Proceedings of the 1st International Workshop on

Semantic Digital Archives

co-located with the 1st International Conference on Theory and Practice of Digital Libraries (TPDL 2011), formerly known as European Conference on Digital Libraries (ECDL) and held on the 29.09.2011 in Berlin, Germany

Editors:

Livia Predoiu, Otto-von-Guericke University of Magdeburg, Germany

Steffen Hennicke, Humboldt University of Berlin, Germany

Andreas Nürnberger, Otto-von-Guericke University of Magdeburg, Germany

Annett Mitschick, University of Dresden, Germany

Seamus Ross, University of Toronto, Canada

CEUR Workshop Proceedings
http://ceur-ws.org
ISSN 1613-0073

Vol-801
urn:nbn:de:0074-801-0

Preface

These proceedings are the result of an exciting workshop held in conjunction with the first international conference on Theory and Practice of Digital Libraries, TPDL 2011, formerly known as European Conference on Digital Libraries, ECDL. The name of the workshop, *Semantic Digital Archives – sustainable long-term curation perspectives of Cultural Heritage* (short: SDA 2011) already provides a first hint towards its general topics and goals: to promote and discuss sophisticated knowledge representation and knowledge management solutions specifically designed for improving Archival Information Systems (AIS) and Archival Information Infrastructures (AII).

Over the past couple of decades, digitally created content has come to permeate all aspects of our lives and the life cycle of these objects is increasingly exclusively digital. A portion of this content can be expected to have enduring value as it delivers insight into the contemporary trends and spirit of its time. Hence, it can be considered being part of our cultural and scientific heritage. This vast corpus needs to be appraised and items of enduring value selected, archived and kept accessible so that it can be made available in response to requests from information professionals, and the general public. Therefore, sustainable long-term curation perspectives for our digital cultural heritage are essential. Digital content poses many socio-cultural and technological challenges which create obstacles to long-term or indefinite preservation. Changing technologies and shifting user communities as well as the increasing complexity of digital content consisting of or being enriched with software and multimedia attachments are only a few examples. Dealing with these challenges was the central theme of the workshop.

The workshop aimed to involve and stimulate discussions between the digital archiving, the digital museums, the digital libraries and the semantic (web) technologies communities. Archives, museums and libraries share a natural bond as all three have a long history of experience with maintaining (storing and retrieving) a large amount of objects, data and information. Hence, there is a lot potential for cross-fertilization between these related fields. Furthermore, libraries already started to adopt semantic web technologies successfully as shown by various workshops and conferences on this topic that recently have emerged. Most remarkably, also a W3C incubator group on *library linked data* has been created. Hence, the workshop aimed at fostering discussions about experiences and best practices of employing semantic web technologies in the library domain yielding so called *semantic digital libraries* in order to inspire and boost the adoption of semantic web technologies in the area of digital archiving as well.

The area of semantic (web) technologies is a broad scientific discipline that focuses on providing promising technical solutions for knowledge representation and knowledge management. It provides knowledge representation languages and management technologies based on a solid artificial intelligence foundation and is supported by appropriate W3C recommendations and a large user community. At the forefront of making the semantic web a mature and applicable reality is the linked data initiative. Using semantic (web) technologies in general and linked data in particular can be expected to mature the area of digital archives as well and technologically tighten the bond between digital libraries and digital archives. Furthermore, digital archives and their users have special requirements that can also inspire semantic (web) technologies research in general.

The workshop was well accepted by the community and was able to attract 23 submissions from which we selected 13 papers with the help of our program committee; giving an overall acceptance rate of 56%. The papers covered a broad range of relevant topics in the area of semantic digital archives, bringing together people from archives, museums, digital libraries and the semantic web as hoped and expected. A lot of different research projects are represented in these proceedings, e.g. the KEEP project (W. Bergmeyer), SHAMAN (J. Brunsman, K. Qian et al.), Semlib (C. Morbidoni et al.), ASPI (C. Cortese and G. Mantegari), Europeana (S. Hennicke et al.) and EUscreen (J. Oomen and V. Tzouvaras). Some of the papers that have been presented during the workshop are very data-oriented and focus on a specific kind of data to be preserved, maintained or kept accessible, like computer games (W. Bergmeyer), metadata on products in a company (J. Brunsmann), digital libraries in general (C. Morbidoni et al.), archival data in general (C. Cortese and Mantegari, S. Mazzini and F. Ricci, S. Hennicke et al.) and pictures of museum items (T. Wray and P. Eklund). Other papers focus on a general approach like the paper by A. Schröder et al. who present a novel and promising approach for semantic hierarchical storage management. Another example for a paper that focuses on a general approach is the paper by Kai Eckert who proposes a basic linguistic indexer for digital libraries.

The workshop started with an invited talk on *The KEEP emulation framework (W. Bergmeyer)* which is also contained as publication in this volume. In this publication, W. Bergmeyer presents the KEEP (Keeping Emulation Environments Portable) project which is a research project of the European 7th Framework Programe. During the workshop, a demo of the KEEP emulation software framework has been shown. This talk brought the general trend of emulation as a preservation strategy which is currently the method of choice when preserving software tools and multimedia systems into the discussions of the workshop. Afterwards, focusing on hardware as well, a *semantic extension of a hierarchical storage management system for small and medium-sized enterprises (A. Schröder et al.)* has been discussed. Since such a system saves costs, capacity and access time, it can be especially useful in large digital archiving frameworks and infrastructures in order to distribute, store and retrieve semantically coherent archival data.

In the submission about the *semantic exploration of archived product lifecycle metadata under schema and instance evolution (J. Brunsmann)*, J. Brunsmann brings a new view into the discussion since he considered a holistic approach for maintaining the life cycle of linked data describing obsolete product ideas within a company archive. Hence, he introduces an interesting application field for semantic digital archives.

The paper *Towards a semantic data library for the social sciences (T. Grotton et al.)* brings a very interesting preliminary approach for a linked library data infrastructure for statistical data in the social sciences into the discussion. Although this work does not consider digital archiving and is on a very preliminary state, it provides an insight into a statistical semantic digital library infrastructure and hence stimulated the discussion on semantic digital libraries versus semantic digital archives. More information on semantic digital libraries is provided by the paper on *introducing the Semlib project: semantic web tools for digital libraries (C. Morbidoni et al.)* which describes an annotation system for digital libraries. The proposed system adds user interaction to digital libraries via annotation and provides semantic structure to such annotations as well.

The paper *LOHAI: Providing a baseline for KOS based automatic indexing (K. Eckert)* proposes a free, open source and easy to use indexer tool for KOSs. This tool can provide the fundament on which to build more ambitious tools; although it has been developed for digital libraries, it can be used in other contexts like digital archiving contexts as well.

The publication on *extending the digital archives of italian psychology with semantic data (C. Cortese and G. Mantegari)* discusses an approach for implementing a semantic digital archive using CIDOC CRM for ontology modeling. Similarly, the paper on *EAC-CPF Ontology and Linked Archival Data (S. Mazzini and F. Ricci)* presents a topic that is relevant for digital archiving. More specifically, the development of an ontology is described that corresponds semantically to the EAC-CPF schema which is an archival standard for modelling and describing individuals, families and corporate bodies that create, preserve, use and are responsible for and/or associated with records in a variety of ways. A related topic is discussed in the submission about the *conversion of EAD into EDM linked data (S. Hennicke et al.)* as it deals with integrating archival finding aids into the portal of the Europeanna project which is an ambitious european project aiming at integrating data and information of museum, archives and libraries in one semantic web enhanced portal.

With *Concepts and Collections: A case study using objects from the Brooklyn Museum (T. Wray and P. Eklund)*, an interesting approach for a browsing framework for digitised cultural collections based on Formal Concept Analysis has been presented. This framework has also been evaluated with a case study using real data of the Brooklyn museum which nicely demonstrates that appropriate NLP techniques can be used to extract formal contexts from textual resources.

By the paper *Publishing Europe's television heritage on the web (J. Oomen and V. Tzouvaras)*, the first results of the European project EUScreen that deals with aggregating television heritage from European television archives for the European digital library Europeana have been presented.

Another very interesting paper is *A security contextualization framework for digital long-term preservation (K. Qian et al.)* as it is concerned with semantic security policies for digital archives. The approach extends the OAIS standard with security related features. Hence, an often neglected but crucial aspect in digital archiving is considered when establishing policies and infrastructures for long-term preservation.

The submission on *DA-NRW: A distributed architecture for long-term preservation (M. Thaller et al.)* presents an ongoing project that aims at creating a digital archive or long-term repository for the German state of North-Rhine Westphalia. This system will have a kind of sandwich position having to ingest data of depositors being archives and act as a pre-aggregator for portals like the Deutsche

Digitale Bibliothek or Europeana. A similar topic is dealt with by the paper on *RDFa as a lightweight metadata interoperability layer between repository software and LOCKSS (F. Ostrowski)* as it considers the extension of the LOCKSS framework with RDF and ontologies using SPARQL endpoints.

We would like to thank all members of the program committee for supporting us in the reviewing process. Altogether, the diversity of the papers in these proceedings represent a multitude of interesting facets about the new, exciting and promising research field of semantic digital archives and semantic digital archiving infrastructures. Hence, these proceedings provide a good and conclusive overview of the current research in this area.

December, 2011

Livia Predoiu,
Steffen Hennicke,
Andreas Nürnberger
Annett Mitschick
Seamus Ross

Organization

Program Chairs

Livia Predoiu	Otto-von-Guericke University of Magdeburg, Germany
Steffen Hennicke	Humboldt University of Berlin, Germany
Andreas Nürnberger	Otto-von-Guericke University of Magdeburg, Germany
Annett Mitschick	University of Dresden, Germany
Seamus Ross	University of Toronto, Canada

Program Committee

Sören Auer	University of Leipzig, Germany
Kai Eckert	University Library of Mannheim, Germany
Armin Haller	CSIRO, Australia
Stijn Heymans	SemanticBits, USA
Pascal Hitzler	Wright State University, USA
Yannis E. Ioannidis	University of Athens, Greece
Christian Keitel	State Archive of Baden-Württemberg, Germany
Thomas Lukasiewicz	University of Oxford, UK
Knud Möller	DERI Galway, Ireland
Vit Novacek	DERI Galway, Ireland
Johan Oomen	Netherlands Institute for Sound & Vision, Netherlands
Jacco van Ossenbruggen	VU University Amsterdam, Netherlands
Daniel Pitti	University of Virginia, USA
Andreas Rauber	Vienna University of Technology, Austria
Thomas Risse	L3S Research Center Hannover, Germany
Sebastian Rudolph	Karlsruher Institut für Technologie, Germany
Francois Scharffe	University of Montpellier, France
Michael Seadle	Humboldt University of Berlin, Germany
Marc Spaniol	Max-Planck-Institut Saarbrücken, Germany

Additional Reviewers

Thomas Low	Otto-von-Guericke University of Magdeburg, Germany
Magnus Pfeffer	HdM Stuttgart, Germany

Table of Contents

Invited Contribution

Hardware and Product Lifecycle Support

Linked Data Infrastructures and Ontologies

Digital Libraries and Museums

Archiving Frameworks and Infrastructures

The KEEP Emulation Framework

Winfried Bergmeyer

Computerspielemuseum, Berlin, Germany
(bergmeyer@computerspielemuseum.de)

Abstract. As part of the overall KEEP Project the task of the Emulation Framework (EF) is to provide the emulation environments required for this purpose. In a very simple and comprehensible way the users can employ emulations for the representation and the performance of both digital objects in obsolete data formats and applications for antiquated computer systems. The virtual reconstruction of original playback environments can reproduce the original look-and-feel-qualities.

By means of this approach audiences unfamiliar with the very concept of emulations can employ them in a private context as well as in an institutional framework. Thus a digital object stored in an archive can be rendered in that digital environment best suited to it without the time-consuming procedure of having to equip the computer on which it is to be run with emulation-specific configurations.

The range of applications is huge: In addition to allowing games to be played with an original atmosphere it encompasses e. g. access to data in obsolete formats or data migration with original software. The decisive factor from the point of view of conservation is that the original stream is being employed. There are almost no limits as to the scope of emulators that can potentially be used inside EF. The first release will support the following platforms: x86, Commodore 64, Amiga, BBC Micro and Amstrad/Schneider CPC.

Keywords: KEEP, Emulation, Emulation Framework, Virtual Machine, Transfer Tool

1 Emulation as a concept of conservation

The advantages of emulation as an alternative form of preservation (if compared to data migration) are numerous: Any strategy of preservation has to guarantee the permanence of the upkeep of the digital object as well as that of its accessibility[1].

[1] In 1999 Jeff Rothenberg declared: "*The best way to satisfy the criteria for a solution is to run the original software under emulation on future computers*". Jeff Rothenberg (1999). Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library and Information Resources – Washington,

Migration, on the other hand, ensures a long-term usability only at the expense of changing the data formats: a text document in wordperfect format .wpd e. g. will be conveyed into one that promises enduring availability such as .rtf or .txt. But this entails a loss of original information. Although you could say that the original file is being preserved, it cannot be rendered any longer for want of the original environment. It is therefore in the strict sense no longer accessible.

A number of repeated transfers can entail in the course of time significant modifications, which are bound to impair the content as well. Highly complex and proprietary formats such as objects in CAD may already suffer losses with the first migration. In the worst case they may not be saved in other formats. Compiled programs can only be migrated by means of a recompilation. For this purpose the uncompiled code will be required which is often not available any more. Thus this procedure cannot be performed when dealing with computer games or with commercial software.

But the approach of preserving the hardware on a permanent basis is not feasible. We can observe both the hardware components of the computer as well as the peripheral devices for input-, output- or reading functions become rare or defunct. Enormous costs would ensue if the spare parts had to be produced again.

Employing emulation, on the other hand, allows us to appropriate the original bitstream without taking recourse to migration. The digital objects are rendered on the virtually provided original platform. This reproduction of an original look-and-feel can be important for quite a few objects, e.g. for computer games, digital art objects or in the case of poetry presented with a sophisticated typography. In guaranteeing permanent access to these objects emulation is a substantial advance in terms of quality assurance. Emulation may yield yet another advantage. The original environment as in the case of an obsolete data bank management system can be employed for the initial transfer of the data in other formats. Thus the correct processing of the source data is ensured.

2 Transfer process

Maintaining the original bitstream is part and parcel of any responsible form of long-term preservation. But compared with conventional media such as books, painting or prints digital data prove to be incredibly frail. Even an apparently small amount of informational loss (like the scratched surface of a CD) can invalidate all the data. Damage to a painting or the pages torn in a book do not prohibit the usage of the object or at least those parts left intact. With digital data this is not possible.

In the course of recent decades various technologies for storing data have been employed. This means that a lot of different reading devices are needed for the transfer. In addition to 3.5“ and 5.25“ disk drives (for which there are still reading devices) there were less popular formats such as 8“- or even 3“-disks. Furthermore we

DC: Council on Library and Information Resources, p VI. http://www.clir.org/pubs/reports /rothenberg /pub77.pdf.

have still some readers for tape-based storage system (like DLP- or LTO-tapes as well as audio cassettes). But the situation is much more complicated for reel-to-reel tapes or microcassettes. In the case of early computer game consoles cartridge systems were the most popular form and they represent a serious challenge for transfer processes.

Thus the very process of data transfer into a new storage system lies at the heart of any concept dealing with conservation. We have to distinguish between the saving of single files (texts, pictures, tables) and that of preserving the entire data carrier (like the cartridges of obsolete gaming consoles). For certain areas of application it is essential to transfer the medium in its entirety into a virtual image, which then in turn can be embedded as a data carrier (a disk, a CD-ROM or a hard disk) in the emulator. These images have one huge advantage. They can be integrated into a systematic long-term archive that can be stored in a system of hard disks.

3 Copy protection

Copy protection systems for computer games prove a double problem for the long-term preservation. On the one hand you cannot always successfully bypass the copy protection. Thus for example a systematic copy protection involving a dongle (i. e. a protection that requires certain hardware elements for the program to work) cannot be transferred into useable images without changing the original program code. (This is the problem with many CD-ROMs and DVDs.) Similarly substantial knowledge about the mechanisms of copy protection is required to produce working images of older protected floppy disks. And it is regrettably extremely difficult to identify a copy protection before starting the transfer. In the case of CD-ROMs and DVDs a variety of tools able to identify many protection systems is available. But in the case of floppy disks no such instruments have been created for data carrier of systems like C64 or Amiga.

And furthermore this bypassing is still an offence even if performed by an institution working under the premiss of conserving these objects of cultural value[2]. (There is different legislation in other countries.) In the case of the Computerspielemuseum we are technically in the position to circumvent many of those copy protection systems. However, this would represent a violation of existing German laws. We thus need legal changes in the future to ensure the preservation of the digital objects[3].

[2] *„Die Gedächtnisinstitutionen können ihren gesetzlichen Aufgaben im digitalen Umfeld nur eingeschränkt nachkommen. Eine effektive und umfassende digitale Langzeitarchivierung ist ihnen rechtlich nicht möglich. Es droht eine digitale Amnesie des Kulturellen Gedächtnisses.“* Digitale Langzeitarchivierung als Thema für den 3. Korb zum Urheberrechtsgesetz. Urheberrechtliche Probleme der digitalen Langzeitarchivierung (2011). Paper of the German competence network for digital preservation (nestor). http://files.d-nb.de/nestor/berichte/nestor-Stellungnahme_AG-Recht.pdf.

[3] Eberhard Hilf, Christian Keitel, Kai Naumann, Martin Iordanidis, Christina Bankhardt, Sven Vlaeminck, Reinhard Altenhöner, Sabine Schrimpf, Natascha Schumann (2010): Sozio-ökonomische Erfolgsfaktoren für die Langzeitarchivierung in Deutschland. nestor-

4 Emulators

The history of digital emulators commenced some 30 years ago. And interestingly enough it all started with a gaming console, ColecoVision produced by Coleco. There was an adapter available on the market allowing to play games initially developed for and to be run on the Atari 2600. And in 1985 Atari, in order to prove the capacity of the 68000-CPU and the TOS-systems software, in turn developed a Z80-emulator, which allowed to run software originally designed for the CP/M-operating system. This meant that a vast range of software was available for that computer system.

Presently a huge number of emulators exist for different consoles and platforms. In the majority they have been created inside the community of the retrogamers and for their purposes. But with more and more devices in circulation that are both portable and equipped with sophisticated graphics such as the IPhone or the IPad we find a growing number of emulators included for commercial products.

One current argument against the appropriation of emulators for the purpose of long-term preservation points to the fact that emulators are still tied to specific platforms. A reliable and permanent availability is not guaranteed as long as the introduction of new operating systems or new hardware technologies will necessitate a new porting of the emulators. In addition to recent attempts to develop Java-based emulators[4], we have the Dioscuri-Project at the Koninklijke Bibliotheek (Den Haag) which presents the first emulator designed to operate independently from specific platforms[5].

These are some of the obstacles on the path for using emulation as a fully fledged instrument for the purpose of preservation. But since keeping the digital tradition alive cannot be achieved without emulation, it is essential to develop concepts and technologies which will facilitate it and make it available on a permanent basis.

5 KEEP (Keeping Emulation Environments Portable)

Keeping Emulation Environments Portable (KEEP) is a medium scale research project started on 1 February 2009 co-financed by the 7th Framework Programme's ICT-3-4.3 Digital libraries and technology-enhanced learning priority[6]. The overall aim of the project is to facilitate universal access to our cultural heritage by developing flexible tools for accessing and storing a wide range of digital objects. Although primarily aimed at those involved in Cultural Heritage, such as memory institutions and games museums, the KEEP Emulation Services can also serve the needs of a wide range of organisations and individuals because of its universal approach.

Positionspapier zum Abschlussbericht der Blue Ribbon Task Force on Sustainable Digital Preservation, http://files.d-nb.de/nestor/berichte/nestor_Stellungnahme_BRTF.pdf.

[4] For example the Amstrad Emulator "JavaCPC" on Java basis. http://sourceforge.net/projects/javacpc /files/

[5] http://sourceforge.net/projects/dioscuri/

[6] http://www.keep-project.eu

The following institutions are partners in this cooperation:

- Bibliothèque nationale de France (Paris) project coordinator
- Koninklijke Bibliotheek (Den Haag)
- Deutsche Nationalbibliothek (Frankfurt)
- University of Portsmouth
- Tessella
- Joguin SAS
- European Games Developer Federation
- Computerspielemuseum (Berlin)

5.1 The KEEP Transfer Tool Framework (TTF)

The development of a transfer tool framework (TTF) is to simplify and to automatize the workflow. Our main objective is to design a concept for the appropriate elements for the workflow, to compile the collections of meta-data required for this purpose and to mould these results into a project allowing long-term preservation. Existing systems for data transferal have been evaluated with regard to their suitability for the TTF. Simultaneously we have analyzed existing copy protection systems.

Part of the workflow is to monitor the transfer process. A failed attempt of transfer or a dysfunctional image may be due to one of the following reasons:

- The reading process was not performed without mistakes.
- A copy protection previously unperceived has undermined the process.

A further obstacle for the preservation is the multitude of data formats for images or the number of formats supported by those emulators currently available. In the field of computer games a lot of special formats have been established as the standards for the individual systems. These in turn have different potential to bypass copy protection.

These are a number of problems you can encounter while working on a transfer. The process of automatically transferring larger quantities of data media from a diversity of operating systems (which by necessity can be the only form of procedure that larger institutions can afford to work with) is still something of a challenge and something to strive for.

Thus we are preparing to embed the TTF in the Planets Interoperability Framework[7], which allows integration of the transfer workflow into an already existing framework for digital preservation. A KEEP study on the integration of the KEEP Transfer Tool with the Planets Interoperability framework states, that it is feasible from a technical and legal perspective. For this purpose a new API for the integration of the services and tools for the TTF has to be developed. After that the infrastructure within the Planets IF can be used for the transfer workflow.

[7] http://www.planets-project.eu/docs/reports/Planets_IF-D11_ConsolidatedReleaseDocumentation.pdf

5.2 The Keep Virtual Machine (KVM)

As one can see, the very name of the project, KEEP (i.e. Keeping Emulation Environments Portable) emphasizes the importance of the permanence accorded to the emulation environments. We are currently working on developing a virtualizer that can serve as a platform for many emulators already available. Portability and flexibility are the core requirements for Keep Virtual Machine (KVM). The immediate advantage is that it prevents the concept of stacked emulation, which is where multiple emulators each translate a platform for a specific era for another emulator. Stacking emulated computers creates a dangerous stack of dependencies that is vulnerable for errors and mistakes.

The KVM is based upon the concept of a virtual layer, based on the idea of Jeff Rothenberg[8]. The KVM implementation is defined by the selection of a base sub-machine, directly implemented by way of an emulator or a dynamic compiler written for a support machine, as well as a work sub-machine, whose complexity is higher or equal to that of the base machine, and for which applications are written. The virtual processors are a set of sub-machines that vary in complexity and efficiency, structured from the simplest to the most complex, featuring in that order: KVM0, KVM1, KVM2, KVM3 and KVM4. The most basic commands allow easy adjustment to fit in new platforms by means of simple and elementary operations, whereas you will find complex command structures at the top level that can communicate with systems such as Linux, Java or with the emulators themselves. These have to be customized initially to be able to cooperate with the KVM. Afterwards it will only be the other way round: the KVM has to be matched with new platforms. This will yield an extremely efficient virtualizer, whose emulators once matched can be employed for a long time. This will mean a major progress for a safe and sound appropriation of emulation technologies for the purpose of conservation.

5.3 The KEEP Emulation Framework (EF)

The theme central to this workshop is a presentation of the KEEP Emulation Framework. Employing technologies of emulation in institutions so far means a high amount of work input before anything is available for the end-user, e. g. somebody reading a book in a library or visiting an exhibition in a museum. To provide a variety of user-friendly emulators for different digital objects entails the following: Firstly the relevant meta-data have to be stored in the digital archive and secondly the different emulators have to be implemented and maintained. A process of automatization for the recognition and the verification of the data and the allocation of the appropriate

[8] Rothenberg, J., Preservation of the Times, The Information Management Journal, March/April 2002, Vol 36, No. 2, pp. 38-43. ISSN 1535-2897, available at: http://www.panix.com/~jeffr/Prof/Pubs/DigitalLongevity/arma.paper.from-journal.pdf

emulation environment will inevitably facilitate the process of issuing them to the end-user and eliminate a number of intermediate administrative procedures.

The KEEP EF has the resources to perform exactly this very process of automatization. Via a defined interface a request for information will be addressed to the digital archive of an institution and will activate the EF. The data-file (a single file or the image of a data-carrier) are identified by means of a format-specific registry and then the appropriate emulation pathways show up on the screen. Selecting a specific pathway will start the emulator and the software package necessary for performing this task.

The EF software consists of three parts: a Core Application, Software Archive and Emulator Archive. The Core EF is the technical heart of the system, performing the automatic characterization of file formats, selecting the required software and automatically configuring the emulation environment. It has a simple GUI to interact directly with the user. For selecting the software and emulator, the Core interacts with external services such as technical registries containing file format classifications, the Software Archive that contains software captured in predefined disk images and the Emulator Archive that contains the emulators available for the EF.

The Core EF, Software Archive and Emulator Archive are developed by Tessella with support from the National Library of the Netherlands (Koninklijke Bibliotheek, KB). The Core GUI is developed by the National Library of the Netherlands.

The software.

For the last few weeks KEEP EF has been made available on the Source Forge Portal for free download9. Java was chosen as the development language because of familiarity, widely supported libraries and overall portability; for the internal database H2 was chosen because of the small footprint and integrated web-interface. An installation program simplifies the installation process. These are the system requirements:

Processor	X86 32/64 bit 1.5 GHz or faster
Memory	At least 2 GB of memory
Disk space	200 Megabytes if free available space for the base install Depending on the number of emulators and software images, from 1 GB upwards
Operating system	Linux or Windows with compatible JRE and network support
Java Runtime Environment (JRE)	Oracle(Sun JRE version 1.6 or higher / compatible

The download package contains the following emulators:

- Qemu (x86)
- Dioscuri (x86)

[9] http://sourceforge.net/projects/emuframework/

- WinUAE /Amiga)
- VICE (C64)
- JavaCPC (Amstrad/Schneider)
- BeebEm (BBCmicro)

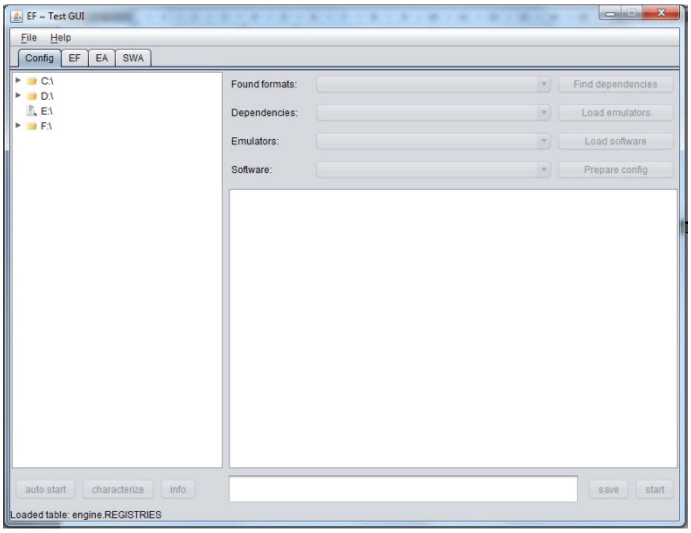
And the following format registries are included:

- Pronom/PCR
- UDFR

We have to emphasize that this package may only contain open source software. Thus you will neither find an operating system such as MS Windows operating system nor an application like wordperfect to come along with the bundle. Via the administrator's interface, however, the system can be expanded and modified to the individual requirements.

Functionalities for the end-user.

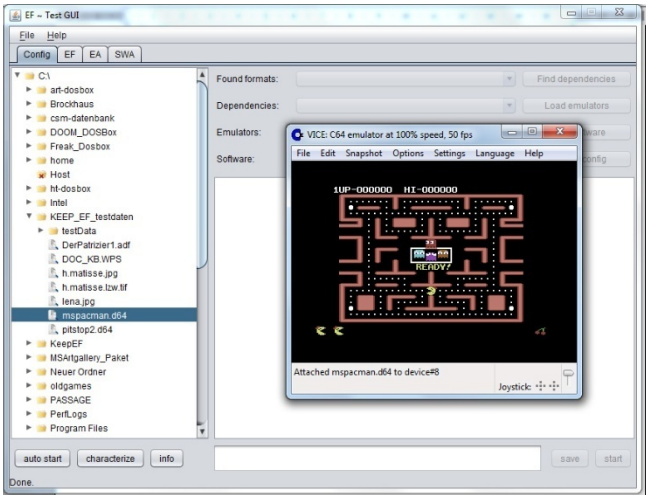
Before we will proceed to explain both the technical and the administrative features, we will briefly describe the capabilities of the system as well as the functionalities relevant for the end-user.



Loaded table: engine.REGISTRIES

Fig.1. Java-based EF GUI

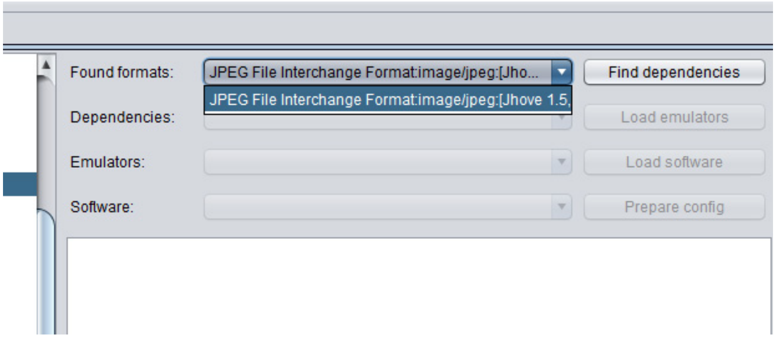
This is still an experimental graphical user interface to give a first impression of the functionality. At the end of the project there will be a more effective interface, especially for the administration. With the actual java-based Gui the user can select a digital object in the left column. To describe the next steps in the procedure, we will employ a d64-image of a game designed for the Commodore64 (*Ms Pacman*). Having selected the object by clicking on the icon, the lower region of the screen presents three different buttons (*auto start*, *characterize* and *info*). If you click on the button for auto start, a window will pop up presenting the VICE-emulator and the game.



Attached mspacman.d64 to device#8

Fig.2. Commodore64 Emulator VICE with *Ms Packman* running

This intermediate step can be omitted. You can directly start the emulation from the digital archive. But in addition to this automatic procedure the user can select between several emulation pathways (if they are defined by the administrator).
To illustrate this approach we will select a JPEG-file: If you activate the button „characterize", information in the right upper corner states that JHove has successfully identified the data format to be JPEG.



Found formats: JPEG File Interchange Format:image/jpeg:[Jho...

JPEG File Interchange Format:image/jpeg:[Jhove 1.5,

Dependencies:

Emulators:

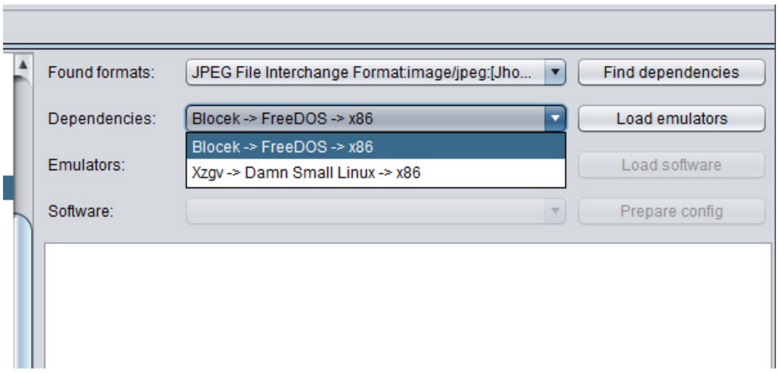
Software:

Fig.3. Format identification

In the next step the current dependencies will be shown, i. e. which emulation paths have been allocated to the data format by the administrator. In this case the following two are available:

- the program Blocek under FreeDos on a x86 system

- the program XzgV under Damned small Linux on a x86 system



Found formats: JPEG File Interchange Format:image/jpeg:[Jho...

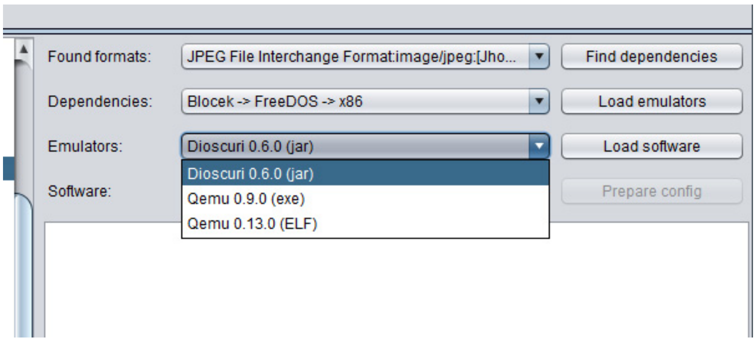
Dependencies: Blocek -> FreeDOS -> x86

Blocek -> FreeDOS -> x86
Xzgv -> Damn Small Linux -> x86

Emulators:

Software:

Fig.4. Dependencies



Found formats: JPEG File Interchange Format:image/jpeg:[Jho...

Dependencies: Blocek -> FreeDOS -> x86

Emulators: Dioscuri 0.6.0 (jar)

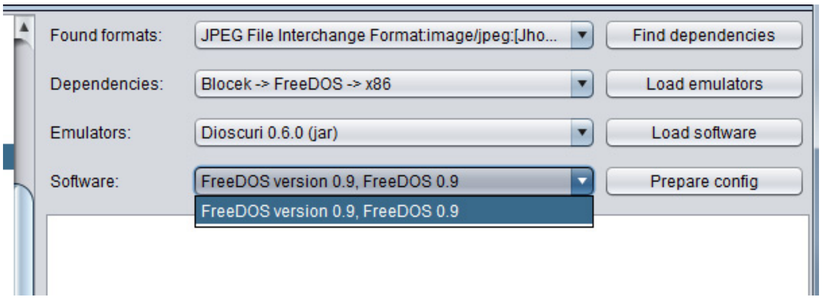
Dioscuri 0.6.0 (jar)
Qemu 0.9.0 (exe)
Qemu 0.13.0 (ELF)

Software:

Fig.5. Available emulators

Now the user can select one of the paths and henceforth the emulators available, in this case the choice is between Dioscuri and Qemu.

The next step consists in opting for the appropriate software package. We see FreeDos being offered in connection with the Blocek-application.



Found formats: JPEG File Interchange Format:image/jpeg:[Jho...

Dependencies: Blocek -> FreeDOS -> x86

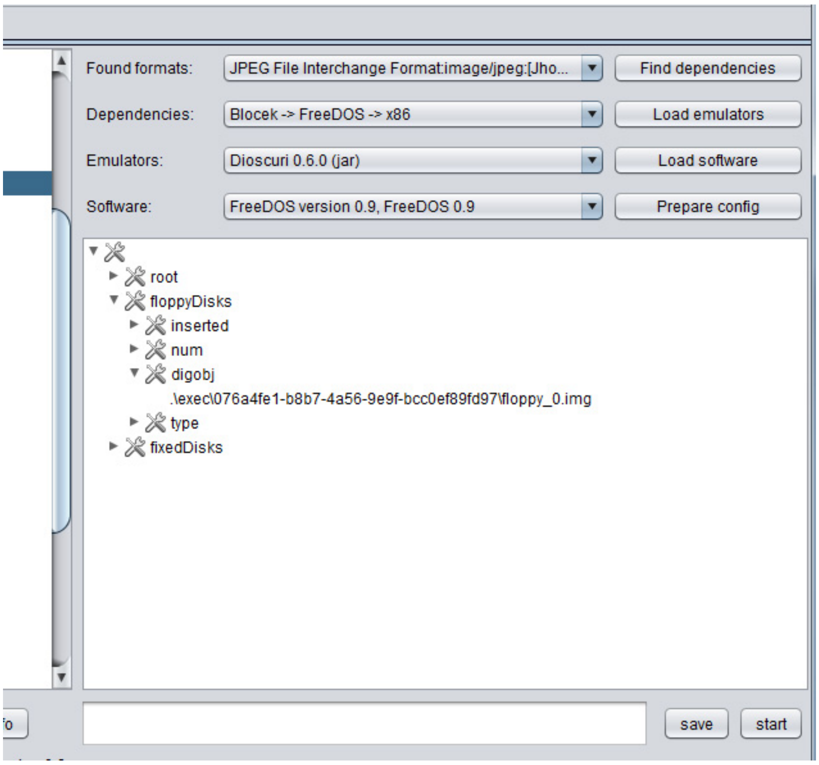
Emulators: Dioscuri 0.6.0 (jar)

Software: FreeDOS version 0.9, FreeDOS 0.9

FreeDOS version 0.9, FreeDOS 0.9

Fig.6. Available software

Finally the configuration is being prepared. This entails mounting file as a disk (drive A). The system mounts every object with less than 1.2 MB as a disk, any file larger than that will be tied in automatically as a hard disk.



Found formats: JPEG File Interchange Format:image/jpeg:[Jho...

Dependencies: Blocek -> FreeDOS -> x86

Emulators: Dioscuri 0.6.0 (jar)

Software: FreeDOS version 0.9, FreeDOS 0.9

```
▼
  ▶ root
  ▼ floppyDisks
    ▶ inserted
    ▶ num
    ▼ digobj
        .\exec\076a4fe1-b8b7-4a56-9e9f-bcc0ef89fd97\floppy_0.img
    ▶ type
  ▶ fixedDisks
```

Fig.7. Available emulators

Once you are on the level of the emulation environment, this integration is necessary in order to have access to the object. By clicking on the button *start* the process of emulation will commence. When the program *Blocek* has been started on the Dos-Interface, the JPEG-file on the disk in the drive may be opened.

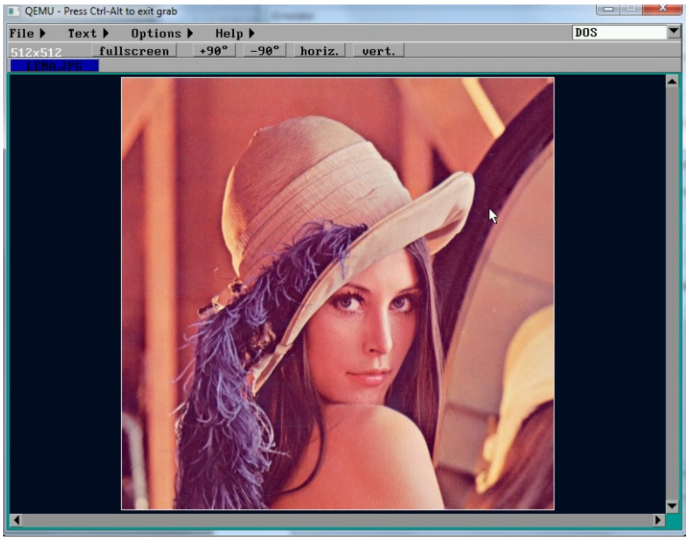


Fig.8. JPEG-image rendered with *Blocek*

By means of this selection the user can navigate and start and test different emulation environments. So let us now take a look at the EF from its technological and adminstrative side.

5.4 Administration area

The concept of the EF is based upon employing three different archives which may operate on different servers. The EF-Core-Archive needs to be installed on the very computer requested to present the emulation. But the archives for both the emulators themselves and the appropriate software have only to be connected with it by means of LAN or WLAN. This allows for a central administration of the emulators and the software for a number of computers or even for a network in an institution. We anticipate two groups of users: administrators and end-users.

In addition to the two format registries contained in the download package it will be possible to create your own registries and to integrate them. Thus individual solutions to fit the requirements of a particular institution can be created.

By means of this GUI the current preferences may be inspected and modified. But due to its complexity we will present only some aspects of the configuration.

Emulator Archive.

The administrator is enabled to integrate new emulators, if the ones available do not fit the standards or cannot emulate the hardware necessary for the presentation. The emulator itself will be integrated as a blob. There may be the need to both define new image formats and to register the linking with the new emulator.

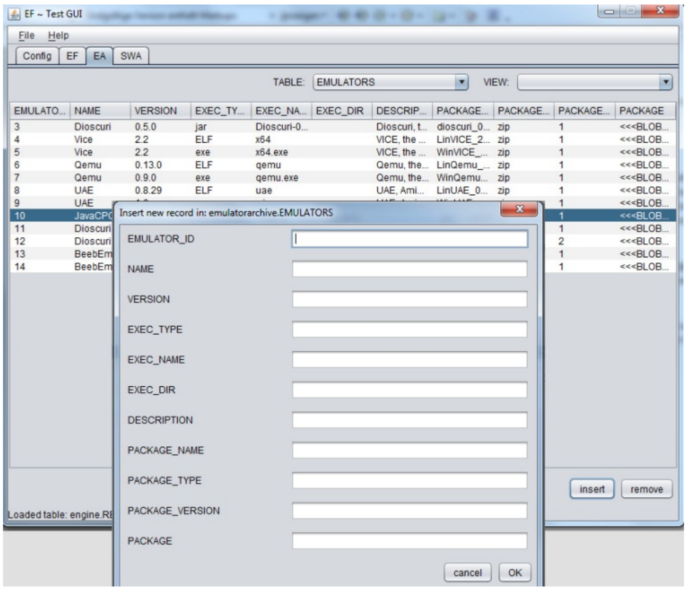


TABLE: EMULATORS VIEW:

EMULATO...	NAME	VERSION	EXEC_TY...	EXEC_NA...	EXEC_DIR	DESCRIP...	PACKAGE...	PACKAGE...	PACKAGE...	PACKAGE
3	Dioscuri	0.5.0	jar	Dioscuri-0...		Dioscuri, t...	dioscuri_0...	zip	1	<<<BLOB...
4	Vice	2.2	ELF	x64		VICE, the ...	LinVICE_2...	zip	1	<<<BLOB...
5	Vice	2.2	exe	x64.exe		VICE, the ...	WinVICE_...	zip	1	<<<BLOB...
6	Qemu	0.13.0	ELF	qemu		Qemu, the...	LinQemu_...	zip	1	<<<BLOB...
7	Qemu	0.9.0	exe	qemu.exe		Qemu, the...	WinQemu...	zip	1	<<<BLOB...
8	UAE	0.8.29	ELF	uae		UAE, Ami...	LinUAE_0...	zip	1	<<<BLOB...
9	UAE								1	<<<BLOB...
10	JavaCPC								1	<<<BLOB...
11	Dioscuri								1	<<<BLOB...
12	Dioscuri								2	<<<BLOB...
13	BeebEm								1	<<<BLOB...
14	BeebEm								1	<<<BLOB...

Loaded table: engine.R

Insert new record in: emulatorarchive.EMULATORS

EMULATOR_ID

NAME

VERSION

EXEC_TYPE

EXEC_NAME

EXEC_DIR

DESCRIPTION

PACKAGE_NAME

PACKAGE_TYPE

PACKAGE_VERSION

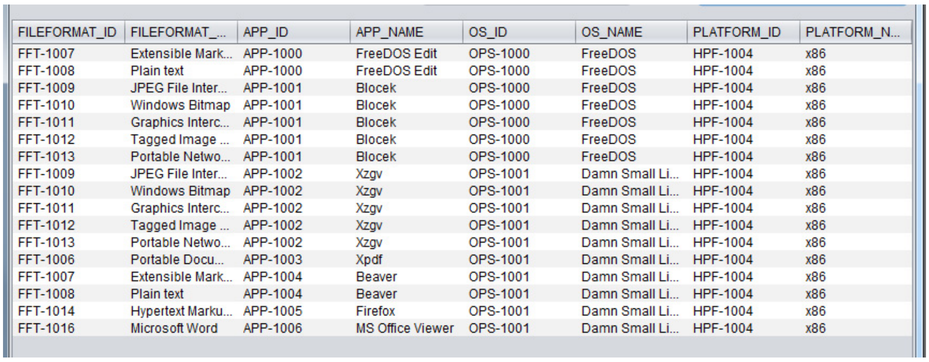
PACKAGE

Fig.9. EF GUI with form for adding new emulators

Software Archive.

If there is an additional demand for an operating system and/or an application (word processing, CAD-programs, database systems) for the emulation environment to work beyond the mere emulation of the hardware, additional software packages will be required. These, too, will be stored as blobs in the database. So additional entries for applications, operating systems, platforms and file formats have to be made.

If all entries are correct, the verification of the new emulation pathway can commence via the viewing of the pathways.



FILEFORMAT_ID	FILEFORMAT_...	APP_ID	APP_NAME	OS_ID	OS_NAME	PLATFORM_ID	PLATFORM_N...
FFT-1007	Extensible Mark...	APP-1000	FreeDOS Edit	OPS-1000	FreeDOS	HPF-1004	x86
FFT-1008	Plain text	APP-1000	FreeDOS Edit	OPS-1000	FreeDOS	HPF-1004	x86
FFT-1009	JPEG File Inter...	APP-1001	Blocek	OPS-1000	FreeDOS	HPF-1004	x86
FFT-1010	Windows Bitmap	APP-1001	Blocek	OPS-1000	FreeDOS	HPF-1004	x86
FFT-1011	Graphics Interc...	APP-1001	Blocek	OPS-1000	FreeDOS	HPF-1004	x86
FFT-1012	Tagged Image ...	APP-1001	Blocek	OPS-1000	FreeDOS	HPF-1004	x86
FFT-1013	Portable Netwo...	APP-1001	Blocek	OPS-1000	FreeDOS	HPF-1004	x86
FFT-1009	JPEG File Inter...	APP-1002	Xzgv	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1010	Windows Bitmap	APP-1002	Xzgv	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1011	Graphics Interc...	APP-1002	Xzgv	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1012	Tagged Image ...	APP-1002	Xzgv	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1013	Portable Netwo...	APP-1002	Xzgv	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1006	Portable Docu...	APP-1003	Xpdf	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1007	Extensible Mark...	APP-1004	Beaver	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1008	Plain text	APP-1004	Beaver	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1014	Hypertext Marku...	APP-1005	Firefox	OPS-1001	Damn Small Li...	HPF-1004	x86
FFT-1016	Microsoft Word	APP-1006	MS Office Viewer	OPS-1001	Damn Small Li...	HPF-1004	x86

Fig.10. List of available emulation pathways

6 Resumee

The EF will simplify the appropriation of emulators for the usage of original digital objects substantially. And the considerable range of possibilities the administrators are provided with means that customized solutions for emulators (and hardware platforms) and software and for individual format registries can be integrated into a framework allowing for a high level of automatization.

In connection with the two big tasks – the transfer (TTF) and the hardware independency of the emulators (KVM) – which both are busily worked on, the KEEP-Project has come up with a solution how to simplify emulation and employ it as a feasible procedure for long-term conservation that in itself is also built to last.

