Proceedings of the 2nd International Workshop on **Semantic Digital Archives**

held in conjunction with the
16th Int. Conference on Theory and Practice of Digital Libraries (TPDL)
on September 27, 2012 in Paphos, Cyprus.

http://sda2012.dke-research.de

Edited by

Annett Mitschick,
Technische Universität Dresden, Germany, annett.mitschick@tu-dresden.de

Fernando Loizides,
Cyprus University of Technology, fernando.loizides@cut.ac.cy

Livia Predoiu,
Otto-von-Guericke University Magdeburg, Germany, livia.predoiu@ovgu.de

Andreas Nürnberger,
Otto-von-Guericke University Magdeburg, Germany, andreas.nuernberger@ovgu.de

Seamus Ross,
University of Toronto, Canada, seamus.ross@utoronto.ca

September, 2012

Preface

The 2nd Workshop on Semantic Digital Archives (SDA 2012) builds upon the success of the previous edition in 2011, held in conjunction with the International Conference on Theory and Practice of Digital Libraries, TPDL 2011 (formerly known as European Conference on Digital Libraries, ECDL). Organized as full-day workshop, SDA 2012 aims to advance and discuss appropriate knowledge representation and knowledge management solutions specifically designed for improving Archival Information Systems. The main objective is to have a closer dialogue between the technical oriented communities with people from the (digital) humanities and social sciences, as well as cultural heritage institutions in general in order to approach the topic from all relevant angles and perspectives. This workshop is an exciting opportunity for collaboration and cross-fertilization.

Intending to have an open discussion on topics related to the general subject of Semantic Digital Archives, we invited contributions that focus on one of the following topics:

- Ontologies & linked data for digital archives and digital libraries (incl. multimedia archives)
- Semantic search & semantic information retrieval in digital archives and digital libraries (incl. multimedia archives)
- Implementations and evaluations of semantic digital archives
- Theoretical and practical archiving frameworks using Semantic (Web) technologies
- Semantic or logical provenance models for digital archives or digital libraries
- Visualization and exploration of content in large digital archives
- User interfaces for semantic digital libraries and intelligent information retrieval
- User studies focusing on end-user needs and information seeking behavior of end-users
- Semantic (Web) services implementing the OAIS standard
- Logical theories for digital archives
- Knowledge evolution
- Information integration/semantic ingest (e.g. from digital libraries)
- Trust for ingest & data security/integrity check for long-term storage of archival records
- Semantic extensions of emulation/virtualization methodologies for digital archives
- Semantic long-term storage and hardware organization tailored for digital archives
- Migration strategies based on Semantic (Web) technologies

We received submissions covering a broad range of relevant topics in the area of semantic digital archives. With the help of our program committee all articles were peer-reviewed. These proceedings comprise all accepted submissions which have been carefully revised and enhanced by the authors according to the reviewers' comments.

These papers were joined by an invited keynote by *Andreas Rauber* (Vienna University of Technology, Austria). In *Digital Preservation in Data-Driven Science: On the Importance of Process Capture, Preservation and Validation* he points out the necessity of capturing and documenting processes (in addition to the context of data objects), especially in e-Science and business settings, and presents an approach for process preservation and verification upon later re-execution.

The paper *Entity Extraction and Consolidation for Social Web Content Preservation (S. Dietze et al.)* presents an approach to extract and consolidate information from archived social Web content in order to facilitate semantic search of Web archives. The work was developed in the EC-funded Integrating Project ARCOMEM.

With *Do we need metadata? - An On-line Survey in German Archives* Marcel Ruhl presents the revealing results of a survey among German archives regarding the use of metadata standards for the annotation of audiovisual media.

The paper *Automatic Classification of Scientific Records using the German Subject Heading Authority File (Ch. Wartena & M. Sommer)* introduces an approach to assign subject classifications to records without using machine learning techniques but by the application of the German Subject Heading Authority File (SWD).

In *Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries (M. Grassi et al.)* the authors propose an annotation system, developed in the context of the Semlib project, which provides the user with the ability to annotate distributed resources, i.e. multimedia content published on the Web, using an extension of the Open Annotation Collaboration (OAC) ontology.

The paper *Towards a Recommender System for Statistical Research Data (D. Bahls et al.)* presents the conceptual ideas and the system architecture of a case-based recommender system for statistical data used in scientific research, and discusses possible similarity measures and notification services.

With *A method and guidelines for the cooperation of ontologies and relational databases in Semantic Web applications* L. Bozzato et al. showcase a methodology for mapping relational data to an ontology structure to support SPARQL queries and inference and to take advantage of the representation possibilities offered by both data models.

A critical issue when developing RDF-based semantic archives is the right choice of an appropriate large-scale storage solution for the data. *Yet Another Triple Store Benchmark? Practical Experiences with Real-World Data (M. Voigt et al.)* presents the experimental setting and the results of extensive performance tests of state-of-the-art RDF stores using non-synthetic RDF datasets.

Finally, the paper *Implementing CIDOC CRM Search Based on Fundamental Relations and OWLIM Rules (V. Alexiev)* proposes an approach to provide a higher-level perspective on RDF data by mapping complex sub-graph patterns to simpler, more abstract descriptions using OWLIM rules. The author presents an implementation of the concept regarding search with the CIDOC Conceptual Reference Model.

We sincerely 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 exciting and promising research field of semantic digital archives and semantic digital archiving infrastructures.

We would also like to thank Sun SITE Central Europe for hosting these proceedings on http://ceur-ws.org.

September 2012

A. Mitschick, F. Loizides, L. Predoiu, A. Nürnberger, and S. Ross

Program Committee

Vassilis Christophides	Foundation of Research & Technology - Hellas, Greece
Kai Eckert	University Library of Mannheim, Germany
Armin Haller	CSIRO ICT Centre, Australia
Steffen Hennicke	Humboldt-Universität zu Berlin, Germany
Stijn Heymans	SRI International, USA
Pascal Hitzler	Wright State University, USA
Christian Keitel	State Archive of Baden-Württemberg, Germany
Birger Larsen	Royal School of Library and Information Science, Denmark
Thomas Lukasiewicz	University of Oxford, UK
Mathias Lux	Klagenfurt University, Austria
Knud Möller	Talis, Birmingham, UK
Kai Naumann	State Archive of Baden-Württemberg, Germany
Jacco van Ossenbruggen	VU University Amsterdam, Netherlands
Andreas Rauber	Vienna University of Technology, Austria
Thomas Risse	L3S Research Center, Hannover, Germany
Sebastian Rudolph	Karlsruher Institut für Technologie, Germany
Mike Salampasis	Alexander Technology Educational Institute of Thessaloniki, Greece
Herbert van de Sompel	Los Alamos National Laboratory Research Library, USA
Marc Spaniol	Max-Planck-Institut Saarbrücken, Germany
Manfred Thaller	University of Cologne, Germany

Table of Contents

Digital Preservation and Metadata

Structuring and Recommendation

Semantic Technologies and Ontologies

Digital Preservation in Data-Driven Science: On the Importance of Process Capture, Preservation and Validation

Andreas Rauber[1]

Department of Software Technology and interactive Systems
Vienna university of Technology
Favoritenstrasse 9-11, 1040 Vienna, Austria
rauber@ifs.tuwien.ac.at

Abstract. Current digital preservation is strongly biased towards data objects: digital files of document-style objects, or encapsulated and largely self-contained objects. To provide authenticity and provenance information, comprehensive metadata models are deployed to document information on an object's context. Yet, we claim that simply documenting an objects context may not be sufficient to ensure proper provenance and to fulfill the stated preservation goals. Specifically in e-Science and business settings, capturing, documenting and preserving entire processes may be necessary to meet the preservation goals. We thus present an approach for capturing, documenting and preserving processes, and means to assess their authenticity upon re-execution. We will discuss options as well as limitations and open challenges to achieve sound preservation, specifically within scientific processes.

Keywords: Digital Preservation, Processes, Context, eScience

1 Introduction

Digital preservation (DP) traditionally has a predominantly data-centric view on both its operations as well as the objects it is dealing with. Digital objects considered for preservation are usually (turned into, as far as possible) self-contained, static objects such as images resulting from scans, classical document-style objects, data sets, but also information packages containing software and other in principle dynamic objects as encapsulated files. Furthermore, objects are usually "removed" from an operational life-cycle into an archival life cycle, ingested into designated repositories for long-term maintenance, from which they are removed and re-inserted into a potentially new life-cycle when needed in an operational manner again, only to be re-ingested as new objects after completion of their new life as now new archival objects. Consequently, also cost estimation and investments are largely based upon aggregated data item-level information, adding up processing costs, storage costs and others.

We claim that this traditional view is hitting severe limitations as we are observing a set of interesting changes in the preservation community: Most importantly, preservation is expanding beyond the traditional cultural heritage

community. While originating in the sciences, recognizing the need to maintain the investment made into data collected electronically (as evident by the early and strong commitment of institutions such as NASA, leading to the infamous OAIS model) the key drivers, expertise and know how and development has come from and taken place in the cultural heritage community. The key characteristic of this community is the dedication to preservation of information as a, or even the, primary mission, resulting in a holistic understanding of the scope of the problem and its long-term implication beyond individual technical or organizational issues. Yet, more recently we see a range of other communities facing the need for digital preservation: back to the origins of DP, science as a whole is becoming increasingly dependent on data as the core facilitator in virtually all scientific disciplines, leading to trends defined as data-driven science, e-Science, Big Data [9], the Fourth Paradigm [6], and others. But even beyond cultural heritage and science communities, both of which have been involved in DP for a long time, we find entirely new players / customers in need of DP solutions, many of them coming from a range of industrial backgrounds, and with a quite diverse set of motivations. These may range from specific legal / compliance requirements, via somewhat more ambiguous risk mitigation desires to serving dedicated business needs. Thus, while in principle being similar to the cultural heritage sector, there are some interesting challenges stretching beyond the ones encountered in more traditional settings.

This paper starts in Sec. 2 with a loose collection of observations on changes in the DP community, highlighting three areas of focus that we deem important, namely a shift towards risk management, viewing DP in the context of e-Governance frameworks, and the shift towards the preservation of processes rather than data objects. This will be followed by a more detailed look at two key aspects, namely process preservation and a framework for evaluating the quality of re-execution of preserved processes in Sections 3 and 4, respectively (largely adopted from [16]), before providing a brief summary in Section 5.

2 Implications of Changes in Stakeholder Communities

The observed expansion in stakeholder communities has some interesting implications for the DP community: first of all, we are experiencing yet another clash of languages and cultures: after partially successfully consolidating the viewpoints of archival and library communities, merging the viewpoints of the museums community and even succeeding in getting computer science to listen and communicate on an increasingly shared level of mutual understanding, DP is currently being recognized, interpreted and contributed to by a whole range of new key players with completely diverse interpretation of the concepts widely accepted in the traditional DP community. Long-term may be as short as 7 years, preservation and loss is not necessarily measured on the level of the need for maintaining an object, but as best effort vs. risk trade-off, specifically not happening "at all costs" wherever possible, with deletion-as-early-as-permissible being a key factor,

Where DP is serving a business purpose, objects are rarely being perceived as frozen and deposited into an archive. Rather, they need to be maintained in an operational environment. Catch phrases such as business continuity capture a lot about the thinking behind this. More importantly, however, they also help to identify approaches and solutions to serve these (but also more traditional) DP needs that stem from different backgrounds: life cycle management of entire IT environments, redundancy, (IT) security, e-governance structures and methods have been developed, deployed, tested, customized and improved over long time spans in these communities. These solutions may prove valuable contributions to the DP community at large.

This integration of yet another heterogeneous set of communities poses severe challenges to a rather tightly-knit network of DP researchers and practitioners that has just started to evolve into a very young community of its own, with it's own jargon, events, and commonly understood basis of generic concepts. Yet, it also offers huge benefits. Apart from contributing new competences and tested solutions that can be adapted to serve more generic DP needs, it also broadens the basis amongst which to share the costs of research and development in DP. It allows us to integrate know-how from different groups and grows the market of where this know-how can be used.

In terms of new areas of activity, at least three major trends can be observed:

Preservation as a cost/benefit trade-off: The primary focus so far has been on preserving objects because they need to be preserved. Current thinking seems to be moving towards obsolescence as a risk, and DP as a risk mitigation strategy. While this concept is anything but new to the DP community, the key difference is on the juxtaposition of risk and benefit, and a much more pronounced and explicit willingness to sacrifice availability of objects when the investments necessary to maintain them would likely outweigh the business benefits or legal sanctions. Although the same principles are applicable in traditional settings as well, this focus shift forces a more explicit formulation of benefits/-value and a clearer specification of risk and costs, especially for more immediate/short term actions. This also will call for the application of existing frameworks for risk identification as well as cost/benefit estimation, with potentially strongly diverging valuations in the non-heritage domains.

DP as capabilities and maturity evaluations: Another important shift we may see coming is a shift from DP being something happening in a data archive or repository setting, i.e. a designated institution or department handling "old" objects, and being audited on its performance via any set of audit and certification routines to ensure proper preservation at the highest possible level. Rather, we may view DP as a set of capabilities that an institution has, as part of many other operational capabilities, and that are integrated with more routine processes. We are thus currently investigating opportunities of integrating DP in an eGovernance framework. Reference models such as TOGAF [17] and COBIT [8] provide a well-tested basis for designing and managing DP activities alongside other routine IT capabilities. It allows us to manage DP capabilities in the framework of IT governance, benefiting from the well-structured concepts

and processes in this domain, defining drivers, constrains and controls [2]. An example of how to model maturity levels for preservation capabilities is provided in [1].

Preserving processes rather than (only) data: The third, major area of activity currently, and probably the most relevant with respect to semantic technologies, relates to the shift from preserving static objects to the preservation of entire process chains. This shift is motivated by two core considerations: first of all, in many new settings, it is not static artifacts, but the need to be able to re-run processes in an authentic manner that are the key DP requirement. While the preservation of data as documentation may be sufficient to provide evidence about processes having been run they are no replacement for preserving the actual process.

But even when the focus is on the actual data, it may be advisable to preserve the process chain the data was subjected to. In an e-Science setting, while preserving the data is an essential first step for any sustainable research efforts, the data alone is often not sufficient for later analysis of how this data was obtained, pre-processed and transformed. Results of scientific experiments are often just the very last step of the whole process, and to be able to correctly interpret them by other parties or at a later point in time, also these processes need to be preserved.

Thus, specifically in an e-Science setting, preserving processes together with the data helps us to meet two goals at the same time: on the one hand, the processes are essential aspects of representation information, allowing to trace the various (pre-)processing steps applied to the data, any bias that might have been introduced, or errors stemming from faulty processing. On the other hand, it also allows us to re-run these processes on new data, learning how models and views evolved, and to discover discrepancies from earlier analyzes.

We thus need to go beyond the classical concerns of Digital Preservation research, and consider more than the preservation of data. The following section takes a closer look at some of the activities centering around process preservation, covering both process capture as well as evaluation of re-executed processes – and the resulting requirements at the level of process capture. We will choose examples from the e-Science/Data curation domain as an exemplary setting. The following sections are largely adopted from a position paper summarizing our considerations to data quality aspects in data curation for a recent workshop by the National Science Foundation [16].

3 Capturing Processes

Curation of business or E-Science processes requires capturing the whole context of the process, including enabling technologies, different system components on both hardware and software levels, dependencies on other computing systems and services operated by external providers, the data consumed and generated, and more high-level information such as the goals of the process, different stakeholders and parties. The context of information needed for preserving processes

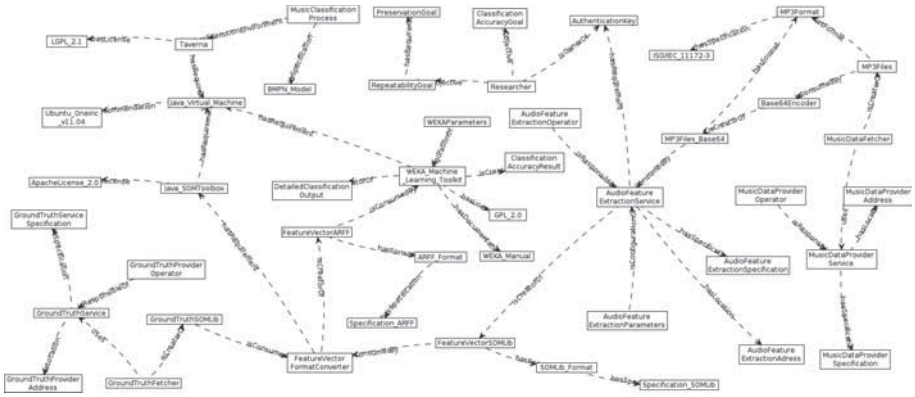


Fig. 1: Context Model of musical genre classification process [12], top-level view

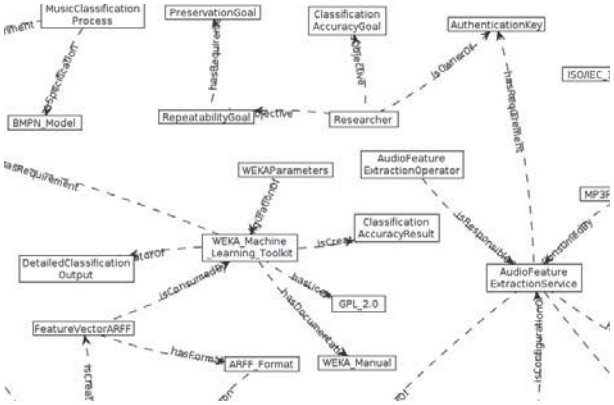


Fig. 2: Excerpt of Context Model of musical genre classification process

is considerably more complex than that of data objects, as it not only requires dealing with the structural properties of information, but also with the dynamic behavior of processes. Successful curation of an eScience process requires capturing sufficient detail of the process, as well as its context, to be able to re-run and verify the original behavior at a later stage, under changed and evolved conditions. We thus need to preserve the set of activities, processes and tools, which all together ensure continued access to the services and software which are necessary to reproduce the context within which information can be accessed, properly rendered and validated.

To address these challenges, we have devised a context model to systematically capture aspects of a process that are essential for its preservation and verification upon later re-execution [12]. The model consists of approximately 240 elements, structured in around 25 major groups. It corresponds to some degree to the representation information network [10], modeling the relationships

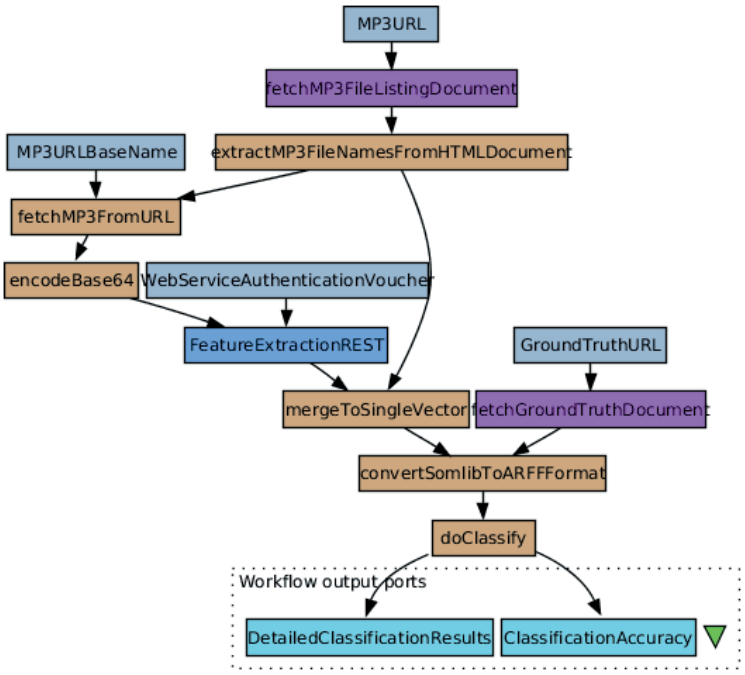


Fig. 3: Musical genre classification, including fetching of data, modelled in the Taverna workflow engine [11]

between an information object and its related objects, be it documentation of the object, constituent parts and other information required to interpret the object. This is extended to understand the entire context within which a process, potentially including human actors, is executed, forming a graph of all constituent elements and, recursively, their representation information. Specific emphasis is given to the identification of distributed components of a process: identifying, for example, external web services is essential to ensure that a process can be preserved, as specific measures need to be devised to ensure their availability in the future. Approaches include an integration of a web service into the process, legal agreements on the preservation of web services (at specified versions) by the service provider, deposit regulations using ESCROW regulations [7], and others. The model is implemented in the form of an ontology, which on the one hand allows for the hierarchical categorization of aspects, and on the other hand shall enable reasoning, e.g. over the possibility of certain preservation actions for a specific process instance. While the model is very extensive, it should be noted that a number of aspects can be filled automatically – especially if institutions have well-defined and documented processes. Also, not all sections of the model are equally important for each type of process. Therefore, not every aspect has to be described at the finest level of granularity.

