperability Framework from a Semantic Social Software Perspective. *Proc. of Int. Symp. on Electronics and Telecommunications ETC 2010, Timisoara, November 2010, pp. 403-406,* IEEE DOI=10.1109/ISETC.2010.5679307.

[27] Veja, C., Weber, G., Hagedorn, G., Giurgiu, M. 2010. MediaWiki Interoperability Framework for Multimedia Digital Resources. *Proc. Of IEEE 6th Int. Conf. on Intelligent Computer Communication and Processing, Cluj-Napoca, Romania. August 26-28. 321-328*. IEEE DOI= 10.1109/ICCP.2010.5606419.

[28] Veja, C., Hagedorn, G., Weber, G., Giurgiu, M. 2010. Metadata Repository Management using the MediaWiki Interoperability Framework A Case Study:The KeyToNature Project. *Proc. Of Int. Conf. eChallenges e-2010, 27-29 October. Warsaw, Poland, IEEE, ISBN: 978-1-4244-8390-7.*

[29] Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.. 2007. Semantic Wikipedia. In: *Journal of Web Semantics. 5, 251-261. Elsevier.*

[30] T. A. S. C. Vieira, M. A. Casanova, and Ferrao, L. G. 2004. An Ontology-Driven Architecture for Flexible Workflow Execution, at WebMedia and LA-Web.

[31] Pop, F.-C., Cremene, M., Vaida, M.-F. Riveill, M. 2010. Natural Language Service Composition with Request Disambiguation. ICSOC. 670-677

[32] Lenzerini, M. 2002. Data Integration: A Theoretical Perspective, In: *PODS '02 Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,* ISBN:1-58113-507-6

[33] Chan, L. M., Zeng, M. L. 2006, Metadata Interoperability and Standardization – A Study of Methodology Part I. In: *D-Lib Magazine*, June 2006, 12/6, ISSN 1082-9873.