7 References

1. Jeffrey van der Hoeven, Dirk von Suchodoletz: Emulation: From Digital Artefact to Remotely Rendered Environments. The International Journal of Digital Curation - Issue 3, Volume 4, (2009)
2. Karsten Huth: Probleme und Lösungsansätze zur Archivierung von Computerprogrammen – am Beispiel der Software des Atari VCS 2600 und des C64. Berlin, (2004), http://www.digitalgamearchive.org/data/news /Softw_Preserv_ huth.pdf.
3. Bram Lohman, Jeffrey van der Hoeven: Building the emulator - A behind-the-scene look at development. Koninklijke Bibliotheek (KB), (2006), http://www.kb.nl/hrd/dd/dd_projecten/slides/eem_kbnatessella_jvdhoeven _blohman.pdf .
4. Lorie, Raymond: the UVC: a method for preserving digital documents – proof of concept, (2002), http://www-5.ibm.com/nl/dias/resource/uvc.pdf
5. Jeff Rothenberg: Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation: A Report to the Council on Library and Information Resources – Washington, DC: Council on Library and Information Resources, (1999), http://www.clir.org/pubs/reports /rothenberg /pub77.pdf.
6. Jeff Rothenberg. Using Emulation to Preserve Digital Documents. Koninklijke Bibliotheek, the National Library of the Netherlands The Hague, (2000)
7. Dirk von Suchodoletz. Funktionale Langzeitarchivierung digitaler Objekte - Erfolgsbedingungen des Einsatzes von Emulationsstrategien. Freiburg im Breisgau, (2008), http://nbn-resolving.de/urn:nbn:de:0008-2008070219.
8. Randolph Welte. Funktionale Langzeitarchivierung digitaler Objekte – Entwicklung eines Demonstrators zur Internetnutzung emulierter Ablaufumgebungen. Freiburg im Breisgau, (2008)
9. Randolph Welte Klaus Rechert, Dirk von Suchodoletz. Emulation Based Services in Digital Preservation. Proceedings of the 10th annual joint conference on Digital libraries, (2010)

10. Hilde van Wijngaarden, Jeffrey van der Hoeven. Modular emulation as a long-term preservation strategy for digital objects. Koninklijke Bibliotheek, the National Library of the Netherlands The Hague, The Netherlands, (2005)

A Semantic Extension of a Hierarchical Storage Management System for Small and Medium-sized Enterprises

Axel Schröder, Ronny Fritzsche, Sandro Schmidt, Annett Mitschick and Klaus Meißner

University of Dresden, 01187 Dresden, Germany, 2011
{axel.schroeder, Ronny.Fritzsche, Sandro.Schmidt, annett.mitschick, klaus.meissner}@tu-dresden.de,
WWW home page: http://www.mmt.inf.tu-dresden.de

Abstract. The number of company data deposited in hierarchical storage management systems heavily increases. Thus, new approaches are necessary to keep track of a data pool. This paper introduces a semantic storage extension (SSE) for existing hierarchical storage management systems that allows them to exploit semantic relations between files and use them for a more efficient and more intelligent data management. Our approach enhances traditional hierarchical storage management systems regarding migration, deletion, and retrieval operations by making use of semantic relations between files and contextual knowledge. Thereby a predictive file management is possible, which contributes to an increasing system performance and a better user experience. To this end, the SSE uses extracted features of documents to define relations between them and also offers the possibility to specify additional knowledge by a domain expert.

Keywords: semantic, hierarchical, storage, management

1 Introduction

At present, the amount of digital data stored by companies doubles year by year [15]. To save costs, companies increasingly use (hierarchical) storage management systems (SMSs). Those systems distribute data between different storage technologies and deposit information in a cost-optimized way. A SMS typically divides the storage environment into three tiers (Figure 1). The performance tier utilizes very fast and also expensive storage technologies like solid-state drives or SAS[1]/FC[2] hard disks. The capacity tier usually consists of SATA RAID systems which offer lower costs per gigabyte, but also suffer from an higher access time. The archive tier uses long time archiving technologies (WORM[3]) like optical

[1] Serial Attached SCSI
[2] Fibre Channel
[3] Write Once Read Many

jukeboxes or tape libraries. This tier offers the lowest costs per gigabyte and the highest capacity, but it also has a very high access time. So the SMS has to find a tradeoff between costs, capacity and access time. It has to distribute its data in an optimized way. We concentrate on three typical operations of a SMS: *migration*, *retrieval* and *deletion*. Migration means the movement of a file to a slower storage tier (e.g., from the performance to the capacity tier) and retrieval means the opposite, the movement to a faster storage tier. However, current SMSs treat files individually and do not recognize and use semantic relations (e.g same author, topic, accessdate, ..) between them. Furthermore, about 80 % of all data in a company is stored unstructured and contentwise unorganized [4], what soon becomes problematic for a useful cost optimization. Also only about 2 % of all stored data is used in daily business [17]. This means only these 2 % of all available data needs to be provided within the SMS's performance tier. The use of inherent and additional semantic information and semantic relations between files is one key to a more efficient and more intelligent way of storage management. There are other approaches to achieve a better performance, like monitoring user behaviour for a anticipatory data management, which are not focussed in this paper. We are concentrating on the aspect of using semantics in SMS. However, current SMS do not or only rudimentarily use those semantic relations. Treating this offers huge potential to improve management algorithms in order to achieve better performance and user experience.

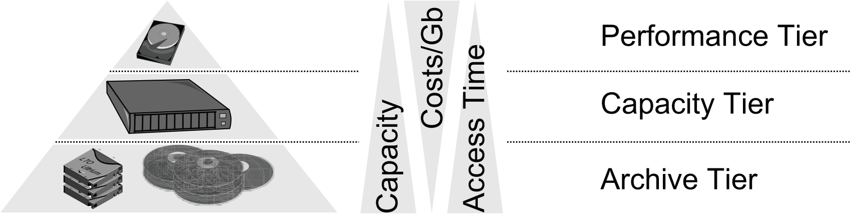


Fig. 1. A typical three tier architecture of a hierarchical storage management system.

This paper introduces a semantic-aware software component as extension for a SMS that allows the use of semantic information to optimize the way data is stored. We call it *Semantic Storage Extension* (SSE). The SSE can be integrated into existing SMS (see Section 3) and enables them to recognize semantic relations between documents, which may be used for distributing digital information. Initially, Section 2 will illustrate current developments and research projects in this field of study. Hence, requirements for the SSE will be derived, which are fundamental for its design, described in Section 3. This section also illustrates the functionality of the SSE and points out its interactions with other components like the SMS. Section 4 summarizes the results of this paper and illustrates both current and further work in this research area.

2 Related Work

Modern file systems like NTFS[4] already use a wealth of metadata in documents. However, these are primarily used for describing files. When trying to find similar information between different documents, the limitations of those file systems become apparent very fast. There is a wide range of papers which deal with semantics in documents (e.g. to improve file searching) [6, 8, 9, 12, 14]. It is also possible to enrich a rulebook of a document management system in order to choose appropriate file handling strategies [2]. Other approaches (e.g. [5]) try to separate the metadata from files to achieve a better system performance.

In contrast to all this work, we focus on improving the document management in a SMS. Thus, the SSE tries to help on managing files cost-orientedly according to their actuality and relevance. To achieve this goal, new semantic relations should be used. There are only few papers addressing semantic associations between documents and attempting to use those for managing files. The next section outlines some selected work of this specific research area.

2.1 Semantic Information in File Systems

To overcome the limitations of current file systems and SMSs, TagFS [1] annotates files with keywords. This is done automatically as well as user-controlled. For example, the keywords could contain different metadata, names of folders in a document path or other manually added terms. Subsequently these tags can be used to filter from a set of files. Thereby restrictions of the documents path are avoided. By using tags that reflect folder names of its original path, it is possible to filter directly for files within subfolders, without knowing their exact location (e.g., *../pictures/vacation* or *../vacation/pictures* then means the same).

Another approach for mapping semantic information in file systems was created in 2003 by introducing semantic vectors [7]. Thereby the metadata of files are converted into vectors, which subsequently span a common feature space. This leads to two results: On the one hand, duplicates easily become visible and on the other hand, strong dependencies between several documents can be found through vectors that are very close together.

There is another approach outlined in [16]. It illustrates how to capture external events in an ontology and to link them with the data that is related to this event. By doing so, the data pool is enriched with additional knowledge and files get indirectly linked with each other via events. For example, if files are created or modified during a phone call, they become related to this event inside the ontology. Thus, these documents have an implicit relation to all other files related to this event. The authors clarify that the correct linking of files with events needs a longer training process. But they also underline that this approach finally works very precisely. This idea allows a SMS to cluster data based on their real connections. Another advantage is that the users also benefit from this concept. They get the possibility to retrieve data by recalling specific events

[4] New Technology File System

(meetings, phone calls, ...). So this approach closely resembles human thinking. The navigation through a data pool is no longer based on the place (Where is information stored?), but on events (Why and Whereby?). A disadvantage is, that the relevance of files is not captured. So this approach has great potential for user-based files searched, but still needs additional improvements to support internal SMS operations.

2.2 Semantic Information in Additional Systems

There is a need for more research in the analyzed areas. The generation and processing of additional information creates a greater workload. For example, the approach of [16] introduces a mechanism to gather events. Corresponding to the size of a company, this concept can require extensive upgrades in the existing IT infrastructure. However, a long-term influence on current file systems is only possible, if such an approach is enforced as a well-established practice in a company. Also, it has to be carefully considered, if the achieved advantages justify the higher resource requirements. At this point, most publications only provide theoretical estimations or smaller field tests. In [17], studies about extensive scenarios are realized and performance and flexibility of those approaches are evaluated. They especially indicate, that modern DBMS[5] (in this case MySQL[6]) are not optimized for a very large number of metadata. Also many of the investigated concepts are only partially applicable for ubiquitous use. For example, some approaches require additional computing power to permanently extract metadata and analyze them, which then can be used to derive semantic relations. Usually it is very difficult to realize event gathering on external devices (e.g., fax) and connect them to documents, because there are no standardized interfaces to catch these events. So the integration of the illustrated approaches in current file systems and the combination of different research concepts is tricky. Thus, a solution is needed, which has enough potential compared to conventional methods and also presents an additional value abreast them.

3 The Semantic Storage Extension

This section illustrates the design for the SSE in detail. In Section 3.1, necessary requirements for the SSE are described. They lead to a software architecture outlined in Section 3.2. The following sections show the functionality of a semantic service component (Section 3.3) which is needed for feature extraction, illustrate necessary modifications inside the SMS (Section 3.4) and finally describe the structure of the SSE (Section 3.5).

[5] Database Management Systems

[6] http://www.mysql.com/

3.1 Requirements

As shown in the introduction, the data pool of SME[7] grows exponentially. Thus, it is important that a semantic extension for a storage management system (SSE) works performantly even after years. In contrast to traditional SMS which just relocate files in regarding their own context, the SSE should offer a foundation to enable the SMS to preemptively relocate related files as well. Furthermore, metadata should be managed centralized and independently from their documents. Similar to the approach of [5], who suggests separating metadata, two advantages follow. Metadata can be accessed faster and the number of read accesses on the actual storage media decreases, which means a longer lifetime [11]. In respect of the limitations of DBMS, metadata should be stored in an ontology [17]. Current DBMS often only support data mining and clustering. An ontology allows a more expressive description of semantic relations and metadata. It offers additional possibilities for reasoning mechanisms to infer semantic knowledge that is not explicitly modelled. Furthermore, the approach of gathering and processing events offers a huge potential [16]. The SSE should link documents with external knowledge to not only manage data about their structure and content, but about their origin and meaning. By doing so, the SMS would be able to not only provide relevant data for fast access, but presenting other data that is semantically close as well.

In order to function as an extension, the SSE has to work autonomously. It should enhance the underlying SMS with minimal modifications, but never be able to interfere with the SMSs basic functionality. This means the SSE should be available for the SMS through its own interface as an independent component.

Another requirement is a centralized metadata storage [5]. It can be used to look up information about the data pool in one place and use it to quickly get a link to related files. The metadata should also be managed centrally inside the SSE. Here it is especially necessary to pay attention to the consistency of information regarding the data pool in the SMS.

The heterogeneity of the data pool requires various and complex extraction algorithms. Metadata should not only be extracted from current file formats, but new file formats should be supported in the future. The complexity and extensibility of those extraction processes requires a complex treatment and was not focused yet. A semantic service component (SSC) has to provide these extraction features. The SSC has to be able to extract all necessary metadata, information from the file system, semantic information of single documents and save them into an ontology. In addition, a mutable set of policies inside the SSE is needed. Through these policies it should be possible to capture additional knowledge and connect it with the ontology. This additional knowledge is not implicitly available and can not be derived through extraction algorithms from the data pool itself. To avoid sophisticated, technical modifications as described by [16], a domain expert should become the informal interface between the company's processes and the SSE. So the SSE operates semi-automatically, whereby existing

[7] Small and Medium-sized Enterprises

information are extracted automatically and additional knowledge is generated manually. An advantage of this procedure is the possibility to consider different processes and requirements of a company.

Another requirement for the SSE is, that it just provides advisory functions for the SMS. The SMS may consult the SSE, but must never lose control over its tasks and responsibilities. The SMS consults the SSE by requesting semantically related documents to a given source document or a set of source documents. Furthermore, it should be possible to inform the SSE about which tasks should be performed with the source file within a request (e.g., deletion, migration, ...), in order to support a decision. To achieve these requirements the SMS needs the ability to query the SSE and correctly interpret its answer.

3.2 Software Architecture

The requirements in Section 3.1 lead to a software architecture that is shown in Figure 2. Our approach concentrates on hierarchical SMSs, but is also usable for otherSMS that use equal operations on files (e.g., migrating them between different storage tiers). The figure shows that the SMS communicates with the SSE over a dedicated interface. Thereby, requests for documents are sent to the SSE, which searches for existing semantic relations to other documents. This interface is also used to inform the SSE about every modification in the data pool, to ensure consistency to its ontology. Additionally, the SSE communicates with the SSC, which extracts all relevant information and manages them in an ontology. In the context of this analysis, a controlled data access to the SMS's files takes place, where the SMS stays in charge and provides access only to files, that are needed for the current update. Furthermore, the system policies for getting semantic associations are administrated in a decentralized way. The following sections describe the realization of this architecture in detail.

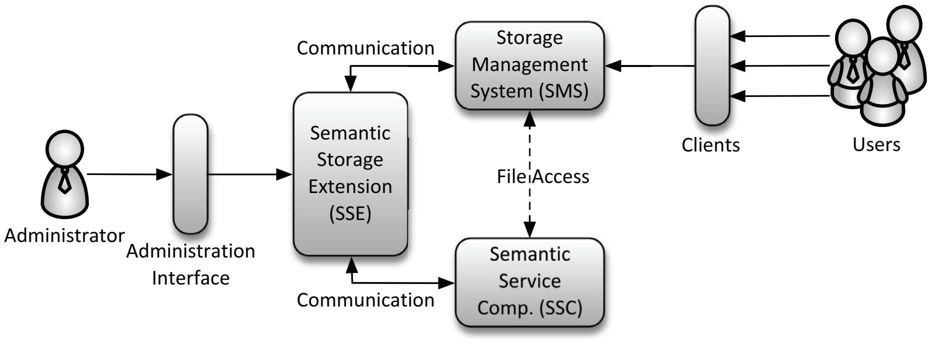

Communication
Storage Management System (SMS)
Clients
Users
Semantic Storage Extension (SSE)
File Access
Administrator
Administration Interface
Semantic Service Comp. (SSC)
Communication

Fig. 2. The software architecture

3.3 Functionality of the Semantic Service Component

The SSC needs access to the data pool to analyze documents in the SMS. To decide which data should be analyzed and stored, the storage system has to specify these documents, whereby two procedures are possible. The first one is a complete analysis. This is normally triggered when the SSE is activated for the first time or if a full consistency check should take place. The second procedure is a partial analysis, which is called at runtime, whereby only modified, removed or new documents of the SMS are examined.
During the analysis, the SSC has to extract all available information regarding files and save them in an ontology. This information can be categorized as follows:

1. File system information (e.g., resident attributes like filename or -type)
2. Metadata (e.g., ID3[8], TEI[9], EXIF[10])
3. Implicit semantic knowledge (e.g., CBIR[11], face recognition or audio analysis)

Another important requirement is the extensibility of the SSC to integrate new extraction algorithms for future file formats or metadata standards.

The available information about documents is very heterogeneous regarding their attributes and parameters. Different identifiers sometimes got the same meaning (e.g., *author/creator*, *creation/date of creation*). This leads to a lack of interoperability. Therefore, navigation or search in the knowledge base is very expensive. To achieve a good performance with an increasing data pool, we broke down all semantic information to four fundamental dimensions: *Person*, *Place*, *Topic* and *Moment*. Figure 3 illustrates the structural layout (schema) of information for a specific document inside the ontology. All extracted information is reduced to these four semantic concepts which are instantiated at runtime and stored in the ontology. So they represent a very compact document knowledge base. The ontology schema allows to ask about the *who*, *where*, *what* and *when* in context of a document.

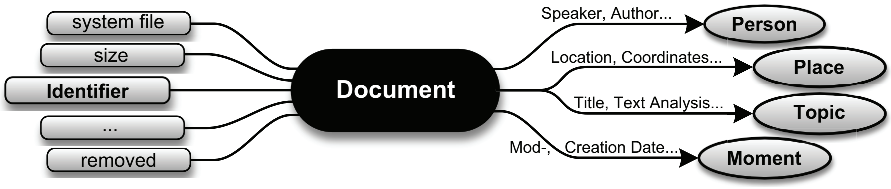


Fig. 3. Used ontology schema to describe a document's context information

[8] Identify an MP3 (metadata for audio files), http://www.id3.org
[9] Text Encoding Initiative (description of text documents), http://www.tei-c.org
[10] Exchangeable Image File Format (metadata for images), http://www.exif.org
[11] Content Based Image Retrieval

Figure 3 shows the semantic concepts and below them extracted feature (*Speaker*, *Location*, ...) which led to the specific concept. For example, an extracted *speaker* leads to a designated *person*and a document *title* leads to a specific *topic*. Furthermore, each document has basic properties which are derived from the file system (*system*, *size*, ...). To differentiate between removing and irrecoverable removing, the property *removed* marks a document as removed. So it can recovered until is was irrecoverably deleted.
We use the K-IMM[12] system [10] as a demo SSC. It especially offers most of the functionality we specified. In particular the extraction mechanism for file system information, metadata and content based information retrieval. Additionally, K-IMM has a modular structure, which easily allows the extension with new extraction mechanisms. Furthermore, K-IMM stores generated knowledge in an ontology and uses the ontology schema described in Figure 3.

3.4 Modifications in the Storage Management System

To interact with the SSE, the SMS needs some modifications. Basically it has to be able to request the SSE for semantically related documents. Also it is necessary that the SMS interprets the answers. To ensure consistency of the ontology it is important that modifications in the data pool are immediately delivered to the SSE. Last but not least, a common interface is essential to realize those tasks. The following subsections illustrate the modifications which have to be done.

Requesting the SSE According to the requirements (Section 3.1) the SMS can use two different types of requests. We call them *simple request* and *directed request*. Simple requests only pass the identifier (ID) of a file in the SMS. Each file that has stored metadata records got an ID (see Figure 3) that is used to link a file to its metadata inside the ontology. Simple requests are used to get a list of semantically related documents to a source document without any additional knowledge. The second type are directed requests which have a second parameter. This informs the SSE about the planned action for the source file (inside the SMS). Currently our concept supports the three core operations: migration, retrieval and deletion. The aim of a directed request is not only to receive a list of semantically related documents, but also to receive a recommendation for a given action.

Another important aspect are concurrency issues. To fulfill the requirements, the SMS should never wait for a response from the SSE. It has to be guaranteed that the SMS can process its tasks without depending on the SSE. The SMS just gets the advice to request recommendations from the SSE before starting a planned task and then to integrate the response to its workload to optionally re-schedule future actions.

Between all files in the storage system, there are at least weak semantic associations. For example, all files are stored in the same file system, that are

[12] Knowledge through Intelligent Media Management

managed by the same SMS. Since the SSE also considers weak semantic bindings, a response list could be very large. In the worst case, the response set contains all files of the storage environment ordered by their relevance. At this point another parameter for requests is introduced. This one is optional and is used to limit the size of the response list to a request. This threshold parameter represents the maximum disk space consumption in megabyte for related files. For example, if only 2 GB of data could be retrieved from a specific storage tier, the value of the parameter has to be 2048. So the response list only contains as many related documents as the SMS can use.

Response Interpretation Responses of the SSE basically contain a sorted list of file identifiers (FIDs). The order and their interpretation is influenced by the type of request. Usually, first listed FIDs have a closer semantic relation to a source document than FIDs with a lower rank.

In case of a directed request the SMS initially examines the recommendation of the SSE. This is constructed as a boolean value. *True* stands for an approval and *false* for a rejection. The detailed interpretation of the response can be categorized as follows:

Directed request for a migration: As long as the SSE approves the planned migration (true), it can take place. In this case no sign has been found that the source file belongs to the active data pool. So it would be recommendable for the SMS to migrate other documents from the response list, which are also most likely redundant. This helps to predictively move unused data into a slower storage tier to save costs. If the SSE responds with a rejection (false), the response list has to be interpreted in the opposite way. The source document (and its semantically related documents) seems to be relevant and active in the company. Thus, a migration to a slower storage tier is not advisable.

Directed request for a retrieval: A file retrieval works in an opposite way. For example, a rejection (false) of a specific document means that this one does not have (many) active or relevant relations to other files. Therefore it is unnecessary to move this file to a faster storage tier. Also semantically relevant documents most likely do not need to be retrieved.

Directed request for a deletion: This case requires an additional treatment by the SMS. Also the significance of the response list has to be handled with care. If the SSE calculates a high relevance or actuality for the requested file, its response contains a rejection (false). Like the other requests, this recommendation counts for all files in its response list as well.

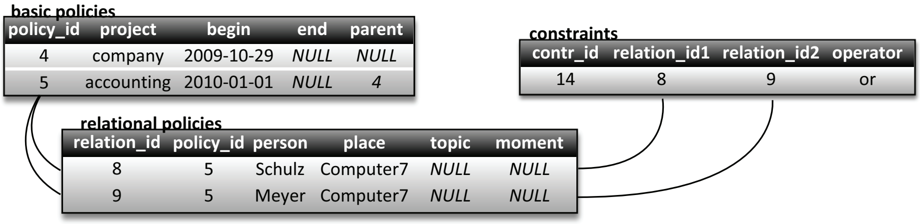
Modifications in the Data Pool As described in Section 3.3, we differentiate a partial and a complete analysis. At the first activation of the SSE and also at the regular consistency check of the ontology, the SMS initiates a complete analysis, whereas the modification of single informations just invokes a partial analysis.

For a well-structured delivery of necessary data to the SSE, the SMS needs an additional module. This module provides all affected documents as separate data streams. Because of the complexity of the data pool, all data streams should be processed iteratively during the analysis. Every data stream contains a reference to following data stream (chained list). The SSE can process them step-by-step until all modified documents, which are affected by this update, are processed. To clearly identify a data object, an FID has to be embedded into the data stream.

3.5 Design of the Semantic Storage Extension

In this section, we introduce the structure of the underlying policies and show how to formalize explicit semantic knowledge. Next, the process of reasoning will be explained, particularly, how the SSE determines related files by semantic bindings. Concluding, this subsection will show how response lists are generated and structured.

Policies With policies, the domain expert should be able to specify any knowledge about the processes and the structure of a company. A DBMS is used to store this knowledge. Below, the structure of the database tables is illustrated by the following example: “If documents were modified at Computer7 by Mrs. Schulz or Mr. Meyer, they belong to the accounting.” To express this knowledge, the domain expert can use two types of policies: *basic policies* and *relational policies* (Figure 4).



basic policies

policy_id	project	begin	end	parent
4	company	2009-10-29	*NULL*	*NULL*
5	accounting	2010-01-01	*NULL*	*4*

constraints

contr_id	relation_id1	relation_id2	operator
14	8	9	or

relational policies

relation_id	policy_id	person	place	topic	moment
8	5	Schulz	Computer7	*NULL*	*NULL*
9	5	Meyer	Computer7	*NULL*	*NULL*

Fig. 4. A simple policy example

Basic policies gather all processes which have temporal boundaries. These processes are called *projects*. *Policy 4* implies that the project *company* started on 2009-10-29. Furthermore, *Policy 5* defines the project *accounting* which started on 2010-01-01. Both projects have no defined ending. Also the *accounting* is part of *company.*

The second type of policies are relational policies which are stored in a separate table (Figure 4). They associate existing projects with persons, places, moments and topics. *Policy 8* declares that all documents which contain the person *Schulz* and were created or modified on *Computer7* belong to accounting (*policy id 5*). *Policy 9* describes the same for *Meyer* and *Computer7*.

To minimize the amount of policies, a third table (*constraints*) is introduced (Figure 4). In this table, dependencies between relational policies can be expressed by logical operators. In our example, constraint 14 describes a relation between policy 8 and 9 and combines them by an *OR*-operator. This means that only one of those policies needs to be fulfilled for a document to be matched to the accounting.

Procedure for Getting new Semantic Information Figure 5 shows how the SSE tries to find semantically related documents and which communication is necessary between the involved components. The SMS requests the SSE and inform it about the ID, and optional about the planned operation, of a source file. Initially this ID is used to query the SSC for all information on this file. The request itself is created by the SSE and formulated in SPARQL [13]. How the returned knowledge is structure, is described in Figure 3. Next, the SSE checks if existing relational policies match and if projects may be associated. For example, if a source document is related to a place *Computer7* and a person named *Meyer*, then already one of the two policies matches. As both policies are *OR*-linked, the file would be assigned to the project *accounting*.

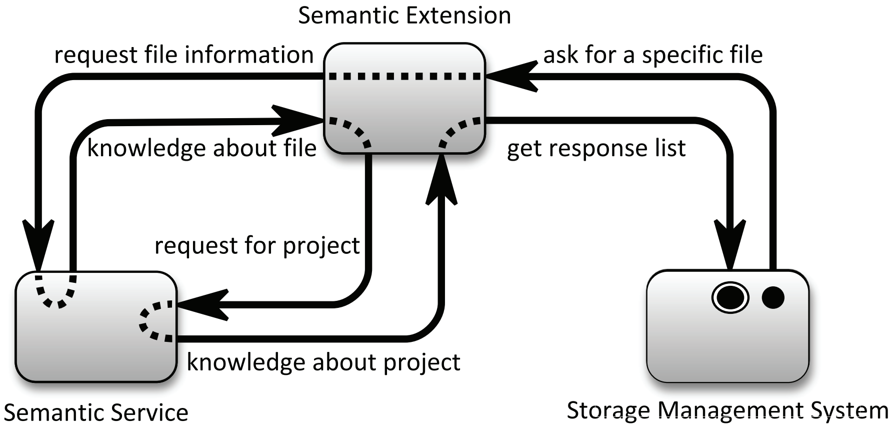


Fig. 5. Workflow for getting semantically related documents

If all projects which can be assigned to a given document were found, the SSE executes another SPARQL query against the SSC. This query determines the IDs of all files which can be associated with the found projects. For the project *accounting* this means that all IDs of files are requested which are related to at least *Meyer* and *Computer7* or *Schulz* and *Computer7*.

Thereafter the results are converted into data objects which are separated into different sets. A data object not only contains the ID of an affected file, it also contains different properties (file size, ...), attributes (system, ...) and the last access time. Each set represents documents with a simple semantic relation to the source file.

The next step is a calculation of the degree of relationship of the determined files. All data objects of the different sets are merged into one response set. Data objects that contain the property "removed" (Figure 3) are skipped. If a data object is not already available in the response set, it is added with a counter initialized with 1. This counter represents the number of determined semantic relations (degree of relationship). If a data object is already available, it is not repeatedly added. Instead the counter is incremented by 1. The result of this calculation is a set of data objects that contains different degrees of relationships for documents regarding their binding to a source file.

Before the response list can be generated, a recommendation has to be prepared, if requested. If the request of the SMS contains a deletion as planned action, the attributes of the source document are checked again. If there is a "system" attribute, the recommendation will be *false*. Otherwise, all associated projects are considered. Those projects have a specified period of time. If one of them is active (the end point is in future), the source document is also classified as active. In this case the recommendation for deletion or migration is set to *false*. Otherwise, if all related projects are inactive (the end point is in the past) the recommendation is set to *true*, which means a migration or deletion of related files is possible. If the planned action is a retrieval, the recommendation is inverted (Section 3.4).

Generating Response Lists If the request of the SMS contains a threshold parameter (Section 3.4), the SSE makes sure that the sum of all file sizes in the response list does not exceed this value. A response list is constructed as a dual sorted list of data objects with an optional recommendation flag (Figure 6). Each data object represents a semantically related document to a source file. The dual sorting offers the SMS an additional benefit on its processing. Furthermore, the order depends on the planned action. Figure 6 illustrates an example for a directed response list in case of a planned migration. The data objects themselves are labeled from *FID1* to *FID9*. The primary order is accomplished in accordance with the number of semantic bindings for each data object (curved brackets) and in the case of a migration the secondary order is accomplished with the file size, starting with the biggest file. So in case of a migration, if documents own the same semantic binding, the bigger ones are migrated first. The advantage is that just a few operations are necessary to create enough space and that a lot of small files can remain on the fast storage tier. This is particularly useful if no threshold value was submitted.

Response lists for simple requests are sorted the same way but do not contain a recommendation bit. In case of a retrieval, the secondary order is based on the file size, starting with the smallest one. This is done to retrieve as many actual

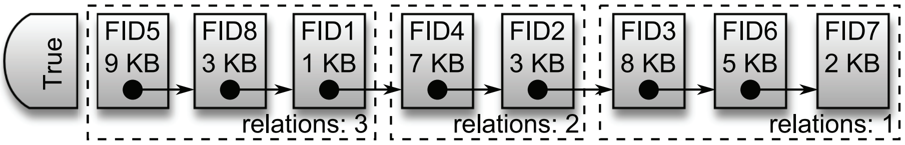


Fig. 6. A sample response list for a planned migration

documents as possible. However, on a deletion, the SSE does not consider file size. Here, the secondary order is based on the last-access-time-property of data objects. Thus, older files are listed first.

4 Conclusion and Further Work

A great amount of data will force software designers to implement more efficient and scalable algorithms to optimize storage solutions. One way to reach this goal is given by semantic web technologies.

As we have showed in Section 2, there is a great lack of publications about using semantic technologies in SMSs. Current approaches do not regard relations between documents. Only simple hierarchies are used for classifying files and folders.

In this paper, we introduced the so-called Semantic Storage Extension (SSE) (Section 3). The SSE is a software component which can be integrated in an existing SMS with just minimal effort. To enable a semantic information extraction, this architecture is completed by a SSC which handles extracted information by using a specialized ontology schema to describe documents in a semantic way. As K-IMM is used as a SSC, it is easy to add new extraction plug-ins to analyze future document formats and enable information retrieval in the way it is necessary for the SSE. Through inference algorithms that are provided by the SSC new relations between indexed documents can be found which cannot be derived by methods like data mining and clustering. Using this knowledge, the SSE can advise the SMS on planned actions for a collection of files. To fulfill the requirements of different domains, we developed an additional policy-based approach to enable a domain expert to describe the application domain in detail. With this architecture a SMS can decide on actions like migration, retrieval and deletion by using semantic knowledge.

For future work we plan to improve the implementation of this architecture in the project HSM [3]. Thereby, we want to perform tests to proof our theoretical evaluation and to show that the benefit and the performance for a SMS increases. Also, a detailed analysis need to be made on security issues (like authentication and authorization), sorting of given answers and the way files are weighted in response lists. At least the handling of concurrence issues between the SMS operations and operations of the SSC and SSE needs to be improved.

References

1. Bloehdorn, S., Görlitz, O., Schenk, S.: TagFS - Tag Semantics for Hierarchical File Systems. Proceedings of the 6th International Conference on Knowledge Management I-KNOW'06 (2006)
2. Chang, W., Masinter, L.: System and method of determining and recommending a document control policy for a document (2008), http://www.freepatentsonline.com/y2008/0059448.html
3. Fritzsche, R.: HSM-Projekt (2011), http://www.mmt.inf.tu-dresden.de/Forschung/Projekte/HSM/
4. Gnasa, M.: Fraunhofer IAIS: Text Mining und Information Retrieval (2009), http://www.iais.fraunhofer.de/4862.html
5. Hackl, G., Pausch, W., Schönherr, S., Specht, G., Thiel, G.: Synchronous Metadata Management of Large Storage Systems. Proceedings of the Fourteenth International Database Engineering & Applications Symposium pp. 2–7 (2010)
6. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall (2008)
7. Mahalingam, M., Tang, C., Xu, Z.: Towards a semantic, deep archival file system. The Ninth IEEE Workshop on Future Trends of Distributed Computing Systems pp. 115–121 (2003)
8. Mierswa, I.: Beatles vs . Bach : Merkmalsextraktion im Phasenraum von Audiodaten. Tagungsband der GI-Workshop-Woche Lernen - Lehren - Wissen - Adaptivitat (2003)
9. Mitschick, A.: Ontologiebasierte Indexierung und Kontextualisierung multimedialer Dokumente für das persönliche Wissensmanagement. Dissertation, University of Dresden (2010)
10. Mitschick, A.: Ontology-based Indexing and Contextualization of Multimedia Documents for Personal Information Management Applications. International Journal on Advances in Software 3(1), 31–40 (2010)
11. Neuroth, H., Oß wald, A., Scheffel, R., Strathmann, S., Huth, K.: nestor Handbuch - eine kleine Enzyklopädie der digitalen Langzeitarchivierung Version 2.3 (2010), http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf
12. Orio, N.: Music Retrieval: A Tutorial and Review. Foundations and Trends® in Information Retrieval 1(1), 1–96 (2006)
13. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF (2008), http://www.w3.org/TR/rdf-sparql-query/
14. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
15. Troopens, U., Erkens, R., Müller-Friedt, W., Wolafka, R., Haustein, N.: Storage Networks Explained. John Wiley & Sons Ltd., Mainz, 2 edn. (2009)
16. Weippl, E.R., Klemen, M., Linnert, M., Fenz, S., Goluch, G., Tjoa, A.M.: Semantic Storage : A Report on Performance and Flexibility. Lecture Notes in Computer Science 3588, 586–595 (2005)
17. Xiong, M., Jin, H., Wu, S.: FDSSS: An Efficient Metadata Management Scheme in Large Scale Data Environment. Fifth International Conference on Grid and Cooperative Computing Workshops pp. 71–77 (Oct 2006), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4031532

Semantic Exploration of Archived Product Lifecycle Metadata under Schema and Instance Evolution

Jörg Brunsmann

Faculty of Mathematics and Computer Science, University of Hagen, D-58097 Hagen, Germany
joerg.brunsmann@fernuni-hagen.de

Abstract. The product lifecycle spans from idea generation, design, manufacturing and service to disposal. During all these phases, engineers use their tacit knowledge to fulfill their tasks. If engineers retire or leave a company, their embodied knowledge also resigns. To circumvent such loss of important company's intellectual property, the engineer's knowledge is captured as linked data and then used as annotation for product lifecycle data models. To enable the reuse of data not only in the near-term, the product data and its annotated metadata are ingested into special long-term archives. However, achieving full preservation of semantically enriched product data requires the consideration of the linked data lifecycle which includes the evolution of schemas and instances. Such conceptualization and terminology changes pose the threat of semantic obsolescence of archived product data. Therefore, this paper describes dedicated metadata preservation functionality which respects knowledge evolution of the linked data lifecycle.

Keywords: Product lifecycle management; Linked data; Long-term preservation; Metadata; Schema evolution.

1 Introduction

Products are designed, manufactured and operated with complex, collaborative and knowledge intensive processes using tools provided by product lifecycle management (PLM) systems. During all PLM phases various actors create a large amount of heterogeneous digital product data. Automatically and manually captured metadata expressed as RDF based linked data [1] is used to annotate product models. In order to provide meaning for metadata, it is described by domain schemas which themselves are expressed in the RDF schema language which provides a vocabulary indicating how elements are to be interpreted as classes and properties.

When a product line reaches its end of life, many manufactured physical products (e.g. airplanes) might still be in operation for several coming decades in which the availability and understandability of the associated product data and metadata has to be guaranteed [13]. Due to the following legal and business reasons, the annotated product data models have to be archived and preserved for later reuse in several product lifecycle phases by various actors:

- an innovation lab engineer reuses ideation metadata to search for similar ideas that were rejected or not realized
- a design engineer reuses collaborative design rationale metadata for a product variation in order to avoid design mistakes
- an engineer compares the fuel consumption of simulated engine runs and the actual fuel consumption to validate the simulation model parameters
- a newly employed engineer reuses previously conducted and archived social search knowledge
- an engineer reuses service experiences and knowledge which is expressed as metadata for product improvements [6]
- an engineer reuses metadata which was inferred from sensor data for process improvements
- an accident investigator exploits project and provenance metadata during accident examination for social network or project organization knowledge
- a service mechanic searches spare parts according to an archived product part specification (product catalogue) which is described by metadata

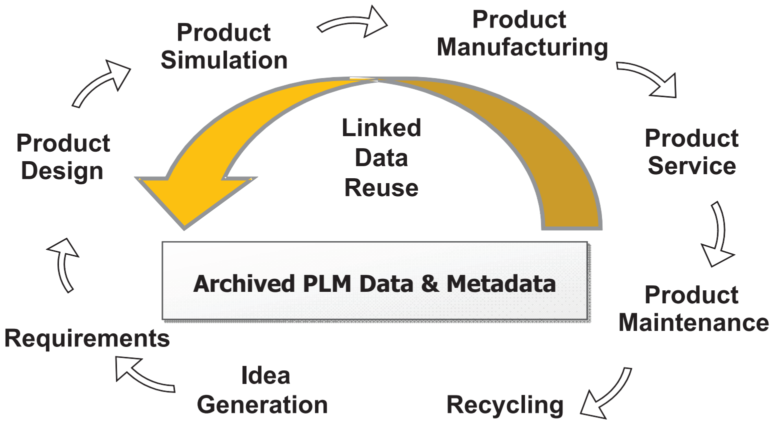


Fig. 1. The product lifecycle and linked data reuse.

Archived linked data is reused anticlockwise in the same or in a previous PLM phase (Figure 1). Linked data reuse is the last phase of the linked data lifecycle which is different from the product lifecycle. The linked data lifecycle spans from creation, annotation, archival up to the final goal of reuse. The reuse of linked data can be cumbersome, because vocabularies and linked data instances evolve due to changes in real-world phenomena. This knowledge evolution might lead to the loss of interpretation and traceability of archived data. Therefore special functionality is needed to preserve metadata under schema and instance evolution.

The remainder of the paper is structured as follows. The next section provides a characterization of the linked data lifecycle in the context of archival of semantically annotated product data models. Section 3 proposes a semantic digital archive system architecture that respects knowledge evolution and Section 4 describes an example scenario of domain schema evolution. The last section describes future work.

2 Linked Data Lifecycle

Linked data instances conform to vocabularies that make common domain knowledge explicit and usable for machines and humans. Vocabularies are expressed as schemas that enable interoperability of systems, actors, tools as well as interoperability with the future. Therefore linked data is suitable for expressing knowledge that is created during the lifecycle of a product. This section describes an idealized lifecycle of linked data in the context of semantic digital archives including the phases of capturing, annotation, archival, evolution, preservation, exploration and reuse.

2.1 Capturing and Annotation

The creation of linked data is done either automatically or manually. Automatic metadata extraction must be executed in real time because it cannot be recreated later on (e.g. simulation run with specific model parameters or metadata for project meetings). Manual metadata capturing has been implemented for the PLM environment ARAS Innovator [5] where a user is able to browse a schema and select a linked data instance as annotation of a product data model entity. This paper, does not consider the important automatic and manually metadata extraction in more detail.

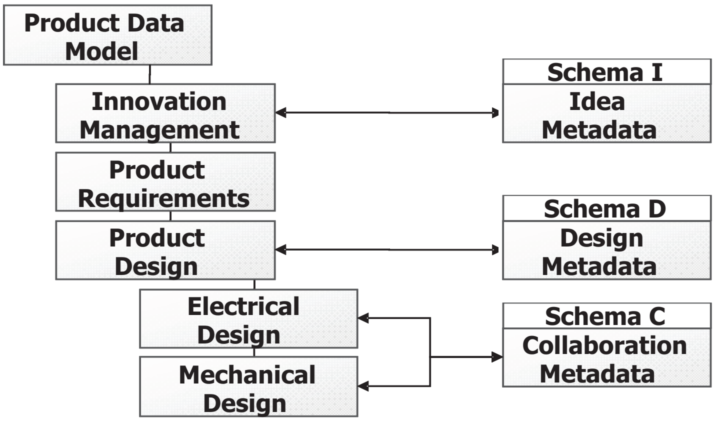


Fig. 2. An annotated product data model

Although linked data is data on its own, it can be regarded as metadata when it is used as annotation of other data. For example, product data model entities of different product lifecycle phases can be annotated with linked data (Figure 2). The product data model has entities (product part, 3D file, requirements document etc.) that describe the different PLM phases and these entities are annotated by linked data instances that conform to domain schemas. The annotated product data model is held in special repositories while the metadata can be stored and maintained external to the PLM repository. The metadata is referenced by using a unique URI and conforms to independently evolving domain schemas.

2.2 Archival

The time of archival of the product data model depends on the product lifecycle. When a product reaches its end of life, the product data model and its annotations are ingested into a long-term archive. Product lifecycle metadata is maintained by *External Systems* (ES) including collaboration capturing tools, MCAD and ECAD applications, design rationale capturing tools, etc. Figure 3 depicts the workflow execution in a PLM environment that includes long-term preservation functionality.

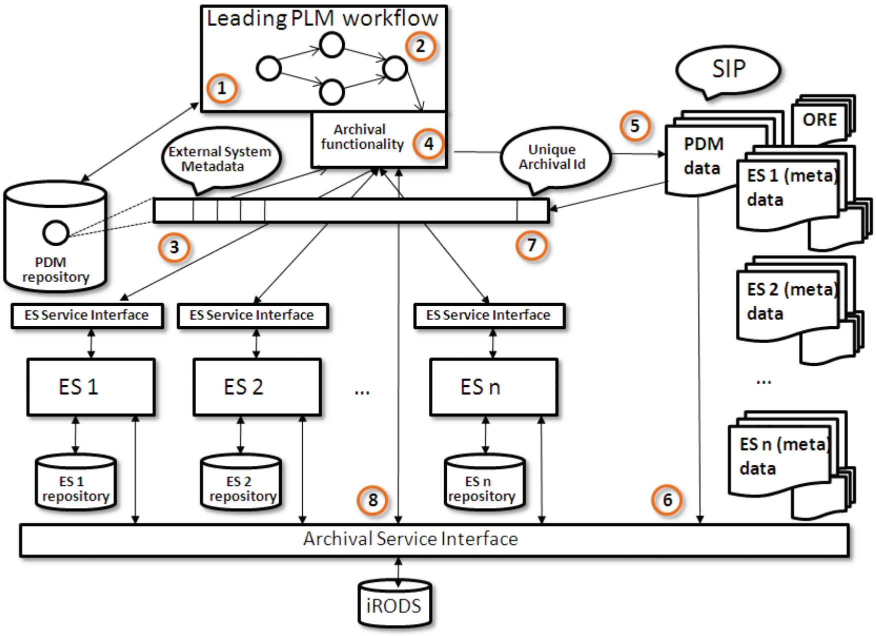


Fig. 3. Integration of long-term preservation functionality into PLM processes.

First of all, the PLM workflow (1) execution triggers the archival functionally (2) at special point in time (end of life, release for production). Because the extensible product data model contains references to external systems (3) that implement a special service interface, the archival functionality iterates over the connected external systems and collects all product relevant metadata (4). The collected data and metadata is then aggregated into an OAIS based SIP (Submission Information Package) [9] described by OAI-ORE based packaging information (5). The SIP processing also includes a normalization of data and metadata. The data normalization transfers a proprietary product data model into a standard product data model (e.g. PLCS or PLM/XML). The metadata which conforms to an external schema can be semantically normalized into metadata conforming to an archive local schema which makes preservation more controllable. In addition, metadata can also be syntactically normalized (e.g. N3, KIF). The whole product data collection is then ingested into a long-term archive (6). The long-term archive returns a unique id for the ingested product data which is stored in the product data repository (7). The product data model might be deleted from the active repository. Finally, by using the long-term archive access interfaces, product data models can be queried and accessed (8).

2.3 Evolution

Linked data model real world domains which are continually changing especially in the engineering realm due to technology innovations and knowledge explosion. The data instances and their associated schemas must reflect these changes. New versions of existing schemas are generated or new domain schemas are being invented. Such semantic heterogeneity poses a threat for archived linked data and also archived queries may become invalid. Therefore, the software application EVO (Evolving ontologies) has been implemented that allows semi-automatic generation of schema and instance mapping when a new version of a schema is generated. In addition, the tool allows detection of mapping inconsistencies during editing schema updates and it allows capturing the rationale and the provenance of schema updates. Finally, the visualization (timeline widget) of schema elements updates and instances is possible. Since the mappings are stored in the same named graph as the schema and the instances are described by a dedicated vocabulary, they are operational and can be exploited during preservation.

2.4 Preservation

In the engineering domain, the preservation of CAD data [8] and the implementation of format registries [4] are of great importance. While these aspects are topics in other research projects, the preservation of metadata (e.g. product categorization ontologies [12]) is as important as the preservation of data but has not been considered in great detail [2]. Therefore during evolution of data and schemas, mappings are generated (see above) which can be used to preserve metadata. The preservation of metadata must include migration functionality as OAIS extension [10] which requires that an operational change set is identified during schema and instance evolution. The change set can be pulled or pushed upon request from administration. After retrieval, the change set can be stored or executed immediately. The change sets can also be used to migrate archived SPARQL queries. After migration, metadata and queries conform to a new version of the same metadata schema or to another domain vocabulary.

2.5 Exploration and Reuse

When an archived product data model is accessed, it is likely that the archive consumer does not have an idea what has been ingested. The consumer only knows the goal of his archive exploration ambition. By using domain schemas and data instances to annotate a product model, it can be easier understood by future archive consumers. When using an archive in this standalone fashion, archived schemas might be used for exploration. However, semantic archives can also be integrated in daily business workflows (e.g integrated as an active repository into the PLM processes). Then, the integration of semantic digital archives faces the problem of evolving domain vocabularies. Fortunately, schema mappings that have been generated during the evolution phase can be exploited for *query mediation* during exploration and for *metadata transformation* during reuse in a contemporary environment.

Query mediation performs searches via other schema versions or on other domain schemas by rewriting incoming SPARQL queries without migrating metadata. Backward vertical query mediation finds archived instances that conform to a contemporary schema while forward vertical query mediation finds contemporary instances that conform to an archived schema. Horizontal query mediation finds instances based on equivalent classes and properties of other vocabularies.

Metadata transformation carries metadata which conforms to schema X into metadata conforming to schema Y upon request by the consumer during archive access. Both schema X and Y describe the same domain but with different conceptualizations modeled by different schema engineers.

2.6 Summary

The sections above described the different phases of the linked data lifecycle and their connection to long-term archival functionality (see also Figure 4).

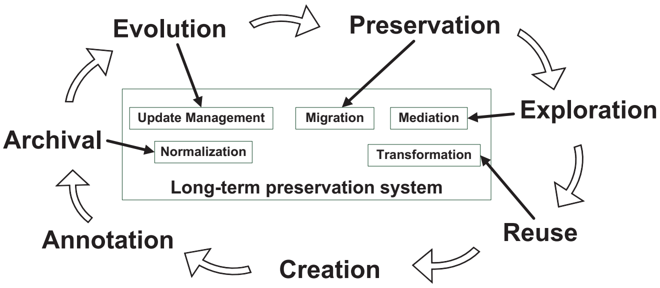


Fig. 4. The linked data lifecycle and long-term archival functionality.

During the pre-ingest phase, metadata is created and being used as annotation by a producer (e.g. engineer). At specific points in time, an engineer or administrator will syntactically and semantically normalize and archive the annotated product data models. A domain schema engineer is responsible for processing changes to data and schema. During this evolution, special tools collect operational change sets and push them to the archive or they are pulled by the archive. Upon request of administration, metadata is migrated within the archive or an archive consumer will explore the product data model by browsing the domain schemas and executing queries that might be mediated due to knowledge evolution. Finally, the archived metadata can be transformed during access of an archive consumer so that the metadata conforms to contemporary vocabulary. The transformation can be regarded as the creation of new metadata and the lifecycle starts from the beginning.

3 Semantic Digital Archive System Architecture

The previous section described the phases of the linked data lifecycle in the realm of the archival and preservation of product lifecycle data models. This chapter unites the functionality needed for handling the linked data lifecycle into a semantic digital archive system architecture that respects knowledge evolution. Figure 5 shows a system architecture of a semantic digital archive that is integrated into the daily business workflow. The architecture can be easily adopted for specific business workflows (e.g. PLM processes, library domain processes).