Figure 2 provides an overview on the concrete instances and their relations identified as relevant aspects of the process context for a music classification process. The process basically represent a typical machine learning experiment in music information retrieval (MIR), where features are extracted from audio data and used as the basis for training a classifier system, sorting pieces of music into different musical genres. For a detailed description of this process and on how to make it fit for preservation, refer to [11]. An excerpt is provided in Fig. **??**, showing, for example, the links to the preservation goal specifications, some of the modules involved (the AudioFeatureExtractor and the WEKA Machine Learning toolkit) and their respective contexts (licenses (GPL_2.0), file format and their link to specifications, manuals, etc.

To move towards more sustainable E-Science processes, we recommend implementing them in workflow execution environments. For example, we currently use the Taverna workflow engine [13]. Taverna is a system designed specifically to execute scientific workflows. It allows scientists to combine services and infrastructure for modeling their workflows. Services can for example be remote web-services, invoked via WSDL or REST, or local services, in the form of predefined scripts, or user-defined scripts.

Implementing such a research workflow in a system like Taverna yields a complete and documented model of the experiment process – each process step is defined, as is the sequence (or parallelism) of the steps. Further, Taverna requires the researcher to explicitly specify the data that is input and output both of the whole process, as well as of each individual step. Thus, also parameter settings for specific software, such as the parameters for a machine learning tool or feature extraction, become explicit, either in the form of process input data, or in the script code.

Figure 3 shows an example of the music classification experiment workflow modeled in the Taverna workflow engine. We notice input parameters to the process such as the URL of the MP3 contents and the ground truth, and also an authentication voucher which is needed to authorize the use of the feature extraction service. The latter is a bit of information that is likely to be forgotten frequently in descriptions of this process, as it is rather a technical requirement than an integral part of the scientific process transformations. However, it is essential for allowing re-execution of the process, and may help to identify potential licensing issues when wanting to preserve the process over longer periods of time, requiring specific digital preservation measures.

During an execution of the workflow, Taverna records so-called *provenance data*, i.e. information about the creation of the objects, on the data transformation happening during the experiment. Taverna uses its proprietary *Janus* format, an extension on the Open-Provenance Model[14] that allows capturing more details. Such data is recorded for the input and output of each process step. It thus allows to trace the complete data flow from the beginning of the process until the end, enabling verification of the results obtained. This is essential for being able to verify system performance upon re-execution, specifically when

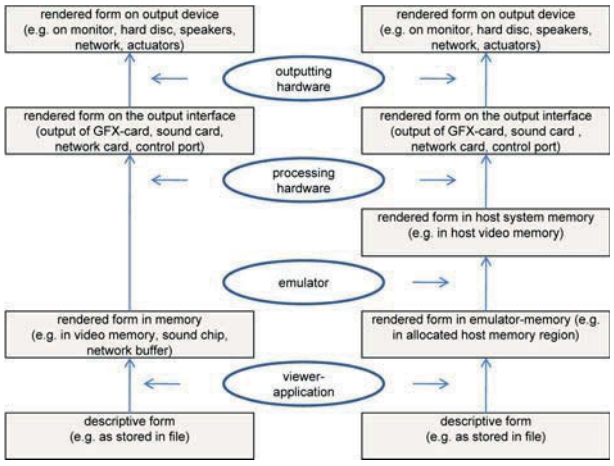


Fig. 4: Different forms of a digital object in a system's memory. On the left the layers in an original system are shown, on the right the layers in the system hosting the emulator are shown. [3]

any component of the process (such as underlying hardware, operating systems, software versions, etc.) has changed.

4 Evaluating Process Re-Execution

A critical aspect of re-using digital information in new settings is its trustworthiness, especially its authenticity and faithful rendering (with rendering being any form of representation or execution and effect of a digital object, be it rendering on a screen, an acoustic output device, or state changes on ports, discs etc.). Establishing identity or faithfulness is more challenging than commonly assumed: current evaluation approaches frequently operate on the structural level, i.e. by analyzing the preservation of significant properties on the file format level in case of migration of objects. Yet, any digital object (file, process) is only perceived and can only be evaluated properly in a well-specified rendering environment within which faithfulness of performance need to be established. In emulation settings, this evaluation approach is more prominently present, yet few emulators support the requirements specific to preservation settings. We thus argue that, actually, migration, emulation and virtually all other approaches to logical/structural data preservation need to be evaluated in the same way, as they are virtually no different from each other as all need to be evaluated in a given rendering/performance environment. [5].

We also devise a framework for evaluating whether two versions of a digital object are equivalent [3]. Important steps in this framework include (1) a description of the original environment, (2) the identification of external events influencing the object's behavior, (3) the decision on what level to compare the

two objects, (4) recreating the environment, (5) applying standardized input to both environments, and finally (6) extracting and (7) comparing the significant properties on suitable levels of an object's rendering. Even though the framework focuses mostly on emulation of environments, the principles are also applicable specifically for entire processes, and will work virtually unchanged also for migration approaches, when complex objects are transformed e.g into a new file format version.

An essential component of the framework is the identification at which levels to measure the faithfulness of property preservation, as depicted in Figure 4. A rendered representation of the digital object has to be extracted on (a) suitable level(s) where the significant properties of the object can be evaluated. For some aspects, the rendering of an object can be performed based on its representation in specific memories (system/graphics/sound card/IO-buffer), for others the respective state changes at the output port have to be considered while for yet others the actual effect of a system on its environment needs to be considered, corresponding to delineating the boundaries of the system to be evaluated. (Note that identity on a lower level does not necessarily correspond to identity at higher levels of the viewpath - in some cases significant effort is required to make up for differences e.g. on the screen level when having to emulate the visual behavior of cathode ray screens on modern LCD screens [15].) An example of applying this framework to the evaluation of preservation actions is provided in [4]

A key challenge in this context will be to come up with a comprehensive model of what information to capture for specific types of processes and preservation requirements, as well as guidelines on how to do this.

5 Conclusions

This paper presents a loose collection of some trends observed in digital preservation research, specifically a shift toward a more risk/cost/benefit based approach to DP, the framing of DP in IT governance principles, and specifically the necessity to preserve entire processes rather than only data. With the growing importance of preserving entire processes rather than sets of homogeneous, static (and usually quite simple) objects, the requirements on techniques to capture, document and reason across increasingly complex sets of context meta-information is growing. This requires new approaches to context capture and representation.

Still, the considerations above cover only a small subset of the quite significant research challenges that continue to emerge in the field of digital curation. We thus strongly encourage the community to contribute to an effort of collecting and discussing these emerging research questions in a loosely organized form. To this end, following the Dagstuhl Seminar on Research Challenges in Digital Preservation[1], a Digital Preservation Challenges Wiki[2] has been created, where we invite contributions and discussion. As a follow-up to the Dagstuhl seminar,

[1] http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=10291
[2] http://sokrates.ifs.tuwien.ac.at

a workshop on DP Challenges[3] will be held at iPRES 2012 in Toronto focusing on the elicitation and specification of research challenges.

References

1. C. Becker, G. Antunes, J. Barateiro, and R. Vieira. A capability model for digital preservation: Analysing concerns, drivers, constraints, capabilities and maturities. In *8th International Conference on Preservation of Digital Objects (IPRES 2011)*, Singapore, November 2011.
2. C. Becker, G. Antunes, J. Barateiro, and R. Vieira. Control objectives for dp: Digital preservation as an integrated part of it governance. In *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, New Orleans, Lousiana, US, October 2011.
3. M. Guttenbrunner and A. Rauber. A Measurement Framework for Evaluating Emulators for Digital Preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 2012.
4. M. Guttenbrunner and A. Rauber. Evaluating an emulation environment: Automation and significant key characteristics. In *Proceedings of the 9th conference on Preservation of Digital Objects (iPRES2012)*, Toronto, Canada, October 1–5 2012.
5. M. Guttenbrunner and A. Rauber. Evaluating emulation and migration: Birds of a feather? In *Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries*, Taipei, Taiwan, November 12–15 2012.
6. T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
7. K. Hobel and S. Strodl. Software escrow agreements. In *Tagungsband des 15. Internationalen Rechtsinformatik Symposions IRIS 2012*, pages 603–610, 2012.
8. IT Governance Institute. *COBIT 4.1. Framework – Control Objectives – Management Guidelinces – Maturity Models*. 2007.
9. S. Manegold, M. Kersten, and C. Thanos. Special theme: Big data. *ERCIM News*, (89), January 2012.
10. Y. Marketakis and Y. Tzitzikas. Dependency management for digital preservation using semantic web technologies. *International Journal on Digital Libraries*, 10:159–177, 2009.
11. R. Mayer and A. Rauber. Towards time-resilient mir processes. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal, October 8-12 2012.
12. R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving scientific processes from design to publication. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, LNCS, Cyprus, September 2012. Springer.
13. P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble. Taverna, reloaded. In *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, pages 471–481. Springer, June 2010.
14. L. Moreau, J. Freire, J. Futrelle, R. E. Mcgrath, J. Myers, and P. Paulson. *Provenance and Annotation of Data and Processes*, chapter The Open Provenance Model: An Overview, pages 323–326. Springer, 2008.

[3] `http://digitalpreservationchallenges.wordpress.com/`

15. G. Phillips. Simplicity betrayed. *Communications of the ACM*, 53(6):52–58, 2010.
16. A. Rauber. Data quality for new science: Process curation, curation evaluation and curation capabilities. In *Workshop notes for the UNC/NSF Workshop Curating for Quality*, Arlington, VA, September 10–11 2012.
17. The Open Group. *TOGAF Version 9*. Van Haren Publishing, 2009.

Entity Extraction and Consolidation for Social Web Content Preservation

Stefan Dietze[1], Diana Maynard[2], Elena Demidova[1], Thomas Risse[1], Wim Peters[2,]
Katerina Doka[3], Yannis Stavrakas[3]

[1] L3S Research Center, Leibniz University, Hannover, Germany
{dietze, nunes, demidova, risse}@l3s.de
[2] Department of Computer Science, University of Sheffield, Sheffield, UK
{diana, wim}@dcs.shef.ac.uk
[3] IMIS, RC ATHENA, Artemidos 6, Athens 15125, Greece
katerina@cslab.ece.ntua.gr; yannis@imis.athenainnovation.gr

Abstract. With the rapidly increasing pace at which Web content is evolving, particularly social media, preserving the Web and its evolution over time becomes an important challenge. Meaningful analysis of Web content lends itself to an entity-centric view to organise Web resources according to the information objects related to them. Therefore, the crucial challenge is to extract, detect and correlate entities from a vast number of heterogeneous Web resources where the nature and quality of the content may vary heavily. While a wealth of *information extraction* tools aid this process, we believe that, the *consolidation* of automatically extracted data has to be treated as an equally important step in order to ensure high quality and non-ambiguity of generated data. In this paper we present an approach which is based on an iterative cycle exploiting Web data for (1) targeted archiving/crawling of Web objects, (2) entity extraction, and detection, and (3) entity correlation. The long-term goal is to preserve Web content over time and allow its navigation and analysis based on well-formed structured RDF data about entities.

Keywords. Knowledge Extraction, Linked Data, Data Consolidation, Data Enrichment, Web Archiving, Entity Recognition

1 Introduction

Given the ever increasing pace at which Web content is constantly evolving, adequate Web archiving and preservation have become a cultural necessity. Along with "common" challenges of digital preservation, such as media decay, technological obsolescence, authenticity and integrity issues, Web preservation has to deal with the sheer size and ever-increasing growth rate of Web content. This in particular applies to user-generated content and social media, which is characterized by a high degree of *diversity*, heavily *varying quality* and *heterogeneity*. Instead of following a *collect-all* strategy, archival organizations are striving to build *focused archives* that revolve around a particular *topic* and reflect the diversity of information people are interested in. Thus, focused archives largely revolve around the *entities* which define a topic or

area of interest, such as persons, organisations and locations. Hence, extraction of entities from archived Web content, in particular social media, is a crucial challenge in order to allow semantic search and navigation in Web archives and the relevance assessment of a given set of Web objects for a particular *focused crawl*.

However, while tools are available for information extraction from more formal text, social media affords particular challenges to knowledge acquisition. These are detailed more explicitly in Section 3. This calls for a range of specific strategies and techniques to *consolidate*, *enrich*, *disambiguate* and *interlink* extracted data. This in particular benefits from taking advantage of existing knowledge, such as Linked Open Data [1], to compensate for, disambiguate and remedy degraded information. While data consolidation techniques traditionally exist independent from named entity recognition (NER) technologies, their coherent integration into unified workflows is of crucial importance to improve the wealth of automatically extracted data on the Web. This becomes even more crucial with the emergence of an increasing variety of publicly available and end-user friendly knowledge extraction and NER tools such as DBpedia Spotlight[1], GATE[2], Open Calais[3], Zemanta[4].

In this paper, we introduce an integrated approach to extracting and consolidating structured knowledge about entities from archived Web content. This knowledge will in the future be used to facilitate semantic search of Web archives and to further guide the crawl. This work was developed in the EC-funded Integrating Project ARCOMEM[5]. Note, while temporal aspects related to term and knowledge evolution are substantial to Web preservation, these are currently under investigation [24] but out of scope for this paper.

2 Related Work

Entity recognition is one of the major tasks within information extraction and may encompass both NER and term extraction. Entity recognition may involve rule-based systems [13] or machine learning techniques [14]. Term extraction involves the identification and filtering of term candidates for the purpose of identifying domain-relevant terms or entities. The main aim in automatic term recognition is to determine whether a word or a sequence of words is a term that characterises the target domain. Most term extraction methods use a combination of linguistic filtering (e.g. possible sequences of part of speech tags) and statistical measures (e.g. tf.idf) [15] and [16], to determine the salience of each term candidate for each document in the corpus [23].

Data consolidation has to cover a variety of areas such as enrichment, entity/identity resolution for disambiguation as well as clustering and correlation to consolidate disparate data. In addition, link prediction and discovery is of crucial importance to enable clustering and correlation of enriched data sources. A variety of

[1] http://spotlight.dbpedia.org
[2] http://gate.ac.uk/
[3] http://www.opencalais.com/
[4] http://www.zemanta.com/
[5] http://www.arcomem.eu

methods for entity resolution have been proposed, using relationships among entities [7], string similarity metrics [6], as well as transformations [9]. An overview of the most important works in this area can be found in [8]. As opposed to entity correlation techniques exploited in this paper, text clustering of documents exploits feature vectors, to represent documents according to contained terms [10][11][12]. Clustering algorithms measure the similarity across the documents and assign the documents to the appropriate clusters based on this similarity. Similarly, vector-based approaches have been used to map distinct ontologies and datasets [2][3]. As opposed to text clustering, entity correlation and clustering takes advantage of background knowledge from related datasets to correlate previously extracted entities. Therefore, link discovery is another crucial area to be considered. Graph summarization predicts links in annotated RDF graphs. A detailed survey of link predictions techniques in complex networks and social network are presented by [4] and [5], respectively.

3 Challenges and overall approach

ARCOMEM follows a use case-driven approach based on scenarios aimed at creating focused Web archives. We deploy a document repository of crawled Web content and a structured RDF knowledge base containing metadata about entities detected in the archived content. Archivists can specify or modify crawl specifications (fundamentally consisting of selected sets of relevant entities and topics). The intelligent crawler will be able to learn about crawl intentions and to refine a crawling strategy on-the-fly. This is especially important for long running crawls with broader topics, such as the financial crisis or elections, where entities are changing more frequently and hence require regular adaptation of the crawl specification. End-user applications allow users to search and browse the archives by exploiting automatically extracted metadata about entities and topics.

Fundamental to both crawl strategy refinement and Web archive navigation is the efficient extraction of entities from archived Web content. In particular, social media poses a number of challenges for language analysis tools due to the degraded nature of the text, especially where tweets are concerned. In one study, the Stanford NER tagger dropped from 90.8% F1 to 45.88% when applied to a corpus of tweets [17]. [19] also demonstrate some of the difficulties in applying traditional POS tagging, chunking and NER techniques to tweets, while language identification tools typically also do not work well on short sentences. Problems are caused by incorrect spelling and grammar, made-up words, hashtags, @ signs and emoticons, unorthodox capitalisation, and spellings (e.g duplication of letters in words for emphasis, text speak). Since tokenisation, POS tagging and matching against pre-defined gazetteer lists are key to NER, it is important to resolve these problems: we adopt methods such as adapting tokenisers, using techniques from SMS normalisation, retraining language identifiers, use of case-insensitive matching in certain cases, using shallow techniques rather than full parsing, and using more flexible forms of matching.

Entity extraction and enrichment is covered by a set of dedicated components which have been incorporated into a dedicated processing chain (Figure 1) which handles

NER and consolidation (enrichment, clustering, disambiguation) as part of one coherent workflow.

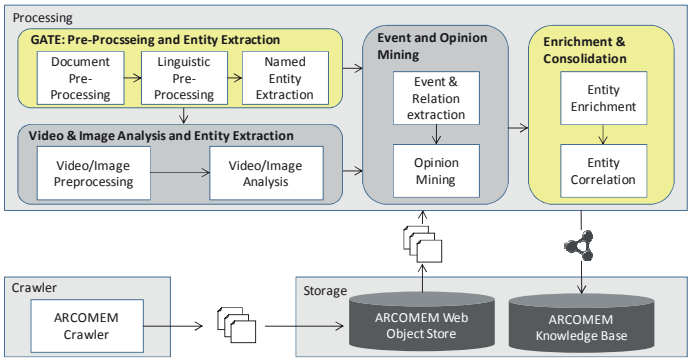


Fig. 1. Entity extraction and consolidation processing chain

The ARCOMEM storage composed of the *object store* and the *knowledge base* handles (a) binary data, in the form of Web objects, which represent the original content collected by the crawler; and (b) semi-structured data, in the form of RDF[6] triples (Web object annotations). Storage is based on a distributed solution that combines the MapReduce [9] paradigm and NoSQL databases and is realised based on HBase[7] (see also [25]). The *ARCOMEM data model*[8] provides an RDF schema to reflect the informational needs for knowledge capturing, crawling, and preservation (see [20] for details).

Within the ARCOMEM model, "entity" encompasses both traditional Named Entities and also single and multi-word terms: the recognition of both is done using GATE tools. GATE has been chosen over other NLP tools primarily for its coverage, extensibility and flexibility: it has a wide range of NLP components, which are easily modifiable for the demands of the project, unlike tools such as OpenCalais and DBPedia Spotlight which are more limited in scope. While extracted data is already classified and labelled as a result of the extraction process, it is nevertheless (i) heterogeneous, i.e. not well interlinked, (ii) ambiguous and (iii) provides only very limited information. This is due to data being extracted by different components and during independent processing cycles, since the tools in GATE have no possibility to perform co-reference on entities generated asynchronously across multiple documents. For instance, during one particular cycle, the text analysis component might detect an entity from the term “Ireland”, while during later cycles, entities based on the term “Republic of Ireland”' or the German term “Irland” might be extracted, together with, the entity “Dublin”. These would all be classified as entities of type *Location* and correctly stored in the data store as disparate entities described according to the data

[6] http://www.w3.org/RDF/
[7] Apache Foundation; The Apache HBase Project: http://hbase.apache.org/
[8] http://www.gate.ac.uk/ns/ontologies/arcomem-datamodel.rdf

model. Thus, *Enrichment and Consolidation* (Fig. 1) follows three aims: (a) *enrich existing entities* with related publicly available knowledge; (b) *disambiguation*, and (c) identify *data correlations* such as the ones illustrated above. This is achieved by mapping isolated entities to concepts (nodes) within reference datasets (*enrichment*) and exploiting the corresponding graphs to discover correlations. Therefore, we exploit publicly available data from the Linked Open Data cloud which offers a vast amount of data of both domain-specific and domain-independent nature (the current release consists of 31 billion distinct triples, i.e. RDF statements[9]).