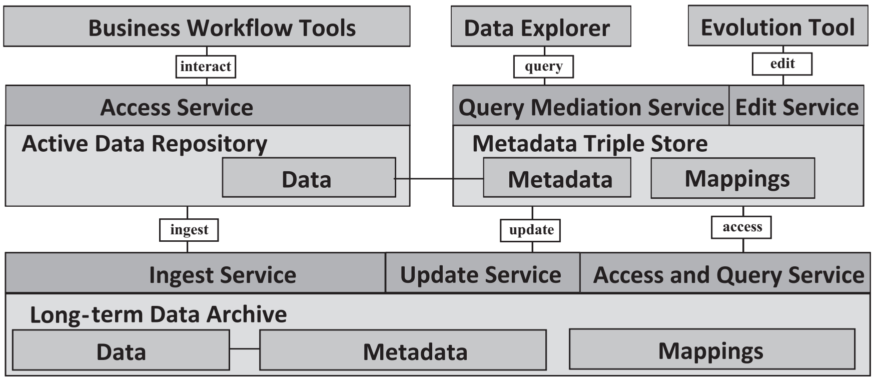


Fig. 5. Semantic digital archive system architecture respecting knowledge evolution.

The system architecture is made of three different layers:

Tool layer: the first layer contains the workflow tools (e.g. CAD design software) and a special data explorer tool that allows accessing the repository and the archive via browsing of metadata schemas. Finally, the evolution tool allows editing the metadata and the associated schemas. While editing the metadata, mappings are maintained semi-automatically.

Active (meta)data repository layer: the second layer contains the data repository and a triple store which holds the metadata. While the workflow tools interact with the data repository, the data explorer is able to query both the repository as well as the metadata repository because the metadata references the active data repository via annotations. By querying and finding metadata, product data model entities can be explored.

Archive layer: the third bottom layer contains the long-term archive functionality. The data from the active repository and the metadata is ingested into the long-term archive on demand when specific points in time of the business workflow are reached. The long-term archive also contains an access and query service that allows the data explorer to access the archived metadata. Finally, an update service is able to accept operational updates from the metadata triple store.

4 Example Usage Scenario

This section illustrates a knowledge evolution scenario (modeled as schema update) from the early ideation phase of a product lifecycle. Assuming, an engineer works for an innovation lab and he has to produce innovative and commercially attractive consumer electronic products. Innovation management software allows maintaining the idea semantics and visualization. Although nearly all of the ideas are not realized they are still important company intellectual property and therefore they are archived.

An idea contains among title, descriptive text, visual illustrations and creation date also the ideas' business category. The business category is a semantic annotation that includes concepts like Beauty Beverage Appliances, Shaving & Grooming, Kitchen Appliances, Sleep and Television. The following schema definition reflects the given scenario (namespaces prefixes are not shown).

RDFS definition excerpt of a product ideation vocabulary in the 1970s

```
:BusinessCategory a rdfs:Class .

:Television a :BusinessCategory .

:Idea a rdfs:Class .

:ThreeDIdeaFromThe70s a :Idea .

:hasBusinessCategory a rdf:Property ;
    rdfs:domain :Idea ;
    rdfs:range :BusinessCategory .

:ThreeDTVIdeaInThe70s :hasBusinessCategory :Television.
```

The schema defines the classes *BusinessCategory* and *Idea* and a property (*hasBusinessCategory*) that connects an idea with a business category. In addition, two instances are defined as *Television* (a business category) and a 3D TV related idea from the 1970s. Then, the schema and the instances are archived. Due to technology innovations, the business category *Television* has evolved into several categories (Figure 6), including the new class *ThreeDTV*.

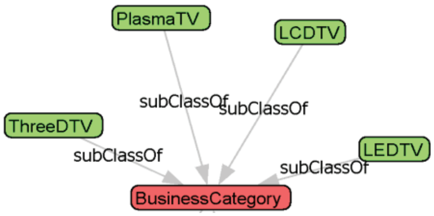
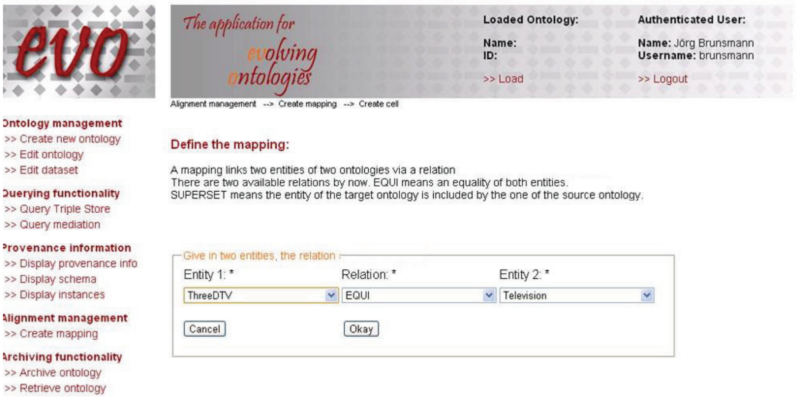


Fig. 6. Example schema evolution.

To prevent semantic obsolescence of archived ideas that contain the business category *Television*, the EVO tool allows defining a mapping between the newly introduced business categories and the previously defined category *Television*. Figure 7 shows the definition of the mapping between *ThreeDTV* and *Television*.



Name: Jörg Brunsmann
Username: brunsmann

Define the mapping:

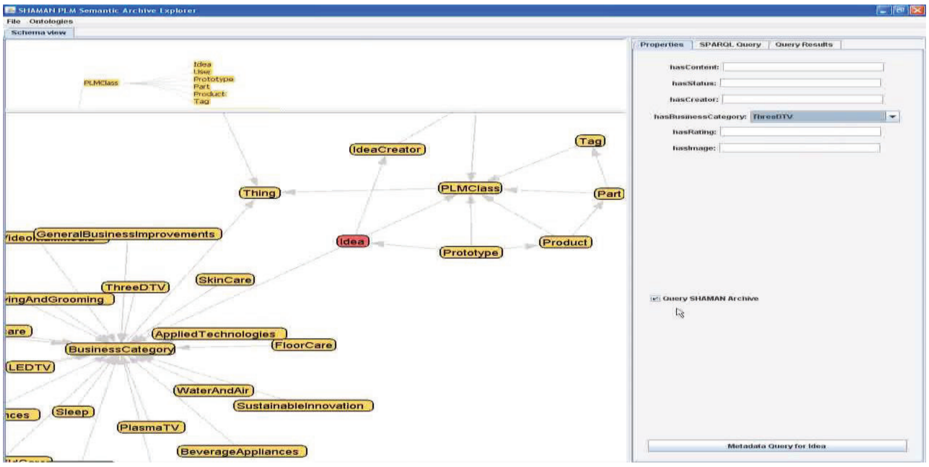
A mapping links two entities of two ontologies via a relation
There are two available relations by now. EQUI means an equality of both entities.
SUPERSET means the entity of the target ontology is included by the one of the source ontology.

Give in two entities, the relation

Entity 1: *	Relation: *	Entity 2: *
ThreeDTV	EQUI	Television

Fig. 7. Mapping definition between the ‘ThreeDTV’ and ‘Television’ business category.

Now, the engineer has a 3D TV related product innovation idea and he remembers that 3D TVs were already envisioned in the 1970s. The engineer wants to explore the active idea repository as well as the long-term archive in parallel because he don’t want to reinvent the wheel or the same idea was probably already rejected for some reason or the engineer wants to get inspirations by studying similar ideas.



hasContent:
hasStatus:
hasCreator:
hasBusinessCategory: ThreeDTV
hasRating:
hasImage:

☑ Query SHAMAN Archive

Fig. 8. Semantic exploration of archived product data under knowledge evolution.

The engineer uses a special semantic exploration tool, to search for ideas (Figure 8). This tool shows on the left-hand side a domain schema which has been loaded from the archive or from the repository as a two dimensional graph. By selecting a class from the schema, its properties are displayed on the right-hand side. For example, the list of business categories can be selected from a drop down box. Since the schema has evolved, the *Television* business category is not available any more (only *ThreeDTV*). Fortunately, a checkbox can be used to indicate that the search should also be executed in the archive. By doing so, the previously defined mappings between the business categories can be exploited so that archived ideas conforming to *Television* category are also part of the result set.

5 Summary and Outlook

While [3] described a high level integration of long-term archival functionality into PLM processes, this paper derived a semantic digital archive system architecture respecting knowledge evolution by investigating the linked data lifecycle. First, metadata is created and used as annotation. Then, during archival, metadata is syntactically and semantically normalized before it is ingested. Upon request, the metadata and queries can be migrated within the archive. While searching for archived data, the incoming queries can be mediated without migrating the metadata. Finally, during reuse the metadata can be transformed to contemporary schemas. The migration, transformation and mediation functionality depend on operational change set that have been collected during linked data evolution.

Future work includes the evaluation of the linked data lifecycle not only for semantic archives but also for single web sites and the whole web of data. Also, the preservation functionality is currently being implemented as standalone prototype application. The integration as metadata services into an OAIS archive has to be done. In addition, a three dimensional interface for browsing and understanding archived schemas can be evaluated. Finally, annotated RDF [11], multidimensional RDF [7] or the approach described in [14] can be explored for archival of instance evolution.

Acknowledgments. This paper is supported by the European Union in the 7th Framework within the IP SHAMAN [3].

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web & Information Systems. Vol. 5, Issue 3, pp 1-22 (2009)
2. Brunsmann, J., Wilkes, W.: State-of-the-art of long-term archiving in product lifecycle management. International Journal on Digital Libraries, Special Issue on Persistent Archives (2011)
3. SHAMAN Project: SHAMAN Homepage. htp://shaman-ip.eu (2009)
4. KIM Project. Knowledge and Information Management Grand Challenge Project. http://www-edc.eng.cam.ac.uk/kim/(2009)
5. Aras Innovator Homepage: http://www.aras.com (2011)

6. Brunsmann, J., Wilkes, W., Brocks, H.: Exploiting Product and Service Lifecycle Data. 8th International Conference on Product Lifecycle Management, Eindhoven, Netherlands, July 11-13 (2011)
7. Gergatsoulis, M., Lilis, P.: Multidimensional RDF. In Proc. 2005 Intl. Conf. on Ontologies, Databases, and Semantics (ODBASE), Vol. 3761, Springer, 1188–1205 (2005)
8. LOTAR Project. Long-Term Archiving and Retrieval. http://www.lotar-international.org (2011)
9. CCSDS Reference model for an Open Archival Information System (OAIS). Blue Book, Consultative Committee for Space Data Systems. Also published as ISO 14721:2003. http://www.ccsds.org/documents/650x0b1.pdf (2002)
10. Brunsmann, J.: Product Lifecycle Metadata Harmonization with the Future in OAIS Archives. International Conference on Dublin Core and Metadata Applications, The Hague, Netherlands, September 21-23 (2011)
11. Lopes, N., Polleres, A., Straccia, U., Zimmermann, A.: AnQL: SPARQLing Up Annotated RDFS. In The Semantic Web - ISWC 20010, 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11 (2010)
12. Hepp M., Leukel J. and Schmitz V.: A Quantitative Analysis of Product Categorization Standards: Content, Coverage, and Maintenance of eCl@ss, UNSPSC, eOTD and the RosettaNet Technical Dictionary (2007)
13. Heutelbeck, D., Brunsmann, J., Wilkes, W. Hundsdörfer, A.: Motivations and Challenges for Digital Preservation in Design and Engineering. First International Workshop on Innovation in Digital Preservation, Austin, Texas, USA, June 19 (2009)
14. McBride, B. Butle, M.: Representing and Querying Historical Information in RDF with Application to E-Discovery. 8th International Semantic Web Conference, Washington, USA, October 25-29 (2009)

Towards a Semantic Data Library for the Social Sciences

Thomas Gottron[1], Christian Hachenberg[1], Andreas Harth[2] and Benjamin Zapilko[3]

[1] WeST – Institute for Web Science and Technologies, University of Koblenz-Landau, Koblenz, Germany
[2] Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] GESIS – Leibniz Institute for the Social Sciences, Knowledge Technologies for the Social Sciences, Bonn, Germany
{gottron, hachenberg}@uni-koblenz.de, harth@kit.edu, benjamin.zapilko@gesis.org

Abstract. Quantitative research in the Social Sciences heavily relies on survey and statistical data. While researchers often put a lot of effort in generating such data, the incorporation and reuse of existing data on the web is far behind its potential. The lack of reuse can be attributed to various deficits in terms of library services, in particular, common exchange formats, annotations with metadata or standard approaches for integrating and merging data sets as well as the lack of an easy approach for searching data records and a lack of publicly available data sets. To overcome such problems which in the past have already been addressed by libraries, we propose a framework for seeking, merging, integrating and aggregating distributed statistical and survey data based on open semantic formats. We present a first prototype implementation as show case for the framework and highlight the benefits for social scientists.

Keywords: Semantic Digital Data Library, Linked Data, Statistics, Data Integration

1 Introduction

Libraries and archives follow a long tradition in surveying, collecting and classifying available knowledge and in providing access to these high quality information resources. With the distributed publishing paradigm of the web, providing such services has grown in complexity, driven by a multiplicity of exchange formats, different terminologies for metadata annotations and missing connections between distributed data sets. However, researchers cannot use distributed data on the web in the same way as they are used to in libraries and archives. One reason is that Digital Libraries and Digital Archives are often still disconnected from each other – not only because of historical and disciplinary reasons, but also because they use different standards and formats.

Research in the Social Sciences often relies on empirical data for studies. The emerging field of "Computational Social Sciences" leverages the possibility of collecting and analysing large-scale datasets to potentially reveal patterns of behaviour of individuals and groups [16]. The necessary data for such an approach is often difficult to find, integrate and process, which is due to a mostly decentralised and historically grown distributed publication and archiving of data in e.g., government agencies, research data centres or universities. Scattered information due to organic growth also occurs on the web at large. To be able to judge the relevance and quality of the data for any upcoming analysis in research, it is important to gain deep insights into both data and especially its documentation. Besides descriptive standard information, the metadata of data used in analysis shall provide extensive information about methodology, sample design, necessary weights or notes on the safe and correct handling of the data concerning privacy and provenance. A lack of metadata annotation complicates the process of data search on the web as well as the comparison of different data sets, e.g., regarding concrete indicators or populations.

While sizable amounts of data useful for research are attainable through the web, the data is published in a large variety of data formats. To process and analyse data, one has to convert data into particular formats of statistic tools, and integrate data from multiple sources. In general, data conversion and integration is not a technical barrier, but the effort spent for conversion is a nuisance, especially for necessary but tedious routine tasks, such as gaining a first insight into the data, or in cases where the expected research gain is minor. All these problems hinder a reuse of available and valuable data resources.

To overcome the challenges that Digital Libraries and Archives are facing with distributed data on the web, we propose a framework for a Semantic Digital Library of Linked Data, which is relevant for research in the Social Sciences. While the framework provides central services for accessing, processing and integration of distributed data sources, their physical storage location remains distributed and will not be collected or hosted by the data library. The difficulties in searching, modelling and annotating distributed data are addressed not only on the metadata level, but also on the directly connected underlying numerical data, which provides researchers an on-the-fly usage of the data in visualisations or for statistical analysis. We present a prototype implementation which demonstrates the automatic aggregation and integration of data using wrappers and a common exchange format.

The rest of the paper is structured as follows. In Section 2 we present a use case of a typical research scenario in the Social Sciences. Section 3 provides related work regarding Linked Data and the use of semantic technologies for processing data. We present existing data formats for modelling statistical and survey data in Section 4. In Section 5 we propose a framework of key modules for a Semantic Digital Data Library. Results of a first prototype implementation are presented in Section 6 and open issues are discussed in Section 7. We conclude and present future work in section 8.

2 GESIS Use Case

As an organisation providing infrastructure for the Social Sciences, GESIS – Leibniz Institute for the Social Sciences[1] offers a wide range of different study series as well as empirical primary data from survey research and historical social research. At the beginning of any research, scientists usually have a first idea what kind of data they

[1] http://www.gesis.org/

will need and which analysis method they would like to perform on the data. For example, a researcher would like to investigate possible correlations in a correspondence analysis of unemployment rate, immigration quota and the subjectively perceived risk of unemployment in Germany. However, the desired data is only available from different authorities. While the researcher can retrieve statistics from German statistical offices, data on attitudes, behaviour and social structure in Germany is part of the German General Social Survey ALLBUS[2], which is archived at GESIS. On the web portals of GESIS, the ALLBUS metadata can be searched, so the researcher can gain insight into the documentation of the data and is able to decide, whether ALLBUS is (completely or partly) relevant to the research interests. For a decision, whether the data is suitable for the intended analysis method, a comprehensive and detailed documentation of the data is essential. Information on e.g., sample design, populations or possible bias and variance has to be provided. In case researchers would like to analyse more than one data set, the individual data sets have to be aligned, i.e., not only technically, but also considering differences in populations or aggregation levels.

Using statistics tools such as STATA[3], SPSS[4] or the R Project[5] might require the data to be converted into application-specific formats. When dealing with different data sets, it has to be clear what dimensions and samples the data is comparable to and thus how data can be matched up. For example, data from ALLBUS has to be aggregated to be comparable to any statistics, because ALLBUS is micro data and therefore determined at individual level due to its origin as survey data. The matching is mostly done manually before importing the integrated data into statistics tools, although some tools can automatically detect comparable dimensions like time or geographic regions. Finally, the research analyses data and defines and executes statistical functions, which depend on the desired analysis method such as multidimensional analysis, time series analysis, correspondence analysis or estimation procedures in complex designs [12][13][17].

After finishing research and data analysis, researchers ought to cite the used data sets in the resulting publications. Referencing the analysed data helps fellow researchers to comprehend the analysis done with the data. Data can be cited and afterwards identified by using a URI (Uniform Resource Identifier) or a DOI (Digital Object Identifier). Newly created data during research obtains an identifier only if it is published afterwards.

3 Related Work

Semantic Data Libraries and Archives address key challenges like information integration and interoperability as well as user-friendly interfaces, all supported by semantic technologies and community interactions [14]. They are the next step and further evolution of traditional digital approaches, which often lack the implementation of Semantic Web and social networking technologies. Considering a

[2] http://www.gesis.org/en/allbus
[3] http://www.stata.com/
[4] http://www.spss.com/
[5] http://www.r-project.org/

Digital Library of distributed data, semantic technologies can facilitate the integration of data from disparate sources.

In recent years the idea of Linked Open Data [3] emerged. Linked Open Data represents a way to expose, share and connect freely available data on the web using Semantic Web standards. The publication of data as Linked Open Data from a technical perspective [2] is based on common standards and techniques which have been developed for years and are established worldwide as fundamental formats and interfaces for publishing data on the web, e.g., URIs, HTTP and RDF. With the standardisation of SPARQL [19], a common technology for querying RDF data has been established. The paradigm of Linked Open Data was well received in the Semantic Web community and has encouraged organisations worldwide to publish data. In recent years a lot of statistics and other numerical data have been published as Linked Data by e.g., government agencies, statistical offices or research organisations. To find available data sources, open data repositories like the Data Hub[6] have been established, where data sets can be described and grouped. Currently a common vocabulary, the Data Catalog Vocabulary[7], for the description of such data sets is under development.

Semantic technologies can aid in the integration and combined querying of data. Both descriptions of a data set (such as author, publication date) and the data set itself (individual observations) can be encoded and interpreted by machines. Thus, the integration is made possible. We present different data formats in the next section in more detail. Both are required: descriptions of the data (e.g., author, responsible organisation) and the data itself (the individual observations). Once data has been published in a uniform base format (e.g., RDF), machine-supported integration is possible. There are several services possible on integrated data, for example keyword search [15] or faceted browsing [22]. VisiNav in particular offers navigation functionality over data integrated from the Web [10]. OLAP clients may be used to perform analysis queries on the integrated data. An overview on semantic web search is given in [21]. Another way to query the data is via SPARQL. The SPARQL plugin for the R Project[8] - an open source software environment for statistical computing - allows for the formulation of SPARQL queries within R and the use of the retrieved Linked Data for statistical calculations.

Retrieving and analysing data on the web is nothing new to researchers in the Social Sciences. Data providers of statistical or survey data are very keen on offering the possibility for browsing, analysing and downloading their data, even if it is only metadata due to privacy restrictions. Examples are ZACAT[9] and SOEPinfo[10]. Both portals offer a wide range of tools for processing, analysing, the visualisation and export of data to different data formats. However, both are restricted to the data holdings of their particular organisation. A web-based application which is more open is GraphPad QuickCalcs[11], a collection of free online services for e.g., statistical calculations based on data manually entered by the user. However, calculations are

[6] http://ckan.net/

[7] http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary/Vocabulary_Reference

[8] http://cran.r-project.org/web/packages/SPARQL/

[9] http://zacat.gesis.org/

[10] http://panel.gsoep.de/soepinfo/

[11] http://www.graphpad.com/quickcalcs/index.cfm

only possible on single numbers and not on entire data sets. As yet, such data analysis tools are not empowered by semantic technologies. However, [9] identify large potential impact in the use of such technologies and available Linked Data for research activities in the Social Sciences.

4 Data Format for Statistical Data

When considering Data Library services for statistical and survey data, the proper format to store and exchange/transform data is a key component. In this section we present the formats which are most relevant to our task.

SDMX (Statistical Data and Metadata Exchange) [20] was established in 2002 by key players in the field of statistical data, such as the World Bank, IMF and the European Central Bank. Paramount was the ability to enable automatic machine-to-machine exchange of data, which requires a self-expressive or self-descriptive metadata model. SDMX defines representations of statistical data and respective metadata annotations, not only for single data items but also for full data sets. The SDMX information model is based on named concepts which are assigned *dimensions* and *attributes*. Dimensions can be grouped into so-called *keys* using *code lists* for available realisations; plain free-text is allowed as well. *Data Structure Definitions* assemble all these components with respect to a specific topic or data source in a well-defined structure. In this way, multidimensional statistical data can be represented by the SDMX information model. As we will elaborate below, parts of SDMX are reused in the definition of the Data Cube metadata model.

SCOVO (Statistical Core Vocabulary) [11] is an RDF-Schema based, lightweight vocabulary for representing statistical data. As such, SCOVO aims for an eased community uptake (since statistical data formats in general are rather complex to use) and promotes the Linked Data publishing principles, which on the one hand require use of RDF and on the other hand include re-usage of existing and well-established vocabularies, such as SKOS (Simple Knowledge Organization System). SCOVO thus fosters extensions both on the schema and instance level. Another important design issue for SCOVO was – in line with SDMX features – the ability to handle as many dimensions as necessary (supporting a multidimensional model). Compared to SDMX's focus on generic and efficient data exchange, SCOVO has weaknesses under this aspect. Being part of the Web of Data and complying to RDF standards as message format enables both self-descriptive data items and generic data exchange. SCOVO consists of mainly three principal classes: *item, dimension* and *dataset*. The first describes a single observation or event. The second describes and identifies the contents of an item whereas the latter is made up of a number of items sometimes, also defining a concept (which is provided as SKOS concept).

The RDF **Data Cube** (QB) vocabulary and its metadata model [5] is another way of representing multidimensional statistical data in RDF following the Linked Data principles (and can be seen as successor of SCOVO). To date, the vocabulary still exists only as a draft but is supposed to become widely accepted in the future due to its various advantages (see also paragraph 5.2). In particular, QB incorporates all the features of SCOVO but goes beyond some of its limitations. The Data Cube vocabulary makes use of relevant parts of the SDMX information model. For the RDF part, QB can use language descriptors of SKOS [18], FOAF [8], VoiD [1] and Dublin Core terms [7]. The metadata model of Data Cube implements the idea of a multidimensional „cube“ where all data points (i.e., observations) are aligned along certain edges and one can cut „slices“ through the cube to get cross-section and low-dimensional data views. QB also has components like *dimension*, *measure* and *attribute* which are all set up in a *data structure definition* class. The semantics of dimensions and attributes are similar to SCOVO or SDMX. Dimensions describe what is observed when considering a single data item whereas a measure describes the overall phenomenon being measured or represented for a single observation. Statistical concepts can also be defined and assigned to a SKOS concept, similar to SCOVO. Furthermore, one can add metadata to data sets using Dublin Core terms or to single observations using the attribute component. Observations are organised in data sets and hold the actual values which are categorized by dimension, measure and attribute, in turn. According to [5] Data Cube is unique in its features compared to SCOVO.

In contrast to aggregated data, so-called micro data in the Social Sciences is described by the **DDI (Data Document Initiative)** [6] metadata specification, which is an international standard describing and maintaining survey data in the social, behavioural and economic sciences. One of the key features of the DDI format is the documentation of the entire research data life cycle, which includes activities on data from the conceptualisation, collection and processing of survey data to their analysis and archiving. The complexity of DDI enables the possibility to document data very extensively, which is necessary for researchers to search and judge data according to relevance and quality. Because micro data is an important basic source for aggregated data, there are crucial similarities and overlaps. However, existing mappings are often undocumented. Since 2009, a working group is defining a detailed mapping between DDI and SDMX. Until now, there is no representation of DDI in RDF, but the process of designing a DDI ontology has begun [4].

5 A Framework for a Semantic Library of Statistical Data

To address the key challenges for semantic library services for survey and statistical data in the Social Sciences, we introduce a generic framework. The framework is composed of modules for identifying and exchanging, searching and integrating, evaluating and publishing data. Thereby we address the main obstacles for reusing statistical or survey data in the Social Sciences, also related to our GESIS use case.

5.1 Common Identifier Format

Identification of data sets, measurements or dimensions is of importance for a variety of reasons. On the data level a unique identifier allows for referencing the data set itself. Referencing is crucial in the context of making data sets citeable in scientific publications, thereby providing valuable metadata about the scientific work. Within the data, the identifiers provide a way to identify the semantics of dimensions, measures and observations. URIs fulfil this requirement and are a core ingredient to semantic web technologies. With respect to integration and aggregation of data sets, in particular the semantics of the dimensions is of interest.

5.2 Common Exchange Format

There are a couple of well-established and proven formats for statistical calculations. Amongst others, Excel spreadsheets, SPSS, SAS, Stata or R native formats are used to carry around data including respective formulas. Unfortunately, these formats are proprietary (locked) and/or in binary format, which makes it difficult to transform data seamlessly from one format to another. Additionally, all these well-known formats do not describe their data in an expressive way, i.e., expressive enough to deliver self-explanatory data via metadata. For the purpose of a data library for the Social Sciences, it is necessary to integrate various heterogeneous data sources and perform calculations directly on data or on aggregated items coming from these sources. To achieve direct calculations, we are interested in self-explanatory or self-descriptive data sources which deliver generic structures which can be semantically processed further on. Thus, we aim for annotated or metadata-enriched data formats which promote easy exchange, integration and annotation using data from many, heterogeneous sources. These requirements are well met by the Data Cube format since it is (a) an open, non-proprietary metadata model in RDF format, (b) widely based on the established SDMX information model and also including other vocabularies, (c) provides a semantic and self-descriptive annotation of the data. Given these advantages it is likely that this metadata model will be supported by established statistics packages or that converter programs will be developed. The advantages of QB foster a thorough adoption by practitioners and facilitate an easy deployment and publication of statistical and survey data. Another advantage of Data Cube is that thanks to its flexibility and simplicity it is easy to convert existing data. In our prototype implementation presented below, we actually use efficient wrapper modules to convert proprietary or other non-semantic formats on-the-fly to the Data Cube vocabulary.

5.3 Retrieval of Statistical Data

The ability to find relevant data sets is a key factor to enable social scientists to make use of existing data sets. Therefore an efficient retrieval module is necessary for search of data being suitable for the respective research topic. Later on in the retrieval process more details about the requested data become evident, for example the granularity of specific dimensions or the frequency of observations. To provide researchers with useful information about a data set, there has to be extensive metadata available. Metadata not only supports the retrieval process itself, but has also to be considered afterwards to be able to evaluate relevance, quality and suitability for the following analysis process. For comparative research the description and attributes of for example different indicators, sample designs and populations have to allow for comparisons to those of other data sets. Eventually, the retrieval module should provide the underlying data itself.

The semantic description of the data also enables more complex search tasks. For instance, if a researcher is interested in the GDPs of European countries, the available data provides these figures in the currency of the corresponding countries and not all of the data might be provided using Euro as a currency. If a second source can deliver the conversion rate, it is possible to combine the data sets and produce the requested information. Beyond the actual retrieval of the data sets, the module will need to provide a simple interaction component to define possible common dimensions by which data sets should flexibly be merged and integrated, i.e., time or geographical areas. Therefore the task of the retrieval module is twofold: retrieve (a) metadata about the datasets (e.g., using taxonomies, as common in libraries - SKOS) and (b) the data sets themselves.

5.4 Data Linking and Integration

The semantic representation and annotation of data allows for services far beyond the simple retrieval and provisioning of data sets. As the semantics of dimensions, values and metrics is explicitly modelled in the data, automatic linking and integration of data is at a researcher´s fingertips.

To correctly join and merge two data sets it is necessary to identify common dimensions, align and map the according values and possibly aggregate some of the data entries. Based on the dimension concept in Data Cube and the possibility for semantic annotation, the identification step can be made without any efforts. Alignment of the values requires some more insights and may be achieved by a more detailed model and description of the data. On data with temporal dimension, for instance, it is necessary to define its resolution and differentiate between hourly, daily, monthly, quarterly or yearly values. Aggregation becomes necessary when there is no direct representation and the data values need to be summed, or averaged. Again the semantic description of the dimension may provide exactly the information necessary to know which aggregation function to apply.

5.5 Preview and Analysis

For any existing or newly created (by the means of linking and integration) data set, the first approach for a social scientist will typically be to take a look at some key characteristics of the data. Therefore, together with the provision of the data itself, the library will present some results of a simple statistical analysis. For existing data sets key characteristics can be pre-computed, for freshly integrated data an overview will be generated on-the-fly. Once more, we benefit from a semantic representation of the data that allows for a better notion of which characteristics will be of interest and which dimensions need to be looked at.

To make an analysis at first glance even easier, data sets should be presented in a graphical form, plotting key indicators over the main or common dimensions of integrated data sets.

5.6 Data Export and Referencing

While the preview and basic analysis can provide first insights into the data it neither can nor is supposed to replace the analysis based on a full statistics application.

Therefore the system needs to allow for exporting the data to enable downstream processing. An export service providing data sets in a selection of common formats (like CSV, Data Cube, or Excel) is crucial to feed into the individual scientific processing pipelines of research groups. Exporters are needed in particular as long as the Data Cube format itself is not supported by all major statistics tools.

As each dataset is compiled based on user-defined parameters and needs, the dataset can be reproduced at any time. Parameters can also be used in a unique identifier to a data set. Thereby data sets can be referenced and cited.

6 Prototype Implementation

The motivation behind our prototype is to investigate further areas of research utilising state-of-the-art technologies. However, we keep the focus on integration and analysis of data since search/retrieval of data on the Semantic Web is an already established field of research. To identify data items and corresponding dimensions, measures or attributes, we use RDF URIs common to the Semantic Web together with data structures defined in Data Cube vocabulary. Data Cube compliant data is generated by on-the-fly wrappers from our IT.NRW data source and by a conversion of data exported from the ALLBUS database. We do not include search capabilities in our prototype for retrieval of data sets since we only process a few data sets. We therefore enable the user to manually select the data sets to be used from a fixed list. For the integration step, all data in Data Cube format is then collected in an RDF memory store and accessed via a SPARQL end-point on top of the RDF store. In our case, we use OpenRDF's Sesame[12] library including a SPARQL interface since the prototype is implemented as a Java-based web application on an Apache Tomcat infrastructure using servlets.

The concrete task for the prototype is to integrate, aggregate and visualize data from two sources, ALLBUS and IT.NRW[13], which has to answer the (exemplary) question to find correlations between the number of votes per party and the people's ratings of economic situation (both personal and national prospect) in the German state of North Rhine-Westphalia. Hereby, ALLBUS provides survey data of individuals rating personal and national economic situation. IT.NRW provides the number of votes per party of elections to the "Bundestag" (the German national parliament) for the state of North Rhine-Westphalia. Figure 1 provides an overview of the architecture for the prototype which can be accessed online[14]. During the implementation phase we came across challenges regarding aggregation using current technologies. Since we use SPARQL 1.0 for querying, aggregation on the query level is not possible (yet) due to lack of functionality in the SPARQL language. Aggregation has to be done on application level or data modelling level. For ALLBUS data we solely aggregate on the data level. We intended to use numbers for the whole state North Rhine-Westphalia, but since we only had data from individuals

[12] http://www.openrdf.org/

[13] http://www.it.nrw.de

[14] http://lod.gesis.org/gesis-lod-pilot/ (note: German only)

from that state we did an upscaling to the whole population. Such processes can be included into metadata in order to reproduce changes on the data.

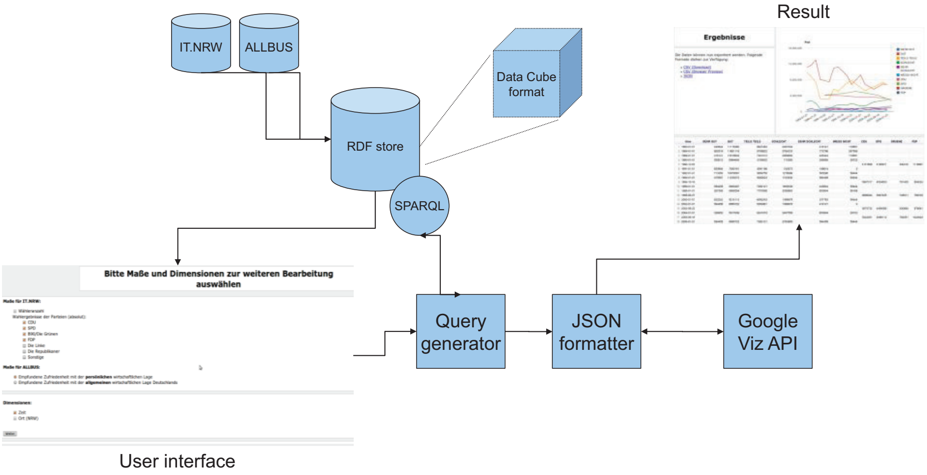

IT.NRW
ALLBUS
Data Cube format
RDF store
SPARQL
Result
Ergebnisse
Bitte Maße und Dimensionen zur weiteren Bearbeitung auswählen
Maße für IT.NRW:
Wähleranzahl
Wahlergebnisse der Parteien (absolut):
CDU
SPD
B90/Die Grünen
FDP
Die Linke
Die Republikaner
Sonstige
Maße für ALLBUS:
Empfundene Zufriedenheit mit der persönlichen wirtschaftlichen Lage
Empfundene Zufriedenheit mit der allgemeinen wirtschaftlichen Lage Deutschlands
Dimensionen:
Zeit
Ort (NRW)
Query generator
JSON formatter
Google Viz API
User interface

Fig. 1. Overview of implemented architecture

Analysis is both done visually and using lightweight calculations on integrated data. For the visualisation, we use the 2D line chart and table component from Google Visualization API which takes data in JSON format. So we transform SPARQL results to JSON just for displaying. Our visualisation allows for time-series analysis of election results in comparison to people’s future prospects by analysing line charts or table data. For an experimental implementation of two statistical methods, calculations of variance and linear regression were integrated [23] on data coming from ALLBUS and Eurostat[15]. Both calculations are performed in Java since SPARQL does not provide for calculations yet. Eventually, data can be seamlessly exported to CSV and JSON for further analysis in e.g., external statistics tools.

7 Open Issues

There are several open issues in the realisation of a large scale Semantic Data Library for the Social Sciences. Some of which are of technical nature on a higher level (relative to the technical details identified in the prototype implementation), others are more related to the research culture of the potential user community.

One rather technical issue is how to deal with privacy. Survey data is anonymised to ensure the privacy of the participants. When merging and integrating data sets these anonymisation efforts can be annulled, as the combination of information allows for identification of individuals. To avoid such problems it is necessary to formalise,

[15] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ via http://estatwrap.ontologycentral.com/

model and describe implications on the kind and type of data sets another data set may be combined and integrated with.

A similar meta-information that is crucial to a valid scientific analysis is the description of any bias present in the data. Statistical data is based on a sample of a larger population. The initial producers of such a data set are typically aware of any sampling bias they might have in the data (over- and underrepresentation of age groups, geographic location, cultural background, etc.). When publishing a data set on a library the knowledge of any bias needs to be preserved, which is of particular importance in a scenario where data sets are integrated and joined, as skewed bias may lead to wrong conclusions (e.g. joining data on perceived job-security and preferences of political parties sampled from different income groups).

To adequately address the issue with biased data as well as to enable (semi-) automatic merging, aggregation and integration of different data sources it is possibly necessary to further extend existing metadata models like Data Cube and/or complement with other vocabularies specifically dealing with data transformation. Bias in statistical data or other limitations of the data in use should have standardised support in terms of vocabulary in metadata models (e.g. descriptive comments are currently supported but lack the advantage of standardized vocabulary for automatic processing). However, more automatic data merging or aggregation needs standardised ways of applying transformation rules to deal with heterogeneous data structure. Here, specific vocabularies/ontologies for data transformation come into play, which is an open research issue.

A less technical issue is rooted in the scientific culture of the Social Sciences. The preparation and curation of data sets is a labour-intensive and time-consuming task. The work invested pays off in the production of high quality papers and an according reward in the sense of scientific reputation in the form of citations. Publishing a data set itself does not create citations (as there is no established process), and thus no scientific reputation. Therefore, data sets are rarely published, as data publication might actually bear the risk that other research groups come up with important findings quicker and thereby exploit the development of the data set without repaying the original work. While this behaviour is a cultural issue in the community of the Social Sciences, a Semantic Data Library which supports citation of data sets might have an impact on the behaviour. If a data set can be cited and thereby provide the authors with scientific credits, they might be less reluctant to publish their data. An issue related to citing data sets is the question of granularity. URIs actually allow for the “deep linking” of individual observations. How to enable fine-grained linkage and referencing with DOIs is an open question.

8 Conclusions and Future Work

We have presented a use case and associated requirements analysis for the publication of and integrated access to data relevant to research in the Social Sciences. During our analysis and discussions with social scientists we have identified the problem of locating relevant data sets, which has to be addressed before more elaborate integration and analysis functionality can be provided. We have presented a prototype implementation of a Semantic Data Library, which differs conceptually from traditional libraries due to a publishing and integration process based on distributed Linked Data. The proposed framework covers the entire life cycle from publication to accessing data via software applications and a web application. Future work includes the addition of more sources to the data collection, better ways for establishing and using mappings between the different data sets, and a live deployment and evaluation of the approach with domain experts. Finally we plan on making the service publicly available on the web to start creating a community around public survey and statistical data sets in the Social Sciences.

9 References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary, http://www.w3.org/TR/void/
2. Bizer, C., Cyganiak, R., Heath, T.: How to publish Linked Data on the Web (2007), http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 5(3), pp. 1--22 (2009)
4. Bosch, T., Wira-Alam, A., Mathiak, B.: Designing an Ontology for the Data Documentation Initiative. In: 8th Extended Semantic Web Conference (2011)
5. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube vocabulary (2011), http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html
6. Data Documentation Initiative (DDI), http://ddialliance.org
7. DCMI Metadata Terms, http://dublincore.org/documents/2010/10/11/dcmi-terms/
8. FOAF Vocabulary Specification, http://xmlns.com/foaf/spec/20100809.html
9. Gregory, A., Vardigan, M.: The Web of Linked Data. Realizing the Potential for the Social Sciences (2010), http://odaf.org/papers/201010_Gregory_Arofan_186.pdf
10. Harth, A.: VisiNav: A system for visual search and navigation on web data. J. Web Sem. 8(4), pp. 348--354 (2010)
11. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: Using Statistics on the Web of Data. In: Proceedings of the 6th European Semantic Web Conference: Research and Applications (Heraklion, Crete, Greece) pp. 708--722 (2009)
12. King, G., Keohane, R., Verba, S.: Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton University Press (1994)
13. Kohler, U., Kreuter, F.: Datenanalyse mit STATA. Oldenbourg (2008)
14. Kruk, S.R., McDaniel, B.: Goals of Semantic Digital Libraries. In: Kruk, S.R., McDaniel, B. (eds.) Semantic Digital Libraries. Springer (2009)
15. Ladwig, G., Tran, D.T.: Linked Data Query Processing Strategies. In: Proceedings of the 9th International Semantic Web Conference (ISWC '10). Springer (2010)
16. Lazer, D., et al.: Computational Social Science, Science: 323 (5915), pp. 721--723 (2009)
17. Schnell, R., Hill, P., Esser, E.: Methoden der empirischen Sozialforschung. Oldenbourg (2005)
18. SKOS Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/
19. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
20. Statistical data and metadata exchange (SDMX), http://sdmx.org/
21. Tran, D.T.: Semantic Web Search – A Process-Oriented Perspective on Data Retrieval on the Semantic Web (2010)
22. Wagner, A., Ladwig, G., Tran, D.T.: Browsing-oriented Semantic Faceted Search. In: Proc. of the 22nd Conf. on Database and Expert Systems Applications (DEXA). Springer (2011)
23. Zapilko, B., Mathiak, B.: Performing Statistical Methods on Linked Data. In: DC-2011: Proc. of the Int. Conference on Dublin Core and Metadata Applications, The Hague (2011)

Extending the Digital Archives of Italian Psychology With Semantic Data

Claudio Cortese and Glauco Mantegari

Lombard Interuniversity Consortium for Automatic Computation (CILEA)
Segrate, Italy

Abstract. ASPI is a project that aims at creating a digital library of historical documents of Italian Psychology and extending it with semantic data. The extension makes it possible to retrieve archival documents not only on the basis of archival metadata, but also according to the connections the documents have with specific activities of researchers, groups and institutions, as well as with more general events in the history of Italian Psychology. The paper provides an overview of ASPI and discusses the approach and workflow we adopted in its development. In particular, ontology modeling according to CIDOC CRM, ontology population and the prototyping of a semantic search and browsing portal based on the ClioPatria platform are introduced.

1 Introduction and Background

Today, cultural heritage represents one of the most promising and challenging areas for the application of the Semantic Web and Linked Data principles and technologies [6] [8]. In particular, digital repositories of historical archives are increasingly paying attention to and taking advantage of the new technologies, especially for what concerns the creation of highly interoperable datasets and the improvement of search functionalities beyond traditional keyword-based approaches.[7].

Our working group has a consolidated experience in the field of digital technologies applied to cultural heritage, and notably in the areas of digital preservation and web-based systems[1]. In 2007, as a part of the "Open Library of Milan" (BAMI) project, we started investigating Semantic Web technologies through creating one of the first semantic digital libraries in Italy [1]. The main objective of BAMI was to offer online access to digitized documents of different libraries and archives held by prominent cultural institutions in Milan. In particular, we focused on a subset of the heritage, which is made up of musical documents of

[1] Since 2004 we have been involved in several projects, and we have been developing the CodeX[ml] system (http://codex2.cilea.it) for the management, preservation, fruition and dissemination of library and archival (meta)data. Today, the system is used by 17 prestigious Italian institutions, which include the Ambrosian Library (Milan), the Conservatorio "Giuseppe Verdi" (Milan), and the State Archives of Milan and Venice.

the 19th century. The semantic dataset we created is based mainly on FRBR[2], the Music Ontology[3], and FOAF[4]. Access to the semantic repository is possible by means of a web portal[5] that makes use of Longwell[6], a faceted browser for RDF datasets developed by MIT. Longwell has been extended in order to offer different search and browsing functionalities, according to different user needs and experiences. In particular, facet-based querying has been integrated with relation browsing, with visual exploration of the RDF graph, and with temporal navigation through an interactive timeline.

Despite the efforts we put in the deployment of a user-friendly system, users (who include archivists, music professionals and more general communities of people interested in the history of music) did not always give positive feedback, especially for what concerns browsing the dataset. For example, in FRBR the concept of "book" is split into four different classes (Work, Expression, Manifestation and Item) whose meaning was difficult to understand by non-specialized users when navigating in the repository. In addition, some users felt slightly uncomfortable with the faceted-browsing approach and the way search results are presented. Nevertheless, BAMI has been altogether a successful project, not only because it offered us the opportunity to test Semantic Web technologies in a real application case, but also because it helped diffusing knowledge of these technologies in the communities of Italian archivists and librarians. Hence, we decided to further investigate the application of the Semantic Web to digital libraries. This has been done with particular reference to the deployment of intelligent retrieval and browsing services built on top of semantic data.

The paper introduces a new project in this area and it is organized as follows: Section 2 introduces the general characteristics of the project and motivates the choice of using Semantic Web technologies. Section 3 describes the approach and the workflow we adopted concerning ontology modeling, ontology population, and the deployment of a semantic search and browsing prototype. Section 4 summarizes the results obtained so far and outlines possible directions for future work.

2 ASPI: The Digital Archives of Italian Psychology

In 2009, a three-year project concerning the creation of a digital repository of archival documents produced by (or related to) the key figures in the history of Italian Psychology was launched. The project is coordinated by the University of Milano-Bicocca[7] and it includes several academic partners[8], each of which is

[2] http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records/

[3] http://musicontology.com

[4] http://www.foaf-project.org

[5] http://bami.cilea.it

[6] http://simile.mit.edu/wiki/Longwell/

[7] "Archivi Storici della Psicologia Italiana" resesarch group (ASPI).

[8] The University of Trieste, the University of Florence, the Catholic University of Milan, the University of Palermo and the University of Turin.

working on the study and cataloguing of important archives that are related to the history of Psychology. The technology partner of the project is the Lombard Interuniversity Consortium for Automatic Processing (CILEA), which is in charge of all the aspects concerning the development of the Digital Library.

The first phase of the project was mostly devoted to the creation of the Digital Library infrastructure, which integrates different applications offering the most important functionalities required by a modern system: accurate metadata creation and ingestion, search and browsing, interoperability and digital preservation according to international standards and protocols.

In particular, the CodeX[ml] system has been used to manage the digitized documents and ensure long-term digital preservation of both the scans and the associated metadata. CodeX[ml] is compliant with the recommendations of the OAIS model [3], and it constantly checks the validity and integrity of data and metadata during and after the ingestion phase[9] in order to prevent bit decay. CodeX[ml] is also able to provide metadata to harvesters according to the OAI-PMH standard[10], therefore enabling full interoperability with other existing repositories. Furthermore, thanks to the integration of the IIPImage server[11], high-resolution scans of the documents in the Tiled Pyramidal TIFF format can be viewed with extreme efficiency.

The AriannaWeb software[12] is dedicated to the browser-based visual navigation of a dynamically generated tree of XML-EAD[13] metadata describing the archival documents.

Finally, a web portal[14] developed with the Typo 3 Content Management System[15] allows the creation of both static and dynamic web pages. These pages provide information about the archival inventories and the historical researches carried out on them.

The Digital Library satisfies the most part of the expectations expressed by the project partners. However, it does not completely meet one of the requirements of the project, i.e. the possibility of retrieving documents on the basis of their relations to specific activities of researchers, groups and institutions, as well as to more general events that are related to the history of Italian Psychology. For example, a user may be interested in archival documents that have been produced by scholars whose activity was influenced by a specific research topic, such as “visual perception”. EAD metadata do not make it possible to answer this kind of query. On the other hand, the unstructured information contained in the pages of the web portal (which may contain relevant data) is not suitable for automated processing. Therefore, we decided to extend the digital archives with structured data that could be linked to the documents, and processed by

[9] Controls on data are done through MD5 checking.

[10] http://www.openarchives.org/pmh/

[11] http://iipimage.sourceforge.net/

[12] http://www.ariannaonline.it/web/15390/11/

[13] http://www.loc.gov/ead/

[14] http://www.archiviapsychologica.org/

[15] http://typo3.org/

machines in an intelligent way, improving search and browsing functionalities. The choice of an approach based on Semantic Web principles and technologies appeared to be the most suitable solution for our needs.

3 Approach and Workflow

Our approach to extending the system with semantic data was based on an intense collaboration with the project partners. A preliminary activity concerned training archivists and researchers in the history of Psychology about the basics of the Semantic Web, and discussing the aspects involved in ontology modeling with them. The result of this activity highlighted the need of a model where the events that are associated with the authors of the documents (such as the affiliation of an author to a particular research institution, or the different interactions between two scholars who share some lines of research) play a central role.

Therefore, we focused our attention on event-centric models and, in particular, on CIDOC CRM[16], an upper-level domain ontology for cultural heritage that is strongly based on an event-centric perspective [5]. To our knowledge, no other domain-specific models having the same characteristics and scope of CIDOC CRM exist. CIDOC CRM was used both to link "contextual data" with the documents, and to provide a semantic description of the archives, as explained in Section 3.1.

In order to allow the project partners to populate the ontology, we built a relational database using PostgreSQL[17]. Data entry is possible through a web-based interface that supports collaborative work between the different research units. We excluded the possibility of using an ontology editor such as Protégé[18] (which also has an extension for collaborative ontology editing[19]) mostly because the archivists and researchers did not feel comfortable with the tool. However, using a relational database was not a big issue, since the database schema has been mapped on the ontology, and data extraction and transformation in CIDOC CRM-compliant RDF have been done through the D2RQ[20] mapping language. The schema of the database and its mapping to RDF are introduced in Section 3.2.

Semantic search and browsing have been implemented with ClioPatria[21], a SWI-Prolog-based platform for Semantic Web applications that is also currently used as a research prototype by the Europeana project[22]. The choice of ClioPatria was motivated by the need to provide efficient means of browsing the semantic dataset, and by the lack of resources to develop our own solution. In

[16] http://www.CIDOC CRM.org/

[17] http://www.postgresql.org/

[18] http://protege.stanford.edu/

[19] http://protegewiki.stanford.edu/wiki/Collaborative_Protege/

[20] http://www4.wiwiss.fu-berlin.de/bizer/d2rq/

[21] http://e-culture.multimedian.nl/software/ClioPatria.shtml

[22] http://eculture.cs.vu.nl/europeana/session/search/

addition, using Prolog for Semantic Web applications offers several advantages, as it is discussed in [13] and [10]. ClioPatria provides different functionalities (such as semantic search, and faceted browsing) that can be easily configured and extended, thanks to the open-source license of the platform. Configuration and customization of ClioPatria according to the requirements of our project are outlined in Section 3.3.

3.1 Ontology Modeling

The ontology, which is based on version 5.0.2 of CIDOC CRM [4], was modeled through the continuous interaction with domain experts.

A fundamental part of the ontology concerns data that extend the digital archives with "contextual" information. These data take into account the following entities:

- Persons: birth, death, research activity, meeting with another person, writing of a book, writing of a paper, creation of a research instrument, participation in conferences, affiliation to a group, affiliation to an institution
- Groups: formation, dissolution, joining a group, disjoining a group, joining an institution, disjoining an institution
- Institutions: formation, dissolution, joining an institution, disjoining an institution, choice of a headquarter
- Gestalts: influence of a topic on one or more research activities

Thanks to the nature of CIDOC CRM, the identification of events and activities characterizing our domain was quite straightforward. Since we decided not to extend the model, we made an extensive use of the "E55 Type" class and the "P2 has type" property to identify different elements that are represented by the same class. For example, the "E7 Activity" class can represent both the participation in a conference and the research activity of a psychologist. Therefore, instances of E7 are associated to types that make it possible to distinguish the different activities and ease the retrieval of relevant data.

The second part of the ontology concerns mapping of some metadata of the archives to CIDOC CRM in order to link them to persons, groups, institutions, and gestalts, and the related events. Our initial intention was to map the entire EAD dataset to CIDOC CRM, following the proposals described in [12] and [11]. We soon realized that the effort required to complete the mapping was beyond the possibilities of the project, especially because of the consistent differences in the structure of the two models, as it is discussed in a very recent work [2].

The EAD elements we took into consideration concern basic metadata of archives, archival partitions, series, and single documents, such as their denomination and the date they were produced.

3.2 Populating the Ontology

In order to facilitate mapping and transformation of relational data in RDF, the database schema has been designed taking into consideration the structure of the ontology.

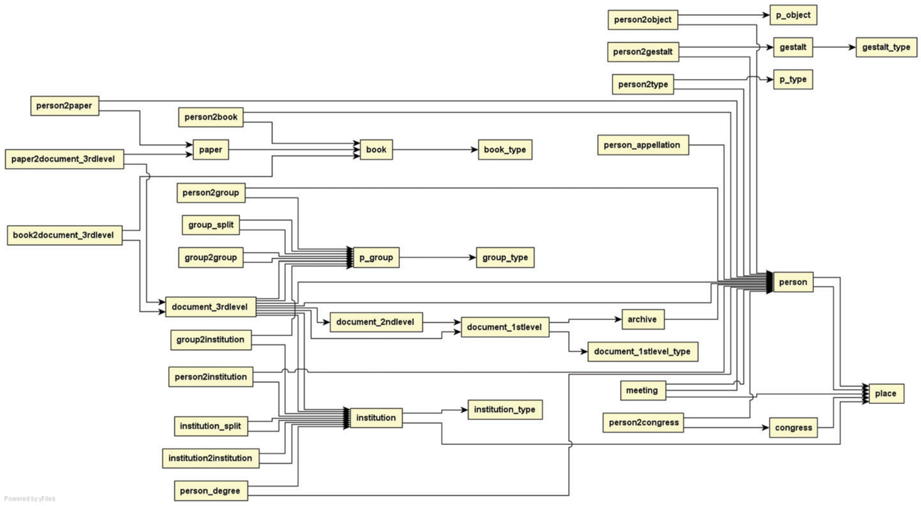


Fig. 1. A simplified representation of the database schema.

The schema (Fig. 1) includes six principal entities: persons, groups, institutions, archival documents, publications, research instruments.

Basic biographical data is represented by a series of entities and relationships that makes it possible to describe psychologists as well as other persons that fall outside the research community but can be considered relevant for the project. These include, for example, a psychologist's relatives or friends who, according to domain experts, may have played an important role in influencing research activities.

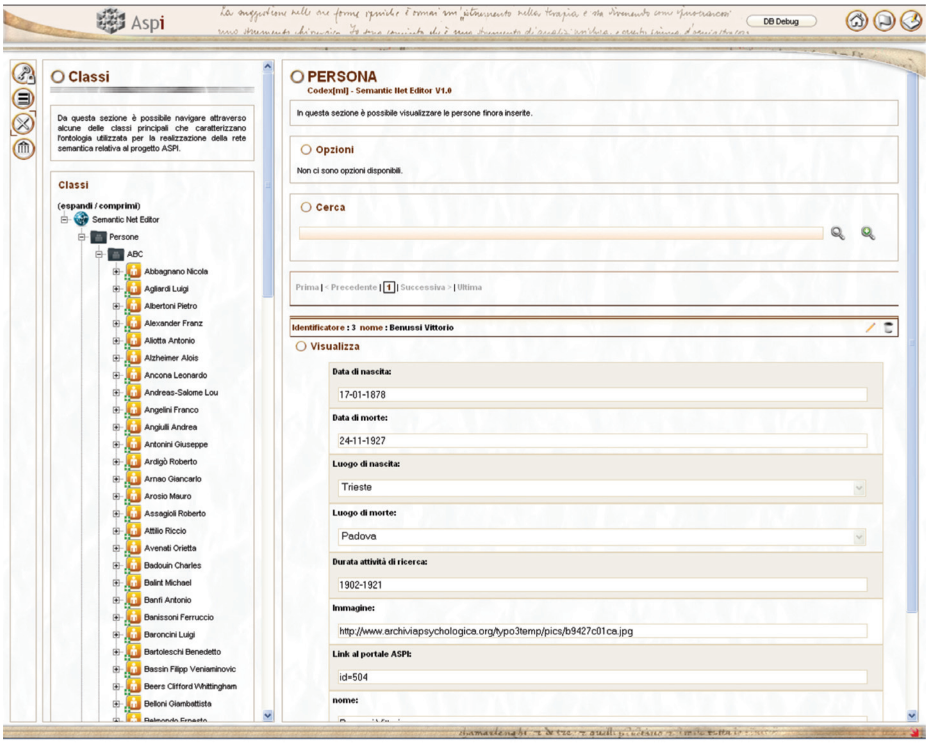
Persons are also connected to the books and papers they have written, and the scientific instruments they have invented.

The structure of the ontology greatly facilitated the development of the database, especially for what concerns the parts of the schema corresponding to events and activities such as conferences, meetings, or groups and institutions dynamics.

A part of the schema is dedicated to archival metadata and it has been populated automatically from the XML-EAD files. Thanks to the database, the documents can be annotated with the names of the persons, groups, and institutions they are related to, as well as with the papers or books for which they represent the draft version.

The web-based interface of the database (Fig. 2) allows an easy and collaborative data entry. Predefined values according to the E55 Type class instances are available in the drop-down lists.

Data extraction and transformation into CIDOC CRM-compliant RDF is very easy and efficient, thanks to the D2RQ platform. The mapping language provided by the platform has been privileged among other solutions [9] because it allows defining the mappings in a very modular and compact way using the RDF


Aspi
DB Debug
Classi
PERSONA
Codex[ml] - Semantic Net Editor V1.0
Opzioni
Cerca
Prima | < Precedente | 1 | Successiva > | Ultima
Visualizza

Da questa sezione è possibile navigare attraverso alcune delle classi principali che caratterizzano l'ontologia utilizzata per la realizzazione della rete semantica relativa al progetto ASPI.

Classi

(espandi / comprimi)

- Semantic Net Editor
  - Persone
    - ABC
      - Abbagnano Nicola
      - Agliardi Luigi
      - Albertoni Pietro
      - Alexander Franz
      - Aliotta Antonio
      - Alzheimer Alois
      - Ancona Leonardo
      - Andreas-Salome Lou
      - Angelini Franco
      - Angiulli Andrea
      - Antonini Giuseppe
      - Ardigò Roberto
      - Arnao Giancarlo
      - Arosio Mauro
      - Assagioli Roberto
      - Attilio Riccio
      - Avenati Orietta
      - Badouin Charles
      - Balint Michael
      - Banfi Antonio
      - Banissoni Ferruccio
      - Baroncini Luigi
      - Bartoleschi Benedetto
      - Bassin Filipp Veniaminovic
      - Beers Clifford Whittingham
      - Belloni Giambattista
      - Belmondo Ernesto

In questa sezione è possibile visualizzare le persone finora inserite.

Non ci sono opzioni disponibili.

Identificatore : 3 nome : Benussi Vittorio

Data di nascita: 17-01-1878

Data di morte: 24-11-1927

Luogo di nascita: Trieste

Luogo di morte: Padova

Durata attività di ricerca: 1902-1921

Immagine: http://www.archiviapsychologica.org/typo3temp/pics/b9427c01ca.jpg

Link al portale ASPI: id=504

nome:

Fig. 2. A section of the web-based interface for data entry.

Notation3 syntax[23]. In creating URIs, we tried to be as compliant as possible with guidelines and recommendations suggested by W3C[24]. The only remarkable limitation of D2RQ with reference to our project is the impossibility of creating hierarchical URIs, which would have instead enhanced human readability and understanding.

The resulting RDF dataset is based on the OWL-DL 1.0 implementation of CIDOC CRM that is known as "Erlangen CRM / OWL"[25]. As of June 2011, our semantic repository is still small (about 45.000 triples) since it is based only on initial data entered by a single project partner. Nevertheless, it is destined to increase progressively along with data entry activities that will be carried out by the other project partners in the next months.

[23] http://www.w3.org/DesignIssues/Notation3.html
[24] http://www.w3.org/TR/cooluris/
[25] http://erlangen-crm.org/

3.3 Enabling Semantic Search and Browsing

Version 1.0 beta 2.5 of the ClioPatria platform[26] was used for the creation of a portal enabling semantic search and browsing on the RDF dataset. Thanks to the administrator web frontend of Cliopatria, the basic aspects involved in RDF management (such as RDF uploading, clearing single statements or the entire repository, and querying) are greatly simplified, and triple storage is managed efficiently. Moreover, the platform is able to provide additional functionalities, such as the evaluation of RDF data quality or alignment checking.

The settings concerning the behavior of the search engine can be configured via the administrator frontend as well, making it possible to obtain in a very short time a fully functional portal for semantic search and browsing of RDF datasets.

Our customization of the ClioPatria semantic portal concerned mostly the layout elements. Beyond extending or overriding the standard CSS files, we made minor changes in the Prolog code in order to modify the parameters that were not directly configurable using the administrator frontend. These include, for example, the removal of links to display options that were not considered relevant for our portal, or the creation of a personalized layout for the home page (Fig. 3). Moreover, we made minimal interventions on JavaScript code in order to manage a few unexpected behaviors of the interface components.

Figure 4 shows the role the semantic portal plays in the overall architecture of the system. Users can search for information either by means of the Typo 3 web portal or by means of the ClioPatria engine. Once the desired document is found, its high-resolution scan as well as its metadata can be visualized respectively in the CodeX[ml] and the AriannaWeb systems.

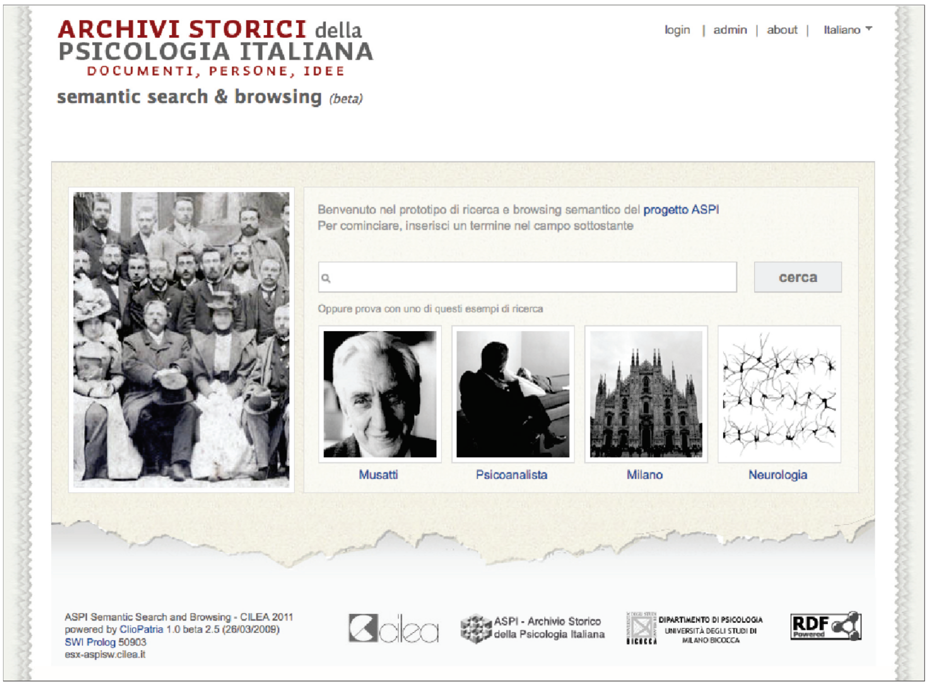
The web portal offers multilingual support with respect to the labels associated to the classes and the properties of the ontology that are shown during search and browsing. English and French versions of the labels were already available, while for Italian we took care of the translation, following the official guidelines provided by the CIDOC CRM working group[27].

Thanks to the semantic portal prototype, search and browsing through the digital library has been considerably extended. For example, now users can search for the name of a research group in the semantic portal and, among the results, see a list of documents that are in some ways related to scholars who, in a certain period of their activity, were affiliated to that group.

If a user search for “visual perception” (see the example query introduced in Section 2), the system displays also a list of the scholars whose activity was influenced by that research topic. Selecting the name of a scholar, users can obtain several data, including a list of the scholar’s documents that are present in the archives. Each item of the list is a hyperlink that leads the user to get more data about that item. Included in these data is a link to the web interface where the image of the document (as well as the images of other documents belonging to the same scholar) can be visualized in high resolution.

[26] The platform we used is based on SWI-Prolog 5.9.3.

[27] http://www.CIDOC CRM.org/translation_guidelines.html


ARCHIVI STORICI della
PSICOLOGIA ITALIANA
DOCUMENTI, PERSONE, IDEE
semantic search & browsing (beta)
login | admin | about | Italiano
cerca
Musatti
Psicoanalista
Milano
Neurologia
ASPI - Archivio Storico della Psicologia Italiana
DIPARTIMENTO DI PSICOLOGIA UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
RDF Powered

Benvenuto nel prototipo di ricerca e browsing semantico del progetto ASPI
Per cominciare, inserisci un termine nel campo sottostante

Oppure prova con uno di questi esempi di ricerca

ASPI Semantic Search and Browsing - CILEA 2011
powered by ClioPatria 1.0 beta 2.5 (26/03/2009)
SWI Prolog 50903
esx-aspisw.cilea.it

Fig. 3. The semantic search and browsing prototype homepage.

4 Conclusions

Extending archival datasets with semantic data represents an important opportunity for the creation of a new generation of digital libraries with improved search and browsing capabilities. Our project shows that encouraging results can be obtained by taking advantage of ready-to-use solutions and applications, and combining them with existing digital library systems. The preliminary feedback we received from the project partners seems to confirm we met their general expectation, i.e. extending the digital library's search and browsing functionalities with the definition of semantic relationships between the archival materials and events in the history of Italian Psychology.

However, the inherent characteristics of the ontology we used and the lack of resources to develop a completely custom presentation layer may limit the usability of the current system.

The event-centric nature of CIDOC CRM, combined with the way the standard ClioPatria interface shows search results, makes it sometimes difficult to easily obtain the desired information. For example, the title of a document created by a particular scholar can be retrieved only passing through a class that represents the activity of writing of that document. Expert users (who represent the main target of ASPI) may get easily familiar with the data structure, while more general and non-expert users may feel disoriented. A more detailed user

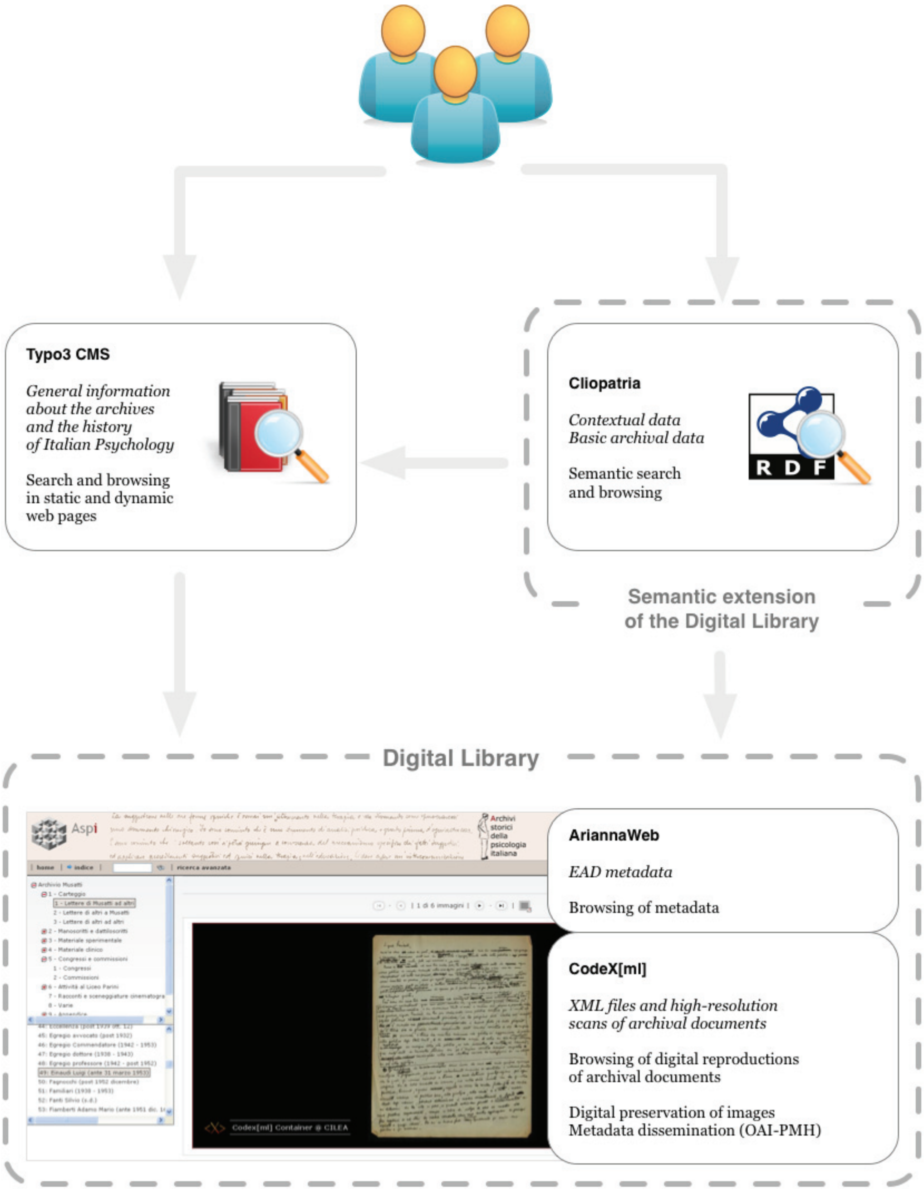

Typo3 CMS
General information about the archives and the history of Italian Psychology
Search and browsing in static and dynamic web pages
Cliopatria
Contextual data
Basic archival data
Semantic search and browsing
R D F
Semantic extension of the Digital Library
Digital Library
AriannaWeb
EAD metadata
Browsing of metadata
CodeX[ml]
XML files and high-resolution scans of archival documents
Browsing of digital reproductions of archival documents
Digital preservation of images
Metadata dissemination (OAI-PMH)
Aspi
Archivi storici della psicologia italiana
Codex[ml] Container @ CILEA

Fig. 4. The components of the system and the principal modalities of search and browsing in the digital repository.

study would help us identify the most critical aspects of the current system and define new strategies for improving the usability of the portal.

From a conceptual point of view we still think that CIDOC CRM represents a suitable model for our domain. Therefore, we are evaluating the possibility of creating a new version of the system based on a completely custom presentation layer hindering the complexity of the ontology. Version 2 of Cliopatria[28] might be a possible solution, since it provides great modularization and offers several JavaScript libraries that can be used for the design of flexible web-based interfaces.

In general, we think that ASPI is a step forward for us if compared to BAMI, especially because it offers improved searching and browsing capabilities that allow exposing the dataset in all its richness while providing a simpler user interface. However, a more detailed evaluation of the project outcomes and an extensive comparison with BAMI will be possible only with a bigger dataset integrating the cataloguing activities of the different research units.

To our knowledge our semantic dataset is the only one available today for the history of Psychology. For this reason, we are willing to define better modalities for sharing our data. To this respect, the creation of a SPARQL endpoint and the alignment of the dataset for Linked Data will be two major improvements we plan for the future, if the project will obtain additional financial support.

References

1. Barbera, M., Cortese, C., Zitarosa, R., Groppo, E.: Building a Semantic Web Digital Library for the Municipality of Milan. In: Mornati, S., Hedlund, T. (eds.) Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies - Proc. 13th International Conference on Electronic Publishing. pp. 133–154 (2009)
2. Bountouri, L., Gergatsoulis, M.: Mapping Encoded Archival Description to CIDOC CRM. In: First Workshop on Digital Information Management. pp. 8 – 25 (2011)
3. CCSDS: Reference Model for an Open Archival Information System (OAIS). Blue book, Consultative Committee for Space Data Systems (2002)
4. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: Definition of the CIDOC Conceptual Reference Model. version 5.0.2. ICOM/CIDOC CRM Special Interest Group (January 2010)
5. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
6. Hyvönen, E.: Semantic Portals for Cultural Heritage. In: Staab, S., Rudi Studer, D. (eds.) Handbook on Ontologies, pp. 757–778. International Handbooks on Information Systems, Springer Berlin Heidelberg (2009)
7. Kruk, S., McDaniel, B. (eds.): Semantic Digital Libraries. Springer (2009)
8. Nixon, L., Dasiopoulou, S., Evain, J., Hyvönen, E., Kompatsiaris, I., Troncy, R.: Handbook of Semantic Web Technologies, chap. Multimedia, Broadcasting and eCulture, pp. 901–965. Springer (2011)

[28] http://cliopatria.swi-prolog.org/home/

9. Sahoo, S., Halb, W., Hellman, S., Idehen, K., Thibodeau Jr, T., Auer, S., Sequeda, J., Ahmed, E.: A Survey of Current Approaches for Mapping of Relational Databases to RDF. Tech. rep., W3C RDB2RDF Incubator Group (2009)
10. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B.: Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. Web Semant. 6, 243–249 (November 2008)
11. Stasinopoulou, T., Doerr, M., Papatheodorou, C., Kakali, K.: EAD mapping to CIDOC/CRM. Tech. rep., Department of Archives and Library Science, Ionian University (2007)
12. Theodoridou, M., Doerr, M.: Mapping of the Encoded Archival Descripition DTD Element Set to the CIDOC CRM. Tech. rep., Institute of Computer Science, Foundation for Research and Technology - Hellas (2001)
13. Wielemaker, J., Hildebrand, M., van Ossenbruggen, J.: Using Prolog as the fundament for applications on the semantic web. In: S.Heymans, Polleres, A., Ruckhaus, E., Pearse, D., Gupta, G. (eds.) Proceedings of the 2nd Workshop on Applicatiions of Logic Programming and to the web, Semantic Web and Semantic Web Services. pp. 84–98

EAC-CPF Ontology and Linked Archival Data

Silvia Mazzini[1], Francesca Ricci[2]

[1] Regesta.exe (Rome, Italy)
smazzini@regesta.com
[2] Istituto per i beni artistici culturali e naturali della Regione Emilia-Romagna (IBC) (Italy)
fricci@regione.emilia-romagna.it

Abstract. The EAC-CPF standard is an XML schema maintained by the Society of American Archivists in partnership with the Berlin State Library used for encoding contextual information about persons, corporate bodies, and families related to archival materials. The main goal of this paper is to demonstrate the feasibility of the application of Semantic Web technology for creating Linked Open Data of descriptions of entities associated with the creation and maintenance of archives. In this paper we present two EAC-CPF ontologies and we provide an in-depth description of all phases of the work, from the study of the standard to the definition of the classes and properties of the two OWL ontologies and a case study of application in authority records of *IBC Archivi* (information system of historical archives in the Emilia-Romagna region).

Keywords. EAC-CPF, ontology, RDF, Linked Open Data, archival description.

1 Introduction

International standards for archival and encoding descriptions are known for a long time in Italy. EAD (Encoded Archival Description)[1] standard has been introduced in Italy early in public and private area, so today many archival description software use EAD schema or offer an XML export for the resources. By now, XML[1] is known as a good standards for semantic interoperability and it is often used for representing archival resources thanks to its simplicity, its flexibility and its capabilities of nesting description particularly useful to archival multi-level description.

Furthermore semantic interoperability is a *sine qua non* for the Semantic Web and today archivists have to deal with the nascent Semantic Web. It is now quite common to use links as means of connecting archival descriptions on the web to other information, in order to increase the information available to users who access archival material on the web.

Increasing development of Linked Open Data in cultural heritage leads to a review of technologies in other areas too, like e.g. the archival domain. We believe that technologies that best introduce archival description background to web of data are RDF [3] and ontologies [4]. In addition to these reasons, we can say that behind the

[1] http://www.loc.gov/ead/

idea to transform the EAC-CPF schema into an ontology and the experiment to "open" eac-cpf authority records as linked open data, there are also:

- the need to describe the resources in a format that can be shared and approved by the international scientific community;
- the choice to use standards allows to process, integrate and deal with data according to standardized rules that are supported by large communities;
- the opportunity to integrate with other web resources described with other standard vocabularies.

Starting from these considerations, we believe that a concrete solution is to use RDF and ontologies, not only as means for representing entities and the relations between the various components of the archival description, but also as an appropriate tool to qualify these relations semantically.

A few simple actions are required to be done in order to describe archival context in a "semantic way". It is necessary to:

1. identify univocally the descriptive resources by means of the URI and preferably use dereferenceable URI;
2. provide descriptions in a standard format so that the resources and their relations can be recognized immediately;
3. include in the descriptions the greatest possible number of relevant links to other information resources.

The current digital environment is clearly oriented towards a more intelligent web, able to support the sharing, enhancement and management of archival information, exploring the meaning of the documents and returning data (and not documents).

Linked Data[2] and ontologies are the technological components on which the passage from Web 2.0 to the Semantic Web is based. However, to make this change a reality, the technological components are not sufficient but it will be necessary for those who publish data on the web to do so in a "open" way, thus contributing to the realization of a truly "open" semantic web.

On the basis of these first premises, the Istituto per i beni artistici culturali e naturali (IBC) of the Emilia-Romagna Region has decided to open up its archival data.

IBC was founded in 1974 and it's the scientific and technical instrument for the Emilia-Romagna regional planning in the field of artistic, cultural and environmental heritage. The Soprintendenza regionale per i beni librari e documentari has been part of IBC since 1983, with the specific task of co-ordinating the regional policy addressed to libraries and archives[3] .

IBC develops the IT facilities that convey archives, libraries and museums data to institutions and the general public, promotes and coordinates the census and the description of archival, book and museum material, grants the readability of specific

[2] http://linkeddata.org/

[3] http://www.ibc.regione.emilia-romagna.it/wcm/ibc/pagine/01chi_inglese.htm

DBs on the web and at present IBC's working on the standards for interoperability through the use of semantic web technologies.

In March 2001 a group of archivists met in Toronto and created a high-level model for the description of individuals, families and corporate bodies that create, preserve, use and are responsible for and/or associated with archival records in a variety of ways. The group has termed the model "Encoded Archival Context - Corporate Bodies, Persons, and Families" (EAC-CPF)[4] to emphasize its important role in archival description and its relationship with the Encoded Archival Description standard.

Since the EACWG meeting in Bologna and the conference “Standards and exchange formats for interoperability among archival information systems” organized by IBC in early May 2008[5], IBC has been committed to the dissemination of EAC-CPF in the Italian context, to promoting knowledge and use of this standard by Italian archivists and archival agencies and to translate in Italian the EAC-CPF tag library[6].

The first step in this direction was the opening of a standard (by publishing an ontology for EAC-CPF in an open format and including parts of other standards within it). Afterwards a second ontology was realized to represent the EAC-CPF records containing the descriptions of archival creators published in *IBC Archivi* (information system of historical archives in the Emilia-Romagna region)[7]. These two ontologies are complementary and closely related because the experience with devising the first one has provided the basis to define the approach for devising and using the second one. In this paper we present:

- the first ontology (described in chapter 2) that is a different formalization of the XML schema of EAC-CPF standard, useful to promote and foster a better comprehension of structure and properties of the standard among Italian archivists;
- the second ontology (described in chapter 3) that was realized to open -by the semantic web- the descriptions of entities (corporate bodies, persons and families) associated with the creation and maintenance of archives;
- an example realized on *IBC Archivi* descriptions (described in chapter 4).

2 EAC-CPF standard Ontology

The EAC-CPF Schema has a fairly simple structure with much less nesting than its relative for archival description EAD: specifies 90 elements and 30 attributes[8]. The structure is designed in such a way as to maintain a division between information controlling the entity and its analytic description.

[4] http://eac.staatsbibliothek-berlin.de/about.html

[5] http://online.ibc.regione.emilia-romagna.it/h3/h3.exe/apubblicazioni/sD:!TEMP!HwTemp!3se2a84aa31d.tmp/d1/FFormDocumento?La.x=;sel.x=NRECORD%3d0000047818

[6] IBC entrusted the italian translation of EAC-CPF tag library to Salvatore Vassallo, under the scientific supervision of Stefano Vitali.

[7] http://archivi.ibc.regione.emilia-romagna.it/ibc-cms/

[8] http://eac.staatsbibliothek-berlin.de/eac-cpf-schema.html

Following an analysis of the relations between elements of the schema and attributes, we thought of proceeding to a first semantic web description of the schema (using OWL) by aiming to create a different formalization of the EAC-CPF standard, to provide a new tool for navigating the schema showing the relations, and pointing to specifications of the official tag library and the diagram of the xml schema for the technical specifications of each element.

The XML schema of EAC-CPF does not present much nesting in the description and, it was fairly simple to convert it into OWL ontology without changing the general settings of the standard and without introducing any new elements. In general, the RDF data model is based on the official schema of EAC-CPF standard. It is not proposed as an alternative standard but quite simply as a different formulation, which is useful for the semantic web and fosters interoperability.

2.1 Classes and properties of the EAC-CPF standard ontology

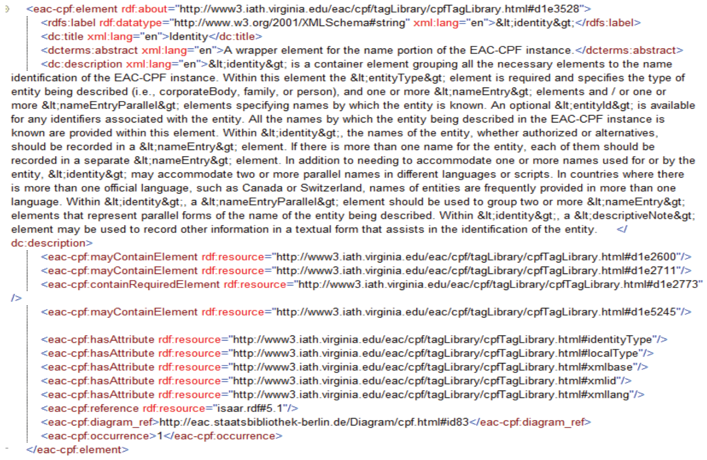
The first ontology describes strictly the domain of the XML schema so that we have created only three *owl classes* (element, attribute and controlled_value) and few properties useful to represent schema's relations.

Table 1. class and properties of ontology

Classes: element, attribute, controlled_value.

Properties: mayContainElement, containRequiredElement, hasAttribute, hasRequiredAttribute, mayContainValue, reference, isElementOf, isRequiredElementOf, isAttributeOf, isRequiredAttributeOf, isControlledValueOf, mayContainDatatype, diagram_ref, occurrence.

Fig. 1 shows an RDF serialization of description of identity element based on the ontology. URIs for the resources are URLs of the element in the tag library official web site.

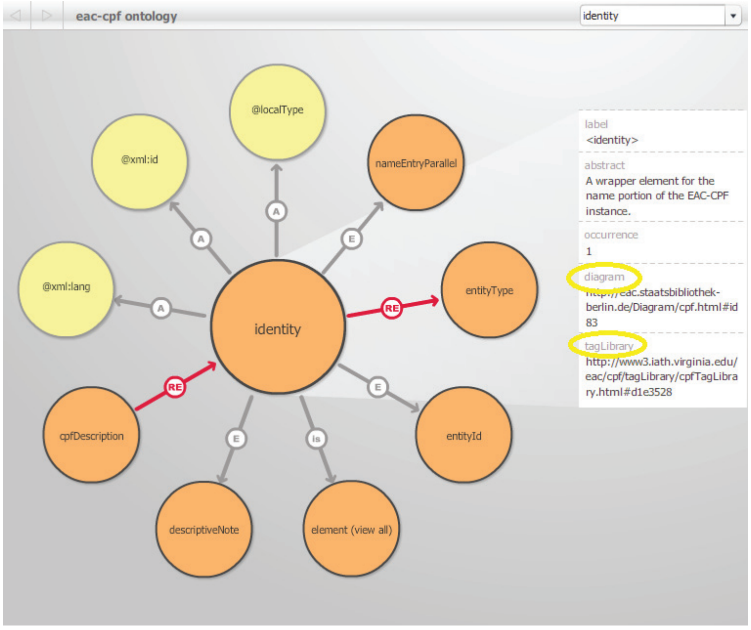


```
<eac-cpf:element rdf:about="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e3528">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string" xml:lang="en">&lt;identity&gt;</rdfs:label>
    <dc:title xml:lang="en">Identity</dc:title>
    <dcterms:abstract xml:lang="en">A wrapper element for the name portion of the EAC-CPF instance.</dcterms:abstract>
    <dc:description xml:lang="en">&lt;identity&gt; is a container element grouping all the necessary elements to the name
identification of the EAC-CPF instance. Within this element the &lt;entityType&gt; element is required and specifies the type of
entity being described (i.e., corporateBody, family, or person), and one or more &lt;nameEntry&gt; elements and / or one or
more &lt;nameEntryParallel&gt; elements specifying names by which the entity is known. An optional &lt;entityId&gt; is available
for any identifiers associated with the entity. All the names by which the entity being described in the EAC-CPF instance is
known are provided within this element. Within &lt;identity&gt;, the names of the entity, whether authorized or alternatives,
should be recorded in a &lt;nameEntry&gt; element. If there is more than one name for the entity, each of them should be
recorded in a separate &lt;nameEntry&gt; element. In addition to needing to accommodate one or more names used for or by the
entity, &lt;identity&gt; may accommodate two or more parallel names in different languages or scripts. In countries where there
is more than one official language, such as Canada or Switzerland, names of entities are frequently provided in more than one
language. Within &lt;identity&gt;, a &lt;nameEntryParallel&gt; element should be used to group two or more &lt;nameEntry&gt;
elements that represent parallel forms of the name of the entity being described. Within &lt;identity&gt;, a &lt;descriptiveNote&gt;
element may be used to record other information in a textual form that assists in the identification of the entity.     </
dc:description>
    <eac-cpf:mayContainElement rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e2600"/>
    <eac-cpf:mayContainElement rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e2711"/>
    <eac-cpf:containRequiredElement rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e2773"
/>
    <eac-cpf:mayContainElement rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e5245"/>

    <eac-cpf:hasAttribute rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#identityType"/>
    <eac-cpf:hasAttribute rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#localType"/>
    <eac-cpf:hasAttribute rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#xmlbase"/>
    <eac-cpf:hasAttribute rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#xmlid"/>
    <eac-cpf:hasAttribute rdf:resource="http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#xmllang"/>
    <eac-cpf:reference rdf:resource="isaar.rdf#5.1"/>
    <eac-cpf:diagram_ref>http://eac.staatsbibliothek-berlin.de/Diagram/cpf.html#id83</eac-cpf:diagram_ref>
    <eac-cpf:occurrence>1</eac-cpf:occurrence>
</eac-cpf:element>
```

Fig. 1. RDF serialization of Identity element

The graph below (fig. 2) shows a visualization of the same element (*identity*) of the standard, its relations with other element of the schema (orange circles) and with attributes (yellow circles); while the color of arrows and the direction clarify the type of relation.



label
<identity>

abstract
A wrapper element for the name portion of the EAC-CPF instance.

occurrence
1

diagram
http://eac.staatsbibliothek-berlin.de/Diagram/cpf.html#id83

tagLibrary
http://www3.iath.virginia.edu/eac/cpf/tagLibrary/cpfTagLibrary.html#d1e3528

Fig. 2. Graph visualization of identity element in Relation Browser

This initial study was concluded last summer with publication of ontology and graph visualization on the web site of the Libray Linked Data Incubator Group[9].

3 EAC-CPF Descriptions Ontology for Linked Archival Data

The work described in chapter 2 was extremely useful as a feasibility study and an effective work tool for archivists, but it could not be used to open the authority records codified with this standard to the world of Linked Open Data. It was necessary to transform the elements of the schema into properties of the ontology and to change the point of view of the description of the model. It was necessary to move from the description of the XSD schema in RDF to the definition of a new model based on the schema (thus maintaining the names of the elements and the attributes). For example, if you write a text in the EAC-CPF tag <bioghist> of an XML file, you mean that the text is a "history of the institution" or a "biography". If you want to obtain the same result in an RDF file, you have to change the xml element <bioghist>

[9] http://www.w3.org/2005/Incubator/lld/wiki/Vocabulary_and_Dataset

into the RDF property eac-cpf:bioghist. In this way, you assign a semantic value to the text itself.

To reach a description of the data model (that could be used for the Linked Archival Data), it was necessary to take a further step: starting from the records describing the authorities, bodies, persons and families of the *IBC Archivi* codified in EAC-CPF, we moved on to the definition of a data model based on the standard, maintaining the names of the elements and the attributes and the relations, but expressing them in RDF. In general, the following basic principles were followed:

- to make the RDF model more explicit, the three typologies of entities (that are included in EAC-CPF schemas as control values for <entityType> element), have become three distinct classes in the ontology: Person, Family, and Corporate Body as subclasses of the more general Entity;
- no new concepts have been added that were not defined in the XML schema;
- if the standard proposes the names of the elements in both the singular and the plural form, in the RDF data model only the singular forms have been maintained, since properties can always be repeated in RDF;
- the elements used in the XML schema to parcel the descriptive information were not used in the data model, aiming to group the information favouring a simpler and more general structure. For example, the element <p> present in almost all the descriptive elements was omitted, as well as the formatting elements (such as span, list, item, level, outline, etc.);
- in the RDF file some information, especially classical descriptive metadata such as title, date and author were duplicated by using other RDF terms that are universally known and used such as Dublin Core and FOAF to allow a natural interoperability with other similar resources;
- to facilitate the linking of external resources and build up the linked archival data, for all those resources for which it was possible to find alternative URIs or alternative information on other websites or with other authorities, the references were added: for example, to link the names of persons to the Virtual International Authority File (VIAF), we used a property of OWL owl:sameAs since this indicates that two URI references actually refer to the same thing - the individuals have the same "identity". The same is true for names of places of birth and death, the property eac:place is not an xml Literal but the URI of a place described in GeoNames database.

3.1 Classes and properties of the ontology

The EAC-CPF schema is made up of two macro sections in which the record control information and the metadata descriptors converge. Therefore in order to reproduce this situation in the EAC-CPF ontology, we created the class *controlArea* and the class *descriptionArea* which contain all the specific information.

The relations between other entities or other resources are managed by a class *relation* which directly points either to other URIs or to resources outside the system.

We introduced the following classes and properties:

Table 2. class and properties of ontology

Classes: entity, person, corporateBody, family, controlArea, descriptionArea, nameArea, language, place, relation.
Properties: authorizedForm, biogHist, control, conventionDeclaration, cpfRelation, cpfRelationType, description, existDates, function, generalContext, languageDeclaration, languageUsed, legalStatus, localTypeDeclaration, maintenanceAgency, maintenanceHistory, maintenanceStatus, mandate, nameEntry, occupation, publicationStatus, recordID, resourceRelation, resourceRelationType, source, structureOrGenealogy

Basically, the graph obtained by the proposed ontology is the following:

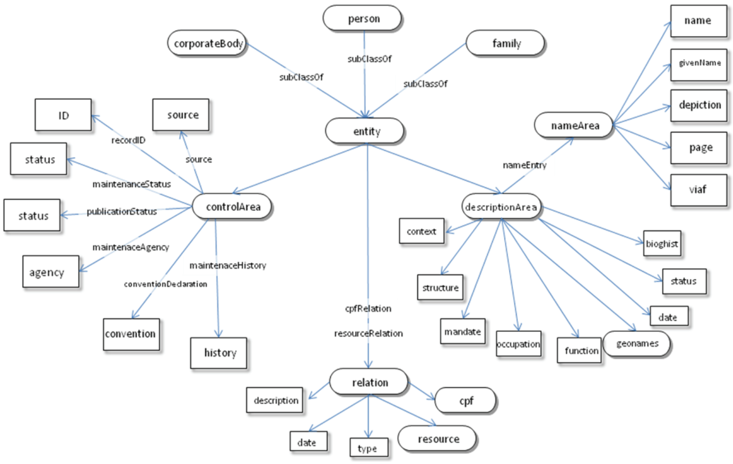


Fig. 3. Graph of the ontology

3.2 External RDF vocabularies references

As far as possible, we have tried to make use of the other popular and widely accepted and supported RDF vocabularies that already exist in the field of cultural heritage and generally in the world of linked data. Besides the Semantic Web languages OWL, RDF and RDFS, we also used the vocabularies: *skos*[10] – Simple Knowledge Organization System, *foaf*[11] – Friend of a Friend, *dc*[12] – Dublin Core, *Bio*[13] - biographical ontology, *Viaf*[14] - The Virtual International Authority File, *Gn*[15] – GeoNames.

[10] http://www.w3.org/2004/02/skos/core#
[11] http://xmlns.com/foaf/0.1/
[12] http://purl.org/dc/elements/1.1/ and http://purl.org/dc/terms
[13] http://purl.org/vocab/bio/0.1/
[14] http://viaf.org/ontology/1.1/#
[15] http://www.geonames.org/ontology#

4 Example

For many years IBC has been experimenting with archival description standards and encoding systems for describing archival institutions, historical archives and creators in the Emilia-Romagna region; actually in *IBC Archivi* the descriptions of 389 archival institutions, 2230 historical archives and 185 creators are published.

This is why we tried to imagine a network (or a graph) which expands slowly but progressively. The graph could show all the resources dynamically connected to it: both the *IBC Archivi* descriptions and the descriptive data opened by other systems and similar environments (libraries, museums, cultural institutions in general, etc.) and recovered thanks to the semantic network.

For example we imagined a map of the Emilia-Romagna region which shows the location of the archival institutions described in the *IBC Archivi*. If we use the GeoNames ontology to reference the institutions locations, automatically the institutions and their archives will be connected to all the other resources referenced in the same place through GeoNames.

In this first test phase, the field of application chosen for this project is the set of descriptive files of the archive producers created in the context of the *IBC Archivi* information system. The authority records of archive producers (about 400, including corporate bodies, persons and families, described in EAC-CPF format) were created on the IBC-xDams platform (a web-based platform for EAD and EAC compliant archive file creation). This is why these descriptions constitute the project's testbed.

A first example was made with the authority record "Andrea Costa"[16], whose papers are kept at the municipal historical archives of Imola and are described using the IBC-xDams platform[17]. The "Andrea Costa" record, in particular, is a suitable case study because it has a fairly analytic description and numerous relations with other archive producers described and with various typologies of resource contained in *IBC Archivi* and in other information systems.

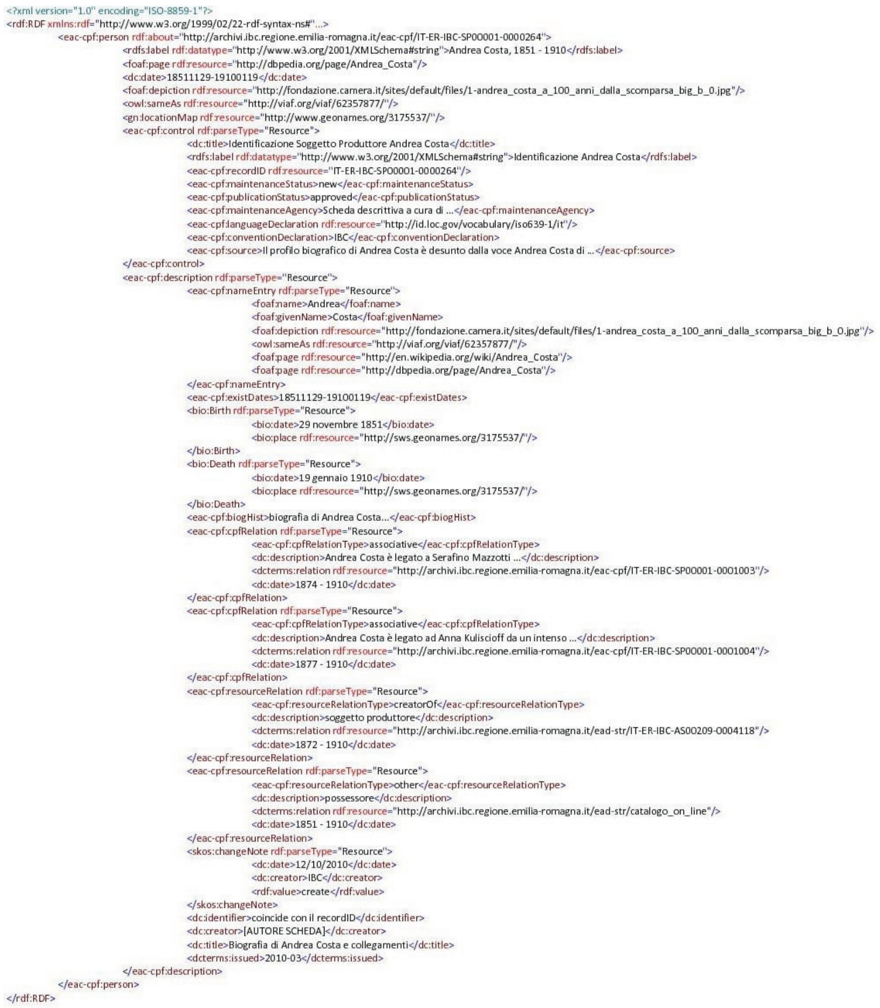
We tried to read the RDF files produced in this way (fig. 4) with an open source faceted browser called *Longwell*[18] created for the Simile project[19]. Faceted navigation adapts well to RDF files precisely because they are not hierarchical files but there only transverse relations between the resources and so it is easy to visualize the data from different points of view or facets; at the same time it is possible to set and remove filters, derived from the properties introduced into the ontology, which allow navigation to be guided and targeted. In this Longwell faceted browser there are some additional small features thanks to the resources which are connected in the RDF. It is possible to visualize on the map the locations that the browser recognizes as such simply because they have already been identified with *GeoNames'* URI and to obtain a graph that best expresses the relations between the resources.