4 Implementation

For entity recognition, we use a modified version of ANNIE [18] to find mentions of *Person*, *Location*, *Organization*, *Date*, *Time*, *Money* and *Percent*. We included extra subtypes of *Organization* such as *Band* and *Political Party*, and have made various modifications to deal with the problems specific to social media such as incorrect English (see [21] for more details). The entity extraction framework can be divided into the following components (GATE component in Fig. 1) which are executed sequentially over a corpus of documents:

- Document Pre-processing (document format analysis, content detection)
- Linguistic Pre-processing (language detection, tokenisation, POS tagging etc)
- Named Entity Extraction: Term Extraction (generation of ranked list of terms and thresholding) & NER (gazetteers, rule-based grammars and co-reference)

For term extraction, we use an adapted version of TermRaider[10]. This considers noun phrases (NPs) as candidate terms (as determined by linguistic pre-processing), and ranks them in order of termhood according to 3 different scoring functions: (1) basic tf.idf (2) an augmented tf.idf which also takes into account the tf.idf score of any hyponyms of a candidate term, and (3) the Kyoto score based on [22] which takes into account the number of hyponyms of a candidate term occurring in the document. All are normalised to represent a value between 0 and 100. A candidate term is not considered an entity if it matches or is contained within an existing Named Entity, to avoid duplication. Also, we have set a threshold score above which we consider a candidate term to be valid. This threshold is a parameter which can be manually changed at any time – currently it is set to an augmented score of 45, i.e. only terms with a score of 45 or greater will be used by later processes.

The entity extraction generates RDF data describing NEs and terms according to the ARCOMEM data model which is pushed to our knowledge base and directly digested by our *Enrichment & Consolidation* component (Fig. 1). The latter exploits (a) the *entity label* and (b) the *entity type* to expand, disambiguate and correlate extracted data. Note that an entity/event label might correspond directly to a label of

[9] http://lod-cloud.net/state
[10] http://gate.ac.uk/projects/arcomem/TermRaider.htmll

one unique node in a structured dataset (as is likely for an entity of type person labelled "Angela Merkel"), but might also correspond to more than one node/concept, as is the case for most of the events in our dataset. For instance, the event labeled "Jean Claude Trichet gives keynote at ECB summit" will most likely be enriched with links to concepts representing the ECB as well as Jean Claude Trichet. Our approach is based on the following steps (reflected in Fig. 1):

S1. *Entity enrichment*
- S1.a. Translation: we determine the language of the entity label, and, if necessary, translate it into English using an online translation service.
- S1.b. Enrichment: co-referencing with related entities in reference datasets.

S2. *Entity correlation* and clustering

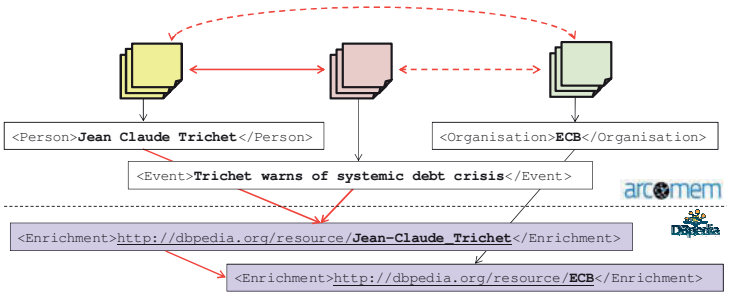
In order to obtain enrichments for these entities we perform queries on external knowledge bases. Our current enrichment approach uses DBpedia[11] and Freebase[12] as reference datasets, though it is envisaged to expand this approach with additional and more domain-specific datasets, e.g., event-specific ones. DBpedia and Freebase are particularly well-suited due to their vast size, the availability of disambiguation techniques which can utilise the variety of multilingual labels available in both datasets for individual data items and the level of inter-connectedness of both datasets, allowing the retrieval of a wealth of related information for particular items. In the case of DBpedia, we make use of the DBpedia Spotlight service which enables an approximate string matching with adjustable confidence level in the interval [0,1]. As part of our evaluation (Section 6), we experimentally selected a confidence level of 0.6 which provided the best balance of precision and recall. Note that Spotlight offers NER capabilities complementary to GATE. However, these were only utilised in cases where entities/events were not in a rather atomic form, as is often the case for events which mostly consists of free text descriptions such the one mentioned above.

Freebase contains about 22 million entities and more than 350 millions facts in about 100 domains. Keyword queries over Freebase are particularly ambiguous due to the size and the structure of the dataset. In order to reduce query ambiguity, we used the Freebase API and restricted the types of the entities to be matched using a manually defined type mapping from ARCOMEM to Freebase entity types. For example, we mapped the ARCOMEM type "person" to the "people/person" type of Freebase, and the ARCOMEM type "location" to the Freebase types "location/continent", "location/location" and "location/country". For instance, an ARCOMEM entity of type "Person" with the label "Angela Merkel" is mapped to the Freebase MQL query that retrieves one unique Freebase entity with the mid= "/m/0jl0g". With respect to data correlation, we distinct *direct* as well as *indirect* correlations. Please note, that a *correlation* does not describe any notion of equivalence (e.g. similar to *owl:sameAs*) but merely a meaningful level of relatedness.

Fig. 2 depicts both cases, direct as well as indirect correlations. Direct correlations are identified by means of equivalent and shared enrichments, i.e., any entities/events

[11] http://dbpedia.org/
[12] http://www.freebase.com/

sharing the same enrichments are supposedly correlated and hence clustered. A direct correlation is visible between the entity of type Person labeled "Jean Claude Trichet" and the event "Trichet warns of systemic debt crisis". In addition, the retrieved enrichments associate the ARCOMEM entities and associated Web objects with the knowledge, i.e., data graph, available in associated reference datasets.



Fig, 2..Enrichment and correlation example: ARCOMEM Web objects, entities/events, associated DBpedia enrichments and identified correlations

For instance, the DBpedia resource of the European Central Bank (*http://DBpedia.org/resource/ECB*) provides additional facts (e.g., a classification as organisation, its members, or previous presidents) in a structured, and therefore, machine-processable form. Exploiting the graphs of underlying reference datasets allows us to identify additional, *indirect correlations*. While linguistic/syntactic approaches would fail to detect a relationship between the two enrichments above (Trichet, ECB) and hence their corresponding entities and Web objects, by analysing the DBpedia graph we are able to uncover a close relationship between the two (Trichet being the former ECB president). Hence, computing the *relatedness* of enrichments would allow us to detect indirect correlations to create a relationship (dashed line) between highly releated entities/events, beyond mere equivalence.

Our current implementation is limited to detect direct correlations, while ongoing experiments based on graph analysis mechanisms aim to automatically measure *semantic relatedness* of entities in reference datasets to detect indirect relations. While in a large graph, all nodes are connected with each other in some way, a key research challenge is the investigation of appropriate graph navigation and analysis techniques to uncover indirect but semantically meaningful relationships between resources within reference datasets, and hence ARCOMEM entities and Web objects.

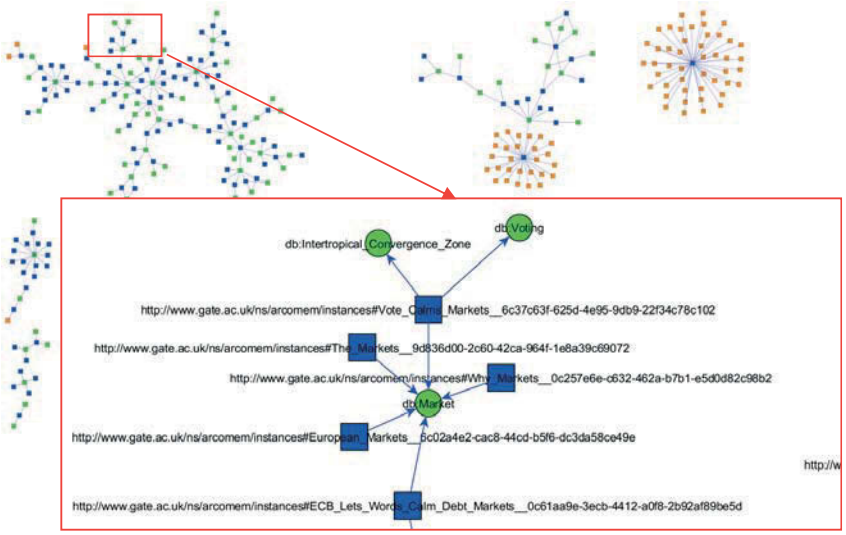
5 Results & evaluation

For our experiments, we used a dataset composed of English and German archived Web objects constituting a sample of crawls relating to the financial crisis[13]. The English content covered 32 Facebook posts, 41,000 tweets and 800 user comments from

[13] Parts of the archived crawls are available at http://collections.europarchive.org/arcomem/.

greekcrisis.net. The German content consisted of archived data from the Austrian Parliament[14] consisting of 326 documents (mostly PDF, some HTML).

Our extraction and enrichment experiments resulted in an evaluation dataset[15] of 99,569 unique entities involving the types *Event*, *Location*, *Money*, *Organization*, *Person*, *Time*. Using the procedure described above, we obtained enrichments for 1,358 of the entities in our dataset using DBpedia (484 entities) and Freebase (975 entities). In total, we obtained 5,291 Freebase enrichments and 491 DBpedia enrichments. These enrichments built 5,801 entity-enrichment pairs, 5,039 with Freebase and 492 with DBpedia.



Fig, 3. Generated ARCOMEM graph and clusters

5.1 Entity extraction evaluation

We have performed initial evaluations on the various text analysis components. We manually annotated a small corpus of 20 Facebook posts (in English) from the dataset described above with named entities to form a gold standard corpus. This contained 93 instances of Named Entities. For evaluating TermRaider, we took a larger set of 80 documents from the financial crisis dataset, from which, TermRaider produced 1003 term candidates (merged from the results of the three different scoring systems). Three human annotators selected valid terms from that list, and we produced a gold standard of 315, comprising each term candidate selected by at least two annotators (221 terms selected by exactly two annotators and 94 selected by all three). While inter-annotator agreement was thus quite low, this is normal for a term extraction task

[14] http://www.parliament.gv.at/

[15] The SPARQL endpoint of our dataset (extracted entities and enrichments) is available at http://arcomem.l3s.uni-hannover.de:9988/openrdf-sesame/repositories/arcomem-rdf?query.

as it is extremely subjective; however, in future we will tighten the annotation guidelines and provide further training to the annotators with the aim of reaching a better consensus.

For the NE recognition evaluation, we compared the system annotations with the gold standard. The system achieved a Precision of 80% and a Recall of 68% on the task of NE detection (i.e. detecting whether an entity was present or not, regardless of its type). On the task of type determination (getting the correct type of the entity (Person, Organization, Location etc.)), the system performed with 98.8% Precision and 98.5% Recall. Overall (for the two tasks combined), this gives NE recognition scores of 79% Precision and 67% Recall. However, the results are slightly low because this actually includes Sentence detection also. Normally, Sentence detection is 100% accurate (or near enough), but in this case, it is subject to the language detection issue, because we only perform the entity detection on sentences deemed to be relevant (in the language of the task and which corresponds to the relevant part of the document - in this case, the actual text of the postings by the users). 26 of the missing system annotations in the document were outside the span of the sentences annotated, so could not have been annotated. Excluding these increased Recall from 68% to 83.9% for NE detection (shown in the table as "NE detection (adjusted)"), and from 67% to 73.5% for the complete NE recognition task (shown in the table as "Full NE recognition (adjusted)").

Table 1. NER evaluation results

Task	Precision	Recall	F1
NE detection	80%	68%	74%
NE detection (adjusted)	80%	83.9%	81,9%
Type determination	98.8%	98.5%	98.6%
Full NE recognition	79%	67%	72.5%
Full NE recognition (adjusted)	79%	82.1%	80.5%

For term recognitions, we compared the TermRaider output for each scoring system with the gold standard set of terms, at different levels of the ranked list, as shown in Figure 4. For the terms above the threshold, we achieved Precision scores of 31% and Recall of 90% for tf.idf, 73% Precision and 50% Recall for augmented tf.idf and 63% Precision and 17% Recall for the Kyoto score. For any further processing, we only use the terms scored by the augmented tf.idf above the threshold.

5.2 Enrichment and correlation evaluation

For this evaluation we randomly selected a set of entity-enrichment pairs. Our evaluation was performed manually by 6 judges including graduate computer science students and researchers. The judges were asked to assign scores to each entity-enrichment pair, with “0” for *incorrect*, and “1” for *correct*. We judge an enrichment as correct if it partially defines a specific dimension of the entity/event, that is, an enrichment does not need to completely match an entity. For instance, enrichments

referring to *http://dbpedia.org/resource/Doctor_(title)* and *http://dbpedia.org/page/Angela_Merkel* and enriching an entity of type Person labelled "Dr Angela Merkel" were both equally ranked as correct. This is due to entities and events being potentially related to multiple enrichments, each enriching a particular facet of the source entity/event. Each entity/enrichment pair was shown to at least 3 judges and an average of their scores was built to alleviate bias. In case an entity label did not make sense to a judge, we assumed that there has been an error in the extraction phase. In this case we asked the judges to mark the corresponding entity as invalid and excluded it from the evaluation.

We computed the average scores of entity-enrichment pairs across judges and averaged the scores obtained for each entity type. Table 4 presents the average scores of the enrichment-entity pairs obtained using DBpedia and Freebase for different ARCOMEM entity types.

Table 2. Enrichment evaluation results

Entity Type	Avg. Score DBpedia	Avg. Score Freebase	Avg. Score Total
Location	0.94	0.94	0.94
Money	0.63	-	0.63
Organization	0.93	1	0.97
Person	0.72	0.89	0.8
Time	1	-	1
Total	**0.84**	**0.94**	**0.89**

Our initial clustering approach simply correlated entities/events which share equivalent enrichments. In total we generated 1013 clusters with 2.85 entities on average, with a minimum of 2 and a maximum of 112 entities. Ambiguous enrichments led to redundant clusters and require additional disambiguation. For instance, a location entity labelled "Berlin" might be (correctly) enriched with *http://rdf.freebase.com/ns/m/0xfhc* and *http://rdf.freebase.com/ns/m/047ckrl* (each referring to a different location "Berlin") requiring additional disambiguation to clean up the clusters. To this end, we exploit graph analysis methods to detect closeness of enrichments originating from the same object. For instance, measuring the relatedness of two location entities "Berlin" and "Angela Merkel" used to annotate the same Web object will allow us to disambiguate enrichments.

6 Discussion and future works

In this paper we have presented our current strategy for entity extraction and enrichment as realised within the ARCOMEM project, aimed at creating a large knowledge base of structured knowledge about archived heterogeneous Web content. Based on an integrated processing chain, we tackle entity consolidation and enrichment as implicit activity in the information extraction workflow.

The results of the entity extraction show respectable scores for this kind of social media data on which NLP techniques typically struggle. However, current work is

focusing on better handling of degraded English (tokenisation, language recognition etc) and especially of tweets, which should improve the entity extraction further. The enrichment results indicate a comparably good quality of generated enrichments. The results obtained from DBpedia Spotlight provided a lower recall, but introduced less ambiguous enrichments due to Spotlight's inherent disambiguation feature. On the other hand, partially matched keywords reduce the precision results. As future work, we foresee different directions to improve quality of the enrichment results. For example, one possibility is to use structured DBpedia queries to restrict entity types, similar to the approach used for Freebase. We also consider the introduction of subtypes of entities to further increase granularity of the types to be matched.

In addition, while preservation of Web content over time has to consider temporal aspects, evolution of entities and terms as well as time-dependent disambiguation are important research areas currently under investigation [24]. While our current data consolidation approach only detects direct relationships between entities sharing the same enrichments, our main efforts are dedicated to investigate graph analysis mechanisms. Thus, we aim to further take advantage of knowledge encoded in large reference graphs to automatically identify semantically meaningful relationships between disparate entities extracted during different processing cycles. Given the increasing use of both automated NER tools and reference datasets such as DBpedia, WordNet or Freebase, there is an increasing need for consolidating automatically extracted information on the Web which we aim to facilitate with our work.

Acknowledgments

This work is partly funded by the European Union under FP7 grant agreement n° 270239 (ARCOMEM).

References

[1] Bizer, C., T. Heath, Berners-Lee, T. (2009). Linked data - The Story So Far. Special Issue on Linked data, International Journal on Semantic Web and Information Systems.

[2] Dietze, S., and Domingue, J. (2008) Exploiting Conceptual Spaces for Ontology Integration, Workshop: Data Integration through Semantic Technology (DIST2008) Workshop at 3rd Asian Semantic Web Conference (ASWC) 2008, Bangkok, Thailand.

[3] Dietze, S., Gugliotta, A., and Domingue, J. (2009) Exploiting Metrics for Similarity-based Semantic Web Service Discovery, IEEE 7th International Conference on Web Services (ICWS 2009), Los Angeles, CA, USA.

[4] Lü, L., Zhou, T.: Link prediction in complex networks: a survey, Physica A 390 (2011), 1150–1170.

[5] Hasan, M. A., Zaki, M. J.: A survey of link prediction in social networks. In C. Aggarwal, editor, Social Network Data Analytics, pages 243–276. Springer,

[6] Cohen, W. W., Ravikumar, P. D., Fienberg, S. E.. A comparison of string distance metrics for name-matching tasks. In IIWeb, 2003.

[7] Dong, X., Halevy, A., Madhavan, J., Reference reconciliation in complex information spaces. In SIGMOD, 2005.

[8] Elmagarmid, A. K., Ipeirotis, P. G., Verykios, V. S., Duplicate record detection: A survey. TKDE, 19(1), 2007.
[9] Tejada, S., Knoblock, C. A., Minton, S., Learning domain-independent string transformation weights for high accuracy object identification. In KDD, 2002.
[10] Boley, D., Principal Direction Divisive Partitioning. Data Mining and Knowledge Discovery, 2(4), 1998.
[11] Broder, A., Glassman, S., Manasse, M., Zweig, G., Syntactic Clustering of the Web. In Proceedings of the 6th International World Wide Web Conference, pages 1997.
[12] Hotho, A., Maedche, A., Staab, S., Ontology-based Text Clustering. In Proceedings of the IJCAI Workshop on \Text Learning: Beyond Supervision", 2001.
[13] Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y., Named Entity Recognition from Diverse Text Types. Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark, Bulgaria, 2001
[14] Li, Y., Bontcheva, K., Cunningham, H., Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Natural Language Engineering, 15(02), 241-271, 2009.
[15] Buckley, C., G. Salton, G., Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, 1988.
[16] Maynard, D., Li, Y., Peters, W., NLP techniques for term extraction and ontology population. In: Buitelaar, P. and Cimiano, P. (eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 171-199, IOS Press, Amsterdam (2008)
[17] Lui, M., Baldwin, T., 2011. Cross-domain feature selection for language identification. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 553–561, November.
[18] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).
[19] Ritter, A., Clark, S., Mausam, Etzioni, O., 2011. Named entity recognition in tweets: An experimental study. In Proc. of Empirical Methods for Natural Language Processing (EMNLP), Edinburgh, UK
[20] Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., Senellart, P., Exploiting the Social and Semantic Web for guided Web Archiving, The International Conference on Theory and Practice of Digital Libraries 2012 (TPDL2012), Cyprus, September 2012.
[21] Maynard, D., Bontcheva, K., Rout, D., Challenges in developing opinion mining tools for social media. In Proceedings of @NLP can u tag #user*generated*content?! Workshop at LREC 2012, May 2012, Istanbul, Turkey.
[22] Bosma, W., Vossen, P., 2010. Bootstrapping languageneutral term extraction. In 7th Language Resources and Evaluation Conference (LREC), Valletta, Malta.
[23] Deane, P. A nonparametric method for extraction of candidate phrasal terms, In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005.
[24] Tahmasebi, N., Risse, T., Dietze, S. (2011) Towards Automatic Language Evolution Tracking: A Study on Word Sense Tracking, Joint Workshop on Knowledge Evolution and Ontology Dynamics 2011 (EvoDyn2011), at the 10th International Semantic Web Conference (ISWC2011), Bonn, Germany.
[25] Weiss, C., Karras, P., Bernstein, A., Hexastore: sextuple indexing for semantic web data management. Proceedings of the VLDB Endowment, 1(1):1008–1019, 2008.