[16] Andrea Costa (Imola 1851-1910) was an Italian socialist activist, he was born in Imola and he co-founded the Partito dei Lavoratori Italiani in 1892

[17] http://www.regesta.com/cosa-e-xdams/

[18] http://simile.mit.edu/wiki/Longwell

[19] http://simile.mit.edu/



```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"...>
        <eac-cpf:person rdf:about="http://archivi.ibc.regione.emilia-romagna.it/eac-cpf/IT-ER-IBC-SP00001-0000264">
                <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Andrea Costa, 1851 - 1910</rdfs:label>
                <foaf:page rdf:resource="http://dbpedia.org/page/Andrea_Costa"/>
                <dc:date>18511129-19100119</dc:date>
                <foaf:depiction rdf:resource="http://fondazione.camera.it/sites/default/files/1-andrea_costa_a_100_anni_dalla_scomparsa_big_b_0.jpg"/>
                <owl:sameAs rdf:resource="http://viaf.org/viaf/62357877/"/>
                <gn:locationMap rdf:resource="http://www.geonames.org/3175537/"/>
                <eac-cpf:control rdf:parseType="Resource">
                        <dc:title>Identificazione Soggetto Produttore Andrea Costa</dc:title>
                        <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Identificazione Andrea Costa</rdfs:label>
                        <eac-cpf:recordID rdf:resource="IT-ER-IBC-SP00001-0000264"/>
                        <eac-cpf:maintenanceStatus>new</eac-cpf:maintenanceStatus>
                        <eac-cpf:publicationStatus>approved</eac-cpf:publicationStatus>
                        <eac-cpf:maintenanceAgency>Scheda descrittiva a cura di ...</eac-cpf:maintenanceAgency>
                        <eac-cpf:languageDeclaration rdf:resource="http://id.loc.gov/vocabulary/iso639-1/it"/>
                        <eac-cpf:conventionDeclaration>IBC</eac-cpf:conventionDeclaration>
                        <eac-cpf:source>Il profilo biografico di Andrea Costa è desunto dalla voce Andrea Costa di ...</eac-cpf:source>
                </eac-cpf:control>
                <eac-cpf:description rdf:parseType="Resource">
                        <eac-cpf:nameEntry rdf:parseType="Resource">
                                <foaf:name>Andrea</foaf:name>
                                <foaf:givenName>Costa</foaf:givenName>
                                <foaf:depiction rdf:resource="http://fondazione.camera.it/sites/default/files/1-andrea_costa_a_100_anni_dalla_scomparsa_big_b_0.jpg"/>
                                <owl:sameAs rdf:resource="http://viaf.org/viaf/62357877/"/>
                                <foaf:page rdf:resource="http://en.wikipedia.org/wiki/Andrea_Costa"/>
                                <foaf:page rdf:resource="http://dbpedia.org/page/Andrea_Costa"/>
                        </eac-cpf:nameEntry>
                        <eac-cpf:existDates>18511129-19100119</eac-cpf:existDates>
                        <bio:Birth rdf:parseType="Resource">
                                <bio:date>29 novembre 1851</bio:date>
                                <bio:place rdf:resource="http://sws.geonames.org/3175537/"/>
                        </bio:Birth>
                        <bio:Death rdf:parseType="Resource">
                                <bio:date>19 gennaio 1910</bio:date>
                                <bio:place rdf:resource="http://sws.geonames.org/3175537/"/>
                        </bio:Death>
                        <eac-cpf:biogHist>biografia di Andrea Costa...</eac-cpf:biogHist>
                        <eac-cpf:cpfRelation rdf:parseType="Resource">
                                <eac-cpf:cpfRelationType>associative</eac-cpf:cpfRelationType>
                                <dc:description>Andrea Costa è legato a Serafino Mazzotti ...</dc:description>
                                <dcterms:relation rdf:resource="http://archivi.ibc.regione.emilia-romagna.it/eac-cpf/IT-ER-IBC-SP00001-0001003"/>
                                <dc:date>1874 - 1910</dc:date>
                        </eac-cpf:cpfRelation>
                        <eac-cpf:cpfRelation rdf:parseType="Resource">
                                <eac-cpf:cpfRelationType>associative</eac-cpf:cpfRelationType>
                                <dc:description>Andrea Costa è legato ad Anna Kuliscioff da un intenso ...</dc:description>
                                <dcterms:relation rdf:resource="http://archivi.ibc.regione.emilia-romagna.it/eac-cpf/IT-ER-IBC-SP00001-0001004"/>
                                <dc:date>1877 - 1910</dc:date>
                        </eac-cpf:cpfRelation>
                        <eac-cpf:resourceRelation rdf:parseType="Resource">
                                <eac-cpf:resourceRelationType>creatorOf</eac-cpf:resourceRelationType>
                                <dc:description>soggetto produttore</dc:description>
                                <dcterms:relation rdf:resource="http://archivi.ibc.regione.emilia-romagna.it/ead-str/IT-ER-IBC-AS00209-0004118"/>
                                <dc:date>1872 - 1910</dc:date>
                        </eac-cpf:resourceRelation>
                        <eac-cpf:resourceRelation rdf:parseType="Resource">
                                <eac-cpf:resourceRelationType>other</eac-cpf:resourceRelationType>
                                <dc:description>possessore</dc:description>
                                <dcterms:relation rdf:resource="http://archivi.ibc.regione.emilia-romagna.it/ead-str/catalogo_on_line"/>
                                <dc:date>1851 - 1910</dc:date>
                        </eac-cpf:resourceRelation>
                        <skos:changeNote rdf:parseType="Resource">
                                <dc:date>12/10/2010</dc:date>
                                <dc:creator>IBC</dc:creator>
                                <rdf:value>create</rdf:value>
                        </skos:changeNote>
                        <dc:identifier>coincide con il recordID</dc:identifier>
                        <dc:creator>[AUTORE SCHEDA]</dc:creator>
                        <dc:title>Biografia di Andrea Costa e collegamenti</dc:title>
                        <dcterms:issued>2010-03</dcterms:issued>
                </eac-cpf:description>
        </eac-cpf:person>
</rdf:RDF>
```

Fig. 4. example of authority record encoded with eac-cpf ontology

5 Conclusion

The experience made with the two ontologies and the testbed on Andrea Costa's records shows that authority records can indeed be the first data to "unlock". In fact authority records by their nature are connection points between different resources. Unlocking authority record of Andrea Costa means connecting not only with his papers, but also with his library, his publications and with other related persons or entities.

We are aware that hard work still needs to be done but according to these first results, the scenario is surprising and, in particular we have to explore all the research directions. In this perspective a future collaborative effort with SNAC project[20] might be useful to share skill, tools and outcomes. At the moment we are working to build a semantic environment[21] for *IBC Archivi* in which users could utilize a SPARQL Endpoint jointly with a reasoning engine and a *linked data api* (ELDA)[22] for navigating resources.

References

1. *Extensible Markup Language (XML) 1.0 (Fifth Edition)* **W3C Recommendation** 26 November 2008, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau eds.
2. Tim Berners-Lee: *L'architettura del nuovo Web*, Feltrinelli, 2001
3. RDF Primer, W3C Recommendation, February 10, 2004, Frank Manola , Eric Miller, eds., http://www.w3.org/TR/rdf-primer/
4. Stefan Decker et al.: The Semantic Web - on the respective Roles of XML and RDF, IEEE Internet Computing, vol.4, 2000

[20] http://socialarchive.iath.virginia.edu/index.html

[21] http://archivi.ibc.regione.emilia-romagna.it/ontology/EAC-CPF/sematic-enviroment.html

[22] http://code.google.com/p/elda/

Conversion of EAD into EDM Linked Data

Steffen Hennicke[1], Marlies Olensky[1], Victor de Boer[2],
Antoine Isaac[2,3], and Jan Wielemaker[2]

[1] Humboldt-Universität zu Berlin, Institut fr Bibliotheks- und
Informationswissenschaft
Dorotheenstrae 26, 10117 Berlin, Germany
{steffen.hennicke,marlies.olensky}@ibi.hu-berlin.de
[2] Vrije Universiteit Amsterdam, Department of Computer Science
De Boelelaan 1081a, 1081 HV Amsterdam, Netherlands
{v.de.boer,aisaac,j.wielemaker}@cs.vu.nl
[3] Europeana, Koninklijke Bibliotheek
Prins Willem-Alexanderhof 5, 2509 LK Den Haag

Abstract. We report on ongoing work in Europeana on the conversion of EAD-XML based archival data to an RDF-based representation using the newly developed "Europeana Data Model" (EDM) ontology. This short paper is based on [4].

Keywords: archive, EAD, semantic web, finding aid, RDF, EDM, Europeana

1 Introduction

The project Europeana[4] was set up as part of the EU policy framework for the information society and media (i2010 strategy) aiming at the establishment of a single access point to the distributed European cultural heritage. Today, Europeana offers access to millions of objects from all kinds of cultural heritage communities. The aim of the current agenda of Europeana is to provide semantically contextualized object representations and new functionality based on an API approach and on integration into the Linked Data context [1]. To enable this vision the "Europeana Data Model" (EDM) has been developed.

A large portion of the cultural heritage metadata that is prospected to be accessible through Europeana is currently described in archival finding aids used by archives across Europe. The "Encoded Archival Description" (EAD) is an XML standard for encoding such finding aids. As such, a method for converting EAD-compliant metadata to EDM will greatly benefit Europeana's goals.

In this paper, we will describe the basic functionality of the EDM and explain some pivotal principles of archival description embodied in EAD-encoded archival finding aids. After this, we elaborate on the EDM-RDF representation of a concrete EAD encoded finding aid. We conclude with the perceived advantages of such a new data representation.

[4] http://www.europeana.eu

2 Europeana Data Model (EDM)

The EDM has been specifically designed to enable Linked Data integration and to solve the problem of cross-domain data interoperability. The EDM builds on the reuse of existing standards from the Semantic Web environment but does not specialize in any community standard [2]. It acts as a top-level ontology consisting of elements from standards like OAI-ORE[5], Dublin Core Terms[6] and SKOS[7] and allows for specializations of these elements. Thus, richer metadata can be expressed through specializations of classes and properties. Some elements were defined in the Europeana namespace, yet contain referrals to other metadata standards. This allows for correct mappings and cross-domain interoperability.

RDF(S)[8] is used as an overall meta-model to represent the data. The ORE approach is used to structure the different information snippets belonging to an object and its representation. It follows the concept of aggregations (`ore:Aggregation`): This concept allows to distinguish between digital representations which are accessible on the Web and thus are modeled as `edm:WebResource`[9] and the provided object, represented as a `edm:ProvidedCHO`.

Furthermore, different, possibly conflicting views from more than one provider on the same object can be handled in EDM by using the proxy mechanism (`ore:Proxy`). The Dublin Core Terms describe the objects. SKOS is used to model controlled vocabularies which annotate the digital objects [5].

3 Archival Description and Finding Aids

Archival finding aids are guides into the archival material that an archive holds in the form of archival collections. Typically, printed versions of finding aids serve archival users in the reading room and archivists in the reference service as the means to identify relevant archival materials.

The archival material in an archival collection is organized into records. A record denotes a group of documents from the archival collection. It does not describe a single information object like a single book in the library domain.

According to the principle *respect des fonds* the description of the internal structure (original hierarchy and ordering) and the external structure (provenance) of an archival collection provides information necessary to understand context and content of the records and to guarantee their authenticity.

A finding aid contains such information in the form of an archival description. This archival description typically consists of several parts arranged in a

[5] "Open Archives Initiative Protocol - Object Exchange and Reuse": `http://www.openarchives.org/ore/` [7.10.2010]

[6] "Dublin Core": `http://dublincore.org/` [7.10.2010]

[7] "Simplified Knowledge Organization System": `http://www.w3.org/2004/02/skos/` [7.10.2010]

[8] "Resource Description Framework (Schema)": `http://www.w3.org/TR/rdf-primer/` [5.09.2011] and `http://www.w3.org/TR/rdf-schema/` [5.09.2011]

[9] The namespace prefix "edm" stands for the Europeana namespace "http://www.europeana.eu/schemas/edm/".

multilevel hierarchy. The top-most part describes the archival collection as a whole. The following descendant parts describe sub-parts of the previous parts with increasing detail.

The leaves of the descriptive tree are about different kinds of unit of records which constitute the smallest parts within the archival description. The smallest parts of the description do not necessarily correspond to the smallest parts of the archival collection. The unit of a record can be, for example, an item which corresponds to one record, or a file with one or more folders of records.

A call number for the unit of records is used to order one or more physical boxes with archival documents (photographs, legal documents, letters, et cetera) from the archive's depot. Typically, searching for archival material in an archival finding aid means identifying call numbers for units of records whose potential relevancy for one's purpose is judged by the contextual descriptions.

The archival descriptions we find in archival finding aids contain huge and rich amounts of contextual and implicit information (especially through inheritance) in order to enable archival users and archivists to efficiently and effectively locate and discover archival material [3].

4 EAD-encoded Finding Aids

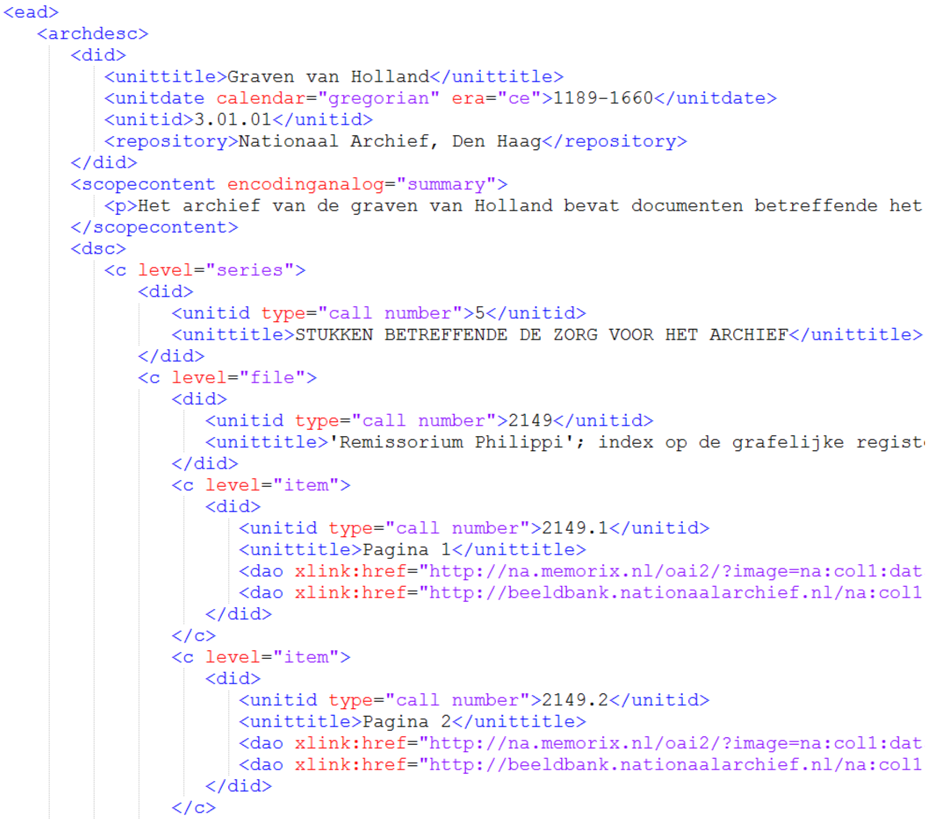
The "Encoded Archival Description"[10] (EAD) standard is the latest and most promising standardization effort for encoding archival finding aids for a digital environment. It provides the infrastructure to accommodate most designs of finding aids. Typically, an institution uses a subset of the full EAD model. Our conversion method is specifically designed for and tested with APEnet-EAD, which is currently developed by the APEnet project[11] within the context of Europeana. We expect, however, that our method is applicable to other EAD dialects with slight modifications to the script. An EAD-document typically contains the description of one archival collection in the form of a finding aid. The `<eadheader>` element[12] contains bibliographic and descriptive information to identify a finding aid document. Its sibling element `<archdesc>` holds information about the archival collection as a whole. In our example, the `<archdesc>` element contains several descriptive metadata fields which hold information about the title of the whole archival fond (`<unittitle>`), the time span the material covers (`<unitdate>`), a call number (`<unitid>`), the name of the repository where the material is kept (`<repository>`), and a summary of the contents (`<scopecontent>`).

Within the `<archdesc>` element, `<c>` elements of different types (classes, series, subseries, files, or items) represent the multilevel hierarchy of the archival description providing the intermediate structure and context for the archival material described in a finding aid. In our example we find a series which contains

[10] "Encoded Archival Description": `http://www.loc.gov/ead/` [7.10.2010]

[11] The "Archives Portal Europe" (`http://www.apenet.eu/`) is a data aggregator for the European archives.

[12] The element has been omitted in figure 1.



```
<ead>
   <archdesc>
      <did>
         <unittitle>Graven van Holland</unittitle>
         <unitdate calendar="gregorian" era="ce">1189-1660</unitdate>
         <unitid>3.01.01</unitid>
         <repository>Nationaal Archief, Den Haag</repository>
      </did>
      <scopecontent encodinganalog="summary">
         <p>Het archief van de graven van Holland bevat documenten betreffende het
      </scopecontent>
      <dsc>
         <c level="series">
            <did>
               <unitid type="call number">5</unitid>
               <unittitle>STUKKEN BETREFFENDE DE ZORG VOOR HET ARCHIEF</unittitle>
            </did>
            <c level="file">
               <did>
                  <unitid type="call number">2149</unitid>
                  <unittitle>'Remissorium Philippi'; index op de grafelijke regist
               </did>
               <c level="item">
                  <did>
                     <unitid type="call number">2149.1</unitid>
                     <unittitle>Pagina 1</unittitle>
                     <dao xlink:href="http://na.memorix.nl/oai2/?image=na:col1:dat
                     <dao xlink:href="http://beeldbank.nationaalarchief.nl/na:col1
                  </did>
               </c>
               <c level="item">
                  <did>
                     <unitid type="call number">2149.2</unitid>
                     <unittitle>Pagina 2</unittitle>
                     <dao xlink:href="http://na.memorix.nl/oai2/?image=na:col1:dat
                     <dao xlink:href="http://beeldbank.nationaalarchief.nl/na:col1
                  </did>
               </c>
```

Fig. 1. An EAD-XML snippet taken from an EAD-encoded finding aid of the Dutch National Archives.

a file which holds two items. All these levels have a call number and a title which are constitutive parts of the contextual description. The two items also link to digital representations `<dao>`, e.g. digital images, of their contents. To suit the original purpose of the finding aids, it is crucial to retain all descriptive and contextual information about records when transforming this structure to RDF.

5 Conversion of EAD-XML to EDM-RDF

Archdesc and each c-level are represented as a "`edm:ProvidedCHO - ore:Aggregation`" cluster (cf. figure 2). Both resources are connected via the property `ore:aggregates`. Their URIs are constructed by concatenating the apenet namespace prefix, the resource type (aggregation-, cho-), the type of the EAD-level (archdesc, series, file, item), and a guaranteed unique identifier, in this case the unitid of the respective c-level. By having a uniform URI creation scheme, objects referring to other objects can be easily represented in RDF by using URIs as objects. This way, if objects are added or metadata is updated, we ensure that existing objects receive their old unique URI while added objects receive a new unique URI. The metadata describing the cultural heritage resource

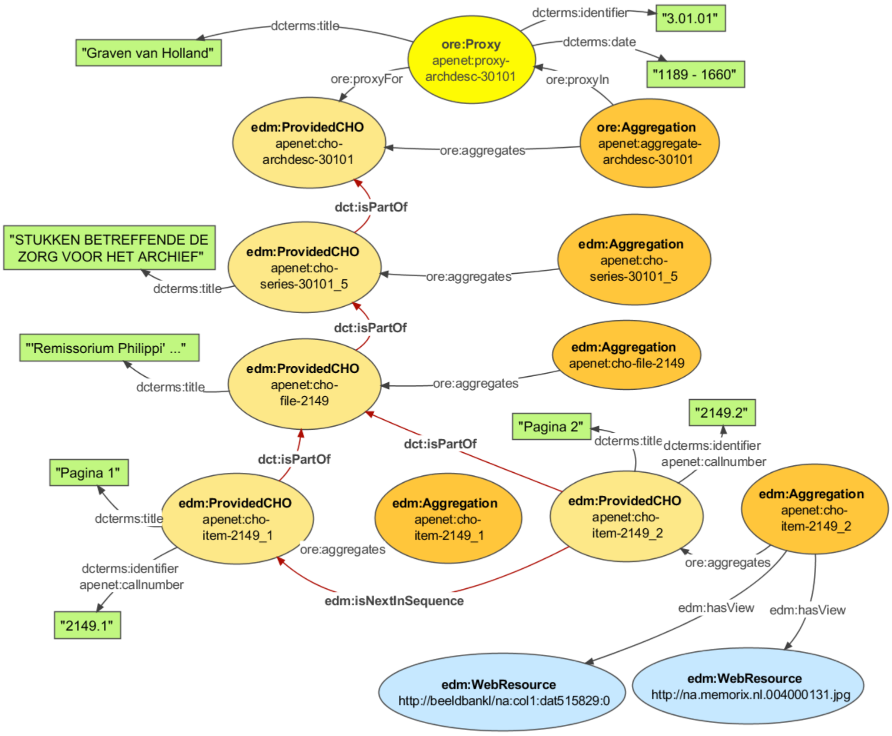


Fig. 2. Parts of the EDM-RDF representation of the EAD-encoded finding aid shown in figure 1.

itself (in our case, those are the different parts of the archival description describing and contextualizing records) can be either attached to the `edm:ProvidedCHO` directly or to a third resource called `ore:Proxy`. The proxy mechanism allows having different views, i.e. descriptions, of one and the same object. In such case a data provider can create proxies and attach the different views to it and thereby keep them distinct. Europeana itself, when doing semantic enrichment, will create proxies in order to retain the original structure and provenance information for the metadata. In our example, the proxy is not necessary as there are no conflicting descriptions. We only created a proxy for the first level (archdesc) to demonstrate the functionality.

The fourth EDM-class shown in figure 2 is `edm:WebResource` which represents associated web pages, thumbnail images or any other web resources and is attached to the aggregation. The URI for such a WebResource is typically the URL provided for the digital representation.

The descriptive metadata fields can be represented in EDM in two ways: In the case where an original field exactly matches a DC Terms property (for example `<unittitle>` and `dcterms:title`), the DC Terms property is used directly. In the case where the match is not exact, an APEnet-EAD property is created in RDF which is specified as being a sub-property of the appropriate DC Terms property: For instance, `apenet:callnumber` is a `rdfs:subPropertyOf` of `dcterms:identifier`, as shown at the two leaves in figure 2. Interoperability at the EDM level is ensured through RDFS semantics by using the sub-property method. The language of the content of descriptive metadata fields can be specified by adding a language tagged-RDF literal as value.

The `edm:ProvidedCHO` resource carries not only descriptive metadata but also properties which are used to relate other objects. During conversion the EAD hierarchy has been translated into a hierarchy relation between the `edm:ProvidedCHO` resources which are connected by `dct:isPartOf` properties. This hierarchy mirrors the XML-structure of the multilevel archival description of the EAD file. At the same time these relations represent, on a more abstract level, the different levels of generality of digital object ”packages” submitted via the EAD file to Europeana. The `dct:isPartOf` properties conceptually reflect the documented structure of the archival material, i.e. the archival collection (archdesc) incorporates a series which has a file which holds two items as parts.

The two c-levels of type item at the bottom in the XML structure are in an intentional and meaningful sequential order. This sequence is expressed by asserting an `edm:isNextInSequence` statement between the resource with title ”Pagina 1” and the resource with title ”Pagina 2”.

6 Discussion: EAD in a Linked Open Data Environment

We demonstrated how an EAD-XML encoded archival finding aid can be modeled in a RDF-based representation. The representation in an RDF graph makes implicit information explicit, for example, the hierarchical and sequential relations between the different parts of the archival description. We aimed at a conversion which produces a RDF representation which stays as close as possible to the original structure of the APEnet-EAD model. This way, we have a conversion template which is feasible for many different variations of EAD encoded finding aids. The method also entails, however, that not all implicit information pieces in the descriptive metadata have been made explicit or have been connected: for example, the unit titles on the different levels remain only indirectly connected to each other. A user being on the level of one of the leafs of the descriptive tree, most certainly needs to know that he is on page 1 (Pagina 1) of the book ”Remissorium Philippi (...)” of the ”counts of Holland”. In order to use this data one needs a special reasoner or a Linked Data browser which brings those information pieces from different levels together.

Another option is to merge intermediate levels of the archival description without digital representations and the descriptive metadata we find there into the leafs of the descriptive tree already during the conversion. In the context of Europeana, each `edm:ProvidedCHO` is an object which can be found via searches. If those objects have no digital representations or only partial descriptions then their information value in the context of Europeana can be questioned. Data providers creating mappings to EDM, on the one hand, have to consider how they want to represent their data in the context of the Europeana information space, but, on the other hand, enjoy great flexibility regarding data modeling with EDM.

At the same time we showed the capability of the EDM to accommodate such a particular archival domain model. The EDM is able to accommodate EAD and other different standards as we demonstrated elsewhere [4]. One of the main reasons to use EDM as the ontology (instead of some specific EAD-RDF model) for an EAD conversion is, that the archival data can now be connected to museum, library and other archival data within the Europeana information space. Contextualization through external resources is now possible. For example, person names can be linked to concepts in a controlled vocabulary like VIAF[13]. Such contextualization allows, for example, to disambiguate meaning or to relate the original object to other cultural heritage objects annotated with the same person name. Europeana is planning to do enrichments for a number of fields like, for example, person or place names.

References

1. Concordia, C., Gradmann, S., Siebinga, S.: Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. IFLA Journal 36(1), 61–69 (2010)
2. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The Europeana Data Model (EDM) (2010), http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf
3. Haworth, K.M.: Archival Description: Content and Context in Search of Structure. In: Pitti, D.V., Duff, W.M. (eds.) Encoded Archival Description on the Internet, pp. 7–26. Haworth Information Press, Binghamton NY (2001)
4. Hennicke, S., Olensky, M., Boer, V.d., Isaac, A., Wielemaker, J.: A data model for cross-domain representation: The "Europeana Data Model" in the case of archival and museum data. In: Griesbaum, J., Internationales Symposium für Informationswissenschaft 12, .H., Hochschulverband für Informationswissenschaft (eds.) Information und Wissen: global, sozial und frei?, vol. 58, pp. 136–147. vwh Hülsbusch, Boizenburg (2011)
5. Isaac, A.: Europeana Data Model Primer (2010), http://version1.europeana.eu/web/europeana-project/technicaldocuments/

[13] "Virtual International Authority File": http://viaf.org/ [16.07.2011]

Publishing Europe's Television Heritage on the Web

Johan Oomen[1], Vassilis Tzouvaras[2],

[1] Nederlands Instituut voor Beeld en Geluid, Sumatralaan 45, Hilversum, the Netherlands
joomen@beeldengeluid.nl
[2] National Technical University of Athens, Iroon Polytexneiou 9, 15780 Zografou, Greece
tzouvaras@image.ntua.gr

Abstract. The EUscreen project represents the European television archives and acts as a domain aggregator for Europeana, Europe's digital library. The man motivation for it is to provide unified access to a representative collection of television programs, secondary sources and articles, and in this way to allow students, scholars and the general public to study the history of television in its wider context. The main goals of EUscreen are to (i) develop a state-of-the-art workflow for content ingestion, (ii) define content selection and IPR management methodology, and (iii) provide a front-end that accommodates requirements of several user groups.

Keywords: Information integration, TV on the Web, Metadata Interoperability, Linked Open Data, Visualization, Europeana

1 Introduction

Providing access to large integrated digital collections of cultural heritage objects is a challenging task. Multiple initiatives exist in different domains. For example, Europeana manages a state-of-the art technical infrastructure to manage the ingestion and management of data from a wide variety of content providers. It aims to give access to all of Europe's digitised cultural heritage by 2025. Europeana focuses on two main tasks (i) to act as a central index, aggregating and harmonising metadata following a common data model [1], and (ii) to provide persistent links to content hosted by trusted sources. The portal currently provides access to 15 million objects, primarily books and photographs; audiovisual collections are underrepresented. However, recent analysis of query logs from the Europeana portal indicated users have a special interest for this type of content. Television content is regarded a vital component of Europe's heritage, collective memory and identity – all our yesterdays – but it remains difficult to access. Even more than with the museum and library collections, the dealing with copyrights, encoding standards, costs for digitization and storage makes the process of its aggregated and contextualized publishing on the Web extra challenging.

In this paper, we will focus in outlining the ingestion workflow; the projects' main technical achievement. In Section 2, we outline the motivation of our work. In Section 3, we elaborate on different components that make up the ingestion workflow.

2 Motivation

The main motivation for our work is to overcome the current barriers and provide a unified access to a representative collection of television programs, secondary sources and articles, and in this way to allow students, scholars and the general public The multidisciplinary nature of the EUscreen project is mirrored in the composition of the socio-technical nature of the consortium; comprising of 20 collection owners, technical enablers, legal experts, educational technologists and media historians of 20 countries. EUscreen represents all major European television archives and acts as one of the key domain aggregators providing content to Europeana.

Several public reports on our work can be downloaded from the project blog. This paper reports on the results of the work performed over the past one and a half years, leading up to the launch of the first version of the portal. Notably, we analyse the design decisions from a Web Science perspective; zooming in on the interplay between user requirements, technical possibilities and societal issues, including intellectual property rights. We will show how EUscreen contributes to a so-called 'Cultural Commonwealth' [2] that emerges by bringing content from memory institutions and the knowledge of its heterogeneous constituency together.

Conceptually, EUscreen is built on the notion that knowledge is created through conversation [3]. Hence, ample attention is given to investigating how to stimulate and capture knowledge of its users. Combining organizational, expert and amateur contributions is a very timely topic in the heritage domain, requiring investigation of the technical, organizational and legal specificities.

The goals of the project are to (i) develop a state-of-the-art workflow for content ingestion, (ii) define content selection and IPR management methodology (35.000 items will be made available), and (iii) design and implement a front-end that accommodates requirements from several user groups. To reach these goals, close cooperation between the different stakeholders in the consortium is essential. For example, the selection policy needs to take in to account the available content, wishes from media historians and the copyright situation. The workflow will need to study the existing metadata structures, should support aggregation by Europeana and provide support for multilingual access.

2.1 Define content selection methodology

In collaboration with leading television historians EUscreen has defined a content selection policy [4], divided into three strands:

1. Historical Topics: 14 important topics in history of Europe in the 20th Century (70% of content);

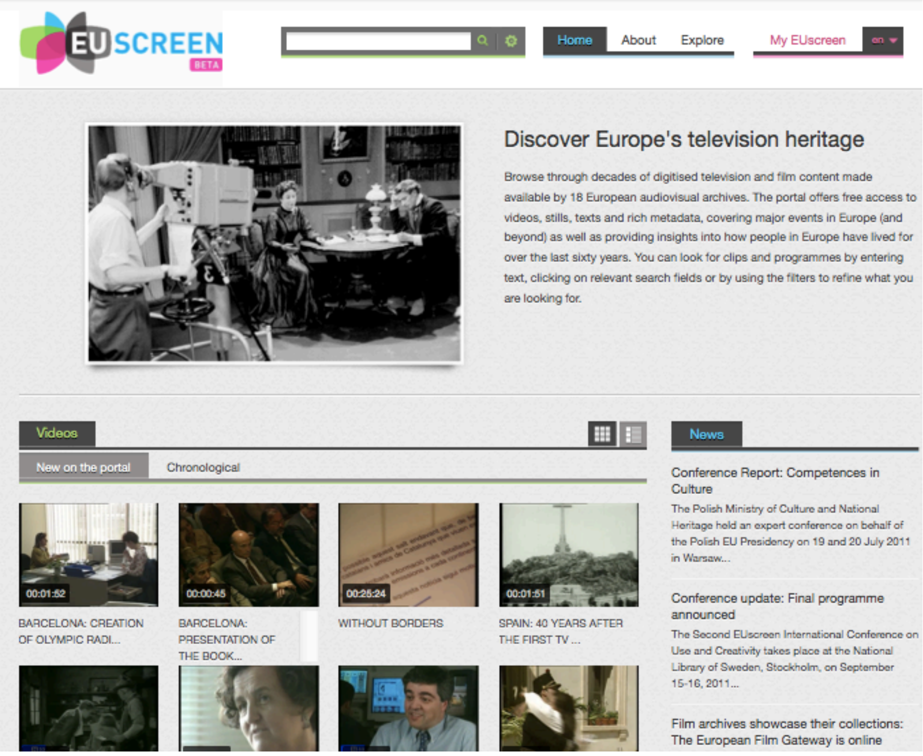
2. Comparative Virtual Exhibitions: two specially devised topics that explore more specialised aspects of European history in a more comparative manner (10% of content – include documents, stills, articles);

3. Content Provider Virtual Exhibitions: Each content provider selects content supported with other digital materials and textual information on subjects or topics of their own choosing (20 % of content).

EUscreen has written a set of guideless regarding management of intellectual property rights. The copyright situation of each and every item is investigated prior to uploading.

2.2 The Front-end

Representatives of the four primary user groups, e.g. secondary education, academic research, the general public and the cultural heritage domain were consulted in order to define user requirements and design front-end functionality. The main challenge for the portal's front-end is to include advanced features for specific use cases without overwhelming the users with a complex interfaces. The Helsinki University of Arts and Design adapted a component-based conceptual model that accommodates this requirement (Figure 1.)



Discover Europe's television heritage

Browse through decades of digitised television and film content made available by 18 European audiovisual archives. The portal offers free access to videos, stills, texts and rich metadata, covering major events in Europe (and beyond) as well as providing insights into how people in Europe have lived for over the last sixty years. You can look for clips and programmes by entering text, clicking on relevant search fields or by using the filters to refine what you are looking for.

BARCELONA: CREATION OF OLYMPIC RADI...

BARCELONA: PRESENTATION OF THE BOOK...

WITHOUT BORDERS

SPAIN: 40 YEARS AFTER THE FIRST TV ...

Conference Report: Competences in Culture

The Polish Ministry of Culture and National Heritage held an expert conference on behalf of the Polish EU Presidency on 19 and 20 July 2011 in Warsaw...

Conference update: Final programme announced

The Second EUscreen International Conference on Use and Creativity takes place at the National Library of Sweden, Stockholm, on September 15-16, 2011...

Film archives showcase their collections: The European Film Gateway is online

Fig. 1. EUscreen Homepage.

Implementation of the front-end services is not done in the traditional way using serverside programming language like php, java or asp. EUscreen implemented a 'server-less' front-end APIs where a javascript/flash proxy system handles the communication with the back-end services. The result will be a front-end system that can be 'installed' on any plain html web server without any need for server-side technologies. This means it can be hosted and moved to any location or multiple locations. It also means partners can use these APIs to integrate parts of the functionality in their own intranet and internet systems using simple 'embed' ideas. This method is gaining more ground, for example companies like Google who provides these types of APIs for services like Google Maps.

3 Metadata Ingestion and Video Playout

The technical standards enabling interoperability form an important dimension of the technical achievements. In order to achieve semantic interoperability, a common automatic interpretation of the meaning of the exchanged information is needed, i.e. the ability to automatically process the information in a machine-understandable manner. The first step of achieving a certain level of common understanding is a representation language that exchanges the formal semantics of the information. Then, systems that understand these semantics can process the information and provide web services like searching, retrieval.

Many different metadata schemas or in a broader sense, sets of elements of information about resources, are being used in this domain, across a variety of technical environments and scientific disciplines. EUscreen has developed an ingestion mechanism providing a user friendly environment that allows for the extraction and presentation of all relevant and statistical information concerning input metadata together with an intuitive mapping service that uses the EUscreen Metadata schema, and provides all the functionality and documentation required for the providers to define their crosswalks. The workflow (Figure 2) consists of four phases, each responsible for specific services to ensure the quality of the ingestion process:

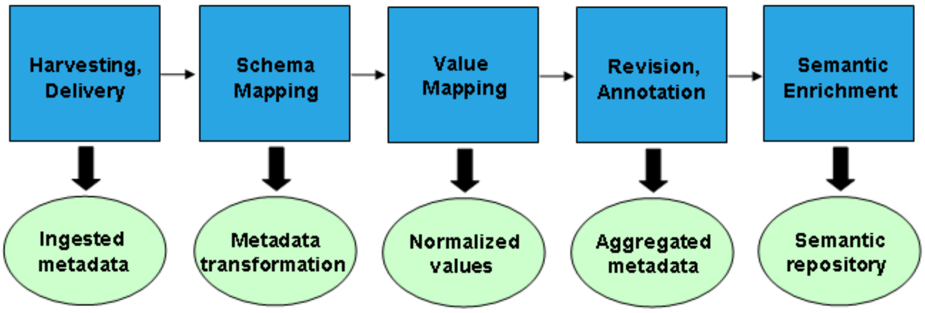


Figure 2. Metadata Ingestion Workflow

The Workflow consists of five steps. The first is *harvesting/delivery*, which refers to collection of metadata from content providers through common data delivery protocols, such as OAI-PMH, HTTP and FTP. The service is implemented as a web service, where authentication is required to perform a series of tasks that correspond to work flow steps. The harvesting service is an application written in the Java and hosted on a web server by the Tomcat servlet engine. Data is imported into a PostgreSQL database in xml format. Once uploaded, the xml structure is parsed and represented in a relational database table.

Second is the *Schema Mapping* that aligns harvested metadata to the common reference model. A graphical user interface assists content providers in mapping their metadata structures and instances to the EUscreen metadata model, using an underlying machine-understandable mapping language. It supports sharing and reuse of metadata crosswalks and establishment of template transformations.

The next step is *Value Mapping*, focusing on the alignment and transformation of a content provider's list of terms to the authority file or external source introduced by the reference model. It provides normalisation of dates, geographical locations or coordinates, country and language information or name writing conventions.

Revision/Annotation, being the fourth step, enables the addition of annotations, editing of a single or group of items in order to assign metadata not available in the original context and, further transformations and quality control checks (e.g. for URLs) according to the aggregation guidelines and scope.

Finally, the *Semantic Enrichment* step focuses on the transformation of data to a semantic data model, the extraction and identification of resources and the subsequent deployment of an RDF semantic repository.

3.1 EBUcore, Solr and Multilinguality

In order to achieve semantic interoperability with external web applications, EUscreen metadata are exported in EBUcore [5], which is an established standard in the area of audiovisual metadata. An extensive evaluation of alternative standards in this area (MPEG7, DCMI, TV Anytime) has been conducted [6] before choosing the EBUcore. EBUCore has been purposefully designed as a minimum list of attributes to describe audio and video resources for a wide range of broadcasting applications including for archives, exchange and publication.

It is also a Metadata schema with well-defined syntax and semantics for easier implementation. It is based on the Dublin Core to maximise interoperability with the community of Dublin Core users. EBUCore expands the list of elements originally defined in EBU Tech 3293-2001 for radio archives, also based on Dublin Core. The metadata is stored in RDF format to improve the search functionality and enable the alignment with external resources.

In EUscreen portal, retrieval is performed using the Solr framework. Solr is an open source enterprise search platform from the Apache Lucene project. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document handling. Providing distributed search and index replication, Solr is highly scalable. Solr uses the Lucene Java search library at its core for full-text indexing and search, and has REST-like HTTP/XML and JSON APIs that make it easy to use from virtually any programming language. Solr's powerful external configuration allows it to be tailored to EUscreen retrieval application without Java coding, and it has an extensive plugin architecture for more advanced customization.

Finally, EUscreen has created a SKOS multilingual thesaurus (15 languages) based on the subject terms of IPTC standard and the geographical places of GeoNames. The baseline of the thesaurus is the *Descriptive NewsCodes vocabulary* from The International Press Telecommunications Council [7]. Translations are made with a software solution for the creation and administration of multilingual thesaurus cslled Thesaurix, as licensed by Joanneum Research. The thesaurus supports multilingual retrieval services and links to open data resources that could be used for enrichment and to contextualise the collection.

3.2 Video Playout

EUscreen requires content providers to provide MPEG 4 part 10 (normally known as H.264). EUscreen advises to encode in a bit rate between 500 and 1000 kb/sec, as this resembles SD quality video. Since the client playback method will be a Flash player with h.264 streaming, EUscreen demands that providers have streaming servers that are capable stream videos to a Flash client. In practice this means using one of the available Flash streaming servers.

This will leave room for the content providers themselves to add HTML5 or Silverlight server programs to create an 100% coverage of the possible technologies.

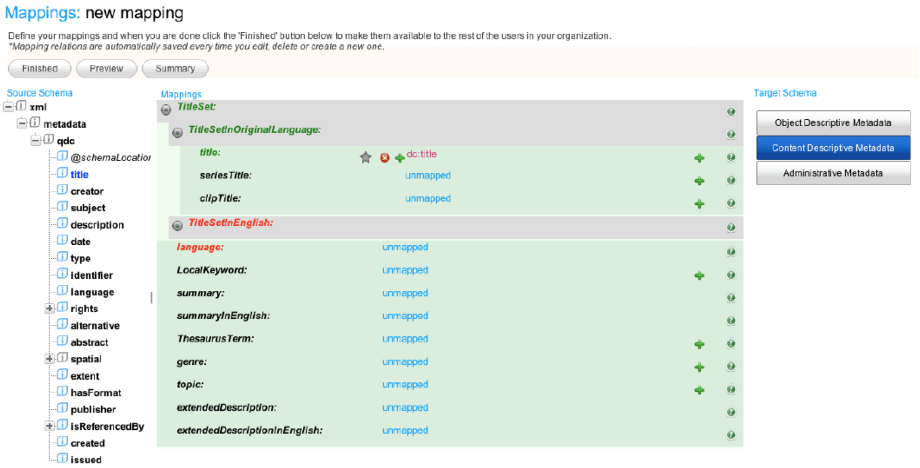
EUscreen supports four scenarios:

1. Content provider transcodes and files are hosted by service provider Noterik
2. Content provider transcodes and the content provider hosts
3. Noterik transcodes and Noterik hosts
4. Noterik transcodes, and the content provider hosts

3.3 The Mapping Tool

Metadata mapping is a crucial step of the ingestion procedure. It formalizes the notion of `crosswalk' by hiding technical details and permitting semantic equivalences to emerge as the centrepiece. It involves a graphical, web-based environment where interoperability is achieved by letting users create mappings between input and target elements. User metadata imports are not required to include the adopted XML schema. Moreover, the set of elements that have to be mapped are only those that are populated. As a consequence, the actual work for the user is easier, while avoiding expected inconsistencies between schema declaration and actual usage.

The structure that corresponds to a user's specific import is visualized in the mapping interface as an interactive tree that appears on the left hand side of the editor (Figure 3). The tree represents the snapshot of the XML schema that the user is using as input for the mapping process. The user is able to navigate and access element statistics for the specific import.



Mappings: new mapping

Define your mappings and when you are done click the 'Finished' button below to make them available to the rest of the users in your organization.
*Mapping relations are automatically saved every time you edit, delete or create a new one.

Mappings	
TitleSet:	
TitleSetInOriginalLanguage:	
title:	dc:title
seriesTitle:	unmapped
clipTitle:	unmapped
TitleSetInEnglish:	
language:	unmapped
LocalKeyword:	unmapped
summary:	unmapped
summaryInEnglish:	unmapped
ThesaurusTerm:	unmapped
genre:	unmapped
topic:	unmapped
extendedDescription:	unmapped
extendedDescriptionInEnglish:	unmapped

Figure 3. Mapping Interface

The interface provides the user with groups of high-level elements that constitute separate semantic entities of the target schema. These are presented on the right hand side as buttons, which are then used to access the set of corresponding sub-elements. This set is visualized on the middle part of the screen as a tree structure of embedded boxes, representing the internal structure of the complex element. The user is able to interact with this structure by clicking to collapse and expand every embedded box that represents an element along with all relevant information (attributes, annotations) defined in the XML schema document. To perform an actual mapping between the input and the target schema, a user has to simply drag a source element and drop it on the respective target in the middle.

The user interface of the mapping editor is schema-aware regarding the target data model and enables or restricts certain operations accordingly, based on constraints for elements in the target XSD. For example, when an element can be repeated then an appropriate button appears to indicate and implement its duplication. User's mapping actions are expressed through XSLT stylesheets, i.e. a well-formed XML document conforming to the namespaces in XML recommendation. XSLT stylesheets are stored and can be applied to any user data, can be exported and published as a well-defined, machine understandable crosswalks and shared with other users to act as template for their mapping needs. Features of the language that are accessible to the user through actions on the interface include:

- string manipulation functions for input elements;
- 1-n mappings;
- m-1 mappings with the option between concatenation and element repetition;
- structural element mappings;
- constant or controlled value assignment;
- conditional mappings (with a complex condition editor);
- value mappings editor (for input and target element value lists).

4 Future Work

The first version of the portal has been launched in August 2011. It is followed by a period of extensive evaluations with end-users. Also, the selection policy will be reviewed. Outcomes of this process will form the basis of the development of the second release, scheduled for early 2012. The major enhancements will be related to the front-end. For instance, EUscreen will support the on-line creation of on so-called virtual exhibitions, consisting of media objects of various archives.

Acknowledgments. EUscreen is co-funded by the European Commission within the eContentplus Programme.

References

1. Isaac, Antoine.: Europeana Data Model Primer. Europeana v1.0 Technical report, http://group.europeana.eu, 2010
2. Scott, B.: Gordon Park's conversation theory: a domain independent constructivist model of human knowing. In: Foundations of Science 6(4):343-360, 2001
3. Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyber infrastructure for the Humanities and Social Sciences.American Council of Learned Societies, 2006
4. Kaye, Linda. D3.1 Content Selection and Metadata Handbook, Euscreen BPN project, euscreen.eu, 2011
5. Evain, Jean Pierre.: European Broadcasting Union Core Metadata, http://tech.ebu.ch/publications, 2009
6. Schreiber, Guus.: D2.2.2 Metadata Models, Interoperability Gaps, and Extensions to Preservation Metadata Standards, PrestoPrime FP7 project, http://www.prestoprime.org/project/public.en.html, 2010
7. Descriptive NewsCodes of International Press and Telecommunication Council. http://www.iptc.org/cms/site/index.html?channel=CH0103, 2010

Introducing the Semlib project: semantic web tools for digital libraries

Christian Morbidoni [a,1], Marco Grassi [b,1], Michele Nucci [c,1], Simone Fonda [d,2], and Giovanni Ledda [e,1]

[1] Semedia Group, Università Politecnica delle Marche, Italy
[a] christian.morbidoni@gmail.com, [b] m.grassi@univpm.it,
[c] mik.nucci@gmail.com, [e] g.ledda@univpm.it
http://www.semedia.dibet.univpm.it/
[2] NET7, Italy
[d] fonda@netseven.it
http://www.netseven.it

Abstract. It is a common opinion that today's digital libraries (DL) can no longer be simple "expositions' of digital objects. Users should no more be passive readers, they need to interact with the library, add their annotations and tags, personalize their experience and collaborate with each other. Web 2.0 technologies, such as social bookmarking and online discussions, are already being applied in DLs to allow users to annotate digital objects. However, the lack of semantic structure of such annotations and a clear social model to share and aggregate community contributions makes it difficult to take full advantage of such collaboratively created knowledge.