Do we need metadata? An on-line survey in german archives

Marcel Ruhl

University of Applied Science Potsdam, Faculty of Information Sciences,
Friedrich-Ebert-Str. 4, 14467 Potsdam, Germany
ruhl@fh-potsdam.de
http://iw.fh-potsdam.de

Abstract. The paper summarizes the results of an on-line survey which was executed 2010 in german archives of all branches. The survey focused on metadata and used metadata standards for the annotation of audiovisual media like pictures, audio and video files (analog and digital). The findings motivate the question whether archives are able to collaborate in projects like europeana if they do not use accepted standards for their orientation. Archives need more resources and archival staff need more training to execute more complex tasks in an digital and semantic surrounding.

Keywords: on-line survey, metadata, archives, audiovisual, semantic web

1 Introduction

Stefan Gradmann said in his inaugural lecture at the Humboldt-University Berlin that digital information objects should be understandable without reading them all [1]. This statement is gaining in weight. Through the power of the Internet more and more information are available, which are crawled with search algorithms. These generate a large number of hits. The resulting amount of information are often unmanageable for the seekers. The location and time independent web resources are accompanied by localized and time-based accessible information resources. This archival material is only accessible to users via analog finding aids. In the context of the semantic web both described kinds of information are information silos. The reason for this is the missing link between the individual stocks. To bring these silos together one requirement is well-formed, standard based metadata [2], [4]. This is also against the background of the development of portals like europeana relevant. For though german archives provide many million on-line metadata objects, it seems that the majority of archives is not deep in this topic. Reason enough to ask precisely what the situation is in the archives. Because it is mandatory for most portals like europeana that the provided metadata describes objects with digital representations like scanned documents [3], this paper focuses on multimedia objects.

The objective was to gain information about metadata for audio-visual objects in archives, how relevant is metadata and how is it used at the moment. The needs and deficiencies of german archives should be determined. To reach these goals an on-line survey was designed.

The paper summarizes the results of an on-line survey from the year 2010. German archives of all branches were asked about their use of metadata. The text is structured as follows. **Section 2 On-line Survey** describes design and execution of the survey. **Section 3 Interpretation and Discussion** summarizes the answers to selected questions. Tables and diagrams are used for visualization. The last **Section 4 Conclusion** lists the main findings and gives a forecast to possible conclusions to change the founded situation.

2 On-line Survey

The survey was conducted from 28.10.2010 until 12.11.2010. It was designed as an open on-line survey. The high number of potential participants spoke for this decision. A paper based survey couldn't be analyzed in appropriate time. Institutions were invited via email to complete the on-line questionnaire. The survey was created with LimeSurvey [5], an open source survey application.

The survey was not personalized, no login or password was needed. The implied problem of multiple attendance was solved by using the possibility of using cookies. So the questionnaire couldn't be completed from one workstation several times.

The participants had to answer 28 questions (one-choice, multiple-choice and free text), from which some were based on the answers of previous ones. Topically the questions covered the archived media (analog and digital pictures, audio and video), used metadata standards and the participation in projects for metadata exchange.

The institutions were chosen from an database provided by [6]. It contains 2733 datasets with addresses from german archives. Not all of them had an email-address, so 2056 institutions from all kinds of archives could be contacted. 191 email-addresses were invalid, so that 1865 Institution could have participate. Within the survey period an reminder email was send on 09.11.2010.

3 Interpretation and Discussion

From the above mentioned 1865 institutions 873 institutions attended the survey, but 485 participants stopped before the end and didn't finish the survey. Altogether 388 complete data sets were created and could be analyzed. The return rate was 46.81% and the drop-out rate was 55.56%, so 20.8% of the potential participants finished the survey.

After answering the first optional question about the name of the institution, the second question was about the archival branch. 18 (4.64%) state archives, 244 (62.89%) city archives, 31 (7.99%) church archives, 4 (1.03%) nobility archives,

25 (6.44%) economy archives, 14 (3.61%) parliament archives, 16 (4.12%) media archives, 29 (7.47%) academy archives and 7 (1.80%) free archives took part.

The following questions were about the media objects the institutions archive. The survey asked separate for analog and digital pictures, audio- and video-objects. See table 1 for details. The participated archives have a huge amount of analog pictures. 64.95% store between 1001 and 100000 objects. The quantity of analog audio- and video-objects is rather low. 41.75% (audio) respectively 44.33% (video) of the participants store between 11 and 100 analog objects. Most of the archives have a little amount of digital media. Especially the number of digital audio- and video-objects is mostly below 100 items. It seems that this kind of archival objects hasn't arrived in the archives yet or that existing analog objects are not broad digitized.

	0	1-10	11-100	101-1000	1001-10000	10001-100000	>100000
analog pictures	4.12%	1.29%	3.35%	13.40%	37.89%	27.06%	12.88%
analog audio	12,63%	18.04%	41.75%	21.91%	3.61%	1.55%	0.52%
analog video	11.34%	18.04%	44.33%	19.85%	4.64%	1.29%	0.52%
digital pictures	14.69%	3.61%	9.28%	22.16%	26.80%	16.75%	6.70%
digital audio	22.94%	19.33%	35.57%	18.56%	1.80%	1.03%	0.78%
digital video	21.13%	24.23%	35.57%	15.98%	2,58%	0.52%	0%

Table 1. How many objects are stored in the archive?

Up to this point, all participants saw the similar questions. The following question number five (Is metadata captured?) was designed to exclude institutions which do not capture metadata from the following block. Answering with NO the participants could not see the following questions. This design was chosen, because it was supposed that this institutions could not answer the questions concerning the metadata elements and used standards. This decision was possibly the reason for the high drop-out rate (see above), because the following sites of the survey were shown empty. The participants had to click several times to get to the last block with the comment field. 227 (58.51%) answered this question with YES and 161 (41.49%) answered with NO.

92.51% of the institutions which capture metadata for audiovisual objects do this manually, 17 participants (7.49%) capture the metadata automatically. After that the survey asked about the used metadata fields for descriptive, administrative, technical and structural metadata for all kinds of objects (analog and digital). The tables 2 and 3 show the fields mostly used. Technical metadata is captured depending on the kind of the object. For pictures the color (148) and the file-format (104), for audio the medium (99) and the material (71) and for video the color (111) and the film-format (95) are the most often used metadata fields. In all three categories many institutions answered, that they don't collect metadata. The exact numbers are for pictures (32), audio (58) and video (49).

The captured fields for structural metadata are the same in all three categories, only the number of mentions differs. Table 4 gives further information.

Pictures (mentions)	Audio (mentions)	Video (mentions)
pictured person (200)	date (134)	date (144)
pictured object (200)	year of publication (127)	year of publication (137)
date (194)	title (119)	title (125)
photographer (186)	description of contents (110)	duration (123)
place (183)	duration (105)	description of contents (120)
description of contents (148)	place (85)	original title (95)
year of publication (143)	keyword (66)	place (94)
title (138)	original title (66)	producer (77)
keyword (117)	speaker (53)	keyword (71)
country (41)	producer (51)	director (52)
genre (35)	original language (30)	actors (41)
language (10)	language (25)	film location (38)
remarks (3)	director (25)	language (31)
event (2)	genre (23)	genre (30)
others (3)	others (23)	others (75)
none (4)	none (35)	none (31)

Table 2. Which descriptive metadata do you collect?

Pictures (mentions)	Audio (mentions)	Video (mentions)
signature (174)	signature (130)	signature (135)
creator (162)	provenience (105)	creator (119)
provenience (156)	creator (98)	provenience (112)
terms of use (82)	terms of use (58)	terms of use (71)
references (71)	retention period (37)	references (40)
retention period (53)	references (32)	retention period (29)
license (45)	availability (30)	availability (27)
others (80)	others (32)	others (49)
none (15)	none (50)	none (43)

Table 3. Which administrative metadata do you collect?

When asked about the use of metadata standards for the annotation of objects, 79.30% (180) answered not to use standards for their guidance. Just 47 institutions (20.70%) used standards like IPTC (19), ISAD-G (16) or EAD (12)[1](see table 5). Furthermore, in this context, the question was asked, what

[1] Multiple-choise was possible.

	Pictures	Audio	Video
Holding	164	118	118
Series	88	61	59
Sequence	37	39	37
Classification	5	2	2
None	36	66	69

Table 4. Which structural metadata do you collect?

reasons were there for non-use of standards. Mentioned reasons were lack of resources, higher priority for classical records and lack of information on standards. In the comment field participants mentioned additional the annotation before the introduction of appropriate standards, the use of its own regulations and the low holding size.

standard	mentions	standard	mentions
IPTC	19	FIAF cataloguing rules	2
ISAD-G	16	RNA	2
EAD	12	EAC-CPF	1
EXIF	9	METS	1
Dublin Core	8	MIDAS	1
RAK (NBM)	7	PND	1
Regelwerk ARD-ZDF	3	GKD	1
in-house guidelines	3	SWD	1
MAB	2	XMP	1
MAB2	2	others	5
MPEG7	2		

Table 5. Which standards are used for annotation?

An major intention of the survey was to find out if archives are participating in projects which focus on the exchange of metadata. The survey shows that 196 institutions are not attending in such projects. 11 institutions provide metadata in the BAM-portal [8] and 6 in europeana [7]. For details see figure 1. The participants which do not participate in projects yet, were asked if it is planned in the future. Here 161 (82.14%) participants answered with NO. Just 35 (17.86%) said YES (see figure 2). Against the background of the intentions of metadata exchange this is an very bad outcome. At this point the answers differ depending on the archival branch. Most of the smaller archives like city archives are not planning the participation in portals. Reasons are the same mentioned above. Institutions with better resources like state archives can afford the commitment of human and financial costs easier.

At the end of the survey respondents had the option to enter their notes and comments into a comment field. Here a large number of participants pointed out

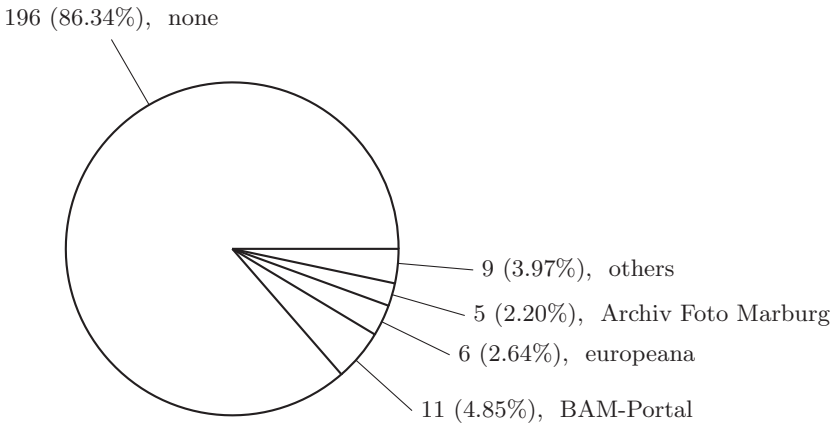


Fig. 1. In which project (metadata exchange) are you participating?

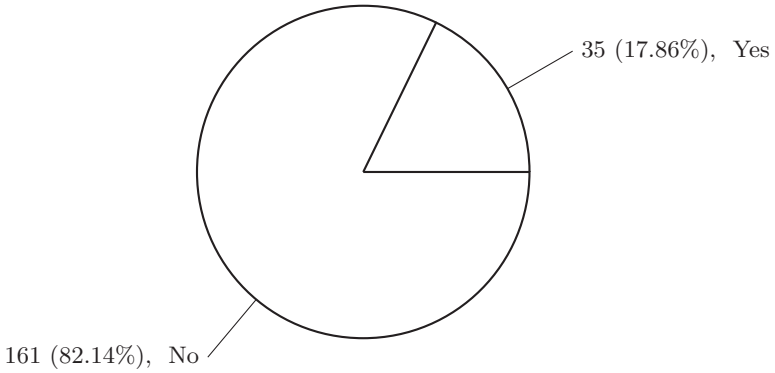


Fig. 2. Is a participation in projects (metadata exchange) planned?

that the archives have to deal with a lack of human and financial resources and therefore no opportunities for the annotation of audiovisual media are present. Archives focus on analog records at the moment. Another frequently mentioned issue is the ambiguity of terms such as *metadata* and *audiovisual*. Here training and understandable guidelines on this topic are demanded.

4 Conclusion

The paper summarized the results of an on-line survey which was executed 2010 in german archives of all branches. The survey found that the issue of metadata for audiovisual objects, metadata standards and their exchange plays a minor role for most of the german archives. Though archives are professionals in making classical records accessible, some archival branches like nobility archives or small city archives somehow can not use this professionalism for archival material with a technical smell. The question has to be asked if, under this results, german archives could be interested in semantic web technologies if they have not the ability to annotate their objects close to accepted international standards and are not planning to share their metadata. This can only be attributed to a lack of knowledge of the subject and its benefits to the archival landscape or to an extreme lack of personnel, temporal and financial resources in the archives. This shortage could be relieved by increased training and advertising of this issue. An other possible solution could be the wider use of pool resources for archival issues like indexing.

References

1. Gradmann, St.: Signal. Information. Zeichen. Zu den Bedingungen des Verstehens in semantischen Netzen. (2008), http://edoc.hu-berlin.de/humboldt-vl/157/gradmann-stefan-3/PDF/gradmann.pdf
2. Hitzler, P., Krötzsch, M., Rudolph, S., Sure, Y.: Semantic Web - Grundlagen. Springer, Berlin (2008)
3. europeana Aggregators Handbook (2011), http://pro.europeana.eu/documents/858566/858665/Aggregators+Handbook
4. Heath, T., Bizer, Ch.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, San Rafael (2011), http://linkeddatabook.com/editions/1.0/
5. LimeSurvey - the open source survey application, http://www.limesurvey.org/
6. Verband deutscher Archivarinnen u. Archivare: Archive in Deutschland, Österreich und der Schweiz. Ein Adressverzeichnis. 20. aktual. Aufl., Ardey, München (2009)
7. europeana.eu, http://europeana.eu/portal/
8. Gemeinsames Portal für Bibliotheken, Archive und Museen - ein Online-Informationssystem, http://www.bam-portal.de/

Automatic classification of scientific records using the German Subject Heading Authority File (SWD)

Christian Wartena and Maike Sommer*

Hochschule Hannover - University of Applied Sciences and Arts
Department of Information and Communication
Expo Plaza 12, 30539 Hannover, Germany
Christian.Wartena@fh-hannover.de
Maike.Sommer@stud.fh-hannover.de

Abstract. The following paper deals with an automatic text classification method which does not require training documents. For this method the German Subject Heading Authority File (SWD), provided by the linked data service of the German National Library is used. Recently the SWD was enriched with notations of the Dewey Decimal Classification (DDC). In consequence it became possible to utilize the subject headings as textual representations for the notations of the DDC. Basically, we we derive the classification of a text from the classification of the words in the text given by the thesaurus. The method was tested by classifying 3826 OAI-Records from 7 different repositories. Mean reciprocal rank and recall were chosen as evaluation measure. Direct comparison to a machine learning method has shown that this method is definitely competitive. Thus we can conclude that the enriched version of the SWD provides high quality information with a broad coverage for classification of German scientific articles.

1 Introduction

Subject classification is one of the major pillars to guarantee accessibility of records in large digital libraries. One of the worldwide most common classification systems is the Dewey Decimal Classification (DDC). The DDC is a universal classification system aiming at representing the entire knowledge of the world. It is used in more than 135 countries and translated into over 30 languages. More than 60 countries use the DDC even for their national bibliography. Apart from the worldwide use the DDC has a second strength: It is administrated by the Decimal Classification Editorial Policy Committee at the Library of Congress. Thus it is updated and developed continuously ([12]).

Classifying records according to DDC is a task that requires carefully reading and understanding of the abstracts and other available meta data as well as a

* This work is partially based on the Bachelor thesis of Maike Sommer ([17]).

detailed knowledge of the DDC class hierarchy. In a number of projects classifiers were built using machine learning techniques ([20], [19]). These approaches are problematic because the DDC-classes are very fine grained. Even in very large repositories, for most classes there are not sufficient training data. Thus a classification can only be made on the highest levels of the DDC hierarchy and even then the sparsity of data poses still a problem for some of the classes ([19]). A second problem is the dependency on training data. Especially, the data that are classified have to be comparable to the data that were used for training. E.g. if the collection to be classifies contains other text types than those used for training, the results might be worse than expected.

The basic principle of text classification based on machine learning is as follows. In the training phase words are given weights indicating how strong they characterize a certain class. During classification these weights are used to guess the most likely class for a text. Instead of determining weights in a training phase we can use a dictionary or thesaurus, if it contains information on the relation between words and the target classes, in our case the classes from the DDC. Recently, a large number of relations between subject headings of the German Subject Heading Authority File (*Schlagwortnormdatei*, SWD) and DDC-classes have been published ([4], [9]). Since most subject headings consist of just a single word or a very short phrase, we can use the SWD as a large lexical resource with a very broad coverage. Now, basically by counting the links of the subject headings found in a text to the DDC-classes we can predict the DDC-class for the text. The disadvantages of this approach are manifest: Weights are just 0 or 1, without any information how indicative a word is for a certain class. Furthermore, only the weights of the words are used and no dependencies between words can be modeled. The method has, on contrary, also the advantage that we need no training data and we directly can classify documents in domains that we did not see before. The success of this approach depends crucially on the quality of the thesaurus used. The main contribution of this paper is, that we show that the SWD is a vary valuable source of information in this respect.

In the following we will describe our approach in more detail and present results on the classification of the German language records from 7 repositories of different German universities. We compare our results with the results given by the Automatic Classification Toolbox for Digital Libraries (ACT-DL) from the University of Bielefeld (`http://clfapi.base-search.net/doc/index.html`), that uses a state-of.the-art machine learning approach. We show that the results are comparable in terms of mean reciprocal rank and in most cases better in terms of recall. The first measure is important with regard to fully automatic classification, the second measure is especially important in an interactive scenario in which the algorithm provides suggestions to a librarian.

The remainder of this paper is organized as follows. In section 2 we discuss related work. In section 3 we present our approach. In section 4 we describe the data we have used in our experiment, the results of which are given in section 5 We conclude with a discussion of results (section 6) and outlook to future work (section 7).

2 Related Work

Waltinger et al. ([19]) treat exactly the same problem that we discuss in this paper, namely classifying English and German scientific abstracts into high-level DDC classes. They use a state-of-the art machine learning approach for text classification. Below we will compare our results of the ontology driven approach directly with the results obtained by their classifier, that is publicly available as a web service.

Various studies consider document labeling or classification with the labels of an ontology, using lexical and structural information from that ontology. Basically, occurrences of ontology concepts in the text are counted and in some manner the information is aggregated to determine the most central or important concepts. Usually these approaches require enrichment of the ontology with additional lexical information, in many cases obtained from WordNet. Examples of such approaches are [18], [14] and [7].

Another approach to using ontologies for text classification is to enrich the representation of the text with features derived from the ontology, like hypernyms or concept labels before applying the classification algorithms. E.g., Scott and Matwin ([16]) add WordNet hypernyms and Bloehdorn and Hotho ([3]) add hypernyms from Wordnet and other ontologies to the representation of the text. The latter authors also try out various disambiguation strategies for words that potentially represent more than one ontology concept. Improvements over the baseline using only the words from the text are in both cases not very convincing.

Addis et al. ([2]) consider text classification rather as a two step process. In the first step WordNet concepts (*synsets*) are extracted. In the second phase an existing mapping from synsets to DDC-categories is used to compute a DDC-classification for the text. However, the authors do not consider this process as their final classifier, but use it only to create text collections to train statistical classifiers on. The two step approach is treated more systematically by Chenthamarakshan et al.([5]), who explicitly distinguish between the process of finding representative concepts on the one hand side and learning a mapping from concepts to document classes on the other hand side. These approaches differ not only in the perspective on the task from those mentioned in the previous paragraph. They are also different because they do not add features to the simple word vector model, but replace the original representation.

In our approach we consider classification as a two step approach as well. Thus the main contribution of this paper is not the method presented, but rather the investigation in the potentials of the German Subject Heading Authority File, that was, to the best of our knowledge, not used for automatic classification before. Since the results turn out to be very competitive, the proposed method might also have practical value for application in libraries.

3 Approach

In our thesaurus based approach, the most relevant Dewey class for a text is determined by the Dewey classification of the words in this text according to the

thesaurus. In our case the thesaurus is the German Subject Heading Authority File (SWD) for which all terms have been related to Dewey classes.