The SemLib project aims at developing a modular and configurable annotation system that can be easily plugged into existing digital libraries in order to allow end-users as well as digital libraries content curators to produce meaningful and customizable aggregations of semantically structured annotations produced by communities. In this paper we introduce the SemLib project, discussing the principles and ideas behind the proposed annotation system, and present a prototypal implementation.

Keywords: Digital libraries, Semantic Web, Ontology, Data Model

1 Introduction

Nowadays, Digital Libraries (DL) are applied in many different contexts ranging from academic institutions to public libraries, archives, museums and industries. Traditionally DLs, as well as Web itself at its beginning, have been based on the expert paradigm according to which experts create content, DL experts provide access to it, and individual users consume it [1]. The advent of Web 2.0 has lead to a Copernican revolution in the Web universe that has pushed users more and more toward its center and transformed them from passive content consumers into primary actors in data and metadata creation. As a result, tagging, linking and commenting resources have become common activities for Web users and a valuable source of metadata that can be exploited to drive resource ranking, classification and retrieval. Annotation creation and sharing in a research context is an established practice since the pre-digital era, therefore its not surprising that in the last years the application of Web 2.0 models has been widely investigated in the context of digital humanities.

One of the ideas at the base of the research and development activities in this field is that user-created annotations, if properly structured and machine-processable, can enrich Web content and enhance search and browsing capabilities. Also allowing users write-access to the collection in DLs can provide users a more engaging experience and "capture diffuse and ephemeral information' [2]. Supporting social annotations has proved to be an enabling feature for scholars to actually benefit from the digital world in their everyday work. Experiments conducted within the Discovery[3] european project have clearly shown that building structured information by annotating Web documents can be a valuable mean of representing aspects of the study process e.g. in e-learning or classroom activities. In [3], authors make a distinction between "social engagement', where users annotate contents for their own purposes (e.g., to better organize study resources), and crowdsourcing, where social engagement is used within groups of users (communities) to "achieve a shared goal by working collaboratively together as a group'. If social engagement has been addressed to a certain extent by modern DLs, they rarely provide support to exploit such collected knowledge to improve libraries metadata, enrich contents, searching and linking different contents together. However, the topic is of high interest and not entirely new to the DLs community, as witnessed by interesting ongoing projects like Digitalkoot[4], which is engaging people through online games, which create different kind of structured contents.

Basing on previous research and developments in Semantic Web oriented collaborative annotations (e.g.: SWickyNotes[5]), the SemLib project[6], shortly presented in section 2, aims at developing a flexible, collaborative annotation system to address single scholars and unregulated user communities as well as curated "authoritative' annotations to incrementally enrich digital contents.

In this paper we discuss the data and social model designed during the project's first phase, presenting a preliminary prototype composed by experimental GUIs to create and exploit annotations and a triple-store based annotation server providing RESTful APIs to create, share and consume them. This paper is organized as follows: chapter 2 shortly presents the SemLib project; chapter 3 provides a brief overview of existing cutting-edge tools for resource annotation; chapters 4 and 5 discuss the annotation system architecture and chapter 6 demonstrates the experimental prototype.

[3] ECP 2005 CULT 038206 project, EC eContentplus programme

[4] http://www.digitalkoot.fi/en/splash

[5] http://www.swickynotes.org

[6] http://www.semlibproject.eu/

2 The SemLib project: use cases and challenges

The SemLib project, funded by the European commission, aims to improve the current state of the art in DLs, through the application of Semantic Web (SW) technologies for data representation and management. One of the main expected outputs of the SemLib project is the design and implementation of an annotation system able to enrich and interconnect digital objects published on the Web, specifically targeting DLs and multimedia archives owned by participating SMEs. As such objects are different, both from technology and from type of provided content points of view, the annotation system has to be designed to be technologically decoupled from the DL (adopting a RESTful architecture), based on established standards in data and metadata representation (such as RDF and Semantic Web ontologies), domain agnostic and adaptable or configurable for a variety of different use cases.

Resources annotation should be supported at different granularity levels in order to enhance resource fruition and interaction. With respect to this requirement, Web standards such as XPointer[7] and Media Fragment URI[8] are being used respectively to unambiguously identify text excerpts in Web pages and subparts of images and audio-video resources. In addition, as digital content can be remixed and replicated inside a DL (e.g. in summary pages or in composite, derivative digital objects), annotations should address not only entire web pages (has it happens for the majority of existing tools), but also small, atomic unit of content, like pictures, single text paragraphs, etc. Also, as SEMLIB aims at addressing different kind of users, they should be allowed to create different types of annotation, structured according to different levels of complexity and provided with diverse expressive flavor and semantics, from natural language comments to semantic tags coming from a restricted vocabulary to full subject-object-value statements based on domain ontologies. Moreover, SemLib should provide tools and models capable of leveraging the process of collaborative and community driven annotation of DLs items. This is an important requirement both for engaging small unregulated end-user communities and for providing effective tools for scholarly communities and DL maintainers to incrementally and collaboratively enrich the quality of metadata (e.g. basing on a crowdsourcing).

The several high level challenges, which have to be tackled in order to accomplish SemLib's goals, can be summarized as follows:

- supporting DLs in aggregating users in communities by providing properly configured tools and uniform domain vocabularies to create interoperable metadata;
- enabling a social model where end-users, as well as content owners, create, share and aggregate annotations into personal, curated "views' of the collective knowledge base;

[7] "XML Pointer Language (XPointer)" http://www.w3.org/TR/xptr/

[8] "Media Fragments URI 1.0" http://www.w3.org/TR/media-frags/

– providing DLs with visual tools and APIs to exploit the collective knowledge base, slice it accordingly to custom policies and make it available to end-users for searching, browsing and studying online content;
– developing annotation GUIs capable of efficently handling the trade-off between the ease of use and the creation/management of meaningful structured data.

3 Related Work

In recent years, several annotation systems have been developed. These allow Web resource annotation providing different approaches and functionalities to be applied in different application scenarios. Some applications have been developed as extensions of popular social bookmarking tools, as Delicious[9] or StumbleUpon[10], that count millions of registered users. Other tools have been more specifically conceived for creating and sharing annotations of digital resources for supporting e-learning, collaborative tasks, such as document reviews or editing, and in general working group cooperation. A complete review of the state of the art tools for Web resources annotation goes beyond the purpose of this work and can be found in [4]. Some of the most interesting applications are now presented and discussed, with regard to SemLib project.

EuropeanaConnect Media Annotation Prototype (ECMAP) [5] is an online media annotation suite based on Annotea [6] that allows users to extend existing bibliographic information about digital items like images, audio and videos. ECMAP allows free-text annotations and semantic tagging, enabling Linked Data resource linkage in the user annotation process, in addition to the possibility to draw user-defined shapes on images, maps and videos. Special support is also provided for high-resolution map images, enabling tile-based rendering for faster delivery, geo-referencing and semantic tag suggestions based on geographic location. ECMAP's annotation system presents several similarities with SemLib, in particular in the overall idea of supporting various types of resources. For this reason, it represents an important reference to identify the basic features that SemLib annotation system should have. LORE (Literature Object Reuse and Exchange) [7] is a tool developed inside the Aus-e-Lit Project "to enable scholars and teachers of literature to author, edit and publish compliant compound information objects that encapsulate related digital resources and bibliographic records'. The OAI-ORE Resource Map[11] is used as the main data model and a specific ontology has been defined to describe the relationships among objects, called LORE Relationship Ontology. The annotation tool provides a graphical user interface for creating, labeling and visualizing typed relationships among individual objects, using terms from a bibliographic ontology. While the user interface is powerful, it probably lacks in simplicity and would not be so straight-

[9] http://http://www.delicious.com/

[10] www.stumbleupon.com/

[11] Open Archives Initiative Object Reuse and Exchange http://www.openarchives.org/ore/0.9/primer#ResourceMap

forward to understand for non-expert users. However, LORE is an interesting source of inspiration, since it presents several conceptual similarities with the SemLib annotation system. One Click Annotator [8] is a WYSIWYG Web editor for enriching content with RDFa annotations, enabling non-experts to create semantic metadata. It allows the annotation of words and sentences, referencing ontology concepts and creating relationships among annotated sentences. The Open Knowledge Foundations Annotator[12] project is developing a Web-based, open-source annotation tool that, from a user interaction perspective, has similarities to SemLib annotation tools. It uses XPath to anchor textual annotations and tags to specific parts of a page, providing also a server-side module for storing annotations represented as JSON data.

The idea of semantic tagging is implemented in Faviki[13], a social bookmarking tool that allows the use of Wikipedia concepts as tags for Web pages. Tags are suggested using auto completion, allowing disambiguation, where the suggested items are ordered by their use frequency. It also proposes tags automatically extracted from the page using Zemanta[14]. Several Web annotation tools exist, which do not make use of structured semantics and handle simple textual annotations. Among those, Diigo[15] (Digest of Internet Information, Groups and Other stuff) is a social bookmarking application, which allows signed-up users to bookmark and tag Web pages. In addition, Diigo allows users to highlight any part of a Web page, attaching sticky notes to it. Diigo provides a simple but interesting annotations sharing model: annotations can be kept private, shared with a group within Diigo or forwarded to someone else with a custom link.

4 Representing semantically structured annotations

Annotations represent a peculiar type of resources that is specifically conceived to add information to other resources. Annotations acquire therefore full significance in relation with the target resource and other contextual information, such as its author, its creation date and the vocabulary terms used. Properly structuring an annotation is therefore necessary at twofold level. On the one hand, an annotation represents an “information container’, whose structured metadata make contextual information explicit. On the other hand, an annotation includes an informative content that expresses a “knowledge bit’ about annotated resources. Such knowledge is strongly domain dependent and, when uniformly structured by means of shared ontologies, can be in turn aggregated and used to increase content accessibility and interoperability.

Several ontologies have been developed in the last few years to provide a generic annotation structure and to improve interoperability among different annotation tools [9] [10]. The Open Annotation Collaboration[16] (OAC) project

[12] http://okfn.org/projects/annotator/

[13] http://www.faviki.com/

[14] http://www.zemanta.com/

[15] http://www.diigo.com/

[16] http://www.openannotation.org/

recently published the first specifications of the OAC data model [11], which at the moment seems to be the most accepted by the Digital Humanities community. In our first implementation the OAC ontology has been adopted and extended. It provides solid support for contextual metadata and for attaching annotations to involved Web resources. Such resources can be entire media objects or fragments (basing on Media Fragments and XPointer). Other ontologies, like the Annotation Ontology[17], mostly used in bio-science community, have similar structure and comparable expressivity. In such ontologies annotations have a payload (body) that represents the user-created informative content. In practice, this is usually a Web page (e.g. a blog entry) or a textual comment.

One of the first issues we had to tackle was how to represent annotations that have an RDF graph as body. Even if this specific case is starting to be discussed within the community, it has not yet been regulated by the OAC specification that makes no assumption on the kind of body an annotation can have. It can be, for example, a plain text or a resource with its own URI. In RDF, there are different methodologies to model such a situation, from standard reification, to Content in RDF [12] or some ad-hoc solutions. As our primary goal is to prove how RDF triples produced by users can be aggregated using flexible criteria, we found it convenient to adopt named graphs to represent semantically structured annotation content. In our model, each annotation has an ”oac:body” that is associated with a named graph, where the informative content is represented in triples. This allow us to exploit standard support for named graphs in SPARQL and in triplestores, thus querying and accessing only little “slices’ of the entire collaborative knowledge graph. As discussed in detail later, this is very important to support personal views and target use cases.

The annotation storage is agnostic with respect to the ontologies used to represent the informative content of annotations. However, communities and DLs would greatly benefit from the uniformity of the data schema and vocabulary used in annotations. Our approach allows DLs to deploy specific configurations of the annotation tools provided, enabling users to transparently adhere to predefined data schemas. A range of pluggable entity spaces (like ontologies or thesauri) can be used in practice to provide users with a shared common vocabulary, enabling effective structured descriptions of any knowledge domain at different levels of expressiveness and with different structures. At the current stage, the annotation tool supports both “open’, relatively flat vocabularies like Freebase (leveraging the reconciliation APIs[18]) and restricted controlled vocabularies and taxonomies, e.g. based on the SKOS model [13]. The following example in N3 syntax shows how an annotation and its informative content are represented in RDF.

Listing 1.1. An annotation example in N3 notation

```
// contextual metadata
ex:ANNOTATION-ID-1 a oac:Annotation ;
  rdfs:label "My test annotation";
  dcterms:created "2011-01-27 10:30:56";
```

[17] http://code.google.com/p/annotation-ontology/wiki/Homepage

[18] http://wiki.freebase.com/wiki/Freebase_API

```
  dcterms:creator ex:ChristianMorbidoni ;
  oac:hasBody ex:ANN-BODY-ID-1 ;
  oac:hasTarget http://example.com/1.htm ;
  oac:hasTarget http://example.com/1.jpg .

ex:ANN-BODY-ID-1 a oac:body ;
  rdfs:comment "This is an optional comment" ;
  semlib:graph ex:graph-ann1 .

// informative content
ex:graph-ann1 {
  http://example.com/1.htm tags:hasTag http://www.freebase.com/view/en/
      pippo_baudo ;
  http://example.com/1.jpg foaf:depicts ex:PippoBaudo ;
  http://example.com/1.jpg ex:is-related-to http://example.com/1.htm.}
```

5 Addressing digital content and fragments

While the system is designed to work on generic web pages, there are some features that pose some requirements on DLs to better handle annotations. Two main issues have emerged from the analysis of the SemLib use cases and previous experiments.

DLs, like other web 2.0 applications, change over time. Presentation can be restyled and content can be re-organized. In addition, the same content (e.g. a page of an essay) can be accessible via different Web location (e.g. a summary page and the whole essay page). If we want annotations to remain consistent in such cases, in particular when they are shared in communities and not under a centralized control, we need a way of unambiguously identify atomic, annotable contents in DL Web pages. For this reasons the annotation system requires DLs to include RDFa tags to wrap atomic content, the granularity being opportunely tuned to address specific needs. Each marked content should have a resolvable URI associated, to which annotations are attached. This allows also for an annotation to be automatically associated to all pages that include the same content, as it might happen, for example, for derivative works.

As it happens for stand-off markup in general, the annotated content can change itself, e.g. typos gets fixed or corrections are made by editors. In such cases, annotations referring to fine granular fragments (e.g. sentences or words) can become invalid or simply no more addressable in the modified version. While editorial changes in some DLs result in new versioned objects, this is not a rule in practice, and preserving annotations through content modifications and revisions can be useful in publication workflows. In SemLib, this issue has not been fully addressed yet, but the model is “tolerant’ to content change. We use XPointers to address DOM documents fragments of the marked content, but we also store the original annotated content, checking for broken annotations and possibly alerting the user when they are shown.

6 Sharing annotations

In our system users collect their annotations in notebooks, which are private by default but can be made public and shared with others. Notebooks are identified by dereferenciable URLs that applications can use to retrieve RDF-encoded annotations and relative metadata in different formats (RDF/XML, JSON, etc.). Being able to collect annotations in different notebooks helps users in organizing their work and in grouping annotations by topic or task, furthermore it allows users to make available to others subsets of their annotations.

Sharing a notebook is as easy as sharing its URL on the web, similarly to what happens for popular file sharing platforms. At the moment our system does not provide a social network itself where notebooks can be shared, rather the idea is that of relying on existing communication tools and social media that users are already familiar with. For example, if users want to share a notebook with a single person (e.g. a colleague), they can send the url via mail. In other cases, where users wants make a notebook of public domain, twitter, facebook or other social media can be used as publishing channels. This simple mechanism is general enough to enable different collaborative scenarios, but has limitation in terms of security: once a notebook is made public, each user that receive or find somewhere its URL can access the annotations. In later versions of the system, in order to better address real world use cases, owners of a notebook will be able to explicitly grant read and write permissions to other users of the annotation system. When a users receive an invitation to view a notebook (e.g. receiving the URL by mail) they simply click on it and, if signed in to the annotation system, they are redirected to the notebook web page where they can “activate' it. Each user has a personal preference page where he/she manage the list of active notebooks. When a notebook is active its content is visible to the user while annotated resources are browsed. In other words, by properly configuring the environment, users will be able to aggregate their and others annotations and explore them as custom semantic graphs.

7 Creating crowdsourced annotation collections

DL owners interact with the annotation system in two ways. On the one hand they deploy custom configuration of the system to deliver domain specific annotation tools to their users, by including Javascript libraries into their Web pages or suggesting shortcuts as bookmarklets to users. Using such annotation tools, communities of users, around single or federated DLs, can transparently produce metadata adhering to agreed schemas and vocabularies. This in turn makes the collectively produced data interoperable with the DL itself.

On the other hand, DLs owners/maintainers can act as content curators. As such, they might want to make their own annotation but also to select relevant end-user contributions, aggregate them and, perhaps implement a proper contribution submission workflow (as it happens, for example, with reviewed publications). This would in turn enable a reward based scenario that can stimulate users to contribute. While SemLib does not implement any specific publication workflow, the intent is that of providing a framework that applications can base on to implement their own. In practice, content curators would act as "power users' of the annotation system. They produce their own annotation as regular user do, and they can copy annotations from users-created notebooks to their own notebooks, preserving authorship and other contextual metadata. Such curated notebooks, along with their informative structured content, can be delivered back to users as trusted/official annotations, or directly imported to enrich the DL. In the first case a properly configured GUI, once embedded in the DL, could show the official annotations distinguishing them from users personal notebooks using some visual effect. In the second case, DLs can use simple REST APIs provided by the system to consume RDF encoded annotations and import them into their own database. Experiments in this directions are being made in SemLib, where some of the involved SME's products are natively based on RDF.

8 Prototypal implementation

At the time of writing, the annotation system implementation has reached a prototype stage and, while collaborative features are still not fully implemented, is supports annotation of generic Web contents. It can be used in any existing Web site without modification to its structure and source-code, it is completely decoupled from the Web sites or DLs to be annotated and can be run by end-users through a dedicated bookmarlet. The system is made of two main macro-components: a client-side and a server-side component. When a user launches the bookmarklet, the client-side component is automatically plugged into the web page the user is currently browsing. The client-side component comprises a set of sub-modules developed in Javascript using the dojo framework [19] to facilitate cross-browser support. The client-side module implements the graphical user interfaces to create and browse annotations as well as modules dedicated to the communication with the server. Among these components the most important are the Fragment Handlers, the Resource Selectors and the annotation composer, called Pundit. Their interactions are depicted in Fig. 1.

During the annotation process, Fragment Handlers and Resource Selectors allow users to import different kind of resources into Pundit, where they can be used to compose structured annotations. Fragment Handlers and Resource Selectors can be configured by the system administrator to use specific vocabularies. Fragment Handlers assist users in selecting parts of content (eg. parts of a web page, parts of images, video frames, etc.) and turn them into actual addressable resources (e.g. using XPointer) to be used into annotations. Fragment Handlers also have the role of resolving resource fragments involved in existing annotations so that they can be highlighted in the page. Resource selectors have a similar role: they allow users to import into Pundit selected terms from

[19] http://dojotoolkit.org/

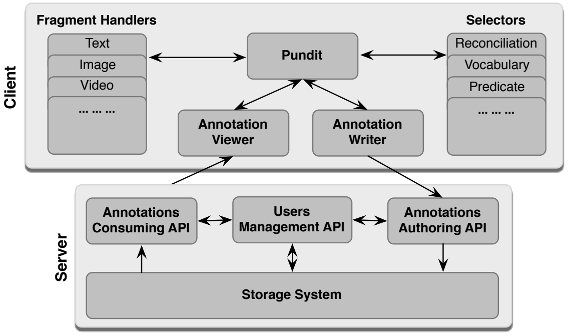
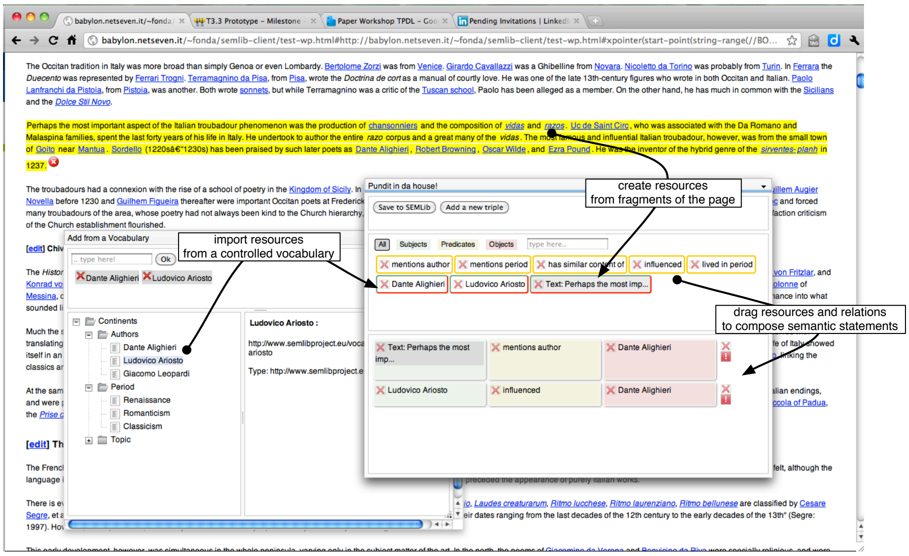


Fig. 1. Simplified architecture of the annotation system

a vocabulary or entity space. Resources are typed, where types are addressable resources as well (as it happens in RDF Schema). The current prototype implements two kind of selectors: one based on the Freebase reconciliation service and one presenting vocabs from a configurable domain taxonomy (e.g. conceptually equivalent to a SKOS vocabulary). Once resources are added to Pundit, users can build structured information in the form of triples (subject, predicate and object), by specifying semantically typed relations that links them, chosen from a predefined, configurable list or RDF properties. Pundit uses domain and ranges of such properties to assist the user and suggest proper relations for different kind of resources. At the current state, the discussed modules can be configured via simple JSON files. However, as the underlaying model is an RDF Schema ontology, such a configuration could be easily extracted from a SPARQL endpoint. This might be useful if the DL exposes its data schema and resources via Semantic Web standard mechanisms such as SPARQL and Linked Data. This point will be addressed later in the project. The screenshot in Fig. 2 shows the prototypal user interface to compose semantic annotations. Users can select fragments of the page and import them into Pundit, where they can be dragged to populate statements. Users can also import resources from provided custom taxonomies (like the simple one in the illustration) or from Freebase, and again use them in annotations.

Once triples have been edited, user can save them to the Annotation Server, which is a modular RESTful web-service. It allows annotation storage, user authentication and management in addition to APIs for annotation authoring, consuming and sharing. Such RESTful APIs, partially inspired by previous works as the Annotea Protocol, allow users to create new notebooks and annotations supporting different data formats (e.g. RDF, JSON, etc.), to browse notebooks and related annotations and to personalize users views by activating public notebooks (e.g. shared by others). Such aggregations of activated notebooks can be then exploited by querying them and retrieving semantic data in the form of RDF triples. A typical use of such querying functionalities is that of retrieving all the RDF statements where a particular web resource (or a fragment of it) is involved. Sub-graphs obtained in this way can be immediately explored with existing Semantic Web aware tools. A prototypal annotation navigator, for example, has been implemented using Simile Exhibit.



The Occitan tradition in Italy was more broad than simply Genoa or even Lombardy. Bertolome Zorzi was from Venice. Girardo Cavallazzi was a Ghibelline from Novara. Nicoletto da Torino was probably from Turin. In Ferrara the *Duecento* was represented by Ferrari Trogni. Terramagnino da Pisa, from Pisa, wrote the *Doctrina de cort* as a manual of courtly love. He was one of the late 13th-century figures who wrote in both Occitan and Italian. Paolo Lanfranchi da Pistoia, from Pistoia, was another. Both wrote sonnets, but while Terramagnino was a critic of the Tuscan school, Paolo has been alleged as a member. On the other hand, he has much in common with the Sicilians and the *Dolce Stil Novo*.

Perhaps the most important aspect of the Italian troubadour phenomenon was the production of chansonniers and the composition of *vidas* and *razos*. Uc de Saint Circ, who was associated with the Da Romano and Malaspina families, spent the last forty years of his life in Italy. He undertook to author the entire *razo* corpus and a great many of the *vidas*. The most famous and influential Italian troubadour, however, was from the small town of Goito near Mantua. Sordello (1220sâ€"1230s) has been praised by such later poets as Dante Alighieri, Robert Browning, Oscar Wilde, and Ezra Pound. He was the inventor of the hybrid genre of the *sirventes-planh* in 1237.

babylon.netseven.it/~fonda/
T3.3 Prototype – Milestone
Paper Workshop TPDL – Goo
Pending Invitations | Linked
babylon.netseven.it/~fonda/semlib-client/test-wp.html#http://babylon.netseven.it/~fonda/semlib-client/test-wp.html#xpointer(start-point(string-range(//BO...
Pundit in da house!
Save to SEMLib
Add a new triple
All
Subjects
Predicates
Objects
type here..
mentions author
mentions period
has similar content of
influenced
lived in period
Dante Alighieri
Ludovico Ariosto
Text: Perhaps the most imp...
Text: Perhaps the most imp...
mentions author
Dante Alighieri
Ludovico Ariosto
influenced
Dante Alighieri
Add from a Vocabulary
.. type here!
Ok
Dante Alighieri
Ludovico Ariosto
Continents
Authors
Dante Alighieri
Ludovico Ariosto
Giacomo Leopardi
Period
Renaissance
Romanticism
Classicism
Topic
Ludovico Ariosto :
http://www.semlibproject.eu/vocab/ariosto
Type: http://www.semlibproject.e
create resources from fragments of the page
import resources from a controlled vocabulary
drag resources and relations to compose semantic statements

Fig. 2. A screenshot of Pundit in action

The storage module defines a completely generic interface, designed to support different kind of storage systems ranging from traditional relational databases to NoSQL databases (eg. RDF triplestores). In the prototype version, the storage is implemented using the Sesame triplestore [20] as this greatly simplifies handling and exporting RDF data. The storage module, besides keeping users annotations, stores also user profiles and related contextual information (e.g.: user's metadata, user's permissions etc.). The Annotation Server supports two single sign-on systems for users authentication, in particular, Open-ID[21] and OAuth[22]. Different authentication systems can be easily implemented developing dedicated plugins. Using single sign-on systems simplifies the integration of the annotation system with existing DL, which may already provide facilities for users authentication.

9 Conclusions

In this paper, we introduced the SemLib project, focusing on the proposed data and social model and explaining how those are expected not only to foster annotation sharing between DL communities and user engagement but also to allow the application of crowdsourcing paradigm in the creation of added value for the DLs. As proof of concept of our ideas, we also presented an early prototype implementation of the system discussing the experimental client-side GUIs for annotation creation and the server's RESTful APIs for annotation storage, sharing and consumption.

[20] http://www.openrdf.org/
[21] http://openid.net/
[22] http://oauth.net/

As SEMLIB is an ongoing project, not all the features here described have been implemented yet, and several challenges are still open in improving annotation creation, visualization and sharing, which will be tackled in future releases of the annotation system. Also, the proposed system will be extensively tested on existing DLs of partner SMEs, which is expected to provide valuable feedbacks and to further boost the development process.

10 Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme managed by REA-Research Executive Agency[23] ([FP7/2007-2013][FP7/2007-2011]) under grant agreement n. 262301.

References

1. DELOS, "The DELOS Digital Library Reference Model: Foundations for Digital Libraries, version 0.96'. November, 2007.
2. R. A. Arko, K. M. Ginger, K. A. Kastens, and J. Weatherley, "Using annotations to add value to a digital library for education'. [Online]. Available: http://www.dlib.org/dlib/may06/arko/05arko.html,
3. Rose Holley, "Crowdsourcing: How and Why Should Libraries Do It?', D-Lib Magazine, The Magazine of Digital Library Research. March/April, 2010.
4. M. Grassi, C. Morbidoni, M. Nucci, "Semantic Web Techniques Application for Video Fragment Annotation and Management', Proceedings of the SSPnet-COST 2102 PINK International Conference on "Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues" pp.95-103. 2011.
5. B. Haslhofer, E. Momeni, M. Gay, and R. Simon, "Augmenting Europeana Content with Linked Data Resources', in 6th International Conference on Semantic Systems (I-Semantics), September 2010.
6. J. Kahan, M. R. Koivunen, "Annotea: An Open RDF Infrastructure for Shared Web Annotations', Proceedings of the 10th international conference on World Wide Web, Page(s): 623-632, 2001.
7. A. Gerber and J. Hunter, "Authoring, Editing and Visualizing Compound Objects for Literary Scholarship', Journal of Digital Information, vol. 11, 2010.
8. M. L. Ralf Heese, "One Click Annotation' in 6th Workshop on Scripting and Development for the Semantic Web, 2010.
9. Marja-Riitta Koivunen, "Annotea and Semantic Web Supported Collaboration". ESWC 2005, UserSWeb workshop. 2005 Marja-Riitta Koivunen, Ralph Swick, Eric Prud'hommeaux
10. "Annotation Ontology' http://code.google.com/p/annotation-ontology/
11. "Open Annotation: Alpha3 Data Model Guide' 15 October 2010 Eds. R. Sanderson and H. Van de Sompel. http://www.openannotation.org/spec/alpha3/
12. "Representing Content in RDF 1.0'. W3C Working Draft 10 May 2011. http://www.w3.org/TR/Content-in-RDF10/
13. "SKOS Simple Knowledge Organization System Reference'. W3C Recommendation. 18 August, 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/

[23] http://ec.europa.eu/research/rea

Concepts and Collections: A case study using objects from the Brooklyn Museum

Tim Wray and Peter Eklund

twray,peklund@uow.edu.au
School of Information Systems and Technology
University of Wollongong
Northfields Ave, Wollongong, NSW 2522, Australia.

Abstract. In this paper, we present a browsing framework for digitised cultural collections. Using a data analysis technique called Formal Concept Analysis (FCA), units of thought can be constructed from a series of objects and their tags. FCA can dynamically generate links in between objects and induce a serendipitous browsing experience using a relatively simple data structure. We evaluate the utility and scalability of our approach to a collection of 15,000 objects from the Brooklyn Museum's collections. We describe how we use natural language processing techniques and external lexical resources to synthesise key terms from museum documentation. We then combine this term extraction process with FCA to effectively demonstrate links between and within collections of objects. In doing so we present a versatile, generalizable term extraction and browsing framework suitable for digital libraries and archives within the art and architecture domain.

1 Introduction

Cultural collections are vast, heterogeneous stores of history that are monumental in their representation of human history and expression. Of particular interest are the philosophical notions on how to best represent knowledge within these collections, beginning from the rigid classification hierarchies that are commonly employed in today's cultural collections to organic, tag-based, associative approaches. Weinberger [1] examines tags as a form of classification, and notes that there are often multiple relationships among objects within a collection, each of which can be meaningful in their own interpretation. He quotes that "trees can be built from leaves" – meaning that sorting and categorisation can be dynamically induced, either from user communities and stakeholders (social tagging) or from the metadata itself without reliance on an imposed classification schema. In effect, sorting, categorising and relating objects can be organic, dynamic and data-driven. When combined with a consistent knowledge representation structure and controlled vocabulary, these relationships can unify multiple, heterogeneous collections.

Large scale cultural heritage projects such as Europeana[1] and Digital NZ[2] are a step in the right direction in unifying and providing accessibility to collections. As a result of projects such as these, there is a large amount of research conducted in making these collections accessible and semantically inter-related. For example, Schreiber et al. [2] investigate approaches towards enhancing and enriching collection metadata and providing semantic annotation and search facilities to large cultural collections. Klavans et al. [3] describe the nuances and challenges of extracting metadata from cultural collections using natural language processing techniques. Work conducted by Trant [4] report on how audiences can contribute new knowledge to collections in the form of social tagging while latter work by Klavans et al. [5] examine how such data could be exploited in order to assist information retrieval and browsing. These authors recognise the importance of deriving meaning from cultural collections. Their research is well aligned with related work in data visualisation, such as the Visible Archive and commonsExplorer projects [6]. Like our project, these works focus on the discovery of patterns and relationships within collections, rather than traditional targeted search.

In our approach, we discover these patterns and relationships by using a data analysis technique called Formal Concept Analysis (FCA). FCA is the mathemetization of conceptual thinking – a way of ordering and relating structured units of thought [7]. A *formal concept* denotes a unit of thought and consists of an extension, the objects that compose that thought, and an intension, the attributes, properties and meanings that apply to all of the objects within the extension. For example, when applied to a collection of works, one may be thinking about "Chinese vases with floral patterns" (the intension, or the attributes) or the actual 17 vases (the extension, or objects). In human conceptual thinking, concepts rarely exist on their own, but rather in relation with many other concepts [8] – as a result neighbouring concepts often play an important role in data analysis and communication. For example, it is inevitable that there would be some sort of link between "Chinese vases" and "vases with floral patterns" – these are *superconcepts* of "Chinese vases with floral patterns", so called because they represent 'broader' concepts with a greater set of objects. Dually, concepts such as "Chinese vases with floral patterns from the Qing dynasty" are *subconcepts* – they provide a more narrow, focused view of the collection. These superconcept-subconcept relationships are one of the core mechanisms in which we use to provide associative links between clusters of objects and as such, it drives our framework for browsing digitised collections.

Over 10 years of research in applied FCA has been dedicated towards new approaches of knowledge discovery within collections. Projects such as ImageSleuth and ImageSleuth2 [9] are precursors to the design of the Virtual Museum of the Pacific [10] in which the current framework is derived from. This research assesses the applicability of the browsing framework towards a large data-set of 15,000 objects from the Brooklyn Museum's collections, using an automated

[1] http://www.europeana.eu/portal/

[2] http://www.digitalnz.org/

term extraction approach to derive the required key terms for analysis. It refines and assesses the applicability of the content-based retrieval component of the framework, and its contribution lies in its applicability to a large, real world data-sets.

The structure of this paper is as follows: Section 2 provides a brief introduction of Formal Concept Analysis as applied in our case study, and its significance as a tool for linking groups of objects. In Section 3, we describe how we extract key terms from the Brooklyn Museum's API in order to provide a suitable data structure for analysis. In Section 4, we describe results of our application of Formal Concept Analysis to those terms, highlighting issues with respect to scalability and complexity along with the results of the description of a prototype collection browser. The paper concludes with a discussion on useful applications and extensions of our work.

2 Formal Concept Analysis

Formal Concept Analysis (FCA) [7] is a core feature of our framework that is used to derive relationships among objects. Central to the theory of FCA is the notion of the *formal concept*, and its resulting algebraic structure, the *concept lattice*. To clarify the theory of FCA, we will use a Pacific collection of objects as an example.

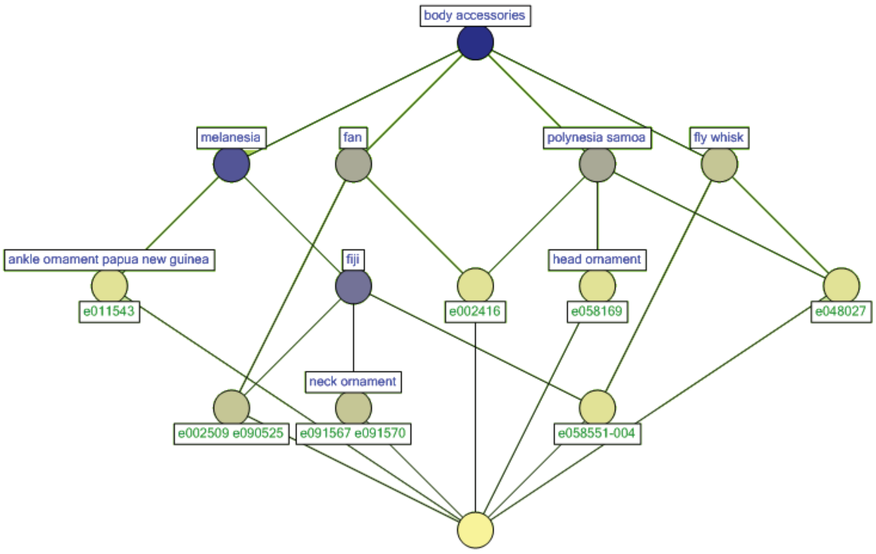


Fig. 1. A concept lattice for a small collection of Pacific objects. Labels above the nodes denote attributes (or tags) and labels below the nodes denote registration IDs from the Museum's content management system

A formal concept (A, B) represents a unit of thought, where A is a set of object identifiers and B is a set of attributes, or 'tags', that describe the objects. For example, the concept "Fijian fans" can be represented by (A, B) where $A = \{\text{e002509}, \text{e090525}\}$ and $B = \{\text{body accessory}, \text{fan}, \text{melanesia}, \text{fiji}\}$. Formal concepts can be ordered and arranged in a specialisation hierarchy. A concept (A, B) is a sub-concept of concept (C, D) if $A \subseteq C$ (or equivalently, $B \supseteq D$). Using this definition, more specific concepts have fewer objects and more attributes. For example: $(A, B) < (C, D)$ where:

$(A, B) = \{\{\text{e002509}, \text{e090525}\}, \{\text{body accessory}, \text{fan}, \text{melanesia}, \text{fiji}\}\}$

$(C, D) = \{\{\text{e090525}, \text{e002509}, \text{e058551-004}\}, \{\text{body accessory}, \text{melanesia}, \text{fiji}\}\}$

The set of all formal concepts, together with the specialisation relation, forms the *concept lattice*. The concept lattice is an algebraic structure that shows hierarchies and relations between formal concepts (Fig. 1). It is derived from the *formal context*, which is a list of objects and their tags, represented as a cross-table (Table 1) and formally denoted as a triple $\mathbb{K} := (G, M, I)$ where G is a set of formal *objects*, M is a set of *attributes* and I is an *incidence* relation between the objects and the attributes.

Table 1. The formal context, or cross table, used to generate the concept lattice in Fig. 1. Note that the core data structure can be expressed as a series of objects and tags.

$\mathbb{K}$	body accessories	fan	head ornament	ankle ornament	fly whisk	neck ornament	melanesia	polynesia	fiji	papua new guinea	samoa
e002509	×	×					×		×		
e090525	×	×					×		×		
e058551-004	×				×		×		×		
e091567	×					×		×	×		
e091570	×					×	×		×		
e002415	×	×						×			×
e002416	×	×						×			×
e058169	×		×					×			×
e058169	×				×			×			×
e011543	×			×			×			×	

In Fig. 1, nodes represent formal concepts. Labels above the nodes represent attributes, (or tags) that describe the object, and labels below the nodes represent the database identifiers of those objects. The set of attributes for a particular formal concept is inferred by gathering all of the attribute labels as one would traverse upwards on the line diagram, starting from the node represented by the formal concept and ending at the top node. For example, based on the interpretation on this line diagram, one can infer that objects 'e002509' and 'e090525' are similar to objects 'e091567' and 'e091570', in that they share common attributes 'fiji', 'melanesia' and 'body accessories' and that they are close to one another. This observation, in part, drives the foundation of the similarity and distance metrics that we use to provide an order ranked list of similar formal concepts for a given object [11].

The similarity metric is a measure based on the number of common objects and the number of common attributes (tags) of two given formal concepts (A, B) and (C, D):

$$similarity((A, B), (C, D)) := \frac{1}{2}\left(\frac{|A \cap C|}{|A \cup C|} + \frac{|B \cap D|}{|B \cup D|}\right).$$

The distance metric is a measure based on the overlap of the objects and attributes of two concepts, normalised with respect to the size of the formal context, where G is the total set of objects and M is the total set of attributes. For two concepts (A, B) and (C, D), the distance metric is as follows:

$$distance((A, B), (C, D)) := \frac{1}{2}\left(\frac{|A \setminus C| + |C \setminus A|}{|G|} + \frac{|B \setminus D| + |D \setminus B|}{|M|}\right).$$

When combined, these two metrics can be used to provide a list of similar formal concepts to a given object, ordered from 'most similar' to 'least similar.' As we are comparing formal concepts, a similarity query can derive both matching and nearby objects (e.g. "An American sculpture that depicts youth") or clusters of objects (e.g. "6 Contemporary sculptures that are made with bronze"). Section 4 of this paper describes how we use these similarity metrics to provide an order ranked list of objects and object clusters from the Brooklyn Museum's collections. However, in order to do so, we need to build the formal context by extracting key terms from the objects.

3 Term Extraction: Building the Formal Context

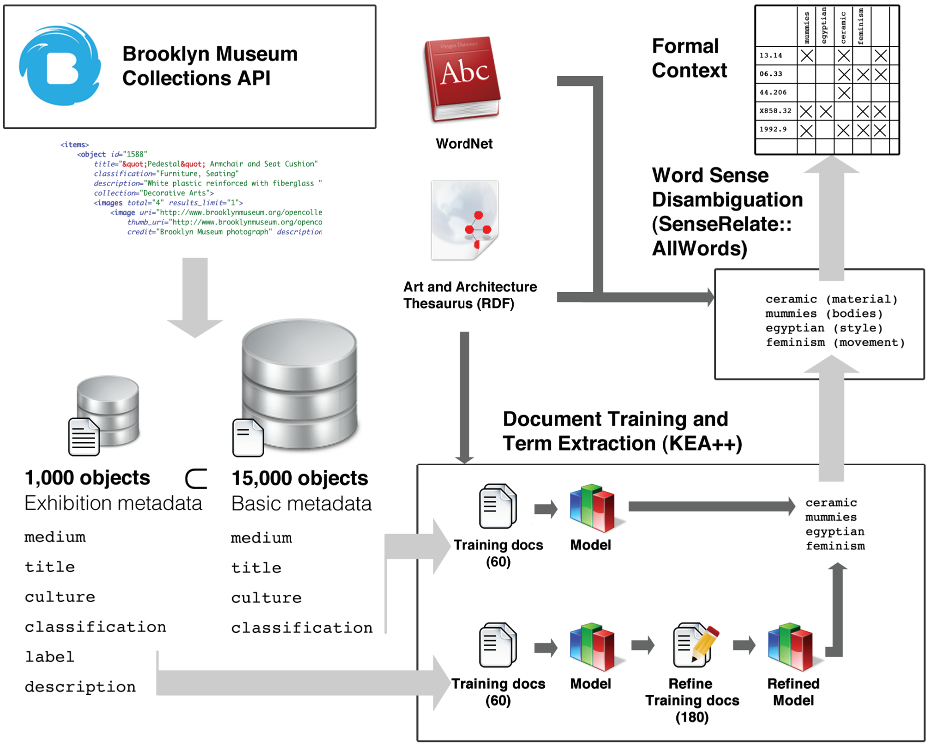
Term extraction algorithms, such as Yahoo's Term Extraction Web Service[3], are commonly employed to assign keywords to documents based on their content. Our term extraction method is built based on the work of Klavans et al. [3] who discuss the application of computational linguistics to museum collections along with current state-of-the-art algorithms developed by Medelyan [12], Frank et al. [13] and Witten et al. [14]. We employ external lexical resources, such as WordNet [15] and the Getty's Art and Architecture Thesaurus[4] to provide semantic background knowledge for the term extraction process. Like many

[3] http://developer.yahoo.com/search/content/V1/termExtraction.html

[4] http://www.getty.edu/research/tools/vocabularies/aat/about.html

natural language processing applications, we employ a pipeline architecture for term extraction, shown in Fig. 2.

We source a collection of 15,000 objects using the Brooklyn Museum's API. These objects are an amalgation of 12 collections from the museum. The completeness of the object records vary considerably – some objects have full descriptions and interpretive labels to a depth and standard typically found within exhibition catalogues and are often procured for exactly that purpose. These descriptions often provide the cultural context of the object, how it was used, where it came from and its significance. Given the time and cost associated with their research, objects of these descriptions would naturally only occupy a small portion of the collection. Therefore, 1000 objects were selected as objects having *exhibition quality metadata*. Likewise, the entire collection of 15,000 objects were documented, in the very least, with notes and details of its medium, title, culture and classification – denoted as *basic metadata*. As the amount of metadata present within an object determines the kinds (and types) of terms that could be extracted from them, we create two instances of our framework to accommodate these two classes.



```
<items>
    <object id="1588"
        title="&quot;Pedestal&quot; Armchair and Seat Cushion"
        classification="Furniture, Seating"
        description="White plastic reinforced with fiberglass "
        collection="Decorative Arts">
        <images total="4" results_limit="1">
            <image uri="http://www.brooklynmuseum.org/opencolle
                thumb_uri="http://www.brooklynmuseum.org/openco
                credit="Brooklyn Museum photograph" description
```

	mummies	egyptian	ceramic	feminism		
13.14	☒		☒		☒	
06.33			☒	☒	☒	
44.206			☒			
X858.32	☒	☒		☒	☒	
1992.9	☒		☒	☒	☒	

Fig. 2. Overview of the term extraction process used to generate formal contexts, shown anti-clockwise from the top-left

To perform the term extraction, we use a program called KEA++. KEA++ has proven to be a high performance extraction program [12] that combines *keyphrase extraction*: identifying features and prominent keyphrases from a document and *keyphrase assignment*: where terms are selected from a controlled vocabulary using a trained model. We employ the Getty's Art and Architecture Thesaurus (AAT) as the controlled vocabulary. The AAT contains over 34,800 unique concepts under 33 hierarchies for describing object categories, materials, activities and functions, styles and periods and other abstract phenomena associated with material culture and artworks. It can be used as a single ontology for unifying disparate collections and digital archives. Where appropriate, we use specific hierarchies to perform term extraction on certain types of data fields. For example, the object's 'medium' data field (shown in Fig. 2) would employ only a sub-section of the thesaurus, mainly the 'physical attributes' and 'materials' hierarchies. This is to reduce the likelihood of a document being assigned an incorrect term due to overstemming (e.g. 'painting (visual works)' was often incorrectly assigned instead of 'paint (medium)'). For basic metadata, each data field ('medium', 'title', 'culture' etc.) was provided with a set of 60 training documents. For the exhibition quality metadata fields, 60 training documents where used to generate a model that produced 180 documents, which were then refined to produce the final model.

For each object record, KEA++ generates a set of candidate terms. However, many of these terms are ambiguous – over 16% of terms extracted from the collections referred to more than one sense within the AAT. For example, the term 'gold' refers to two senses of the word, referring to both the material and the color property of an object. As described by Palmer et al. [16], the common linguistic problem of *word-sense disambiguation* is a particularly challenging one. To solve this problem, we adapt a method proposed by Klavans et al. [3] that uses an external algorithm called SenseRelate::AllWords [17]. This algorithm is a Perl module that identifies the correct WordNet sense of each word in a sentence, using the surrounding text as its context. This AAT sense is then selected by performing a word overlap of the definitions of the AAT record and the WordNet sense – the AAT sense with the highest match is assigned to that word.

Once the terms are extracted and disambiguated, we then use them to construct the formal context. As hierarchical term relationships are naturally embodied within FCA, we exploit the broader-narrow relationships within the AAT to enrich the formal context with parent tags so that for example, 'streetscapes' → 'cultural landscapes' → 'landscapes'. These hierarchical relations complement the similarity and distance metrics described in Section 2 as these metrics favour objects that share attributes with common parents so that for example, 'streetscapes' is notionally similar to 'suburban landscapes'.

The final step is to prune the formal context in order to reduce its complexity. Although FCA is theoretically robust, applications that employ it for data analysis and communication commonly apply a number of techniques to remove extraneous data points while retaining meaningful representation of its information space [18]. It is also necessary to employ these complexity reduction measures given the high computational cost of FCA-based operations with respect to the size of the formal context [19]. While more elaborate approaches for reducing complexity in fully formed concept lattices exist [20] [18], our approach needs to rely on more rudimentary measures of complexity reduction as each similarity / distance operation traverses only part of the data-set as required. We use an approach called *context reduction* – it removes rarely occurring tags, which, despite their 'insignificance', reduces the size and complexity of the formal context considerably. This makes sense as the "aboutness" of the objects are dictated by the attributes that they have in common, rather than the attributes that they don't have in common. We remove tags that do not belong to a threshold percentage of objects, with the threshold value set by default at 0.05%.

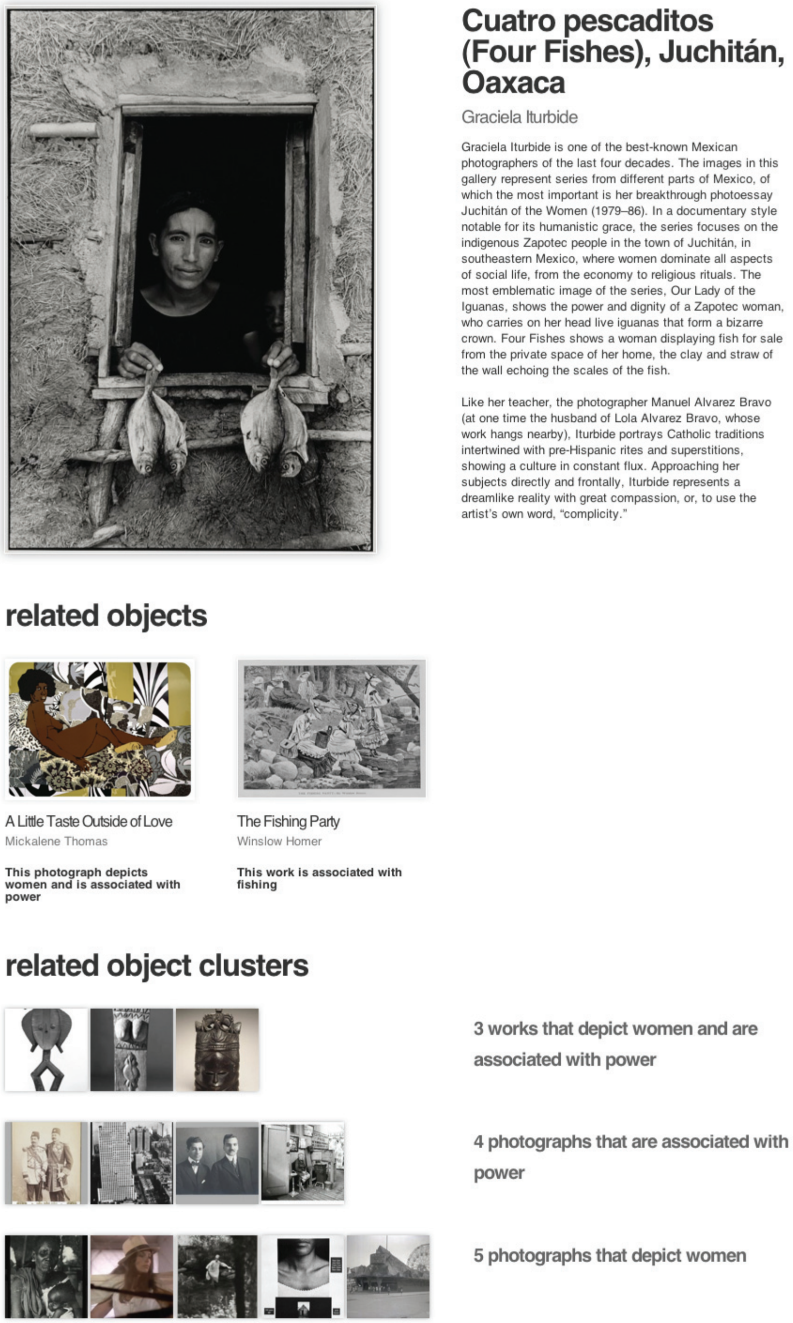
4 Results and Scalability of our Approach

A key design requirement of our framework is to induce an explorative browsing experience by computing the similarities and differences between objects, deriving natural pathways within collections, and highlighting key concepts – showing collections within collections. Furthermore, its implementation needs to be scalable with a performance requirement to suit real time interactive browsing over the Web. To show the results of our work, we have developed a light-weight prototype collection browser[5], shown in Fig. 3. The browser shows a detailed catalogue description, with links to conceptually similar objects and object clusters.

In the example shown in Fig. 3, the extracted terms of the artwork: {photographs, rituals, women, power, fishing} are used to compute the following similar formal concepts, order ranked from most similar to least similar:

- { women, photographs, power }
  2 objects (Similarity: 0.55, Distance: 0.99)
- { women, fishing }
  2 objects (Similarity: 0.45, Distance: 0.99)
- { women, power }
  5 objects (Similarity: 0.30, Distance: 0.99)
- { photographs, power }
  6 objects (Similarity: 0.28, Distance: 0.99)
- { women, photographs }
  7 objects (Similarity: 0.27, Distance: 0.99)

[5] Two prototype collection browsers are publicly available for the two collections:
1,000 objects with exhibition quality metadata:
http://epoc2.cs.uow.edu.au/brooklyn_r_1000_ws/similarity/
15,000 objects with basic metadata:
http://epoc2.cs.uow.edu.au/brooklyn_m_15000_ws/similarity/



Cuatro pescaditos (Four Fishes), Juchitán, Oaxaca

Graciela Iturbide

Graciela Iturbide is one of the best-known Mexican photographers of the last four decades. The images in this gallery represent series from different parts of Mexico, of which the most important is her breakthrough photoessay Juchitán of the Women (1979–86). In a documentary style notable for its humanistic grace, the series focuses on the indigenous Zapotec people in the town of Juchitán, in southeastern Mexico, where women dominate all aspects of social life, from the economy to religious rituals. The most emblematic image of the series, Our Lady of the Iguanas, shows the power and dignity of a Zapotec woman, who carries on her head live iguanas that form a bizarre crown. Four Fishes shows a woman displaying fish for sale from the private space of her home, the clay and straw of the wall echoing the scales of the fish.

Like her teacher, the photographer Manuel Alvarez Bravo (at one time the husband of Lola Alvarez Bravo, whose work hangs nearby), Iturbide portrays Catholic traditions intertwined with pre-Hispanic rites and superstitions, showing a culture in constant flux. Approaching her subjects directly and frontally, Iturbide represents a dreamlike reality with great compassion, or, to use the artist's own word, "complicity."

related objects

A Little Taste Outside of Love
Mickalene Thomas

This photograph depicts women and is associated with power

The Fishing Party
Winslow Homer

This work is associated with fishing

related object clusters

3 works that depict women and are associated with power

4 photographs that are associated with power

5 photographs that depict women

Fig. 3. Screenshot of the prototype collection browser with links to conceptually similar objects and object clusters

From these results, the algorithm derives all objects as a unique set[6], and clusters them according to their member concept. Each similar formal concept consists of the object we are comparing plus member objects of that same formal concept, i.e., the first two results indicate individual objects tagged
{ women, photographs, power } and { women, fishing }, respectively. These objects are presented as 'related objects' as shown in Fig. 3. Other formal concepts are shown as 'object clusters' with a series of thumbnails indicating other focal points of interest within the collection.

We employ natural language labels to describe the objects and object clusters. These labels are generated from the tags from the formal concepts' tags, while their semantics are inferred from their hierarchical membership within the AAT. For example, for a given set of tags { women, photographs, power }, one may assume that they describe photographic works that depict women and are also associated with power, given that 'women' exists in the 'Agents' hierarchy; 'photographs' exists in the 'Visual Works' hierarchy and that 'power' exists in the 'Associated Concepts' hierarchy. Within each hierarchy, its member tags indicate what aspect of an artwork they describe. However, a problem with this approach lies in the inherent ambiguity of whether a term is object-oriented (describing the object itself, its properties) or subject-oriented (describing what the work is about or what it depicts) [21] [22]. In some cases, terms such as 'water' could refer to both a work that is *made with* water or a work that *depicts* water features – an apparent shortcoming of many tag based systems. Currently, the AAT only recognises water in the former sense, and further curation of these sorts of tags may be necessary to prevent these semantic ambiguities.

Performance and scalability are important factors for real world implementation. As theorized by Carpineto and Romano [19], the computational cost of FCA-based operations increases as the size of the formal context gets larger. The results of our performance testing have indicated that dynamically perfoming these computations is unsuitable for a collection of more than 200 objects, with average query times approximating 60 seconds on the full collection of 15,000 objects. To solve this problem, we have adopted a caching method where similar formal concepts are pre-computed and cached with each object record. The system updates these caches as new objects are added, or their tags change.

5 Conclusion and Future Work

We have presented a term extraction and browsing framework as applied to the Brooklyn Museum's collection, using objects and tags as a core data structure. We have also developed a prototype browsing application to demonstrate our framework. It is scalable to a collection of 15,000 objects and it can dynamically generate links to neighbouring objects and object clusters, expressed in natural language. With a focus on *concepts* rather than *objects*, it follows a contemporary

[6] Similar formal concepts have a high overlap of common objects. Based on user feedback, we've adopted a design decision to not show duplicate objects within the UI.

data-driven approach of collections browsing, and it can be suitably adopted for experiments and applications in collections visualisation.

Given that we use a common vocabulary for tagging objects, this work could be extended to cover multiple collections from different institutions with an assessment on if or how our framework could scale, along with how it can adapt to the varying kinds of metadata each collection presents. Tags present a simple and versatile data structure that can be provided or derived from free text. *Semantic tagging* [23] introduces an interesting possibility of solving the previously mentioned semantic ambiguity problem described in Section 4.

Social tagging in museum collections is gaining traction and has proven to add worthwhile community knowledge to museum collections [4] – for example, the Brooklyn Museum provides programs such as "Tag You're It!"[7], and these social tags are commonly used on their website to assist searching and browsing. As an extension of our work, leveraging social meta-data not only closes gaps in museum documentation and opens up interpretation to visitors, but it can also induce dynamic relationships among objects, allowing for a self-evolving and community-driven approach to the display and interpretation of collections.

References

1. Weinberger, D.: Taxonomies to tags: From trees to piles of leaves. Release 1.0 **23**(2) (2005)
2. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B.: Semantic annotation and search of cultural-heritage collections: The multimedian e-culture demonstrator. Web Semantics: Science, Services and Agents on the World Wide Web **6**(4) (2008) 243 – 249 Semantic Web Challenge 2006/2007.
3. Klavans, J., Sheffield, C., Abels, E., Lin, J., Passonneau, R., Sidhu, T., Soergel, D.: Computational linguistics for metadata building (climb): using text mining for the automatic identification, categorization, and disambiguation of subject terms for image metadata. Multimedia Tools and Applications **42** (2009) 115 – 138 10.1007/s11042-008-0253-9.
4. Trant, J.: Tagging, Folksonomy and Art Museums: Results of steve.museum's research. Technical report, University of Toronto (2009)
5. Klavans, J., Stein, R., Chun, S., Guerra, R.D.: Computational Linguistics in Museums: Applications for Cultural Datasets. In Trant, J., Bearman, D., eds.: Museums and the Web 2011: Proceedings, Archives and Museum Informatics (2011)
6. Hinton, S., Whitelaw, M.: Exploring the digital commons: an approach to the visualisation of large heritage datasets. `http://www.bcs.org/upload/pdf/ewic_ev10_s3paper2.pdf` (2010)
7. Wille, R., Ganter, B.: Formal Concept Analysis: Mathematical Foundations. Springer-Verlag, Berlin (1999)
8. Wille, R.: Formal concept analysis as mathematical theory of concepts and concept hierarchies. In Ganter, B., Stumme, G., Wille, R., eds.: Formal Concept Analysis. Volume 3626 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2005) 47–70

[7] `http://www.brooklynmuseum.org/opencollection/tag_game/start.php`

9. Eklund, P., Ducrou, J., Wilson, T.: An intelligent user interface for browsing and search MPEG-7 images using concept lattices. In: Proceedings of the 4th international conference on concept lattices and their applications. LNCS 4923, Springer-Verlag (2006) 1–22
10. Eklund, P., Wray, T., Goodall, P., Bunt, B., Lawson, A., Christidis, L., Daniels, V., Olffen, M.V.: Designing the Digital Ecosystem of the Virtual Museum of the Pacific. In: 3rd IEEE International Conference on Digital Ecosystems and Technologies, IEEE Press (2009) 805–811
11. Saquer, J., Deogun, J.S.: Concept aproximations based on rough sets and similarity measures. In: Int. J. Appl. Math. Comput. Sci. Volume 11. (2001) 655 – 674
12. Medelyan, O.: Automatic keyphrase indexing with a domain-specific thesaurus. Master's thesis, University of Waikato (2005)
13. Frank, E., Paynter, G., Witten, I., Gutwin, C., Nevill-Manning, C.: Domain-specific keyphrase extraction. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, San Francisco, CA, Morgan Kaufmann (1999) 668 – 673
14. Witten, I., Paynter, G., Frank, E., Gutwin, C., Nevill-Manning, C.: Kea: Practical automatic keyphrase extraction. In: Proceedings of the 4th ACM Conference on Digital Libraries, Berkeley, CA, ACM Press (1999) pp. 254 – 255
15. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)
16. Palmer, M., Ng, H.T., Dang, H.T.: Word sense disambiguation: algorithms, applications and trends. In Edmonds, P., Agirre, E., eds.: Text, Speech and Language Technology. Kluwer Academic Publishers, Netherlands (2003)
17. Pederson, T., Kolhatkar, V.: Wordnet::senserelate::allwords: a broad coverage word sense tagger that maximizes semantic relatedness. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Demonstration Session. NAACL-Demonstrations '09, Association for Computational Linguistics (2009) 17 – 20
18. Kuznetsov, S., Obiedkov, S., Roth, C.: Reducing the representation complexity of lattice-based taxonomies. In Priss, U., Polovina, S., Hill, R., eds.: Conceptual Structures: Knowledge Architectures for Smart Applications. Volume 4604 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2007) 241 – 254
19. Carpineto, C., Romano, G.: Concept Data Analysis: Theory and Applications. John Wiley & Sons (2004)
20. Stumme, G., Taouil, R., Bastide, Y., Lakhal, L.: Conceptual clustering with iceberg concept lattices. In: Proceedings of GI–Fachgruppentreffen Maschinelles Lernen'01. Volume 763., Universität Dortmund (2001)
21. Chen, H.: An analysis of image retrieval tasks in the field of art history. Information Processing and Management **37**(5) (2001) 701 – 720
22. Choi, Y., Rasmussen, E.M.: Searching for images: the analysis of users' queries for image retrieval in american history. Journal of the American Society for Information Science and Technology **54**(6) (2003) 489 – 511
23. Marchetti, A., Tesconi, M., Ronzano, F.: Semkey: A Semantic Collaborative Tagging System. In: Proceedings of the WWW Workshop on Tagging and Metadata for Social Information Organisation. (2007)

LOHAI: Providing a baseline for KOS based automatic indexing

Kai Eckert

University of Mannheim
University Library
Mannheim, Germany
eckert@bib.uni-mannheim.de

Abstract. Automatic KOS based indexing – i.e. indexing based on a restricted, controlled vocabulary, a thesaurus or a classification – can play an important role to close the gap between the intellectually, high quality indexed publications and the mass of unindexed publications. Especially for unknown, heterogeneous publications, like web publications, simple processes that do not rely on manually created training data are needed. With this contribution, we propose a straight-forward linguistic indexer, that can be used as a basis for own developments and for experiments and analyses to explore own documents and KOSs; it uses state-of-the-art information retrieval techniques and hence forms a suitable baseline for evaluations. Finally, it is free and open source.

1 Introduction

Intellectual indexing of publications based on a Knowledge Organization System (KOS) – like controlled vocabularies, thesauri or classifications – is still performed to ensure high accuracy in information retrieval. Even with the availability to search the electronic full text, the resolution of synonyms and homonyms by the introduction of a controlled vocabulary – functioning as a common language between creators and searchers of the indexed content – is very important. To close the gap between the subset of publications that are traditionally indexed intellectually – books in libraries, but also selected journal articles in mostly commercial databases – automatic indexing approaches are widely introduced.

The German National Library, for instance, decided, that web publications, while being collected, will not be indexed intellectually, but only by means of automatic processes and search engine technology [9]. A recent workshop focused on automatic indexing, called PETRUS[1] showed that there are mainly two types of approaches: linguistic and statistical approaches. While there are smooth transitions between both, linguistic approaches use techniques from natural language processing (NLP) to process the texts and extract meaningful concepts, while statistical approaches use machine learning techniques to assign concepts based on a manually created training set.

[1] http://www.d-nb.de/wir/projekte/workshop_petrus.htm

Based on the discussions and contributions of the workshop, there is currently a preference for statistical approaches, although the reported quality of the results varies. A mentioned problem was the bias that is introduced by the training set. For example, for recent news articles, the indexer learned that the occurrence of "Nuclear power plant" should lead to "Japan" as a concept to be assigned. More general is the observation that the indexing quality relies on the homogenity of the documents to be indexed. If they vary very much regarding content, style or even length, the quality of the indexing result is affected.

In the semantic web, heterogenous contents are the rule rather than the exception. A lot of different KOSs are widely used to describe all kind of resources[2], but for a real semantic interoperability, we have to be able to match the concepts between different KOSs or to quickly assign concepts of a KOS to a new resource. Albeit with inferior quality, such a bridging is needed to connect all kinds of resources, especially in the area of libraries, archives and museums.

With Maui [4], there exists a statistical indexer that incorporates a lot of NLP techniques, but there is to the best of our knowledge no free and open source implementation for a strictly lingustic indexer that can be used without any training data on arbitrary documents. Especially for the evaluation of "real" automatic indexers, such a simple implementation is useful. There are a lot of additional scenarios where this indexer can be used, be it for experimental services or whenever more sophisticated approaches are just not needed. And of course, as a reasonable baseline, more sophisticated approaches have to outperform it in the first place.

In this paper, we present LOHAI[3], a strictly linguistic indexer that incorporates mainly all these techniques that are state-of-the-art in information retrieval. The development of LOHAI is led by the following motivational thoughts:

Simplicity over quality: While every single step could be improved or replaced by a more sophisticated technique that is already developed and published somewhere, we tried to develop everything as simple as possible. Everything should be easy to use, easy to understand and easy to improve if needed.

Knowledge-poor and without any training: To be usable for arbitrary KOS and documents, the indexer can not rely on any *additional* knowledge sources, however, of course, the KOS itself can and will be used. The indexer must not employ a training step, as there are many settings where no preindexed documents are available and the creation of a training set would be too cumbersome for the user.

With these prerequisites in mind, we compose the indexer as a pipeline with several components, as illustrated in Figure 1 on page 3.

[2] E.g. by means of SKOS, `http://www.w3.org/2004/02/skos/`.

[3] LOHAI is pronounced like Low-High and means something like LOw HAnging Fruits Automatic Indexer, which gives a brief summary about the development process.

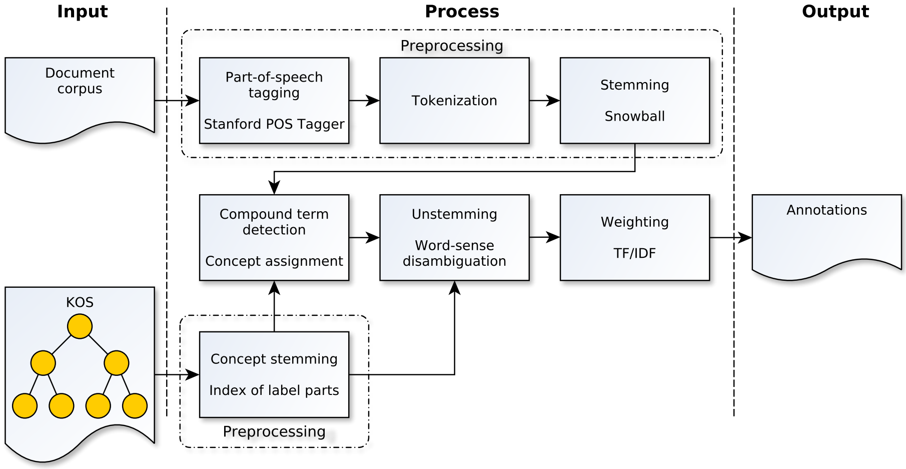


Fig. 1: The indexing pipeline

2 Preprocessing

The preprocessing consists of the several steps:

1. **Part-of-speech tagging:** We use the Stanford Log-linear Part-Of-Speech Tagger[4], as described by Toutanova et al. [10]. Part-of-speech (POS) tagging simply means the identification of nouns, verbs, adjectives and other word-types in a text. To avoid wrong concept assignments like the assignment of the concept "need" (as noun in the sense of requirements) whenever the verb "to need" is used, we only consider nouns (NN[5], NNP, NPS, NNS), adjectives (JJ, JJR, JJS), foreign words (FW) and unknown words (untagged).
2. **Tokenization:** The tokenization splits the text into single terms. The tokenization is performed together with the POS tagging. The result is a list of terms that are further investigated for proper concept assignments. In the tokenization step, there is also a cleaning of the terms included, where everything is truncated that is not a letter, a hyphen or a space. Note, that numbers are truncated, too, as they usually contain no meaning and are generally highly ambiguous. In some domains, this would not be desired, consider for example history or chemistry.
3. **Stemming:** Finally, the single terms are stemmed, i.e. they are reduced to their stem. That way, same terms can be matched, even, if they use different grammatical forms, like "banks" and "bank". We use the English (Porter2) stemming algorithm for the Snowball stemmer [6].
4. **KOS preparation:** This is only performed once per KOS. All concept labels are stemmed by means of the same stemmer that is employed on the

[4] `http://nlp.stanford.edu/software/tagger.shtml`

[5] Tag definitions according to the Penn Treebank tag set [8].

document texts. An index is created that maps the single stems to the corresponsing concepts. Additionally an index of stemmed label parts is created that is used for the identification of compound terms – for instance, “insurance market” would be stemmed to “insur market”, mapping to the corresponding concept, additionally, both stem parts are indexed and mapped to the stemmed compound term.

The preprocessing uses only freely available standard approaches. The POS tagging and the stemming are language dependent; both algorithms employed are implemented for various languages, including English and German. We assume that both the KOS and the documents are in the same language and that only one language is used in the document, so that the appropriate implementations can be used. If the KOS is multilingual and the documents use different language, an additional language detection step has to be employed.

After the preprocessing steps, the actual concept assignment and weighting takes place, as described in the next section.

3 Concept assignment with compound term detection

The general assignment strategy is a pure string based matching: If a stem that belongs to a concept in the KOS appears in the stems extracted from the text, the concept is assigned. In this step, we consider every concept a matching concept that contains a label that has the same stem or contains the same stem in the case of an compound term. Under this assumption, we have to deal with three possibilities that consequently would lead to wrong assignments:

1. A stem could belong to several concepts, including compound term concepts, e.g. “insur” that belongs among others to “insurance” and “insurance market”.
2. A stem could belong to several concepts that have different labels with the same stem (overstemming), like “nationalism”, “nationality” and “nation”.
3. A stem could belong to several concepts that have the same labels with the same stem (homonyms), like “bank” (the financial institution) and “bank” (a raised portion of seabed or sloping ground along the edge of a stream, river, or lake).

Approaches to handle the latter two variants are described in the next section, the first variant is dealt with directly in the assignment phase: The basic assumption is that we want to assign the most specific concept, i.e. in the above example, we would like to assign “insurance market”, but not “market”.

We implement this as follows: Whenever a stem is recognized as a potential part of a compound term, the stem is temporarily stored in a list. When a stem is found that can not be part of a compound, the list is analyzed for contained concepts. In this step, the algorithm simply checks every chain of stems for every starting stem if it corresponds to a compound concept. The algorithm starts with the longest possible chain and stops if a compound is found, thus avoiding the assignment of additional concepts contained in the compound. With this approach, the algorithm has generally a linear runtime with respect to the words contained in the text. Only the parts that potentially contain compounds have to be further analysed with a runtime of $O(n^2)$ with n denoting the number of words within such a part.

4 Unstemming and word-sense disambiguation

Whenever one or more stems could be assigned to more than one concept, we would like to identify a single concept as the correct one in the given context. This task is generally denoted as word-sense disambiguation (WSD). We use two different approaches for WSD, the first being a specific check that tackles the problem of overstemming mentioned above. If this step is not able to disambiguate the potential concepts, the actual WSD is performed.

Unstemming. Overstemming – the reduction of two different terms to the same stem – leads to ambiguous stems that have to be disambiguated during indexing. Consequently, we first unstem the stem, i.e. we go back to the original, unstemmed form of the term, as found in the text. If the unstemmed term corresponds directly to an unstemmed label of a concept, we assign this concept. If there is only one such concept, we finish the WSD step. Otherwise, we continue with the actual WSD, as described in the following.

KOS based word-sense disambiguation. Word-sense disambiguation is a broad field in the area of natural language processing. Leaving the technical issues of overstemming aside, it generally consists of the task to determine the correct sense of a word that appears in a particular context. The variety of possible senses is often based on some background-knowledge, like a thesaurus or other types of KOS. As Manning and Schuetze [2, pp. 229 f.] pointed out, this can be unsatisfactory from a scientific or philosophical point of view, as the definitions in the background knowledge are often quite arbitrary and possibly not sufficient to describe the actual sense of a word in a given context. However, our goal is not the perfect assignment of a sense to a word, our goal is the assignment of the best fitting concept in the KOS.

WSD approaches can be divided in supervised and unsupervised approaches, additionally in knowledge-rich and knowledge-poor approaches [5]. In our setting, we clearly need an approach that is unsupervised – as it has to work without any previously tagged texts – and knowledge-rich – as we clearly have a KOS at hand and of course want to use it to improve the disambiguation quality.

A supervised, knowledge-rich approach would be the adaptive thesaurus-based disambiguation, as presented by Yarowsky [11], where a Bayes classifier is trained on a large document corpus and thus probabilities for the occurrence of specific words in the context of a specific sense are determined.

Yarowsky [13] also proposed an (almost) unsupervised approach that makes use of two assumptions:

One sense per collocation. We assume that words collocated with the word to be disambiguated are unique for the correct sense and would not be collocated with the word for other senses. This basically is the rationale to use the context of a word – usually a window of words before and after the word in question – for disambiguation.

One sense per discourse. We assume that only one sense for a given word is used throughout a whole document. With this assumption, we can make use of any occurrence of the word in the text and thus get a more stable disambiguation result.

Both assumptions have been examined and verified [1, 12]. However, as Yarowsky's approach is not completely unsupervised – a small set of pretagged senses is needed as seed – we only make use of the two assumptions, but use a much simpler approach: Word-sense disambiguation based on a Jaccard comparison (cf. Ramakrishnan et al. [7]).

For this comparison, we define two sets of words: W as the context of an occurrence of the ambiguous word w, and C as the context of a candidate concept c, respectively. We then compute the Jaccard measure as follows:

$$\text{Jaccard}(W, C) = \frac{|W \cap C|}{|W \cup C|} \tag{1}$$

Based on the assumption "One sense per discourse", we assign each occurrence of w the concept c that was mostly assigned in the document, i.e. got in most cases the highest Jaccard value. If only abstracts are available for indexing, this procedure can be further simplyfied by just assuming the whole abstract as the context for each occurrence of w, which leads to the direct assignment of the concept c with the highest Jaccard value.

As context of an ambiguous word w, we either define a window of 100 words before and after the word or just use the whole document in case of short texts, like abstracts. The context of a concept c is defined as the union of all labels of the concept, its direct child concepts, its parent concepts and the direct children of the parent concepts, i.e. its siblings. Other definitions are of course possible, for example the weighting of words and labels depending on the distance to the word or concept, but for our purpose as part of a simple baseline indexer, our approach is sufficient.

5 Weighting

The last step in the indexing pipeline is the weighting of the assigned concepts. As the baseline indexer so far assigns every concept that can be identified by an occurring word, the weighting of these concepts is vitally important to determine which concepts are important and descriptive for the given text and which concepts are only marginally touched. It is also desirable to give concepts a higher weight when they are not used in the majority of documents, because these concepts usually only denote common terms and are not important for the indexing result.

The common approach for this kind of weighting is *tf-idf*, which is based on the term frequency $tf_{c,d}$ of a term (in our case concept c) in a given document d and on the document frequency df_c of a concept c, i.e. the number of documents, where the concept appears:

$$w(c,d) = tf_{c,d} \cdot \log \frac{D}{df_c} \tag{2}$$

D denotes the total number of documents in the indexed corpus. The last term is called inverse document frequency (*idf*), as the overall weight becomes smaller the higher df_c is.

6 Results

To show the weaknesses and strengths of LOHAI, we investigate some of the indexing results. For our experiments, we used the German STW Thesaurus for Economics[6]. A concept in the STW consists of preferred and alternative labels, both in English and in German. For example, there is the concept "Migration theory" with alternative labels "Economics of migration" and "Theory of migration".

Figure 2 shows an example abstract that we indexed. LOHAI produces the output as shown in Figure 3. Additionally, we listed the intellectually assigned concepts by a librarian. It can easily be seen that the characteristics of the results are quite different. But if one takes the weighting into account, it can be seen that there are no wrong assignments with a weight above 0.3. Below that threshold, there are especially common terms that form a concept in the thesaurus and that are either not helpful or wrongly assigned, as "Exchange". "Government", for example, seems to be correct, but is rather a coincidence, as it is assigned due to the verb "govern" in the text – an indication for a mistake during the POS tagging. On the other hand, the very abstract concepts that are assigned by the librarian (besides "Theory") are not found by LOHAI, as the terms do not directly appear in the text in some form.

All in all, the results are very promising, even with a relatively simple approach like ours. Most assignments are correct, even if a human indexer would not assign all of them. The indexing quality correlates with the employed weighting, especially assignments with lower rank often contain more common concepts that sometimes are just wrong. A lot of these mistakes could be avoided if the thesaurus would be more precise about homonyms and would provide additional information to disambiguate them, when necessary. The indexer could be further improved, e.g. common concepts should not be assigned, if more specific concepts down the tree are found in the text (Like "Law" and "Contract Law" above). On the other hand, we wanted to keep it simple. Such adaptions and

[6] Standard Thesaurus Wirtschaft, `http://zbw.eu/stw/versions/latest/about.en.html`. The thesaurus is published and maintained by the German National Library of Economics (Deutsche Zentralbibliothek für Wirtschaftswissenschaften, ZBW)

Title	Contractarianism: Wistful Thinking
Authors	Hardin, Russell
Abstract	The contract metaphor in political and moral theory is misguided. It is a poor metaphor both descriptively and normatively, but here I address its normative problems. Normatively, contractarianism is supposed to give justifications for political institutions and for moral rules, just as contracting in the law is supposed to give justification for claims of obligation based on consent or agreement. This metaphorical association fails for several reasons. First, actual contracts generally govern prisoner's dilemma, or exchange, relations; the so-called social contract governs these and more diverse interactions as well. Second, agreement, which is the moral basis of contractarianism, is not right-making per se. Third, a contract in law gives information on what are the interests of the parties; a hypothetical social contract requires such knowledge, it does not reveal it. Hence, much of contemporary contractarian theory is perversely rationalist at its base because it requires prior, rational derivation of interests or other values. Finally, contractarian moral theory has the further disadvantage that, unlike contract in the law, its agreements cannot be connected to relevant motivations to abide by them.
Journal	Constitutional Political Economy, 1 (2) 1990: 35-52

Fig. 2: Example of a document abstract used for annotation

Constitutional economics
Influence of government
Ethics
Theory

(a) Intellectual indexing

Contract Law (1.21)
Contract (0.76)
Social contract (0.64)
Law (0.51)
Politics (0.37)
Prisoner's dilemma (0.34)
Theory (0.32)
Rationalism (0.24)
Association (0.23)
Exchange (0.20)
Knowledge (0.19)
Government (0.16)
Information (0.12)

(b) LOHAI

Fig. 3: Intellectual indexing vs. LOHAI

improvements are easy to implement, if they are needed. A quantitative evaluation of the results by comparing them to a manually created gold standard is still missing, but former experiments [3] with a comparable indexer showed that such results are nevertheless not very meaningful due to the very different characteristics of such an automatic approach and a trained librarian.

7 Conclusion

To the best of our knowledge, there is no free indexer available that does not require any data preparation step or the creation of some training data. With LOHAI, we developed such an indexer by just using the standard approaches in natural language processing and information retrieval for the single steps in the indexing pipeline. Each and every step could be improved by employing new and more sophisticated approaches, but we intentionally restricted ourselves to the well-understood approaches that are state of the art in information retrieval and natural language processing. All in all, the indexer consists of about 500 lines of code in Java, without the POS tagger and the Snowball stemmer. We showed that the indexer performs quite well and – maybe most important – does not behave like a black box, every assignment is easily understandable. We expect that the indexer would even be usable in serious indexing projects. LOHAI is already successfully employed in SEMTINEL[7], a thesaurus evaluation framework, where it is used to quickly process large document sets for a given thesaurus to determine its concept coverage. LOHAI is not a stupid indexer, it is a baseline indexer. It is free, open source and available at `https://github.com/kaiec/LOHAI`.

References

1. Gale, W.A., Church, K.W., Yarowsky, D.: One sense per discourse. In: Proceedings of the workshop on Speech and Natural Language. pp. 233–237. HLT '91, Association for Computational Linguistics, Stroudsburg, PA, USA (1992), `http://dx.doi.org/10.3115/1075527.1075579`
2. Manning, C.D., Schuetze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
3. Maynard, D., Dasiopoulou, S., Costache, S., Eckert, K., Stuckenschmidt, H., Dzbor, M., Handschuh, S.: D1.2.2.1.3 Benchmarking of annotation tools. Tech. rep., Knowledge Web Project (2007)
4. Medelyan, O.: Human-competitive automatic topic indexing. Ph.D. thesis, University of Waikato (2009)
5. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2), 1–69 (2009)
6. Porter, M.: Snowball: A language for stemming algorithms. Published online. (2001), `http://www.snowball.tartarus.org/texts/introduction.html`

[7] `http://www.semtinel.org`

7. Ramakrishnan, G., Prithviraj, B., Bhattacharyya, P.: A gloss-centered algorithm for disambiguation. In: SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004. ACM, New York, NY, USA (2004)
8. Santorini, B.: Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision). Tech. rep., University of Pennsylvania (1990), http://repository.upenn.edu/cis_reports/570/
9. Schwens, U., Wiechmann, B.: Netzpublikationen in der Deutschen Nationalbibliothek. Dialog mit Bibliotheken 1(1), 10–13 (2009)
10. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of HLT-NAACL 2003. pp. pp. 252–259 (2003)
11. Yarowsky, D.: Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In: Proceedings of the 14th conference on Computational linguistics - Volume 2. pp. 454–460. COLING '92, Association for Computational Linguistics, Stroudsburg, PA, USA (1992), http://dx.doi.org/10.3115/992133.992140
12. Yarowsky, D.: One sense per collocation. In: Proceedings of the workshop on Human Language Technology. pp. 266–271. HLT '93, Association for Computational Linguistics, Stroudsburg, PA, USA (1993), http://dx.doi.org/10.3115/1075671.1075731
13. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics. pp. 189–196. ACL '95, Association for Computational Linguistics, Stroudsburg, PA, USA (1995), http://dx.doi.org/10.3115/981658.981684

A Security Contextualisation Framework for Digital Long-Term Preservation

Kun Qian[1], Maik Schott[1], Christian Kraetzer[1], Matthias Hemmje[2], Holger Brocks[2], Jana Dittmann[1]

[1] Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Germany
{kun.qian,mschott,kraetzer,jana.dittmann}@iti.cs.uni-magdeburg.de
[2] Faculty of Mathematics and Computer Science, University of Hagen, Germany
{matthias.hemmje,holger.brocks}@fernuni-hagen.de

Abstract. Nowadays a growing amount of information not only exists in digital form but was actually born-digital. Digital long-term preservation becomes continuously important and is tackled by several international and national projects like the US National Digital Information Infrastructure and Preservation Program or the EU FP7 SHAMAN Integrated Project. The very essence of long-term preservation is the preserved data, which in turn requires an appropriate security model, which is so far often neglected in the preservation community. To address this problem, we extend the security relevant parts of the Open Archival Information System (OAIS) standard, in which security aspects are underspecified, by a conceptual framework for hierarchical security policy development based on given use-cases for a long-term archival system. The corresponding policies are then distributed and implemented by applying an iterative procedure to turn them into rules before these are then finally enforced. In this paper we describe how to construct a corresponding context model and derive such policies using the iterative approach to assure the system and data security.

Keywords: context model, digital archive, security policies, system security

1 Introduction and motivation

In this paper we perform security-oriented context modelling as well as policy generation, implementation and enforcement focused on the security of a digital long-term preservation environment which currently focuses on archiving texts (i.e. PDF files) and digitised books (i.e. TIFF files). This context model is based on the established OAIS ISO standard [1] as well as our previous work on digital long-term archival system security. In [2] we describe a use-case-centric approach of deriving operations, actions, objects, rights and roles from user-cases and how to employ these for usage within a security model, e.g. an extended version of the Clark-Wilson model [3] including a syntactic-semantic integrity and authenticity verification approach. This extended Clark-Wilson model is in [4] combined with an extended Information Lifecycle Model [5] developed within the EU FP7 SHAMAN integrated project [21] to form a secure preservation framework for images and describing in detail the integrity and authenticity verification processes.

The work described in this paper aims at the development of a concept for implementing and managing security for digital long-term preservation environments of all kinds. The main instrument we foresee for this is the usage of policies, which of course provides the following two challenges for this paper: First, to define a suitable security-oriented context model for archival systems to act as basis for the policy generation. Second, the (security) policies themselves have to be derived from the context model. As a basis to address these two challenges we take use-cases based on the OAIS standard [1] for digital archival systems.

The main scientific contribution of this paper is the conception of a contextualisation framework for context model and policy generation as well as policy implementation and enforcement in the scope of multimedia archival systems and data security.

This paper is structured as follows: Firstly, in section 2 we present the state-of-the-art in methodologies for context modelling, policy generation, implementation and enforcement in security relevant contexts. Then in section 3 we present our concept for contextualisation and policy-based security realisation. At last in section 4 we finally conclude and summarise this work.

2 State-of-the-art

This section introduces the state-of-the-art in methodologies for context modelling, policy generation, implementation and enforcement in security relevant contexts. In the scope of IT security, a good context model reflects the characteristics, the intended application scenarios as well as corresponding threats for a system and allows the design and implementation of policy controlled security mechanisms that enforce the security aspects that are required to protect the application scenarios against the threats.

2.1 Methodologies for context modelling in IT security

Context modelling, differing from other modelling methods, not only describes the entities involved in a system but also explains how the entities are related with each other by revealing their causality and relationships. The design of a context model can either start from the very basic knowledge and be progressed by gradually adding necessary information to achieve proper complexity (“bottom-up”) or start from the vivid representation of the physical world and be progressed by gradually removing redundancy to achieve the proper simplicity (“top-down”). Thus a well-designed context model is at the same time a well balanced compromise between complexity and simplicity, always being sophisticated enough to offer all the necessary details yet still straightforward enough to be understood and applied.

In the field of IT security context modelling plays an essential role in various aspects. To meet the requirement of *confidentiality*, context modelling is for example used to implement various access control policies. For example, Bhatti et al. developed their model for web-services in [6], based on an extended, trust-enhanced version of Role Based Access Control (RBAC) framework that incorporates context-based access control. To recognise better the broader context in which security requests arise, the Task Based Access Control (TBAC) extends the traditional subject/object based access control models by including domains that contain task-based contextual information [7]. To provide security for computing infrastructures in which access decisions may depend on the context. Covington et al. developed a context-aware access control by extending RBAC with the notion of environment roles [8]. For the aspect of *authenticity*, various models have been proposed. In the scope of sensor forensics, Fridrich describes in [20] a simplified sensor output model, which contains the basic elements of the process of digital cameras acquiring images, and then applied the model to derive a maximum likelihood estimator for the sensor fingerprint, which can be used to identify digital cameras. In [9] we propose a context model for microphone recording by describing the involved signal processing pipeline to reveal the influential factors that might be used as characteristics of different microphone to contribute to microphone authentications. More often, when a context model is developed aiming at assisting the construction of an information system, multiple aspects of security issues need to be covered. For instance, the policy model for clinical information systems developed by Anderson in [10] focuses not only on *confidentiality* and *availability* but also *integrity*. The context model described in this paper for the application scenario of a secure digital long-term preservation archive system is explained in detail in section 3. It takes not only *confidentiality*, *authenticity*, *integrity* and *availability* but also *non-repudiation* in consideration.

Currently, there exist three most prominent context modelling approaches [11]: *Object-role based context modelling* originates from attempts to create sufficiently formal models of context to support query processing and reasoning, as well as to provide modelling constructs suitable for use in software engineering tasks such as analysis and design. This approach is generally not applicable for hierarchical structured modelling. *Spatial context modelling* focuses on location information. It is well suited for context-aware applications that are mainly location-based, like many mobile information systems. *Ontology-based context modelling* exploits the representation and reasoning power to describe complex context data that cannot be described by simple languages [12]. It provides formal semantics to context data and thus makes it available to check for consistency of the set of relationships describing a context scenario as well as to recognise that a particular set of instances of basic context data and their relationships actually reveals the presence of a more abstract context characterisation. Compared to simpler approaches, it provides clear advantages in terms of expressiveness and interoperability, which is the reason we use ontology-based context modelling in our framework conception. In our concept the ontology describes the entities in the security system as well as the relationships among the entities, both of which are described by digital long-term preservation use-cases taken from the SHAMAN research project.

Some evaluation criteria on the performance of context models have been proposed in literature. For example, Strang et al. point out that the demands for context modelling include distributed composition, partial validation, richness and quality of information, (in-) completeness and ambiguity as well as level of formality and applicability to existing environments [13]. However these evaluating aspects are rather based on the requirements of context modelling for ubiquitous computing, thus suitable metrics for context models in the scope of IT security are still to be developed. This point is not addressed here but reserved for our future work.

2.2 Methodologies for security policy generation and enforcement

Context models describe entities as well as the relationships among the entities in a system. Good policies, as the systems governing mechanisms, are the foundation of well-secured systems. Simply speaking, security policies define what in the system should be protected [14] to meet different aspects of security requirement depending on the application scenario. Baskerville's approach from [13] can be considered as a functional hierarchy of policies, using a three level division: meta-policies are "policies about policies", which declare plans for creating and maintaining security policies; high-level policies are security policies which are high-level overall plans embracing the general security goals and acceptable procedures; low-level policies are defined information security methods of action that are selected from among alternatives and applied based on given conditions that guide and determine present and future information security decisions. This functional hierarchy increases in granularity from the abstract meta-policies to specific detailed policies, which may be so concrete that they directly demand or prohibit certain implementations or mechanisms. An issue is, if abstract policies are made more specific in a parent-child interactive relationship, this will also refer to the system or, analogously the other way around, distinct parts of the system get their own low-level child policy which is a refinement of a high-level parent policy for this very part. As such, for complex system there may be large numbers of low-level child policies, and many of them may originate from a single parent high-level policy. Thus management of these can become quite complicated as changes of a policy regarding only a special system module can either only be made at a high level, which would require the revalidation of a vast amount of its child policies for every policy referring to this module. Therefore to solve this issue, in extension to Baskerville's scheme [14], we propose the introduction an additional hierarchy level between high-level and low-level policies. This new level, called mid-level policies, is intended to encompass policies that only refer to such larger system modules.

Besides the policy hierarchy considerations, for this paper we also adapt a policy life-cycle model from Baskerville et al. [14] for security focussed policies. The adapted life-cycle contains for each policy the following phases:

Specification of policy requirements: The identification and classification of security objects and subjects are two essential requirements that have to be encompassed by context modelling prior to the policy design and implementation, as the meta-policies should ensure that these requirements become primary features of the security policies. Security objects are the security relevant assets of the system, and security subjects refers to the different entities that have a relevant security connection to the objects. In the context model describing the objects and subjects, also the connections between these (e.g. access levels and types) have to be specified.

Policy design processes: In general, some form of meta-policies should specify the process by which the lower-level policies are generated and enforced. For a complex system, the usage of a hierarchy of policies ensures the required scalability. The granularity of the different security policy levels in this hierarchy should be specified in the design. As mentioned above, we use in this paper a hierarchical policy approach with meta-policies, high-level, mid-level and low-level policies. The policy design process also includes decisions on policy expression languages and policy distribution as well as enforcement. Some policies should be enforced technically with computer technology (i.e. using access control software), some policies should be enforced organisationally, while some other policies should be enforced using personnel-focused mechanisms (like training of users or raising security awareness). Furthermore, the design of policies enforced technically should also consider the intended expression, distribution and enforcement standards (see the remarks below).

Policy implementation: How the policies are to be implemented based on expression languages and standards should also be determined and specified using meta-policies. The implementation also includes policy testing. Here functional evaluations as well as investigations on potential policy conflicts have to be performed. Nevertheless there is a usual problem that the implementation encounters: the policies are expressed in a natural language and thus too complex. In our concept this problem is solved by applying a manual and iterative procedure to turn them to enforceable rules, which are defined as formalised atomic descriptions of specific actions. More details are offered in section 3.2.

Policy enforcement: Boyle et al. developed the Common Open Policy Service (COPS) standard [15], which serves well for typical policy-based systems such as Authentication, Authorisation and Accounting (AAA) systems [16]. Using COPS the policies can be enforced via a three-tier-model: Policies are stored in the Policy Repository (PR), which could be a database, a flat file, an administrative server, or a directory server [17]. A Policy Definition Point (PDP) retrieves the policies from PR, parses and evaluates them and sends necessary commands to policy targets [18], while a Policy Enforcement Point (PEP) communicates directly with the policy targets and gives instructions of performing the policy actions following the received commands [17]. For communication between PDP and PEP, a query and response protocol is developed for exchanging policy information and decisions between them [17]. It is designed to operate reliably and in real time with minimal overhead, thus it provides a dedicated QoS controller for the PEP. Additionally, when necessary, Local Policy Decision Point (LPDP) can be defined between PDP and PEPs. In this case the PEPs take policy decisions from the LPDP for their domain, while the PDP remains the authoritative decision point at all times. Parallel to the enforcement an auditing of the system has to be performed where some monitoring mechanism should detect any failed enforcement attempt or policy conflict. In this enforcement phase also the execution of replacement or termination of policies is performed.

3 Design of our contextualisation framework

Based on the state-of-the-art presented in section 2, here we describe our framework for contextualisation of security for digital long-term preservation. This framework consists of four major functional blocks: *context modelling*, *policy generation hierarchy* with its different stages, *Information Package (IP) processing* and *control*. The *context modelling* block consists of two distinct parts: global (system-wide) and local context modelling. The *policy generation hierarchy* is a hierarchy of stages beginning at the top with the generation of system-wide global policies and ending in the deviation and invocation of rules for IP processing operations. The *IP processing* itself identifies the IPs (or related system data) to be processed and applies the rules as sequences of atomic data processing operations. The *control* block controls the *context modelling* and the *policy generation hierarchy* and acts as a central policy repository as well as a central audit service for the overall system.

Within the following sections we describe these functional blocks in more detail and show how to model context, generate and implement as well as enforce security policies based on use-cases from a data intensive, complex, security-oriented data processing system like an archive for digital long-term preservation.