In order to find all relevant words we stem all words in the text. To be sure that only relevant words are found we restrict our search for thesaurus terms to nouns only. The text analysis is implemented as a GATE pipeline ([6]). For stemming and part-of-speech tagging we use the TreeTagger ([15]). Search for thesaurus terms is implemented by Apolda ([21]).

In the next phase we can determine the class of the text on the basis of the identified occurrences of thesaurus terms. For this phase we keep only unambiguous words, since only these terms give a clear indication of the topic of the text. Usually enough unambiguous terms remain to determine the topic of the text. We consider a word as unambiguous if the word occurs as the label of only one subject in the thesaurus, or if the word is the preferred label of exactly one subject. E.g. the word *Student* occurs 9 times as an alternative label for subjects like *Studentenwohnheim* (student accommodation) or *Auslandsstudium* (study abroad). However, there is one subject that has *Student* as its preferred label. Thus we treat *Student* as a non-ambiguous term representing that subject. The word *Untersuchung* (investigation) in contrast is found 2 times as an alternative label but never as a preferred label. Thus this word is not considered in the following steps. In this way many very general terms are filtered out.

Once all subjects have been identified we count the Dewey classes they are related to. In the (enriched) SWD each word is related to one or more Dewey classes via an anonymous node. For each relation a confidence of correctness is given by an integer between 1 and 4. For our purposes we ignore all links with a confidence level of 1. Given a (non-ambiguous) term occurrence t we let $ddc(t)$ be the set of all DDC-classes that t is related to with a confidence level greater than 1. Most words are related to very specific class in DDC. In order to aggregate occurrence information on a higher level in the DDC-hierarchy we denote for each class c in the DDC-System the broader class at the n-th level as c^n. Since the DDC-system is a strict hierarchy c^n is uniquely defined for each class with a depth smaller than n. E.g. if c is the class 342.0684 then c^2 is 340. Now we can define the contribution of a term t to each DDC-class c as

$$w(t,c) = \frac{|\{c_i \in ddc(t) \mid c_i^n = c\}|}{|\{c_i \in ddc(t)\}|} \tag{1}$$

where n is the hierarchy level of c. Considering a text T as a set of term occurrences we define the weight of a class c for T as

$$w(T,c) = \sum_{t \in T} w(t,c). \tag{2}$$

This gives us almost a ranking of DDC-classes for a text T. Only in case two classes have the same weight we need to specify their ranking. In these cases we order the classes by the order of their first occurrences in the text, where an earlier occurrence implies a higher rank.

4 Data and experimental setup

The experiment we present here was enabled by the results of the CrissCross project conducted by the Cologne University of Applied Sciences (Fachhochschule Köln) in collaboration with the German National Library (Deutsche Nationalbibliothek). In this project a concordance between the German Subject Heading Authority File (Schlagwortnormdatei, SWD) and the DDC was constructed. In other words the subject headings were mapped to notations of the DDC. The SWD is a universal indexing language based on rules, namely the rules for the subject catalog (Regeln für den Schlagwortkatalog, RSWK) and the practice rules for the RSWK and the SWD. In contrast to the DDC there are not that many relations between the subject headings. In accordance with an unpublished study from 2004 almost 87 % of the subject headings do not have associative relations. Furthermore 34% have neither associative nor hierarchical relations. The enrichment of the SWD with DDC notations is helpful in structuring the SWD because it generates hierarchical, equivalence and associative relations through similar DDC notations. We already mentioned that there were not were not many relations between the subject headings before the Criss-Cross project ([11]). Thus a subject heading can be interpreted differently. The project group mostly mapped one subject heading to several DDC notations ([10]). Furthermore, the meaning of a subject heading is often very specific. Therefore the mapped DDC notations are also very specific, which means mappings to a deep hierarchy level. Hence this is called deep level mapping ([11]). In our experiment we only wanted to gain notations up to the second hierarchy level, that can easily be obtained through the DDC hierarchy as explained above. Furthermore, there is much variance to what extent a subject heading fits into a DDC class. To express this distinction, the project group invented four confidence levels (degrees of determinacy) with 1 for the lowest and 4 for the highest congruency. As aforementioned we disregarded all first level relations, because these mappings point to DDC notations with only a small thematic intersection ([1]).

The released version of the enriched SWD (`https://wiki.d-nb.de/download/attachments/34963694/SWD_s_rdf.zip`) has about 188,000 concepts linked to 51,748 DDC-classes. The concepts have preferred and alternative labels. These labels are however labels of subject headings and not intended to be used as a lexical resource for analyzing texts. Some concepts have labels that are very unlikely to appear in running texts. However, in many cases the terms are single words or small phrases that will appear in normal texts.

More problematic are however concepts that have labels that will occur in many texts for which the concept is not relevant. This can be the case with words that have a meaning that is related to some subject area but that also can be used in a more general way. E.g. the word *Zusammenhang* can be used in a general way, meaning *context* or *connection*, but it is also the alternative label of *Zussamenhang in einer Mannigfaltigkeit* (Connectedness in a manifold) that is mapped correctly onto the Dewey class 516.35 (Algebraic geometry). Another class of words causing problems in a similar way are the homographs and homonyms. E.g. the abbreviation *ALS* (for the disease amyotrophic lateral

sclerosis) is a homograph for the very frequent conjunction *als* (as). A number of these homographs can be filtered out, because the different meanings correspond to different parts of speech. As mentioned before we only consider words from the text that were tagged as noun. Another example is constituted by the word *IM* that is an alternative label of the term *Spitzel* (spy), since it is the abbreviation of *Informeller Mitarbeiter* (informal staff), especially for the intelligence department of the GDR. Its homograph *im* is a highly frequent word that is the contraction of the words *in dem* (in the). Furthermore, many auxiliaries and function words are included in the category linguistics. In order to avoid problems with these words we removed all concepts from the class 435 (German grammar) except for the subject headings *rational*, *irrational* and *Gloria* because they are mapped into a second DDC class apart from 435. Also the subject heading Grammis was not removed because it is not a stop word but the abbreviation for grammatical information system (Grammatisches Informationssystem des IDS (*Institut für Deutsche Sprache*, Institute for German language)) Additionally we removed all concepts with a question mark ("?") as preferred label and the following alternative labels: *im* and *in* (as abbreviation of *intelligentes Netz* (intelligent net) as an extension of a telephone network). After that we could use the subject headings as textual representations for the DDC-classes. In sum 314,287 preferred and alternative labels could be used as textual term representations.

Table 1. OAI-Metadata repositories used in this paper. From each repository all records of publications in German language with an abstract available at the date of retrieval were used.

URL	University	#records	date of retrieval
`http://opus.bsz-bw.de/fhhv/oai2/oai2.php`	Hanover UAS	271	2012-05-27
`http://opus.bibl.fh-koeln.de/oai2/oai2.php`	Cologne UAS	254	2012-06-13
`http://opus.bsz-bw.de/fhff/oai2/oai2.php`	Frankfurt am Main UAS	120	2012-06-13
`http://opus.kobv.de/tuberlin/oai2/oai2.php`	TU Berlin	2036	2012-06-13
`http://opus.bsz-bw.de/ubhi/oai2/oai2.php`	Univ. Hildesheim	97	2012-06-13
`http://www.opus-bayern.de/uni-regensburg/oai2/oai2.php`	Univ. Regensburg	790	2012-06-15
`http://opus.bsz-bw.de/phfr/oai2/oai2.php`	Freiburg Univ. of Education	258	2012-06-14

For testing the effectiveness of the proposed classification strategy we have used in the first place the repository of the Hochschule Hannover - University

of Applied Sciences and Arts. This repository supports the Open Archives Initiative Protocol for Metadata Harvesting ([13]). We have classified metadata records of this repository using different fields, like title, abstract and keywords at the first and second level of the DDC-hierarchy. In most realistic scenarios one will have the title and the abstract of a publication that has to be classified, but not keywords. Thus we concentrated on classification using title and abstract. Besides the repository of the Hochschule Hannover, we used 6 more repositories. The repositories were chosen on the basis of the presence of an OAI-PMH interface, the size of the repository and the availability of the required metadata and classification. We selected three repositories from universities (among which one technical university), three universities of applied sciences and one university of education. Details of the repositories are given in Table 1.

Since the SWD is a German resource, we are only interested in publications in German. Thus we have selected from the repositories only those publications that are marked explicitly as written in German. However, most German publications have German and English abstracts. We did not include a language detection but simply assumed that the first abstract is the German one. We did not find any counterexample. The universities have an emphasis on fundamental research and are internationally oriented. Hence they publish mainly in English. The majority of their German publications are PhD-theses. In contrast the universities of applied sciences (UAS) are regional oriented and have an emphasis on knowledge transfer. Moreover, they usually don't have PhD-Students. Thus, they have a lot of publications in German that are intended to inform professionals in industry about new research and developments. The Universities of Education have a position in between with PhD-theses but also a lot of other German publications.

Each of the repositories mentions a subject area of the publication. For all repositories this is a DDC class at the second hierarchy level. However, some of the repositories use the class 004 for computer science. We did not take this exception into account. All records with this label consequently will have a recall and mean reciprocal rank of 0 for every classification method.

5 Results

All analyzed records provide a subject area that is in fact a second level DDC notation. It has to be noted that in many cases there is more than one possible label that could be regarded as true and a more or less arbitrary choice had to be made by the annotators. In fact labels closely related to the ground truth could be considered as correct as well ([8]). Furthermore, on closer inspection of the results, it turns out that in some cases of mismatch, the predicted label is the correct one and the label given by the repository was wrong ([17]). In the following we will nevertheless use these labels as the ground truth and consider only exact matches as being correct.

Since we consider assignment of DDC Notations as a classification task, in which each record should be assigned to exactly one category we have to observe the results for each record and not for each category, like one would do

Table 2. Mean reciprokal rank (MRR) and recall at 5 at first and second DDC level for SWD based classification of OAI Metadata from the SerWiss repository of the Hochschule Hannover using different fields and two classification methods, sci. the SWD-based classification and the ACT-DL classification service.

	SWD Based		ACT-DL	
Fields	MRR	rec@5	MRR	rec@5
title + abstract (DDC level 1)	0.68	0.89	0.67	0.90
title + abstract (DDC level 2)	0.48	0.66	0.39	0.37
title + keywords (DDC level 2)	0.61	0.76	0.32	0.39
title + abstract + keywords (DDC level 2)	0.61	0.77	0.39	0.47

in a retrieval setting. Thus the evaluation presented here differs from the one used in [19] who use the retrieval perspective. Since the algorithm produces a ranked list of results, we use mean reciprocal rank as an evaluation measure. Automatic classification might be used in a setting where a subject librarian is given suggestions for manual classification. Here it would be important that the correct label is always among the top 5 or top 10 results. Thus we also consider the recall at the fifth position in the ranked list (recall@5). Note that for each individual record the recall@5 is always 0 or 1.

Table 2 gives the results for classification of records form the Hochschule Hannover using different fields for both the ACT-DL classification service and the method presented in this paper. Though ground truth labels are given at the second level of the DDC-hierarchy, we can of course also evaluate the results at the first level. These first level results are given on the first line of the table.

Table 3. Mean reciprocal rank (MRR) and recall at 5 at second DDC level for SWD based classification of OAI Metadata of records from 7 German OAI-repositories using two classification methods, sci. the SWD-based classification and the ACT-DL classification service.

	SWD Based		ACT-DL	
Repository	MRR	rec@5	MRR	rec@5
Hanover UAS	0.48	0.66	0.39	0.37
Cologne UAS	0.32	0.44	0.35	0.39
Frankfurt UAS	0.55	0.75	0.49	0.62
TU Berlin	0.41	0.61	0.59	0.66
Univ. Hildesheim	0.25	0.39	0.25	0.29
Univ. Regensburg	0.61	0.80	0.65	0.72
Freiburg UE	0.53	0.75	0.33	0.36

Of course the results using the keywords gives the best results but is of least practical relevance for a library that wants to speed up the process of metadata generation for new publications. In the realistic situation only the abstract is

available, or author provided keywords that might be of less quality than the subject headings assigned by a librarian with in-depth knowledge of the subject authority file. Thus in Table 3, where we compare results for 6 more repositories, we use only title and abstract for classification. All the differences between the two methods are significant at the level of 0.001 according to the Wilcoxon signed rank test.

Finally, we have compared the results for different publication types. These results are given in Table 4.

Table 4. Mean reciprocal rank (MRR) and recall at 5 at second DDC level for SWD based classification of OAI Metadata for 6 most frequent publication types from 7 repositories.

		SWD Based		ACT-DL	
Publ. type	#records	MRR	rec@5	MRR	rec@5
PhD Thesis	2503	0.47	0.66	0.63	0.70
Master thesis	277	0.32	0.44	0.37	0.41
Essay	195	0.59	0.75	0.29	0.36
Monograph	176	0.46	0.62	0.43	0.51
Festschrift	106	0.43	0.76	0.38	0.40
Lecture	84	0.40	0.96	0.03	0.06

6 Discussion

The results of the SWD based approach are similar to those given by ACT-DL, which is rather surprising given the simplicity of our approach. Especially the recall@5 is very good for the SWD based approach as compared to the machine learning method: For 6 out of 7 repositories the recall@5 was even better. The mean reciprocal rank is 3 out of 7 cases better, in 1 case the same and in 3 cases worse. This shows that our method is rather successful in getting the correct label among the best 5 candidates but has difficulties to decide which one to put on top. A detailed analysis of a small subset shows that in many cases the first and the second result have the same weight and the ordering is arbitrary. Here machine learning techniques considering statistical relations between words of different categories or at least using a priori probabilities for the categories could improve the results.

The results split up for different publication types are probably most interesting. Especially, the SWD-based approach is able to outperform the machine learning approach for the less typical publication types. It is likely that there were not many examples of these publication types in the training data for the ACT-DL classifier, while at the same time there is a considerable difference in vocabulary between the publication types. Thus, it should be fairly easy to adjust the classifier by including additional training documents to get better results for

these publication types as well. Nevertheless it shows the weakness of the machine learning approach: it is extremely dependent on the proper composition of the training data. The thesaurus based approach on the other hand, might not reach the best possible results, but is independent of training.

The quality of results that can be achieved with the thesaurus based approach of course depends on the coverage and quality of the thesaurus. In the work presented here we could show that the enriched version of the German Subject Heading Authority File (SWD) is a high quality resource for classifying German scientific records into DDC-classes.

7 Conclusion and future work

We have shown that the SWD with the mapping of SWD-subject headings to DDC classes provides a very valuable resource that can be used for classification of scientific records. With basic methods we could already achieve results that are comparable to results from state-of-the-art machine learning algorithms. There are various possibilities for improvement. In the first place, the SWD is a file of subject headings. Especially for complex or ambiguous concepts subject headings are often formulated in a way that might never be found in a running text. Thus, lexical enrichment might improve the results. Furthermore, we have simply counted the links to DDC classes. This works only, to some extend, if the number of terms per class is well balanced. In general, it does not have to be that case that class from which the most terms are found, is also the most likely class for the text. Here a machine learning approach as proposed by [5] could be used.

Another issue for further research is the reason why for some repositories the SWD-based approach is better, while for others the trained classifier is superior. The difference can partly be explained by the different distribution of text types. The reason might also be hidden in some properties of the repository but also in the policies and habits of the libraries that assign the labels that we have used as ground truth.

References

1. Leitfaden zur Vergabe von DDC-Notationen an SWD-Schlagwrtern (September 2010), `\url{http://linux2.fbi.fh-koeln.de/crisscross/CrissCross_Endg_Grundlagenpapier_Sept2010.pdf}`
2. Addis, A., Angioni, M., Armano, G., Demontis, R., Tuveri, F., Vargiu, E.: A novel semantic approach to document collections. In: Isaías, P., Paprzycki, M. (eds.) IADIS Multi Conference on Computer Science and Information Systems. vol. 2008, pp. 73–85 (2008)
3. Bloehdorn, S., Hotho, A.: Text classification by boosting weak learners based on terms and concepts. In: Rastogi, R., Morik, K., Bramer, M., Wu, X. (eds.) Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), 1-4 November 2004, Brighton, UK. pp. 331–334. IEEE Computer Society (2004)
4. Boteram, F., Hubrich, J.: Specifying intersystem mapping relations: Requirements, strategies and issues. Knowledge Organization 37(3), 216–222 (2010)

5. Chenthamarakshan, V., Melville, P., Sindhwani, V., Lawrence, R.D.: Concept labeling: Building text classifiers with minimal supervision. In: Walsh, T. (ed.) IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. pp. 1225–1230. IJCAI/AAAI (2011)
6. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: A framework and graphical development environment for robust nlp tools and applications. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 168–175. ACL (2002)
7. Gazendam, L., Wartena, C., Brussee, R.: Thesaurus based term ranking for keyword extraction. In: Tjoa, A.M., Wagner, R. (eds.) Database and Expert Systems Applications, DEXA, 10th International Workshop on Text-based Information Retrieval, TIR. pp. 49–53. IEEE (2010)
8. Gazendam, L., Wartena, C., Malaisé, V., Schreiber, G., De Jong, A., Brugman, H.: Automatic annotation suggestions for audiovisual archives: Evaluation aspects. Interdisciplinary Science Reviews, 34 2(3), 172–188 (2009)
9. Hubrich, J.: Crisscross: Swd-ddc-mapping. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare 61(3), 50–58 (2008)
10. Hubrich, J.: Thematische suche in heterogenen informationsrume. In: Bergner, U., Gmpel, E. (eds.) The ne(x)t Generation, das Angebot der Bibliotheken: 30. sterreichischer Bibliothekartag Graz 15. - 18.09.2009. Schriften der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, vol. 7, pp. 234–242. Neugebauer, Graz-Feldkirch (2009)
11. Jacobs, J.H., Mengel, T., Müller, K.: Benefits of the crisscross project for conceptual interoperability and retrieval. In: Gnoli, C., Mazzocchi, F. (eds.) Paradigms and conceptual systems in knowledge organization. Proceedings of the Eleventh International ISKO Conference. pp. 236–241. ERGON-Verlag (2010)
12. Joan, S. (ed.): Dewey Dezimalklassifikation und Register, Dt. Ausg., vol. 1. Sauer, München, 22 edn. (2005)
13. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S.: Open Archives Initiative-Protocol for Metadata Harvesting-v. 2.0. Open Archives Initiative (2002), `http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm`
14. Malaisé, V., Gazendam, L., Brugman, H.: Disambiguating automatic semantic annotation based on a thesaurus structure. In: Hathout, N., Muller, P. (eds.) Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (communications orales). pp. 197–206. Association pour le Traitement Automatique des Langues, Toulouse (2007)
15. Schmid, H.: Improvements in part-of-speech tagging with an application to german. In: Proceedings of the ACL SIGDAT-Workshop (1995)
16. Scott, S., Matwin, S.: Text classification using wordnet hypernyms. In: Use of WordNet in natural language processing systems: Proceedings of the conference. pp. 45–51. Association for Computational Linguistics (1998)
17. Sommer, M.: Automatische Generierung von DDC Notationen für Hochschulveröffentlichungen (2012), Bachelor Thesis
18. Tiun, S., Abdullah, R., Kong, T.E.: Automatic topic identification using ontology hierarchy. In: Gelbukh, A.F. (ed.) Computational Linguistics and Intelligent Text Processing, Second International Conference, CICLing 2001, Mexico-City, Mexico, February 18-24, 2001, Proceedings. Lecture Notes in Computer Science, vol. 2004, pp. 444–453. Springer (2001)

19. Waltinger, U., Mehler, A., Lösch, M., Horstmann, W.: Hierarchical classification of oai metadata using the ddc taxonomy. In: Bernardi, R., Chambers, S., Gottfried, B., Segond, F., Zaihrayeu, I. (eds.) Advanced Language Technologies for Digital Libraries - International Workshops on NLP4DL 2009. Lecture Notes in Computer Science, vol. 6699, pp. 29–40. Springer (2011)
20. Wang, J.: An extensive study on automated dewey decimal classification. Journal of the American Society for Information Science and Technology 60(11), 2269–2286 (2009)
21. Wartena, C., Brussee, R., Gazendam, L., Huijsen, W.: Apolda: A practical tool for semantic annotation. In: Database and Expert Systems Applications, DEXA, 7th International Workshop on Text-based Information Retrieval, TIR. pp. 288–292. IEEE (2007)

Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries

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

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

Abstract. This paper introduces Pundit[3]: a novel semantic annotation tool that allows users to create structured data while annotating Web pages relying on stand-off mark-up techniques. Pundit provides support for different types of annotations, ranging from simple comments to semantic links to Web of data entities and fine granular cross-references and citations. In addition, it can be configured to include custom controlled vocabularies and has been designed to enable groups of users to share their annotations and collaboratively create structured knowledge. Pundit allows creating semantically typed relations among heterogeneous resources, both having different multimedia formats and belonging to different pages and domains. In this way, annotations can reinforce existing data connections or create new ones and augment original information generating new semantically structured aggregations of knowledge. These can later be exploited both by other users to better navigate DL and Web content, and by applications to improve data management.