3.1 Context modelling for complex, security-oriented data processing systems

A complex data processing system usually contains multiple processing entities with different types of relationships among the entities. Therefore the "top-down" modelling approach is not suitable, as achieving a complete and vivid representation as a starting point in context modelling is not feasible under these circumstances. Instead, it is more appropriate to first extract typical tasks (workflows) from use-cases in such systems and then gradually extend these into a fully developed context model. As ontology based context modelling has its speciality in organising complex structured context data, it is reasonable to apply it in the construction of the model for such systems. The resulting ontology describes the entities in the system as well as their relationships. The latter are expressed here in the form of policies and rules.

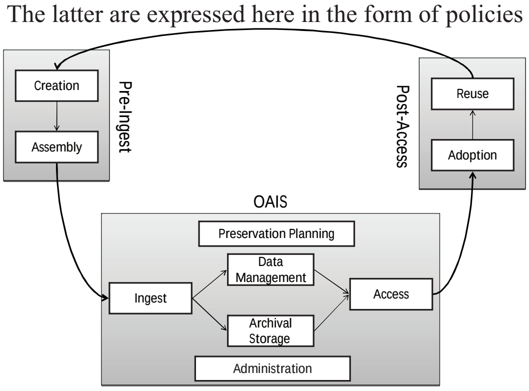


Figure 1. Phases and processes in the Information Lifecycle Model (based on [5])

In digital long-term preservation the basis for context modelling is the OAIS standard [1]. Its functional model describes several processes of an archival system, their tasks and their relationships as well as data items – thus providing a general context of systems for this application scenario (see Figure 1). In these processes typical use-cases are grouped. In the *ingest* data objects that should be preserved are received from a producer and converted into the archives data format. The *archival storage* stores and manages these data objects inside the archival system. The *data management* provides services for the discovery, access to the metadata, and maintaining the referential integrity between data objects. The *administration* process is responsible for the operation of the archival storage, procuring and installing new hardware and software and the organisational enforcement of the policies and standards. The *preservation planning* ensures that the archival storage can fulfil its requirements by observing the technical state-of-the-art and legal requirements and adapting its policies with this regard. *Access*, as the last major process, provides services for consumers to locate and retrieve data objects or information about them.

Within the SHAMAN project Brocks et al. [5] extended the OAIS model by introducing an extended *Information Lifecycle Model*. In this the aforementioned processes from ingest to access are seen as phases of the lifetime of a data object. The information lifecycle model extends this by including the objects "life" before and after its management within an archival. The phase before a digital object enters an archival system is called "Pre-Ingest". This is further divided into the processes of the actual *creation* of the data later to be ingested and its *assembly* into a package supported by the archive. The phase after a digital object leaves an archival system is called "Post-Access" and is also divided into two processes: *adoption* where the received data is unpacked, examined, transformed, displayed or in short all tasks that are needed for repurposing the content and *reuse* where the content is actually exploited. Reuse may also include the re-ingest of this object or a derivation thereof into an archival system, leading to a real life-cycle as shown in Figure 1. Such connection of reuse and creation is especially the case for collaborative environments.

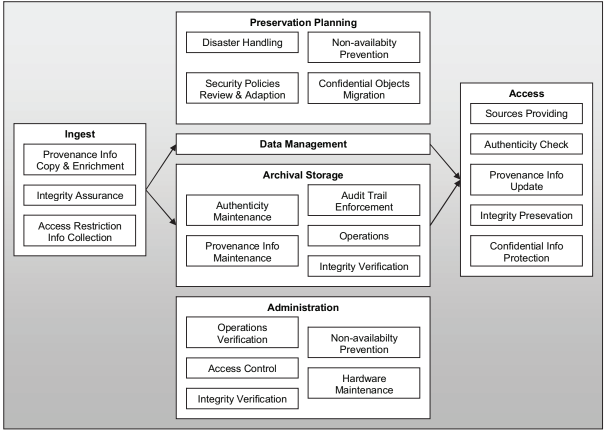


Figure 2. Required extension of the OAIS processes from a security point of view

Within this paper the considerations are limited on the central phase of the Information Lifecycle Model, the archival phases described by OAIS, and security considerations. As the original OAIS ISO standard is lacking detailed information about security requirements, it needs to be enhanced in this regard within this paper. Thus we analyse the archival use-cases provided by the SHAMAN research project and extract the tasks which would have to be considered in addition to the already existing OAIS functional entities. Thereby these extensions, which may not be separate entities but can be incorporated into existing ones, provide better context representation in terms of security. Figure 2 shows just these new tasks focused on risk mitigation for the OAIS processes, whereas the tasks and functional entities of the original OAIS model are omitted in this figure for the sake of clarity. The interested reader may refer to the OAIS documentation [1] for details on these original tasks and functional entities.

The global context model visualised on different levels of detail in Figure 1 (Information Lifecycle Model) and Figure 2 (detailed overview over the security tasks in the OAIS processes) is then used as starting point for policy generation described in section 3.2. Each process (Ingest, Access, etc.) has its own entity taking responsibility of the local context modelling as well as the generation, distribution, implementation and enforcement of policies within its own domain (scope of the process). All entities within this domain can act as the enforcement points for policies.

3.2 Use-case-driven policy generation, implementation and enforcement

Based on the assumption that a "top-down" modelling approach is not suitable, which is explained in section 3.1, it is reasonable to derive policies for complex data processing systems from use-cases. Furthermore, to more vividly represent the complex relationships among the entities in such systems and to implement means of governance or orchestration a hierarchical organisation of the policies is applied.

As mentioned in section 2.2, in this paper we use the three-level approach from Baskerville et al. [14] and enhance it to a four-level policy hierarchy.

Our proposed policy generation starts on a global system level with the most abstract types of policies – meta-policies and high-level policies. The first makes statements about other policies and the second about general security goals and acceptable procedures on a system-wide perspective. Thus they can either be derived from use-cases making policy assertions and from system use-cases, respectively, or come from the general understanding of the system or the application scenario.

Inspired by practice of defining optional LPDPs in the COPS standard [15], we decide to add another layer of mid-level policies in the previous three-layer policy model introduced in [14], for better handling of larger complex system modules. This reflects the fact that many use-cases do not make assertions about the system as a whole, but about certain functionalities. Such use-cases are restricted to larger system modules (in our case equivalent to the OAIS processes) and their domain of functional entities. In the policy generation hierarchy these mid-level policies on the one hand serve as a process-based filter for the use-cases of which a system may have a large amount of, and on the other hand they serve to verify if the high-level policies themselves make sense by not contradicting the existing use-cases (i.e. verify the consistency between global and local context modelling).

The mid-level policies are used to act as the basis for the generation of low-level policies, which provide sufficient information what should be implemented as a rule in the enforcing. In the ideal case these low-level policies should be precise enough to directly derive rules in a formalised language from them.

For the sake of clarity and for the sake of the traceability of the policies origins, a policy derived from a higher policy should have an identifier indicating its parent policies. If high-level policies have an identifier of the format P*x* (with *x* being an unique identifier) their children mid-level policies should have an identifier that includes their parent's identifier (e.g. P*x*-*y*). As policies need to be updated or even removed at certain times, this form of traceability eases the browsing of the hierarchical tree structure of the policies that would be required in these cases.

For highly complex systems there arise some issues for the implementation and enforcement of policies: First, when introducing a new policy into such systems, there could be multiple possible methods to implement it, thus it requires specific analyses (e.g. complexity-based) to identify the optimal method. If the COPS standard for policy management were applied in this case, these considerations would have to be also extended to policy decisions on the selection of PEPs. Second, complex systems are with a high probability also heterogeneous, therefore considerations have to be included on the interoperability, distribution and orchestration of policies and policy descriptions (for instance how to interpret between possible different policy syntaxes used in different parts of a heterogeneous system). Third, due to the quantity and complexity of the policies, it is necessary to develop an assurance and auditing mechanism to make sure that all the policies are enforced properly.

The policies considered here are basically descriptions in natural language of what the preservation system does, which creates barriers for actual enforcement. Thus in our concept, the generated policies are implemented by applying a manual and iterative procedure which turns low-level policies into enforceable rules. The procedure is described as follows:

Create Rules: This turns low-level policies, which define what needs to be done, into rules, which define how the policy is enforced. It analyses the statement in the policy by utilising validation criteria that consist for the significant properties, format validation, organisational- and domain information. Then a sequence of steps is derived, describing specific actions. Each step should be as atomic as possible, ideally performing one action and also verifiable, so it can be considered as one abstract rule. Optionally a rule can comprise sub-rules if one of the steps is too complex to be described as a single action. Therefore the output here is a sequence of abstract rules.

Instantiate Rules: Abstract rules are not executable as they only describe actions in natural language. Therefore it is necessary to derive executable rules from abstract ones. Templates containing the grammar and syntax for rule-engines can be used by a rule instantiation tool to create realisations of the abstract rules. Such tool should also keep track of the realisation process so that it is possible to track from an executable rule back to the abstract rule and then back to the policy. Additionally, similar to Event-Condition-Action (ECA) rules which always have the form of *if…then…else*, the executable rules are formalised as Semantic Web Rule Language (SWRL) [19] rules embedded in the Web Ontology Language (OWL) context representation, thus each rule becomes an executable atomic data processing operation.

Validate Rules: Here it is ensured by validation that the instantiated executable rules are correct implementations of the policies. The functionality of the used validation tools would be defined by the validation criteria, which are the adherence to the global and local context models (developed in 3.1). After a rule passed the validation, it is deployed with records of its deployment time and intended deployment enforcement point in the production system and ready to be enforced.

Once the policies are implemented, i.e. turned into formalised and validated rules describing executable actions, it is easy to enforce them. The COPS standard can be adapted to fit this case. Instead of PR, a Rule Repository (RR) would be used to store the rules. Similar to PDP, Rule Decision Point (RDP) would retrieve the rules from the RR, parse and evaluate them, then send rule decisions to rule targets, which can be either devices or humans to perform the actions. Similar to PEPs, Rule Enforcement Points (REPs) make direct communication with the rule targets and give them instructions on performing the actions following the commands. Depending on the complexity of the system, Local Rule Decision Point (LRDP) can share the responsibility with RDP by feeding REP with detailed rule decisions, while RDP remains authoritative rule point at all times.

3.3 IP processing and Control

In the *IP processing* block a system entity (here equivalent to a rule target) enforces rules on IP from the archival system and/or system data (like search indexes, the user database, etc.). The result of the enforcement has to be communicated by the responsible REP to the central audit service. This central audit is part of the functionality of the *control* block. Besides this audit functionality there are also mechanisms for the storage of the policy tree (all policies are communicated to this repository during the construction of the *policy generation hierarchy*) as well as the policy conflict analysis and conflict resolve. The corresponding OAIS authority responsible for these operations would be the task “Security Policies Review & Adaption” in the process of “Preservation Planning” (see Figure 2). It should keep track of all the policies to ensure they operate properly, especially no policy from one phase conflicts with those from other ones, similar to the responsibility shouldered by policy decision points in the COPS standard.

3.4 Combination of the functional blocks of the framework

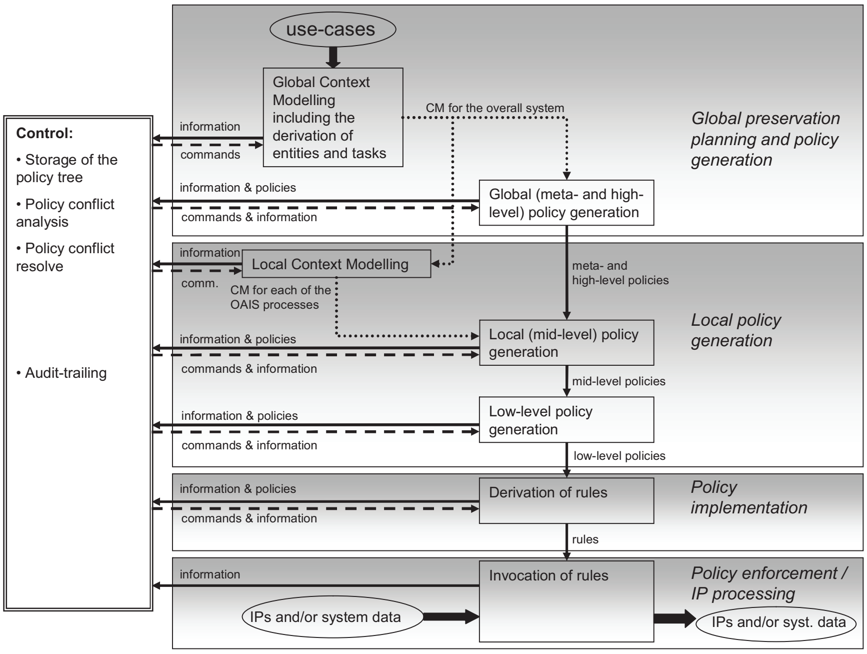


Figure 3. General overview over the security contextualisation framework

Figure 3 extends the low detail description of the contextualisation framework explained in introduction of section 3 by the data, information and control flows described for the four functional blocks in sections 3.1 to 3.3.

In Figure 3 especially the importance of the control block sticks out as a dominant factor. Each context modelling block, the different stages of the policy generation hierarchy and the IP processing communicate their actions to the control block. This is on one hand done to audit all operations for purposes of transaction control and non-repudiation of transactions as. On the other hand this functional block also acts as central policy storage repository and performs policy conflict analysis and resolve.

4 Summary and conclusions

In this paper we have outlined a bottom-up context modelling approach which derives a hierarchical policy structure from given use-cases for a long-term archiving system. An existing concept from literature has been extended accordingly into a complete contextualisation framework to meet the (security) requirements of a digital long-term preservation system.

However, there exist several limitations for our approach that have to be addressed in future work: First, it is hard to evaluate whether the constructed context model (as basis for the policy generation process) is too specific with unnecessary redundancy or too abstract with lack of necessary details, as currently no metrics for the preciseness of such models exist. Furthermore, it is difficult to investigate how vividly the lower level policies reflect the intentions of the high level policies from which they derived, yet the biases (or even conflicts) between could lead to problems in their enforcement.

Acknowledgement

The work in this paper has been supported in part by the European Commission through the FP7 ICT Programme under Contract FP7-ICT-216736 SHAMAN. The information in this document is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

References

1. Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Standards, CCSDS 650.0-B-1, Blue Book (ISO 14721:2003), 2002.
2. M. Schott, C. Kraetzer, J. Dittmann, C. Vielhauer: Extending the Clark-Wilson Security Model for Digital Long-Term Preservation Use-cases, Proc. of Multimedia on Mobile Devices, 2010, SPIE Electronic Imaging Conference 7542, 2010.
3. D. D. Clark, D. R. Wilson: A Comparison of Commercial and Military Computer Security Policies, IEEE Symposium on Security and Privacy, 1987.
4. M. Schott, C. Kraetzer, N. Specht, J. Dittmann, C. Vielhauer, Ensuring Integrity and

Authenticity for Images in Digital Long-Term Preservation, Proc. of Optics, Photonics and Digital Technologies for Multimedia Applications, SPIE Photonics Europe, 2010.

5. H. Brocks, A. Kranstedt, G. Jäschke, M. Hemmje: Modeling Context for Digital Preservation, Studies in Computational Intelligence, vol. 260, pp. 197-226, 2010.
6. R. Bhatti, E. Bertino, A. Ghafoor, A Trust-based Context-Aware Access Control Model for Web-Services, Proc. of the IEEE International Conferences on Web Services, 2004.
7. W. Tolone, G. Ahn, T. Pai, S. Hong, Access Control in Collaborative Systems, ACM Computing Survays, Vol. 37, March 2005.
8. M. Covington, W. Long, S. Srinivasan, A. Dey, M. Ahamad, G. D. Abowd, Securing Context-Aware Applications Using Environment Roles, ACM Symposium on Access Control Model and Technology, ACM, Chantilly, VA, USA, 2011.
9. C. Kraetzer, K. Qian, M. Schott, J. Dittmann, A Context Model for Microphone Forensics and its Application in Evaluations, Proc. of Media Watermarking, Security and Forensics XIII, IS&T/SPIE Electronic Imaging Conference7880, San Francisco, CA, USA, 2011.
10. R. J. Anderson, A Security Polity Model for Clinical Information Systems, Proc. of IEEE Symposium on Security and Privacy, 1996.
11. C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, D. Riboni, A Survey of Context Modelling and Reasoning Techniques, Pervasive and Mobile Computing, Elsevier, 2010.
12. G. Klyne, F. Reynolds, C. Woodrow, H. Ohto, J. Hjelm, M. H. Butler, L. Tran, Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0, W3C Recommendations, W3C, 2004.
13. T. Strang, C. Linnhoff-Popien, A Context Modeling Survey, Proc. of the First International Workshop on Advanced Context Modelling, Reasoning and Management, in conjunction with UbiComp 2004, Nottingham, England, 2004.
14. R. Baskerville, M. Siponen, An Information Security Meta-policy for Emergent Organizations, Logistics Information Management, Volume 15, Number 5/6, 2002.
15. J. Boyle, R. Cohen, S. Herzog, R. Rajan, A. Sastry, The COPS (Common Open Policy Service) Protocol, RFC2748, 2000.
16. C. Rensing, M. Karsten, R. Stiller, AAA: A Survey and a Policy-Based Architecture and Framework, IEEE Network, Vol. 16, 2002.
17. R. Rajan, D. Verma, S. Kamat, E. Felstaine, S. Herzog, A Policy Framework for Integrated and Differentiated Services in the Internet, IEEE Network, Vol. 13, 1999.
18. K. Yang, A. Galis, C. Todd, Policy-Based Active Grid Management Architecture, Proc. of 10th IEEE International Conference on Networks, 2002.
19. I. Horrocks, P. F. Petal-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, SWRL: A Semantic Web Rule Language Combining OWL and RuleML, Member submission 21 May 2004, W3C, 2004.
20. J. Fridrich, Digital Image Forensic Using Sensor Noise, IEEE Signal Processing Magazine, vol. 26, no. 2, 2009.
21. SHAMAN website: http://www.shaman-ip.eu

DA-NRW: a distributed architecture for long-term preservation

Manfred Thaller manfred.thaller@uni-koeln.de, Sebastian Cuy sebastian.cuy@uni-koeln.de, Jens Peters jens.peters@uni-koeln.de, Daniel de Oliveira d.de-oliveira@uni-koeln.de, and Martin Fischer martin.fischer@uni-koeln.de

Historisch-Kulturwissenschaftliche Informationsverarbeitung, Universität zu Köln,
Albertus-Magnus-Platz, D-50923 Köln

Abstract. The government of the German state of North-Rhine Westphalia is considering the creation of a state-wide long-term repository for digital content from the cultural heritage domain, which at the same time will act as a pre-aggregator for the *Deutsche Digitale Bibliothek* and the Europeana. The following describes a software architecture that relies exclusively on existing open source software components to implement a distributed, self-validating repository, which also supports the notion of "executable contracts", allowing depositors a high degree of control over the methods applied to individual objects submitted for preservation and distribution.

1 Introduction: Background

North-Rhine Westphalia, as well as other political entities responsible for the cultural heritage in the public domain, faces the problem that - as of now - few, if any, workable solutions for the preservation of digital content exist. That is true for digital content created by projects within the field of retrospective digitization of cultural heritage, and it is even more true when we look at the safe-keeping of digital content created by public administration or arriving in the public domain through deposition in one of the deposit libraries of the state.

At the same time North-Rhine Westphalia is expected to support the creation of the Europeana, as one of many entities. As Germany has decided to channel its contribution to the Europeana through an intermediate layer, the *Deutsche Digitale Bibliothek*, the original metadata schemas of the content holding institutions have to be converted for both target systems. At the same time, few, if any, memory institutions would be willing to submit the very top quality of their digital holdings to a European (or any other) portal that allows the completely unrestricted use of that material. It is, therefore, necessary to convert the data submitted by the memory institutions to a form that can be distributed completely without restriction.

It was decided to attempt an integrated solution for both problems: a framework is to be developed under the name of *Digitales Archiv Nordrhein-Westfalen (Digital Archive North-Rhine Westphalia)* (DA-NRW) which would allow all memory institutions (archives, museums, libraries) of the state to submit their digital content to a state wide repository, which would follow the OAIS model [2] and specifically:

- Ingest the material into a long-term repository system, which allows for a technology watch, triggering migration if necessary, and other active methods.
- Perform automatic verification of the redundantly stored material between geographically distributed sub-repositories.
- Evaluate user-submitted contracts expressed in XML, describing in detail which of several options for storage as well as distribution to the public are to be provided for that object.
- Derive suitable representations of the administered material, keep them on a server which supports OAI-PMH (cf. [5]) and other protocols to make these representations available to various cultural heritage portals.

The system is to be based upon existing infrastructural institutions from different sectors: the *Hochschulbibliothekszentrum*, the Computing Center of the *Landschaftsverband Rheinland* and the Computing Center of the *Universität zu Köln*. The chair for Humanities Computer Science at the *Universität zu Köln* is responsible for design and implementation of a prototype.

In order to avoid performance and cost problems during the transfer from prototype to production system, and to create a scalable prototype in less than 18 months, the following decisions have been made:

- The system is built according to agile software development rules.
- Only Open Source components are being used.
- The prototype is expected to perform with 200 TB, being scalable without re-design by one order of magnitude, to 2 PB.

At the end of June 2011, after an initial preparatory phase and four months into the core development time of 14 months, a functionally complete pre-prototype is available.

2 Introduction: Overall Architecture

The three participating computing centers, referred to as the nodes of the DA-NRW, are to be understood as independent nodes of a network. The flow of data within each node is directed by an instance of a *content broker*, directing the flow of data from ingest into the archive, on the one hand, and that of derived copies of the data into a presentation area, on the other hand, where these data can be accessed by appropriate harvesters. For a diagram of the component structure of these *content brokers* see figure 1. The individual components will be described in the following sections of this paper.

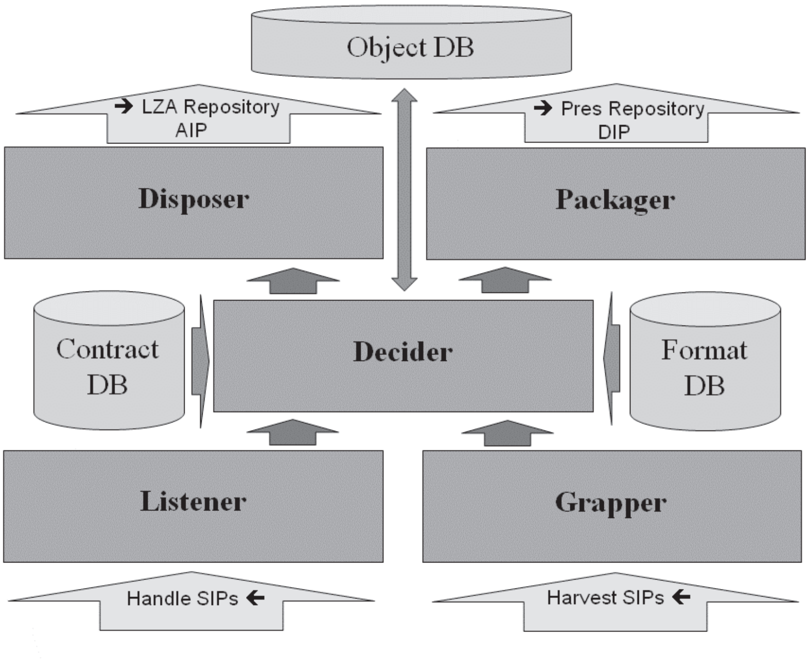


Fig. 1. Component structure of the *content broker*

The individual nodes are bound together by a synchronizer and deliver their data into a presentation component, which is separated from the actual long term preservation components by appropriate firewall techniques.

3 Ingestion methods

One key feature of systems providing long-term preservation is the delivery of digital objects from an institution to a preservation node. In our system this can be accomplished in two different ways.

The first one allows contractors to build *Submission Information Packages* (SIPs) [2] on their own. In this case, however, the structure of the SIPs has to be valid prior to ingestion into the archive. That means the SIPs have to contain structural metadata in a format supported by DA-NRW (e.g. METS[1]). If contractors decide to build their own SIPs, they are also responsible for creating checksums for the package contents in order for the *content broker* to be able to check for consistency.

The second possibility of building valid SIPs is to use the *DA-NRW SIP Builder*. This tool enables users to create SIPs in a very simple manner. In order

[1] cf. `http://www.loc.gov/standards/mets/`

to make the tool available for a wide audience, the *SIP Builder* is written in Java and therefore constitutes a platform-independent application. It provides a graphical user interface for comfortable usage. After choosing a destination path where the SIPs will be created, the user chooses which strategy to use for compiling the SIPs. On the one hand, one can choose a metadata XML file which describes the package structure. The tool then collects the files referenced in the XML. On the other hand, the tool is able to compile valid SIPs from directories taking into account folder structure and contained metadata files.

Another important aspect of the *SIP Builder* is the possibility of declaring contracts in a user-friendly way. Statements generated by the *SIP Builder* are serialized as a machine-readable contract in a PREMIS-based XML (see [1]) format that can subsequently be evaluated by the *content broker*.

4 Content broker

The central part of the architecture is called the *content broker*, a tool written in Java. This component is responsible for manipulating complete information packages in various ways. It does so by executing predefined chains which correspond to use cases such as ingest, retrieval or the migration of information packages. Each chain consists of atomic actions which in turn operate on the aforementioned information packages. Examples for actions are: the self-explanatory 'FormatConversionAction' that converts bitstreams and/or metadata into target formats, or the 'RegisterObjectAction' that registers an information package at the global object database. Administrators can define different chains for different tasks. Chains can be configured in an easily readable XML syntax.

Format conversion and identification are also implemented in a highly flexible manner in the overall design. As far as format identification is concerned, 3rd party software (such as DROID[2], JHOVE[3] or the Linux FILE-Command) can easily be plugged into the the workflow. Format conversion policies can also be configured from a set of XML files.

Migration happens along the same lines. Policies and corresponding conversion routines have to be defined in order to automatically retrieve and convert packages which are marked as containing deprecated formats. At this stage two aspects have to be stressed: first of all, there is the problem of 'marking' formats as deprecated. At present this is done manually, but for the future we plan to use an automatic approach by connecting the system to an automated obsolescence notification system, as currently discussed within some preservation infrastructure projects.

The second aspect refers to the selection of appropriate conversion routines. Here an administrator of a node, or an administrator of the whole DA-NRW system, is requested to choose which conversion routine delivers the best results in terms of quality for long-term preservation. That means it first has to be chosen which target format serves as a long-term preservation format. Once the

[2] cf. `http://droid.sourceforge.net`

[3] cf. `http://hul.harvard.edu/jhove`

format is chosen, the next decision will be which program to use with which parameters to achieve good results.

5 Presentation repository

The architecture of DA-NRW also includes a *presentation repository* that acts as a central service provider for different partnering institutions and interdisciplinary portals – such as Europeana, the *Deutsche Digitale Bibliothek* and the North Rhine-Westphalian portal developed at the HBZ during the course of this project. Also the *presentation repository* can serve as a data source for subject-specific repositories aggregating specialized collections. Finally, small institutional repositories can harvest the central repository in order to implement own applications for the presentation of their data. While doing this they can profit from the format conversions and normalizations that the packages undergo on their way through the services of the digital archive as a whole.

Contractors of the DA-NRW can define if and under which conditions an object will be available through the *presentation repository.* These conditions include restrictions on the quality of the presented material, such as resolution and bit rate, restrictions on the content, e.g. by allowing only specific parts of the data and metadata to be retrieved, as well as time-based statements in order to be able to represent "moving-walls" or the expiration of copyright.

Currently the *presentation repository* is based upon the Fedora Commons Repository Software and supports the dissemination of Dublin Core (DC) and Europeana Semantic Elements (ESE) [4] metadata for every object in the repository. These standards represent a common basis for the heterogeneous objects we have to deal with. However, we are planning to support richer metadata formats in the presentation of objects and are examining ways to make the data available as part of the ongoing efforts to support Open Linked Data.

6 Storage layer

Our basic approach in long-term preservation regarding storage is to synchronize the stored information across at least three different storage locations, technically and geographically independent, across the state of North Rhine-Westphalia. To accomplish this major goal, we decided to use the iRODS (Integrated Rule-Oriented Data System) Data Grid Software [7].

In order to test our system under realistic conditions and with real data at a relatively early stage of development, we chose an iterative approach for the design and realization our project. In terms of iRODS, we implemented the basic features related to the data storage part corresponding to the final stage of the archival workflow after the *content broker* actions have already taken place. So we first focused primarily on the storage capabilities of iRODS. In the upcoming iteration we plan to use iRODS, in particular its "Rule Engine" and its "Micro-services", more intensively in the entire workflow of the archival storage process as well as the ongoing data curation process in the years to come.

The consistency of each digital object will be ensured by physical checksum comparisons and by keeping the minimum number of replicas of each object on the desired nodes after AIPs (Archival Information Packages) being ”put” to the node. These use cases will be implemented using ”Rules”, statements executed on the data by the Rule Engine.

7 Future research

As mentioned in the introduction, the architecture described here is a pre-prototype version which was developed within four months. A ”pre-prototype” means that all major components exist and can be used. However, quite a few details are missing (e.g. the notification of an end user about the results of the ingest process). Furthermore, it means that major processes, which shall run automatically at the end of development, have to be started explicitly at this point.

In the near future we plan to replace a large part of the orchestration of individual services, which has now been rapidly prototyped in Java, by a stronger reliance on iRODS Micro-services. In other words: we plan to shift from just storing data in the iRODS-Grid at the final stage of our existing archival workflow-chain to a more iRODS-centric architecture by making the features of ”Rules” and ”Micro-services” do the major work. This will also ensure computational scalability. The leading design principle in our already developed components was to develop fine-grained actions which are only loosely coupled. These actions can now easily be replaced by or be incorporated into iRODS Micro-services. A lot of research has to be done on the second question, i.e. how Rules can help us build up ”policies” for archived content itself. A major part of our work in the next months will be the usage of iRODS Rules to execute policies on our stored objects.

We are currently also evaluating the use of PREMIS OWL [3] and triple stores for the representation of contracts in RDF (Resource Description Framework). This allows for easier extension of the contract format, reduces the mapping overhead between the XML format and the relational database, and simplifies the organization of machine-processable contracts. We are also investigating different RDF-based variants for wrapping package metadata. One approach might for example be the application of OAI-ORE as an alternative for METS as proposed in [6]. This would allow us to incorporate contract, format, structural and descriptive metadata into one unifying RDF model.

References

1. Brandt, O.: PREMIS. Nestor-Handbuch: eine kleine Enzyklopädie der digitalen Langzeitarchivierung pp. 60–62 (2007)
2. CCSDS - Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System. Blue book. Issue 1 (January 2002)

3. Coppens, Mannens, Evens, Hauttekeete, Van de Walle: Digital long-term preservation using a layered semantic metadata schema of premis 2.0. In: Cultural Hertiage on line. International Conference Florence, 15th-16th December 2009 (2009)
4. Europeana Foundation: Europeana Semantic Elements Specification, version 3.4 edn. (March 2011)
5. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S.: Open archives initiative - protocol for metadata harvesting - v.2.0 (Jun 2002)
6. McDonough, J.: Aligning mets with the oai-ore data model. In: Heath, F., Rice-Lively, M.L., Furuta, R. (eds.) JCDL. pp. 323–330. ACM (2009)
7. Rajasekar, Wan, Moore, Schroeder: iRODS Primer. Morgan Claypool Publishers (2010)

RDFa as a lightweight metadata interoperability layer between repository software and LOCKSS

Felix Ostrowski

Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft
Unter den Linden 6, 10099 Berlin, Germany
felix.ostrowski@hu-berlin.de
http://www.ibi.hu-berlin.de/

Abstract. Semantic Web and Linked Data standards have recently been gaining momentum in the library domain. It seems more likely than not that future systems used in library environments will make increasing use of these standards. This paper outlines possible usage scenarios of semantic metadata in the LOCKSS digital preservation software generally, and the possibilities for metadata interoperability between repository software and the LOCKSS system based on the RDFa standard specifically.

Keywords: digital preservation, repository software, LOCKSS, Semantic Web, metadata, RDFa, interoperability

1 Introduction

Digital libraries and digital archives are closely related. In fact, it is a common misunderstanding that digital library software such as repositories *are* digital archives. If a document is ingested into a repository, it is often thought to be archived. That is not true; it might even be considered a dangerous misconception. At the very bottom of an archival system lies the long-term preservation of bitstreams. Repository software on the other hand are designed for ingest and medium-term access. Of course repositories store bitstreams, too, but this storage usually lives in a typical web server environment. While that environment is hopefully integrated into some sort of backup routine, the storage layer of any common repository software can not be considered to fulfil long term archiving needs in any way.

Commercial long-term preservation systems are built as a whole from the ground up[1] and thus consist of an ingest mechanism that closely resembles that of a repository system and a management layer to control the archive and perform tasks such as format migration. Most interestingly, bitstream preservation is considered a solved problem[12] and is not addressed with the appropriate

[1] Ex Libris' Rosetta[2] and kopal[1] for example focus heavily on ingest, access and management tasks such as format migration.

attention in these systems. Considering the facts that (1) an open source solution for bitstream preservation exists in form of LOCKSS[9], (2) high quality open source repository software to publish documents to the web is available and (3) bitstreams get into a LOCKSS-Network by means of web harvesting, the only missing part to build an OAIS compliant archival system out of these components is a management component[2].

This paper will first briefly discuss the advantages of using semantic metadata in shape of RDF in a LOCKSS environment and then sketch out a possible interoperability of repository software and LOCKSS based on RDFa. It should thus be considered a rather theoretical outline of an archival solution that is based on the loose coupling of existing software solutions by means of semantic metadata. The availability of this data can smooth the way how to add a management component to the environment by facilitating data exchange and integration.

2 LOCKSS and semantic metadata

There are at least three different aspects of how semantic metadata can be of advantage in a LOCKSS environment. Among them are the data management within individual LOCKSS nodes, the integration of metadata present in a LOCKSS network and the interaction of publication platforms such as repositories and the LOCKSS crawler. These three perspectives are briefly discussed next, followed by a more in-depth view on the latter.

Firstly, a generic data model such as RDF allows for a generic database in LOCKSS nodes that can store arbitrary metadata. LOCKSS is at its core a web crawling system that consists of several independent gathering nodes that communicate in order to ensure the integrity of the harvested content. With regards to metadata of any kind, LOCKSS natively used to consider only information at HTTP level, such as content-type and length. It has been extended to support additional metadata by making use of a metadata extraction framework, though. This framework is capable of extracting metadata by analyzing the crawled content and extracting metadata from there. The data is then available within the system and can be further processed. The current version of LOCKSS only includes a relational database with a fixed schema to store descriptive metadata. In context of the LuKII project[5], the need to add a database for technical metadata arose. While the modular architecture of LOCKSS allowed for a relatively easy implementation of an additional metadata manager for technical metadata, it is still tedious to implement such a component for each future data model.

Second off, the integration of metadata accross an entire LOCKSS network or even between LOCKSS nodes and additional services is facilitated. The metadata that is currently being dealt with in the LuKII project[14, p. 4-7], for

[2] The dispute regarding sense and nonsense of prophylactic format migration[13], and other tasks a management component should perform, implies that such a component should not be tied to the other layers too tightly. Discussion of this layer is beyond the scope of this paper.

example, is XML-based, because of which an embedded eXist XML-database has been prototypically added to the LOCKSS daemon and is currently being evaluated. This solves the problem of hard-coded database schemas. Since one goal of the project is testing prophylactic format migration in a LOCKSS network, the metadata from all nodes in the network will have to be integrated at some point to be evaluated by a central preservation management tool. While out of scope of the LuKII project, evaluating the use of a data model that makes it easier to integrate data from multiple sources seems promising for the future. In a distributed environment such as a LOCKSS network, a distributed data model such as RDF appears to be a natural match. On top of that, the management component could apply preservation policies formally expressed in OWL ontologies to identify objects for which actions such as format migration should be taken. Figure 1 outlines this scenario.

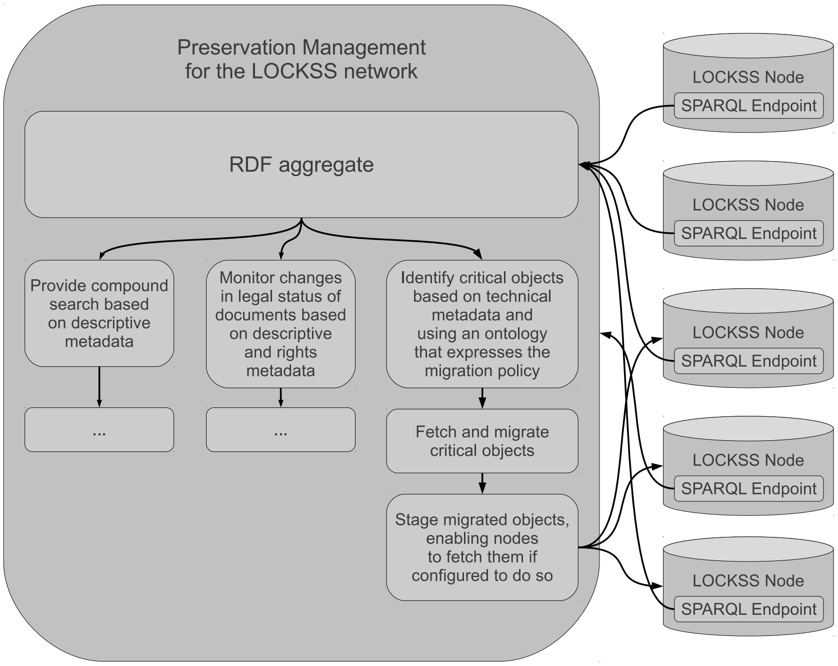

Preservation Management
for the LOCKSS network
RDF aggregate
Provide compound search based on descriptive metadata
Monitor changes in legal status of documents based on descriptive and rights metadata
Identify critical objects based on technical metadata and using an ontology that expresses the migration policy
...
...
Fetch and migrate critical objects
Stage migrated objects, enabling nodes to fetch them if configured to do so
LOCKSS Node
SPARQL Endpoint
LOCKSS Node
SPARQL Endpoint
LOCKSS Node
SPARQL Endpoint
LOCKSS Node
SPARQL Endpoint
LOCKSS Node
SPARQL Endpoint

Fig. 1. SPARQL-based aggregation of metadata from LOCKSS nodes

In the third place, the possibilities of semantic metadata are also interesting from the crawling component's point of view. There is an obvious need for a crawler to know what comprises a complex object that should be archived in order to know which links to follow in the potentially infinte web environment. LOCKSS plugins allow to define these rules for different publication platforms. Currently these specific crawling rules of the LOCKSS crawler are defined on the server side. Vocabularies such as OAI-ORE[6] can be used to more explicitly express these bounds of a complex object (an “article” in the LOCKSS terminology) on the client side, and thus allow for more generic crawling rules in the server side plugin. This would reduce the need for new plugins on the one hand, and allow the specification of an object’s bounds at the authorative place - the publication platform - on the other hand. The remainder of this paper will investigate this third usage-scenario for semantic metadata in a LOCKSS environment. Of course, the usage of semantic metadata for interoperability of LOCKSS and repositories is not limited to a single vocabulary. Ontologies such as [4] can be used, for example, to pass legal metadata from the repository to LOCKSS, thereby enabling a more fine-grained notion of rights-management compared to the generic permission statement[10] mechanism currently used by LOCKSS.

3 Exposing semantic metadata from repositories using RDFa

Besides providing the means to ingest and access documents, common repository software at its core also includes possibilities to expose metadata. The metadata capabilities usually consist of a description page that focuses primarily on human readability on the one hand, and machine-readable interfaces such as OAI-PMH on the other hand. With regards to passing content from a repository to a long term preservation system, two problems arise: The lowest common denominator of expressiveness in OAI-PMH is Dublin Core. This does not include structural information of any kind and thus is not detailed enough with regards to long-term archiving needs. Unfortunately, adding support for additional metadata schemes usually entails non-trivial extensions to the software that need to query the database and implement an additional query interface. Besides that, the machine-readable version of the metadata is exposed at a different URL than the human-readable version. This makes the configuration of a web crawler such as the one in LOCKSS more complicated than necessary.

A lightweight solution for both of those problems can be found in RDF in attributes (RDFa)[11]. The essentials in brief are that this Semantic Web standard enables human-readable versions of websites to be enriched in such a way that they can also be interpreted by machines. This means that it is possible to add machine-readable metadata at the template level of repository software. Boulal et al. conclude that the OAI-ORE vocabulary already mentioned above is qualified “as an interoperability layer because it allows describing scholarly work items in a way that is compatible to the web architecture”[3, p. 9]. Providing the necessary resource maps using RDFa is most likely the easiest way to enable existing repository software to expose complex objects in a machine-readable manner, since human-readable splash pages that describe an object already exist in all common repository software. [7] gives an impression of the necessary modifications. A further advantage of using RDFa is that the machine-readable metadata is available at the same URL as the human-readable version, naturally making it available along with the archived ressource.

4 Processing RDFa in LOCKSS

As mentioned above, LOCKSS is already capable of extracting metadata from the harvested content. The usual extraction procedure is based on extracting the metadata from the human-readable description, which can be error prone. A change in the structure of the page for example can easily result in changes becoming necessary on the plugin-level, even more so changes in the data model that is being used. This is where the advantage of RDFa-enriched HTML-pages becomes evident: the underlying RDF-model stays the same, even when elements are moved around etc., and RDF can be extracted no matter which vocabularies are being used.

With regards to storing metadata extracted from RDFa-enriched web pages, there are several options:

- mapping the metadata to a relational database,
- writing RDF/XML to an XML-database and
- using a triplestore.

While the first possibility enables reuse of the relational database already available in LOCKSS, it would imply unacceptable constraints on the flexibility of the system. Changes in the ontologies used to describe the content in the repository would necessarily imply changes to the database schema. Using an XML-store would be a solution for this, but limits the query possibilities to languages focusing on syntax rather than semantics, such as XPath or XQuery. The usage of a triplestore, ideally with an enabled reasoning component, is the most natural solution for RDF data and provides a powerful SPARQL interface.

The components that have been identified as necessary to enable the storage of RDFa metadata exposed by repository software in LOCKSS using a triplestore are:

- RDFa-enabled repository software that includes semantic markup for structural information,
- a LOCKSS plugin that is able to crawl according to the resource maps exposed by the repository,
- an article iterator that is able to make complex objects internally available,
- an article metadata extractor to extract RDF data about the complex object from the RDFa in the resource map,
- a file metadata extractor to extract and merge RDF data about individual bitstreams from the RDFa in the resource map and optionally in the individual files of an article,
- a metadata manager to add, update and delete RDF data from an embedded triplestore and

– a SPARQL endpoint to expose the metadata.

Figure 2 shows how these elements interact within a LOCKSS node and inbetween the node and a repository that provides the content that is to be archived. For more information on the concepts of plugins, article iterators, metadata extractors and metadata managers in LOCKSS components see [8]. While the work on a MetadataManager for RDF data has already begun as a side project of the author, the other parts are still missing. Once they are in place, the system will need to undergo extensive testing, especially with regards to the performance of the triplestore in a real-word environment.

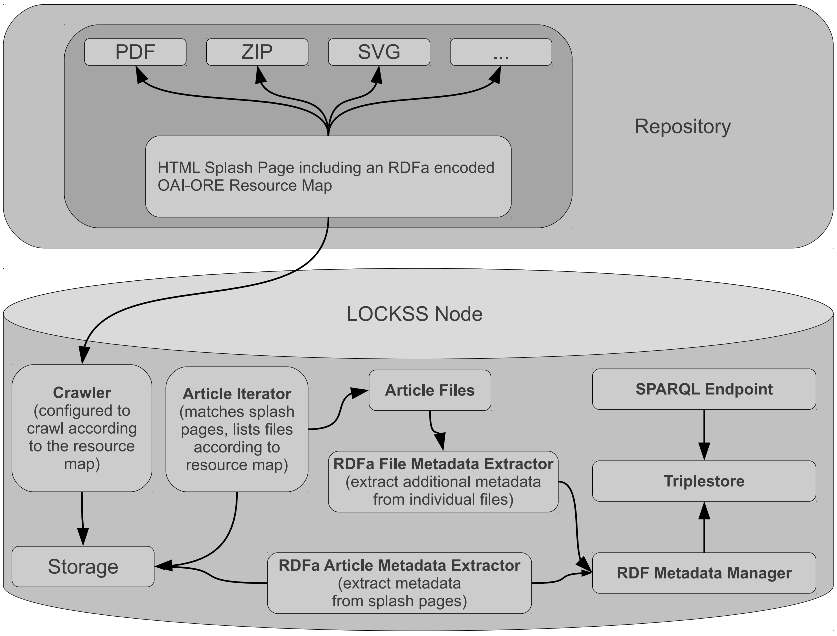

PDF
ZIP
SVG
...
Repository
HTML Splash Page including an RDFa encoded OAI-ORE Resource Map
LOCKSS Node
Crawler (configured to crawl according to the resource map)
Article Iterator (matches splash pages, lists files according to resource map)
Article Files
SPARQL Endpoint
RDFa File Metadata Extractor (extract additional metadata from individual files)
Triplestore
Storage
RDFa Article Metadata Extractor (extract metadata from splash pages)
RDF Metadata Manager

Fig. 2. Interaction of components necessary to store RDF data extracted from RDFa-enriched HTML-pages

5 Conclusion

This has been a rough and purely technical view on a modular digital preservation system that makes use of Semantic Web standards. Question such as "Which metadata belongs into a digital archive?", "Which part of the system is responsible to generate technical metadata?" and "Does descriptive metadata belong into the preservation layer?" remain open. The model sketched out above theoretically makes it possible to add – along with the content it describes – arbitrary, but highly expressive metadata to a modular long-term archiving system.

The possibility to access that metadata through an HTTP-SPARQL-interface provides the means to add a management layer to the system without tying it in too tightly. With the advent of Semantic Web technologies and standards in the library domain, further investigations in this direction seem promising.

References

1. About kopal, http://kopal.langzeitarchivierung.de/index.php.en
2. A New Way of Preserving Cultural Heritage and Cumulative Knowledge, http://www.exlibrisgroup.com/category/RosettaOverview
3. Boulal, Anouar et al.: Report on Enhancing Interoperability between existing Open Access Publication Infrastructures. Draft available at http://www.eco4r.org/downloads/eco4r_report_compoundobjects_draft.pdf (2010)
4. CASPAR Rights Ontology, http://www.casparpreserves.eu/publications/ontologies/RightsOntology.html
5. DFG-Projekt: LuKII (LOCKSS und KOPAL Infrastruktur und Interoperabilität), http://www.lukii.hu-berlin.de/
6. Lagoze, Carl et al.: ORE User Guide - Primer, http://www.openarchives.org/ore/1.0/primer (2008)
7. Lagoze, Carl et al.: ORE User Guide - Resource Map Implementation in RDFa, http://www.openarchives.org/ore/1.0/rdfa (2008)
8. LOCKSS API documentation, http://www.lockss.org/lockssdoc/gamma/daemon/index.html
9. Lots of Copies Keep Stuff Safe), http://lockss.stanford.edu/lockss/Home
10. Making Your Titles LOCKSS Compliant, http://www.lockss.org/lockss/Making_Your_Titles_LOCKSS_Compliant#Permission_to_the_LOCKSS_Software
11. RDFa Primer. Bridging the Human and Data Webs, http://www.w3.org/TR/xhtml-rdfa-primer/ (2008)
12. Rosenthal, David S.H.: Bit Preservation: A Solved Problem? In: International Journal of Digital Curation, vol. 1, no.5. http://www.ijdc.net/index.php/ijdc/article/viewFile/151/224 (2010)
13. Rosenthal, David S.H.: The Half-Life of Digital Formats, http://blog.dshr.org/2010/11/half-life-of-digital-formats.html (2010)
14. Steinke, Tobias: Universal Object Format. An archiving and exchange format for digital objects, http://kopal.langzeitarchivierung.de/downloads/kopal_Universal_Object_Format.pdf (2006)