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

1 Introduction

Since the advent of the digital era, cultural heritage preservation has been increasingly dealing with the conservation and the management of digital contents in Digital Libraries (DLs). These contents can be the digital reproduction of non-digital artefacts and manuscripts or more and more often born-digital multimedia contents. As this amount of data multiplies everyday faster and faster, its proper classification and management is becoming an increasingly complex task but nevertheless more and more crucial to make such information effectively consumable.

[3] www.thepund.it

With such purpose, in recent years, Semantic Web technologies and guidelines have been finding growing application in DL libraries scenario. RDF data model is currently employed by Europeana[4] initiative to aggregate independently provided digital contents. Several DLs have also made their data publicly available over the Web following the Linked Data recipes to join the giant and interconnected knowledge base of the Linked Open Data cloud [1]. Several efforts have also been done to introduce common accepted ontologies and schema for metadata encoding of DL contents, as BIBO[5], OAI-ORE[6] and Europeana Data Model[7].

Since the advent of the Web 2.0, the capability to annotate Web content, even with simple approaches based on plain-text comments or tags, has been growingly recognized as an highly beneficial feature not only for the user, making the navigation a more engaging and profitable experience, but also for the content providers that can leverage on user created metadata to better classify and search their published resources. Nevertheless, in several research scenarios, the annotation of DL contents and more in general of Web resources represents a fundamental activity daily performed by scholars. Also, in most of these cases an higher level of accuracy and granularity is typically required in the annotations to encode information about multimedia resource fragments, such as text excerpts or image regions, according to specific controlled vocabularies.

Most of the existing systems rely on simple textual comments and tags. Such approach is relatively easy to implement and very intuitive for users but it suffers from several issues related with the ambiguity of natural language and limits the accuracy and the efficiency of resource classification and retrieval. The founding idea of this research is that, if properly structured and provided with clearly-defined and machine-processable semantics, annotations can constitute themselves a primary information which can enrich the original contents and provide added value for other users as well as for third party applications. On this line, Semantic Web technologies are employed to foster the flexibility and interoperability of user created annotations, to promote their linkage with the Web of Data and to permit their reuse by other people or applications beyond the context they originated from.

This paper introduces Pundit[8], a novel semantic annotation tool, developed in the context of the Semlib project[9] [1]. Pundit has been conceived not only to permit the annotation of generic Web pages and multimedia resources but to be also specifically tailored to and integrated in existing DLs. Pundit provides support for different types of annotations, ranging from simple comments to semantic links to Web of data entities, to fine granular cross-references and citations. Pundit can be configured to include custom controlled vocabularies

[4] http://www.europeanaconnect.eu/
[5] http://bibliontology.com/specification
[6] http://www.openarchives.org/ore/1.0/primer
[7] http://pro.europeana.eu/edm-documentation
[8] Pundit: http://www.thepund.it
[9] http://www.semlibproject.eu/

and has been designed to enable groups of users to share their annotations and collaboratively create structured knowledge. This paper is organized as follows: Sec. 2 shortly provides a brief overview of related works; Sec. 3 explains the proposed data model for the annotations; Sec. 4 discusses Pundit prototype and its main functionalities.

2 Related Work

Nowadays, Web content annotation has become a common practice users are familiar with. In particular, textual comments and plain tags are supported in several mainstream Web applications like Facebook and Flickr.

In recent years, a growing number of tools have also been specifically created to allow user to annotate digital resources. Some of those found on Semantic Web technologies to improve the efficiency and the productivity of user created annotations. An exhaustive state of the art in Semantic Annotation goes beyond the purpose of this paper and can be found in literature [3], [2]. This section briefly discusses some of the most interesting annotation approaches implemented in the recently developed semantic annotation tools.

Semantic tagging paradigm, which exploits publicly available Linked Data knowledge bases to retrieve unambiguous concept to use in resource tagging, has been implemented in several application. Faviki[10] is a social bookmarking tool that uses DBpedia concepts as tags for Web pages. Zemanta[11] uses natural language processing techniques to automatically extract semantic tags from pages. Europeana Connect Media Annotation Prototype (ECMAP) [7], an online media annotation suite based on Annotea [8], allows to augment textual comment linking Dbpedia resources.

Other tools also allow the use of entities belonging to restricted vocabularies or ontologies in the annotations. One click annotation [9] and CWRC-Writer [10] allow to annotate entities in text excerpts by choosing between predefined categories (as person, location, etc, ...) or creating new ones. LORE(Literature Object Reuse and Exchange)[11], a Mozilla plugin developed inside the Aus-e-Lit Project, allows to annotate Web pages fragment adding textual comments and specifying tags selected from the AustLit thesaurus or entered as free text.

Some annotations tools enable also the creation of more expressive annotations other than textual comments or tags. LORE allows to create the so called "compound objects', by bookmarking Internet resources and describing them using standard terms coming from a bibliographic ontologies. A graphical user interface is provided to create and visualize typed relationships among individual objects based on LORE Relationship Ontologies. CWRC-Writer provides an experimental interface for the creation of subject-object-predicate statements.

If most of these tools focus on the annotation of text, some of those support the annotation of other types of digital items. ECMAP in particular permits also the annotation of maps, video fragments and images.

[10] http://www.faviki.com/
[11] http://www.zemanta.com/

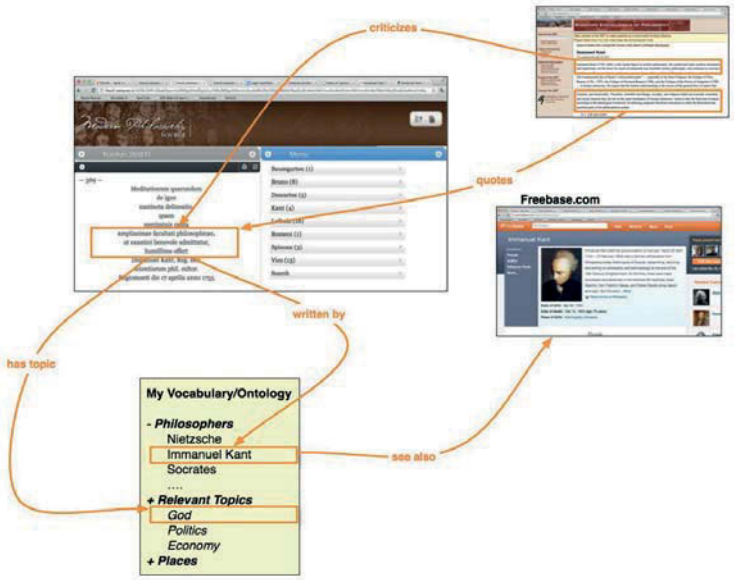


Fig. 1: Creating semantically structured aggregations of knowledge by means of annotations

3 Semantically structured annotations

The main idea in Pundit is that of enabling users not only to comment, bookmark or tag Web pages, but also to create semantically structured data while annotating, thus enriching the so called Web of Data. The ability to express semantically typed relations among resources, relying on ontologies and specific vocabularies, not only enables users to express unambiguous and precise semantics, but also, more interestingly, fosters the reuse of such collaboratively created knowledge within other Web applications. In Pundit annotations contain a set of RDF triples that connects annotated object (e.g. text excerpts) among each other and with entities in the Linked Data Web. Thanks to the nature of RDF data model (where triples can be flexibly combined to form arbitrary graphs) and to the use of URIs as identifiers for both entities and annotated objects, different annotations independently authored by different users, can be combined to form a semantic network that applications can retrieve via SPARQL endpoint and dedicated REST API. The resulting RDF graph is exemplified in Fig. 1.

Annotations acquire full significance in relation with the target resource and other contextual information, such as their author, their creation date and the vocabulary terms used. Such metadata are encoded in RDF relying on the OAC ontology, which provides a framework to represent annotations context in a standard way. The OAC data model uses the `aoc:hasTarget` property to define the target to which the annotation is attached. The target of an annotation can be entire Web pages or media objects, or their fragments (basing on Media Fragments and XPointer standards). Annotations also contain a payload, defined by the `oac:hasBody` property, which represents the user-created informative con-

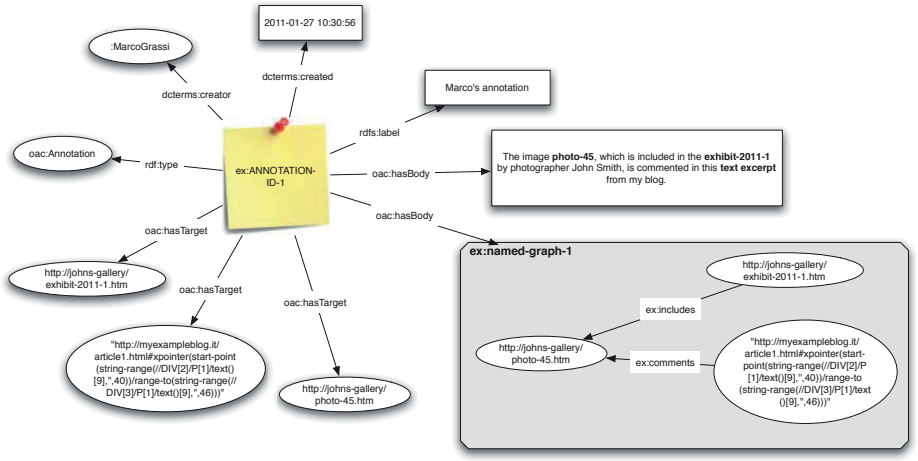


Fig. 2: The representation of an annotation using a named graph

tent. At the time of writing, the OAC (Open Annotation Collaboration) model[12] has been merged with the Annotation Ontology[13] to form the OA (Open Annotation) specification[14] that only recently reached its first stable state and is considered the de facto standard for representing annotations on the Semantic Web. It's worth to remark that, among other news, OA explicitly validates the use of named graphs that has been already put in place in Pundit. At the time of writing, full compliancy with such a specification is currently under development. In Pundit, named graphs are used as "bodies' (using the OA jargon) of annotations, which in our system is composed by RDF triples itself. This allows to keep separated statements belonging to different annotations, while still being able to aggregate them into "composite' graphs and query them using standard SPARQL language. For example, one could query for all the annotations whose target is a specific image and whose author is one (or more) specific user, and then extract all the resources that "comments' the image according to the selected annotations. Fig. 2 illustrates how annotations are represented in our system.

While the OAC ontology is used to represent contextual information, the semantic content cannot be represented based on a fixed ontology. Different users communities operating in specific domains need specific shared vocabularies (ontologies) of terms and relations that they can use in annotations. At RDF data storage level, the system is therefore agnostic with respect to the domain ontologies used in structuring annotation informative semantic content, and specific configuration at application level can be used to build an ad-hoc vocabulary for each community addressed. Pundit supports both "open", relatively flat vo-

[12] http://www.openannotation.org/
[13] http://code.google.com/p/annotation-ontology/
[14] http://www.openannotation.org/spec/core/

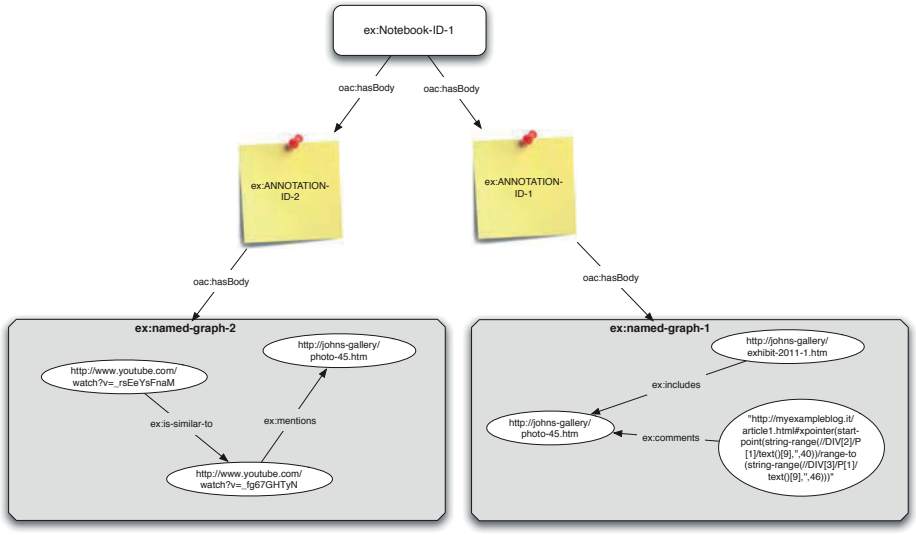


Fig. 3: The RDF representation of a notebook including two distinct annotations

cabularies like Freebase (leveraging the reconciliation APIs[15]) and restricted controlled vocabularies and taxonomies, e.g. based on the SKOS model.

In Pundit, "notebooks" are resources that aggregate a set of annotations so that they can be retrieved and queried. By default, each user has a proprietary notebook where all her annotations are collected. Notebooks have a central role in collaborative annotation. These can in fact have read/write privileges and can be used for giving users control over her annotations, allowing to set them as private or public and to select what notebooks are relevant. More precisely, Pundit supports the concept of "active notebook": when a notebook is active for a given user the annotations in it will be shown by default. As a big number of public notebooks might be available, this mechanism allows a user to restrict the amount of annotations visualized to only those she expressed interest in. While such notebooks management features are fully implemented by the Pundit annotation server, their full support at UI level is still under development.

4 Pundit prototype

Pundit has a client-server architecture. The client-side component comprises a set of sub-modules developed in Javascript using the dojo framework[16] 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. The storage module defines a completely generic interface, designed to support different kinds of storage systems ranging from traditional relational databases to NoSQL databases (eg. RDF triple-

[15] http://wiki.freebase.com/wiki/Freebase_API
[16] http://dojotoolkit.org/

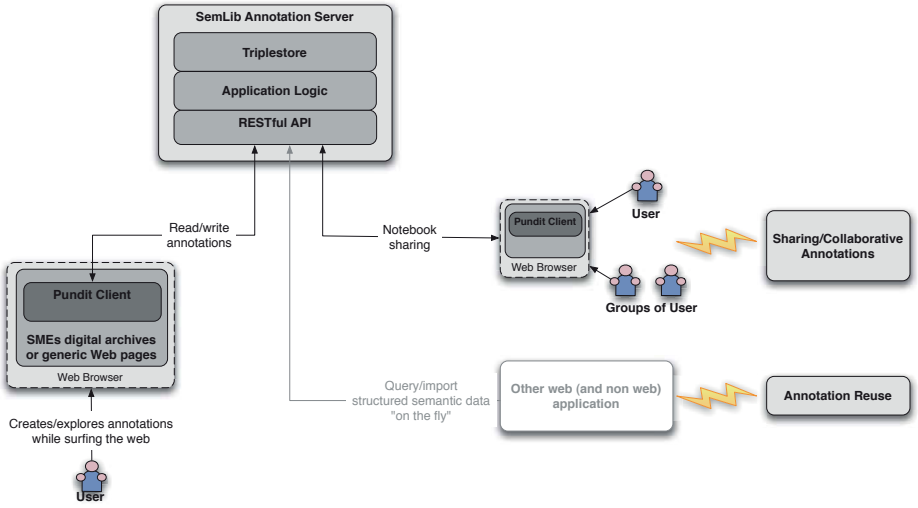


Fig. 4: Simplified architecture of the annotation system

stores). In the prototype version, the storage is implemented using the Sesame triplestore[17] as this greatly simplifies handling and exporting RDF data. The storage module, besides keeping user annotations, stores also user profiles and related contextual information (e.g.: user's metadata, user's permissions, etc.). The Annotation Server supports Open-ID[18] for users authentication with single sign-on. Different authentication systems can be easily implemented developing dedicated plugins. The use of single sign-on approach simplifies the integration of the annotation system with existing DL, which may already provide facilities for users authentication. In the following subsections, some of Pundit main features are discussed.

4.1 Annotations of different multimedia contents and at different levels of granularity

Pundit provides specific *Fragment Handlers* to assist users in selecting and highlighting parts of different contents and turn them into actual addressable resources (e.g. using XPointer or Media Fragments Uri) to be used into annotations. This means that with Pundit it is not only possible to attach annotations to single resource but also to establish semantic relations between different resources fragments also of different type. Also, selected resources can be added to "favourites' (*My Items*) and stored to the server to be displayed also in different pages other than the one in which they have been selected. This is fundamental to create cross-page and cross-domain annotations as discussed in more details in Sec. 4.4. At the time of writing Pundit prototype provides support for text fragment and image selection, while image fragment annotation is currently under

[17] http://www.openrdf.org/
[18] http://openid.net/

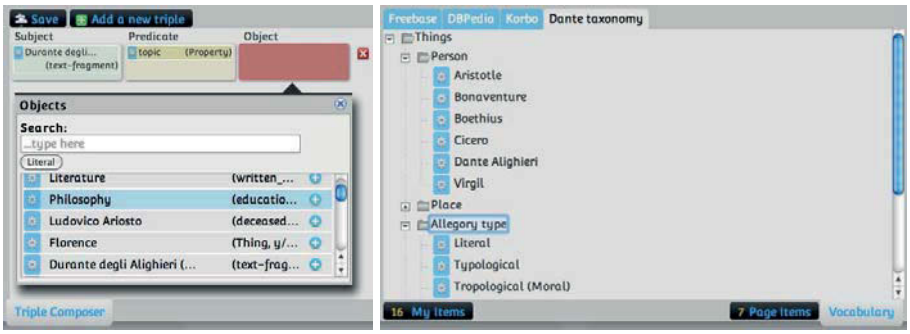


Fig. 5: The Pundit Triple Composer (a) and an example of Pundit taxonomy (b)

development. In addition, the annotation of video and of temporal and spatial video fragments has been already implemented in Semtube prototype[15]: a Web tool for semantic video annotation of YouTube videos, which has been recently developed basing on Pundit client API and Annotation Server.

4.2 Annotations at different levels of complexity and structure

Pundit provides support for different types of annotations, ranging from simple textual comments and semantic tags to semantic statements. Annotations can be created using different GUIs.

The *Comment/tags Panel* allows the user to type a comment and to automatically extract tags from it using Dbpedia Spotlight service. User can remove suggested tags that are not considered relevant or add others using Dbpedia Lookup service.

The *Recognizer Panel*, Fig. 6 is intended to be used when a user wants to mark the occurrence of a specific entity that is mentioned in text. Once one or multiple words have been selected, the recognizer searches in a set of different sources (including custom taxonomies, Freebase, DBpedia and Wordnet) and suggests matching entities. Once "recognized', entities mentioned in the text are semantically disambiguated and enriched with structured data.

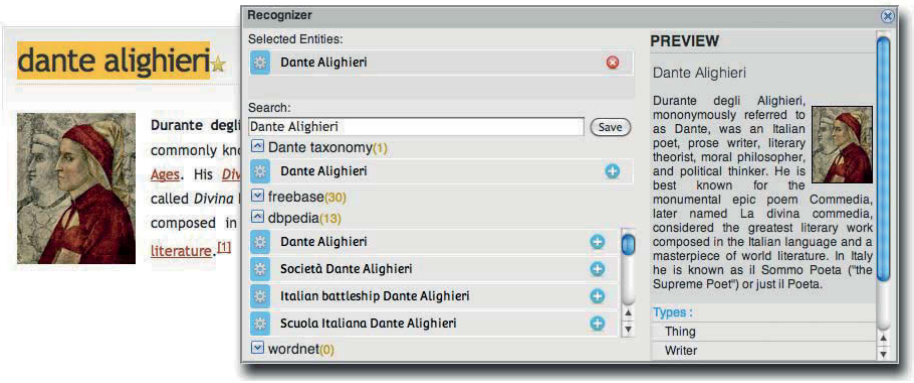


Fig. 6: The recognizer panel in action

Fig. 7: A screenshot of Pundit in action

Finally, the *Triple Composer* is the most expressive way of creating structured data, providing a specific GUI for editing semantic statements (triples) in the form of subject-object-predicate. All kinds of items (selected text, taxonomy entries and web of data resources) can be used in statements and put in relations by choosing from a customizable set of predicates. Statement can be create both dragging and dropping items or choosing between suggested items as shown in Fig. 5.a).

Fig. 5.b) shows the taxonomies tab. The ability of customizing Pundit with domain specific taxonomies is an important feature of Pundit. Digital Library maintainers can add custom taxonomies "on the fly" just by adding a simple markup to their pages, linking to a JSON file containing the taxonomy.

Fig. 7 shows the overall prototypal user interface to compose semantic annotations and to display contextually created annotations.

4.3 Named Content

DLs, like other Web 2.0 applications, change over time. Presentation can be restyled, changing page layout and mark-up, and content can be re-organized and moved to different pages. 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). In order to grant annotation consistency in such cases, in particular when they are shared in communities and not under a centralized control, it is not sufficient to attach annotations to the Web page.

To overcome this issue, Pundit relies on specific page mark-up. Compliant digital libraries can benefit from a more intelligent behaviour by using the simple named content specification (documented on the web site) to mark-up atomic portions of their content as exemplified in Fig. 8. Each marked content should

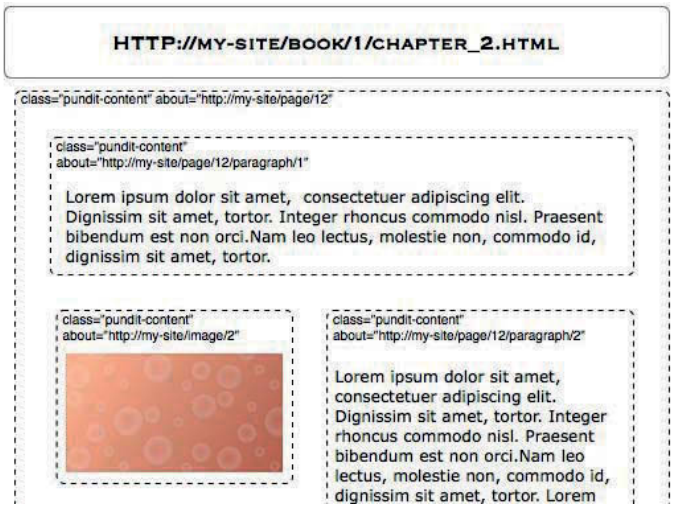


Fig. 8: Using Named Contents to allow annotations to be attached to content

have a resolvable URI associated, to which annotations are attached. In this way, annotations regarding the same content, but created in different pages, can automatically be merged and consistently displayed in all the pages where such content appears.

4.4 Cross-page and cross-domain annotations

Cross-pages annotation constitutes a key feature of the proposed annotation system that captures the distributed nature of the Web, in which information is often spread between different sources and can be augmented linking and referencing additional information beyond the boundaries of single Web site or DL. Properly structured annotations can allow weaving a semantic net in order to interconnect and merge fragments of information into a unique knowledge base. For example, an expert of literature can augment the information about Dante Alighieri appearing on a Web page of a DL with text excerpts of the Divine Comedy taken from another Web source. The implementation of such feature requires the system be able to:

- create annotations on every Web page
- create relations between different resources (as text fragment, images, etc...) belonging to different pages.

The former requirement is supported by the availability of the application as a bookmarklet, which allows running the annotation system in every Web pages injecting the required javascript. With such purpose, particular care has been required in protecting Pundit css and variable namespace, in order to avoid clashes that could result in page style and layout alteration as well as in application malfunctioning.

Regarding the latter requirement, it's worth to remark how RDF data model is perfectly suitable to cope with it, being in fact specifically conceived to create

statements that connect two resources by means of a property. From an implementation point of view, such requirement is fully fulfilled by means of the *Triple Composer* and by the *MyItems* mechanism, described in the previous subsection. These allows, for example, a user to add an image to *My Items* later assert that a text excerpt selected on another page describes the image.

5 Conclusions

In this paper, Pundit annotation system, which at the time of writing has reached its first stable release, has been introduced. Pundit data model leverages on OAC annotation model and further extends it to fully support the embodiment of semantic statements in the annotation payload by means of named graphs. This provides high flexibility to annotate and interconnect heterogeneous resources over the Web and to be potentially applied in every application scenario. Pundit prototype enables the creation of semantically rich annotations at high granularity levels. These allow to interconnect different resources distributed over the Web and augment original information generating new semantically structured aggregations of knowledge. These can in turn be exploited both to provide user with a more engaging and productive experience in consuming DL and Web content, and effectively reused by other applications.

Compared with other existing semantic annotation tools, Pundit not only provides support all the main annotation approaches introduced by others tools (textual comments, semantic tagging, named entities recognition and the use of taxonomies and ontologies) but enable also more expressivity and flexibility in annotations. In particular, it allows the creation of semantic statements that enable to put in link resources, resource fragments, named entities and vocabulary resource according to semantically defined relations.

In addition, differently by other tools, Pundit has been conceived to provide specific support for annotation sharing, relying on the mechanism of notebooks to aggregate relevant information and make these available both to other users and third party applications by means of a dereferenciable URI and to be easily consumed by means of RESTfull API.

A user evaluation of the tool has been conducted for the video annotation prototype. The obtained results can be found in [15] and are driving the current Pundit development. Further user evaluations are going to be performed on the continuation of the development.

6 Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme managed by REA-Research Executive Agency[19] ([FP7/2007-2013][FP7/2007-2011]) under grant agreement n. 262301.

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

References

1. C. Morbidoni, M. Grassi, M. Nucci, "Introducing SemLib Project: Semantic Web Tools for Digital Libraries". International Workshop on Semantic Digital Archives 15th International Conference on Theory and Practice of Digital Libraries (TPDL). 29.09.2011 in Berlin.
2. Andrews, P., Zaihrayeu, I., Pane, J., "A classification of semantic annotation systems. Semantic Web Journal". Online Available: http://www.semantic-Web-journal.net/content/classification-semantic-annotation-systems
3. V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semant., 4(1), January 2006.
4. R. A. Arko, K. M. Ginger, K. A. Kastens, and J. Weatherley, "Using annotations to add value to a digital library for education'.
5. Rose Holley, "Crowdsourcing: How and Why Should Libraries Do It?', D-Lib Magazine, The Magazine of Digital Library Research. March/April, 2010.
6. 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.
7. 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.
8. 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.
9. Markus Luczak-Rsch, Ralf Heese, Adrian Paschke, "Future Content Authoring', In Nodilities The Magazine of the Semantic Web, Issue 11, pp. 17-18, 2010.
10. G. Rockwell, S. Brown, J. Chartrand, S. Hesemeier, "CWRC-Writer: An In-Browser XML Editor' - Digital Humanities 2012 Conference Abstracts. University of Hamburg, Germany. July 1622, 2012
11. A. Gerber and J. Hunter, "Authoring, Editing and Visualizing Compound Objects for Literary Scholarship', Journal of Digital Information, vol. 11, 2010.
12. M. L. Ralf Heese, "One Click Annotation' in 6th Workshop on Scripting and Development for the Semantic Web, 2010.
13. M. Koivunen, R. Swick, E. Prud'hommeaux "Annotea and Semantic Web Supported Collaboration'. ESWC 2005, UserSWeb workshop. 2005
14. "Open Annotation: Alpha3 Data Model Guide' 15 October 2010 Eds. R. Sanderson and H. Van de Sompel. `http://www.openannotation.org/spec/alpha3/`
15. M.Grassi, C. Morbidoni and M. Nucci. A Collaborative Video Annotation System Based on Semantic Web Technologie. In press: Cognitive Computation. Springer-Verlag, Berlin Heidelberg (DOI: 10.1007/s12559-012-9172-1)

Towards a Recommender System for Statistical Research Data

Daniel Bahls[1], Guido Scherp[1,2],
Klaus Tochtermann[1], and Wilhelm Hasselbring[2]

[1] Leibniz Information Centre for Economics (ZBW), Kiel, Germany
[2] Software Engineering Group, Kiel University, Germany

Abstract. To effectively promote the exchange of scientific data, retrieval services are required to suit the needs of the research community. A large amount of research in the field of economics is based on statistical data, which is often drawn from external sources like data agencies, statistical offices or affiliated institutes. Since producing such data for a particular research question is expensive in time and money—if possible at all— research activities are often influenced by the availability of suitable data. Researchers choose or adjust their questions, so that the empirical foundation to support their results is given. As a consequence, researchers look out and poll for newly available data in all sorts of directions due to a lacking information infrastructure for this domain. This circumstance and a recent report from the High Level Expert Group on Scientific Data motivate recommendation and notification services for research data sets.
In this paper, we elaborate on a case-based recommender system for statistical data, which allows for precise query specification. We discuss required similarity measures on the basis of cross-domain code lists and propose a system architecture. To address the problem of continuous polling, we elaborate on a notification service to inform researchers on newly avaible data sets based on their personal request.

Keywords: Research Data Management, Semantic Digital Data Library, Linked Data, Statistics, Recommender Systems, Case-Based Reasoning

1 Introduction

At present, efforts are being made to pick up research data as bibliographic artifacts for re-use, transparency and citation. Data publications will be submitted to digital archives and registered in central catalogs which lays the ground for information services to support the scientific community in finding relevant data. Since every scientific discipline brings its own challenges in this endeavor, specific solutions are required, so that valuable, and hence accepted, services can be offered to the scientific community [1]. The High Level Expert Group on Scientific Data recommends to provide data recommendation services that suggest relevant research data to the individual scientist [2]. This appears to be particularly applicable in the domain of economics where research activity is influenced

by the availability of statistical research data sets.[3] Researchers adjust to what data is available and adapt research questions so that the empirical foundation can be given.

As a consequence, researchers look out and poll for newly available data in all sorts of directions due to a lacking information infrastructure for this domain. They exchange news on newly available data at conferences, at meetings or simply at lunch time or during coffee-break. They also revisit websites of data agencies, repositories and familiar institutes to run their personal portfolio of keyword-based queries on regular web search engine interfaces—trying to express their request for specific data sets. Although best practice at present, this strategy seems effortful and insufficient in returning a complete list of relevant data sets. This picture was shared with us in interviews we have conducted with researchers in economics.

Having catalogs of registered research data sets puts us in a good position to address the above problem and develop well-conceived search tools and services for our scientific community. Besides the fact that the catalog itself lays the ground for a more organized search, this paper tries to address the following two aspects of the identified problem:

1. Phrasing several queries with different keywords and filters of all kinds to cover the range of relevant data sets.
2. Continuous polling at regular time intervals.

The remainder of the paper is structured as follows. We review related work and decide on our approach in Section 2. Section 3 concludes the findings and formulates the functional requirements for our proposed system. Since we follow a case-based recommendation approach, we examine case base and case structure in Section 4 and elaborate on a similarity measure design on the basis of common code lists subsequently in Section 5. We propose a system architecture in Section 7. Finally, we close with conclusions and outlook in Section 8.

2 Related Work

In the domain of statistical research data, one main difficulty is given by data protection and usage rights, so that uploading entire data collections to an independent repository causes legal problems. This is one of several reasons why we have decided to use Semantic Web technologies for the data model, which are strong in fine-grained referencing and in dealing with distributed data sources. In particular, we use the RDF Data Cube Vocabulary (QB), which integrates the SDMX standard[4] and is increasingly recognized in the domain of statistics [3]

[3] A large amount of research in the field of economics is based on statistical data, which is often drawn from external sources like data agencies, statistical offices or affiliated institutes. Producing such data for a particular research question is expensive in time and money—if possible at all.

[4] Statistical Data and Metadata eXchange Language `http://sdmx.org/`

[4] [5] [6]. A more detailed argumentation and an overall vision for our research is given in [7].

There are several different types of recommender systems for which a comprehensive overview can be found at [8]. Especially in e-commerce environments, *collaborative filtering* has established as a common technique. Online stores like amazon[5] recommend products on the basis of similar user profiles, following the idea that one might be interested in the products that other users with similar interest patterns have purchased. While this technique can be applied irrespective of the kind of items operated on, it demands large amounts of usage data from a sufficient number of users in order to produce meaningful recommendations. This initial overhead is known as the *cold-start problem* and usually requires user acceptance long before the value of item recommendation can be experienced.

Another technique makes use of the items' digital content[6] which we refer to as content-based recommendation systems [8]. Typically, items are mapped onto a vector space model where distances between them can be calculated using common mathematical means. This technique has established particularly in the context of textual items, where means like[7] are frequently used. However, this approach again depends on an initial set of usage data. While collaborative filtering compares patterns among user profiles, content-based retrieval is based on usage history of a single user and suggests similar items according to what she or he found useful or not useful earlier.

A third system type is based on background knowledge and calculates recommendations merely on the basis of a given user query and domain-specific preference knowledge encoded in the form of rules or specifically designed similarity measures. The approach therefore does not build on usage data at all and thus is not affected by the cold-start problem. Since usage data on statistical research data sets is not easily available to us and difficult to acquire in sufficient quantity, we find this approach most suitable for our domain. The amount of statistical research data is tremendous, and the amount of usage data required scales accordingly if we plan to include all available data sets for recommendation. In addition, recommending data sets that are similar to the ones used previously may not be helpful in the scientific domain, where researchers often work on various projects simultaneously or change their research area when moving to another organization. The above described systems tend to recommend older items, because usage statistics on newer ones build up slowly[8]. While these drawbacks do not apply for knowledge-based recommenders, another advantage is their strength in explaining results, so that users can understand why a particular recommendation was considered relevant. Furthermore, a lot of background knowledge for statistical data is available and has even been formalized

[5] urlhttp://www.amazon.com

[6] be it metadata, a textual description or the digital item itself like for example in document retrieval scenarios

[7] Term Frequency - Inverse Document Frequency

[8] also known as the time-span problem

in SDMX[9], DDI[10], code lists and the RDF Data Cube Vocabulary (QB), which also encourages a knowledge-based approach.

Knowledge-based recommender systems are typically constraint-based or case-based [8]. While the former uses rule sets and constraint resolvers to produce recommendations, the case-based approach uses specifically designed similarity measures that shall reflect the user's understanding of utility [9]. Eventually, we have chosen to follow a case-based approach on the grounds of positive experiences in earlier projects. As a consequence, a research data set is considered and may be referred to as a *case* in the following. Cases in general can be represented textually, as a feature vector, or as a structured representation [10]. The cases according to our RDF-based data model are already in structured shape, which gives reason to choose a structured CBR[11] approach over a textual, feature-based or other.

Common data repositories do not yet offer recommendation features and focus on providing full text search interfaces and filtering features. Text search algorithms often yield scores that allow for relevance ranking and are applied on textual fields of the respective underlying metadata model. Search criteria given for the more structured part of the model[12] are usually filtered on, meaning that all unmatched items are removed from the ranking [11]. A typical implementation imposes this rather technical and limited viewpoint on the user who switches back and forth modifying query phrase and parameters to cover the whole spectrum of possibly interesting search results, simply to deal with the limitations of such rigid interface[13]. It is to say that these issues are difficult to overcome, and most retrieval algorithms incorporate stemming, query expansion and other strategies while targeting a yet simple interface which certainly is another important design goal. Our aim is to get a clear picture of the user needs first which needs no further editing once specified clearly. Every item that matches the query entirely would be considered a perfect match, and therefore the approach performs like the common ones. In addition, however, the system should be able to find near matches and offer further means of knowledge discovery, which is a more high-level approach in the first place.

3 Functional Requirements

To address research objective 1, the system must provide an interface that allows for precise specification of a data request, enabling researchers to pinpoint to the perfect data set regardless of whether such data exists. This can be done on the basis of the RDF Data Cube Vocabulary which provides a wide range of predicates and attributes to formulate precise queries. The system further needs

[9] Statistical Data and Metadata eXchange Language `http://sdmx.org/`
[10] Data Documentation Initiative `http://www.ddialliance.org/`
[11] Case-based reasoning—or case-based recommending in our case
[12] e.g. creation date, size, country of origin or other domain-specific fields
[13] rephrasing query terms, resetting date ranges, size parameters, geo location and other

to know what aspects of the query are of greater, and what aspects are of lesser significance, which can be handled with the help of user-defined weights [12].

The second objective can be achieved through a notification service that sends out updates on newly available data to the individual user whenever estimated relevant.

An understanding of utility must be encoded in the system, so that data not perfectly matching the user's description can be estimated whether it yet may interest the user. In case-based recommender systems, such knowledge is encoded in similarity measures that are used to determine an estimated degree of utility of a particular case under a given query. Such measures must be designed carefully and must not make assumptions on user preferences where no foundation is given. Case-based recommenders in principle can be applied for our research objectives. However, the value of this approach depends on the question whether meaningful similarity measures can be implemented, which will be investigated in Section 5.

4 Case Structure

CL_OBS_STATUS	Status of an observation with respect events such as the ones reflected in the codes composing the code list.
CL_CONF_STATUS	Coded information about the sensitivity and confidentiality status of the data.
CL_DECIMALS	Gives information on the number of decimal digits used in the data.
CL_FREQ	Indicates the "frequency" of the data (e.g. monthly) and, thus, indirectly, also implying the type of "time reference" that could be used for identifying the data with respect time.
CL_SEX	Provides information on the gender.
CL_TIME_FORMAT	Time Format as written in the SDMX-EDI and SDMX-ML messages; these codes (based on the ISO 8601 standard) indicate the type of time references used in the data. The numeric codes below (203, 102,,702) are used only in the SDMX-EDI messages; and the alphanumeric codes (P1DPT1M) only in the SDMX-ML messages.
CL_UNIT_MULT	Unit Multiplier; indicates the magnitude in the units of measurements.
CL_AREA	Reference area and/or counterpart area; geographical areas, defined as areas included within the borders of a country, region, group of countries, etc.
CL_CURRENCY	Provides code values for currencies.

Table 1. Cross-domain code lists as provided by the SDMX consortium

Since we use the RDF Data Cube Vocabulary to organize statistical research data, the number of available attributes to describe research data sets is very

large, also because RDF-based descriptions are per se extensible, which might be made use of when dealing with long tail research data of individual researchers. Hence, we only review some of the common attributes in order to assess the value of this approach. Table 1 gives an overview to the Cross-Domain Code lists issued by the SDMX consortium [13] [14]. First of all, we need to clarify the notion of a case and how to map the RDF data to a case base. Figure 1 illustrates the structure of a case, and where the SDMX code list attributes are located.

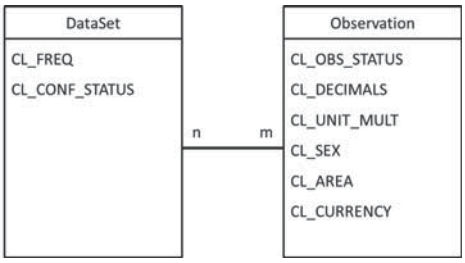


Fig. 1. The case structure

This structured representation suggests to apply the *local-global-principle*, which is an established paradigm in the CBR domain [9]. Local similarity measures are used to determine similarities on attribute level, while global similarity measures aggregate the resulting values on object level. There are two types of objects: DataSet and Observation, and thus, two global similarity measures are needed. Instances of DataSet are the items to be retrieved or recommended, while instances of Observation make for a large portion of its actual content. Because of the *n,m* relation between DataSet and Observation, we need measures for dealing with multiple values [15]. In the following, we write *sim(q,c)* to denote the similarity function of a query value q and a case value c, whereas both variables q and c are elements of the respective attribute's value range.

5 Similarity Measures

5.1 Local Similarity Measures

CL_CURRENCY specifies the currency used in a data set. If the user explicitly queries for data sets with Euro as a currency, only such data sets should be considered suitable. Suggesting that a data set using USD would be more useful

than one using CHF has no basis.[14] Thus, the similarity measure for such type should be totally uninformed:

$$sim(q,c) = \begin{cases} 1, & \text{if q=c} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

We find a different situtation for the area code, where groups of countries and regions build up taxonomies. Whether data about Bavaria is useful when Germany was specified in the query depends on interpretation: It may be interpreted as "data about any region in Germany is fine" or "data about Germany on the national level is needed". Sophisticated user interfaces would be required for disambiguation, and we rather try to bypass this problem and approach a more vague but generic measure. This is supported by the consideration that even with a more precise query, the utility of a data set on other regions still remains hard to assess in general. When a data set on Bavaria is requested and a data set on Brandenburg is given, one may argue that merely the data on Bavaria represents the population a researcher wants to do research on, and any other are simply unsuitable. In contrast, one may argue as well that both regions are siblings in the sense that Germany is the subsuming parent, and a similarity value above zero appears resonable, as the data set may still reflect some of the features the researcher is after, while a data set on Idaho (USA) may not be suitable anymore, and yet another dataset on Chengdu (China) cannot be used at all. Several techniques are available for implementing a taxonomy-based similarity measure. One generic option is given to calculate a value based on the length of the shortest path. A more specific option depends on the actual query semantics and needs further consideration and discussion.

Other codes have ordered range sets. The CL_DECIMALS code list denotes the number of decimals used in the data. It seems reasonable to assume that any data providing a higher number of decimals suits just as well as the data queried for, since numbers can always be rounded. In contrast, a smaller number of decimals than requested should be assumed as less suitable, as it means a lack in precision. And since the degree of precision decreases proportionally with the difference between case and query value, it suggests a typical *more is better* similarity measure:

$$sim(q,c) = \begin{cases} 1, & \text{if } q \leq c \\ \frac{q-c}{d}, & \text{otherwise} \end{cases} \quad (2)$$

where d denotes the maximal difference between query and case, which is ten in this case.

A similar case is found for the code list CL_FREQ. Quarterly data can be aggregated from monthly or daily data. But if monthly data is requested, and quarterly data is given, the request is not perfectly met. Such data might yet be more useful than yearly data, so that a similar measure like the above could be reasonable. While cl_decimals was based on numbers, cl_freq is symbolic.

[14] U.S. Dollars (USD), Swiss franc (CHF)

Therefore, we could define an order and map frequency symbols to integers, so that a similar function as the above can be applied.

For the free text fields as listed in [13], the measure should be based on common techniques like TF-IDF[15] or n-gram. However, it must be ensured that the value is normalized, so that resulting similarity values can be set in relation to the ones of other attributes when aggregating in the global measure.

5.2 Aggregation of Similarity Values

A so-called global similarity measure is used to aggregate the results from the attribute-level similarity calculations. As we are still in a stage of considerations, there is no point in arguing whether weighted means, Euclidean or other types of aggregation is the right method to choose. We state, however, that the measure should enable user-defined weights in the query, as it allows the researcher to emphasize on the one or other parameter.

To complete the similarity measure, we further need to specify how multiple values are dealt with. For example, the researcher may request data on the geographic locations France, Germany and United Kingdom. The utility of a data set that represents the populations of England and France may then be calculated by finding best partners for every requested country[16] and build minimum, maximum or average for the overall similarity value of the geographic attribute. Which strategy to choose depends on the particular attribute and should be examined carefully in evaluation with end users [15].

5.3 Undefined values

There are some special cases we need to consider. When a query specifies a value for a particular attribute, for which the case compared does not provide any value, the measure must yield some value as well. For instance, if `male` is specified for `gender` in the query, and the attribute value is not given in the case, utility should be considered zero, because the user explicitly stated that represented population should be male. It appears reasonable to take this as the default measure. However, if the user specifies `free` for confidentiality status, and the case does not provide any information in this regard[17], the data set is not necessarily unsuitable. A reasonable way to deal with this issue could be to simply ignore this attribute in the global similarity function.

Another special situation occurs when a researcher requests data that contains values of some currency, but she does not want to specify more precisely on it. She is certain that monetary values must be part of the data while the currency unit itself is subordinate. One way to cater for this is to introduce a special value * and let sim(*,c)=1 for any case c.

[15] term frequencyinverse document frequency
[16] best with respect to the local similarity measure
[17] due to incomplete annotation

6 Notification Service

Due to the impression that empirical research is quite data-driven, and researchers need to conintuously look out for new data sets in order to stay up-to-date, we want to make some considerations on a notification service[18]. As our approach was to capture the researcher's request for data in high precision, we are in a position to test incoming data sets for relevance and send out messages[19]. One strategy in this regard would be to notify about every data set that meets a user-defined similarity threshold. From experience, however, similarity values tend to accumulate in a particular range, which is highly dependent on the similarity measure design and the respective user query, and thus, it may be difficult to be provide a specific threshold value. In that sense, the values calculated with the help of the similarity measure should rather be regarded as scores that give means for a ranking. To bypass this problem, we suggest to send out such rankings of newly registered data sets on a regular basis as per user settings. The user may then take a closer look at the top matches and estimate their utility individually.

7 Proposed Architecture

The recommender component is considered part of a larger digital archive system that manages statistical research data. Figure 2 gives an illustration of the entire system architecture, where three main components depict the relevant parts of the recommender system.

The case retrieval engine requires access to the data base that contains the data sets and the similarity measures which should be contained in a separate data base as to allow for independent editing whenever administrative review is needed. The archived research data usually is maintained in its specific data format, which in our case is based on RDF[20]. If the retrieval engine is implemented using sequential similarity calculation, the data repository can be accessed as a case base directly, since no further indexing is required. This, however, leads to long computation times in case of a large case base. For more efficient retrieval, more optimized methods like Case Retrieval Nets [16] should be considered, which builds its own data structure from case base and similarity measures. Therefore, the recommender component needs to be notified whenever there are updates on the data repository or similarity measures.

The notification service needs access to the users' notification queries and their e-mail addresses, which are stored in the user preferences data base. The component should be notified whenever updates occur on the data repository, so that new data sets can be tested for relevance immediately.

[18] cf. Google Alerts

[19] whenever a relevant candidate is detected or collated, per e-mail, twitter or other channel

[20] Resource Description Framework

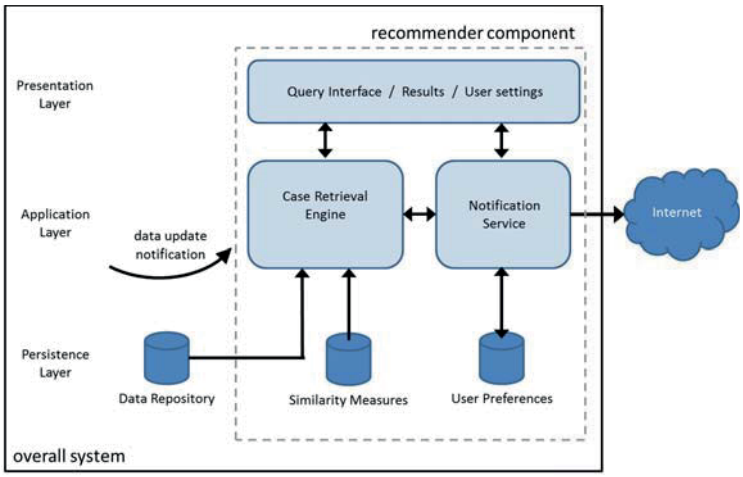


Fig. 2. System architecture

A detailled discussion on the user interface exceeds the scope of this paper. However, it must provide for query specification, display of results and configuration of user settings regarding the notification service as discussed in Section 6. Furthermore, we suggest to integrate explanation features in order to provide transparency to the user on how results were retrieved. A generic interface design and implementation can be found at [15], and some idea on how a query interface particularly designed for statistical research data using the RDF Data Cube Vocabulary can be found at [7].

8 Conclusion and Outlook

We have examined some of the common code lists for statistical data with respect to their specification and found indicators that motivate a particular similarity measure design. For some of them we were able to reason a specific design, whereas other code lists are difficult to make assertions on and suggest rather uninformed measures. Eventually, a final assessment on utility of a particular data set can only be done by the researcher. A similarity measure can only approximate a common sense of utility [9]. It easily fails due to limited query expressiveness and inability to interpret its actual semantics and the actual user needs. One option to overcome this problem is to allow for customization and personalization of similarity measures. Whenever a user is presented with unexpected results, an explanation may be given and the user may give feedback on the similarity measure. Since structural CBR systems in general are easily equipped with explanation support and customization of similarity measures [17] [18], some of the open similarity design questions could be answered by the individual user within a particular research context. However, ordinary measures for dealing with multiple values and the application of user-defined weights in

the aggregating function enable a more gradual scaling of retrieval results with respect to user needs.

Another common practice in empirical research is the use of proxy variables, where some data highly correlates with other. Such information could be useful for recommending relevant data. A similarity measure could again be extended to make use of such relations if represented in the data model.

The proposed recommender system is based on the RDF Data Cube Vocabulary. The user is therefore in a position to specify precisely on the kind of data needed, and the system has the required means to assess suitability of available data sets. In addition, provided the measure reflects a reasonable understanding of utility, the introduced notification service helps researchers keep up to date and thus, both research goals defined in Listing 1 were met. Nevertheless, an evaluation is yet to be carried out, which is subject of future work. With further progress on a research data management infrastructure and the continuing exchange with the scientific community, we will get a clearer picture on the applicability of this approach.

Eventually, a prototype is needed in order to gain feedback from the research community we are addressing, which we consider implementing as we proceed with the reasearch on a data management infrastructure.

References

1. Feijen, M.: What researchers want - a literature study of researchers' requirements with respect to storage and access to research data (February 2011)
2. Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D.R., Los, W., Marinucci, M., Romary, L., Van de Sompel, H., Vigen, J., Wittenburg, P., Giaretta, D.: Riding the wave: How Europe can gain from the rising tide of scientific data. European Union (2010) Final report of the High Level Expert Group on Scientific Data: A submission to the European Commission.
3. Gottron, T., Hachenberg, C., Harth, A., Zapilko, B.: Towards a semantic data library for the social sciences. In: SDA'11: Proceedings of the International Workshop on Semantic DigitalArchives. (2011) in Preparation.
4. Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J.: Semantic statistics: Bringing together sdmx and scovo. In Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M., eds.: LDOW. Volume 628 of CEUR Workshop Proceedings., CEUR-WS.org (2010)
5. Miloevi, U., Janev, V., Spasi, M., Milojkovi, J., Vrane, S.: Publishing statistical data as linked open data. In: Proceedings of the 2nd International Conference on Information Society Technology, Information Society of the Republic of Serbia (2012)
6. Halb, W., Raimond, Y., Hausenblas, M.: Building Linked Data For Both Humans and Machines. In: WWW 2008 Workshop: Linked Data on the Web (LDOW2008), Beijing, China (2008)
7. Bahls, D., Tochtermann, K.: Addressing the long tail in empirical research data management. In: 12th International Conference on Knowledge Management (I-KNOW '12), Graz, Austria, ACM (2012) in Preparation.

8. Burke, R.: Recommender systems: An introduction, by dietmar jannach, markus zanker, alexander felfernig, and gerhard friedrich. International Journal of Human-Computer Interaction **28**(1) (2012) 72–73
9. Bergmann, R., Richter, M.M., Schmitt, S., Stahl, A., Vollrath, I.: Utility-oriented matching: A new research direction for case-based reasoning. In: In professionelles Wissensmanagement: Erfahrungen und Visionen. Proceedings of the 1st Conference on Professional Knowledge Management. Shaker. (2001) 264–274
10. Bergmann, R., Kolodner, J., Plaza, E.: Representation in case-based reasoning. Knowl. Eng. Rev. **20**(3) (September 2005) 209–213
11. Bridge, D., Göker, M.H., McGinty, L., Smyth, B.: Case-based recommender systems. Knowledge Engineering Review **20** (September 2005) 315–320
12. Richter, M.M.: Case based reasoning and the search for knowledge. In: Proceedings of the 7th industrial conference on Advances in data mining: theoretical aspects and applications. ICDM'07, Berlin, Heidelberg, Springer-Verlag (2007) 1–14
13. Guidelines, S.C.o.: Annex 1: cross-domain concepts 2009. Area (2009) 1–47
14. Guidelines, S.C.o.: Annex 2: cross-domain code lists 2009. Area (2009)
15. Stahl, A., Roth-Berghofer, T.: Rapid prototyping of cbr applications with the open source tool mycbr. In Althoff, K.D., Bergmann, R., Minor, M., Hanft, A., eds.: ECCBR. Volume 5239 of Lecture Notes in Computer Science., Springer (2008) 615–629
16. Lenz, M.: Case retrieval nets as a model for building flexible information systems (1999)
17. Roth-Berghofer, T.R.: Explanations and case-based reasoning: Foundational issues. In Funk, P., Gonzlez Calero, P.A., eds.: Advances in Case-Based Reasoning. Volume 3155 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2004) 195–209
18. Bahls, D., Roth-Berghofer, T.: Explanation support for the case-based reasoning tool mycbr. Proceedings of the TwentySecond AAAI Conference on Artificial Intelligence July 2226 2007 Vancouver British Columbia Canada (2007) 1844–1845

A method and guidelines for the cooperation of ontologies and relational databases in Semantic Web applications

Loris Bozzato[1⋆], Stefano Braghin[2], and Alberto Trombetta[3]

[1] Data and Knowledge Management Unit, Fondazione Bruno Kessler, Trento, Italy
[2] School of Computer Engineering, Nanyang Technological University, Singapore
[3] Dip. di Scienze Teoriche e Applicate, Univ. degli Studi dell'Insubria, Varese, Italy

bozzato@fbk.eu, s.braghin@ntu.edu.sg, alberto.trombetta@uninsubria.it

Abstract. Ontologies are a well-affirmed way of representing complex structured information and they provide a sound conceptual foundation to Semantic Web technologies. On the other hand, a huge amount of information available on the web is stored in legacy relational databases. The issues raised by the collaboration between such worlds are well known and addressed by consolidated mapping languages. Nevertheless, to the best of our knowledge, a best practice for such cooperation is missing: in this work we thus present a method to guide the definition of cooperations between ontology-based and relational databases systems. Our method, mainly based on ideas from knowledge reuse and re-engineering, is aimed at the separation of data between database and ontology instances and at the definition of suitable mappings in both directions, taking advantage of the representation possibilities offered by both models. We present the steps of our method along with guidelines for their application. Finally, we propose an example of its deployment in the context of a large repository of bio-medical images we developed.

1 Introduction

Ontology-based knowledge representation systems are well known to be successful in representing complex and heterogeneous information. In particular, recently, Semantic Web tools and systems permit to build and reason over ontologies providing logically founded representations and even increasing possibilities for data size. Moreover, the growing interest and availability of Semantic Web ontologies opens the possibility to reuse known data sources and, above all, to share and integrate information between systems.

On the other hand, the vast majority of data is nowadays stored in relational databases, so tools and techniques bridging ontology-based repositories and relational databases are needed in order to effectively deploy the potential provided by ontology-based representations. Significant efforts have been

⋆ This work has been realized while the first and second author were working at Univ. degli Studi dell'Insubria, Varese, Italy.

made to make possible to provide translations between ontologies and relational schemas in order to easily *publish* readily available database data: however, there is no accepted way on how to use such tools to let cooperate an existing relational database system with a paired ontology based system. For example, to the best of our knowledge, there is no method supporting the decision on what to represent and how to map information in both directions by using already available mapping languages and tools (such as D2R [5,6], Virtuoso RDFview[4] and Sponger[5] just to name a few).

In this work we propose our experiences in the collaboration of an ontology-based knowledge base and a legacy relational database under a single application. In particular, believing in the fact that the problems of this setting can be quite common, we try to generalize the approach that we chose in our case to a general method for the integration between an ontology and a relational database schema, when deployed together in a Semantic Web-based application. We refer to the definition of *method* provided in the context of knowledge re-engineering [17]: a set of "orderly processes or procedures used in the engineering of a product or performing a service". More precisely, we define a sequence of steps that an application designer may follow in order to decide *how* and *what* to map between an ontology and a relational database schema.

We point out that we do not aim at defining a novel mapping language between ontologies and relational schemas. Rather we aim at a method for deciding what are, loosely speaking, the relationships occurring between the ontology and the relational database, e.g. to decide what data stored in the relational database may be fruitfully published as RDF or on what data apply the inference tools proper to ontologies, that is, how to distribute data between both repositories in order to take advantage of the capabilities of the two representations. We have deployed our method and guidelines during the implementation of a large image database currently in use at a veterinary institute (namely, *Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna*, *IZSLER* for short[6]) serving a large user base distributed over more than fifteen sites in northern Italy. In the following we will briefly introduce the structure of the system that lead us to the formulation of this method: the *Imm@base system*, a repository of bio-medical images supporting advanced classification-based functionalities.

1.1 Motivating scenario

The definitions of the method and of the guidelines described in this paper have been carried out in the context of a project to satisfy the necessity of a major italian veterinary institute, the previously mentioned IZSLER. The requirement was to create a repository of biomedical pictures to be annotated with semantic information from well-known biomedical taxonomies, such as ICTVdb[7] and

[4] http://virtuoso.openlinksw.com/
[5] http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VirtSponger
[6] http://www.izsler.it
[7] http://www.ictvdb.org/

NCBI[8]. Moreover, the institute required a further classification of the pictures according to the medical cases they refer to. Such information is stored in a legacy RDB system, called DARWIN, which can not be modified for legal and pragmatic reasons. As an example of the firsts, the information stored in the database is used to quantify the refund which several farmers are entitled of in case of epidemics.

The architecture of the resulting application is shown in Fig. 1. According to the present architecture, the user interacts with the application through a web interface developed in PHP. Such interface provides, in a comprehensive way:

(i). a guided procedure to upload new pictures, properly annotated with metadata retrieved from the ontology database, which – we remind – contains both semantic data from the domain ontologies and a semantic technology-based representation of the data contained in the legacy DARWIN system used by the veterinary institute.
(ii). a web form for retrieving pictures and medical cases matching complex criteria defined by the user.

All the semantic data is retrieved via both ad-hoc and dynamically composed SPARQL queries used against a Joseki end-point. The domain ontologies data and the annotated pictures are stored in a PostgreSQL database while the DARWIN system uses of a SQLServer database. The first database has been created using the tools provided by the Jena API while the others are connected by means of D2RQ [5,6] mappings.

As it is easy to understand, the proposed architecture asks for the definition of a clear policy of cooperation between the semantic and relational repositories. After summarizing the related works on such cooperations, we introduce our solution, presented as a general method specification.

2 Related works

Several works address the issue of generating Semantic Web content from data stored in traditional databases. Such works can be classified, according to the chosen approach, in three categories:

(i). Annotation of the data extracted from databases with informations tracking how data have been obtained,
(ii). Mapping of the database model to an ontology,
(iii). Generation of an ontology related to the relational model of the DB.

The first approach works with the so-called DeepWeb [9] only and requires the database model to be public [8].

The second approach consists in mapping the database models to a given ontology by means of a mapping language in order to provide access to the content of the database as if it were a "semantic repository" [19]. Examples of

[8] `http://www.ncbi.nlm.nih.gov/`

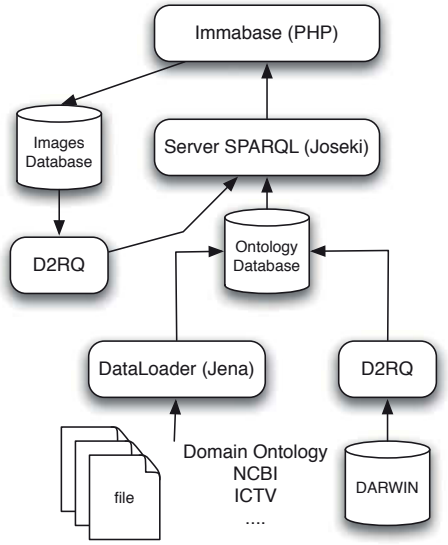


Fig. 1. The architecture of Imm@base web-application

such approach are D2RQ [5,6] and R2O [4]. The first one takes advantage of a proprietary mapping language, derived from the Jena assembler language, to allow the user to incorporate domain semantics in the mapping process. R2O, instead, is a XML based declarative language to express mappings between RDB elements and an ontology. The mappings realized with R2O can be used to "detect inconsistencies and ambiguities" in mapping definitions. A more detailed analysis of mapping languages and tools can be found in [14], where the authors also introduce interesting guidelines about how to further develop such mapping languages. In our work we take advantage of languages provided by works like [6] but proposing a more general methodology for the co-existence of databases and ontologies.

The third approach consists of the semi-automatic generation of an ontology from the database schema [15,16]. Such approach typically uses reverse-engineering techniques to generate the ontology from the database schema, like the ones we describe in Section 3.1, and to migrate the mapped data from the database creating ontological instances based on the tuples.

Moreover, several works have been presented with respect to the development of tools and algorithms to automatically match and merge ontology schemas, such as [1,3,13,18] (refer to [12] for a more detailed discussion). Such techniques may be used as tools for the identification of the common schema and for the definition of the mapping among the distinct repository which will be defined using the proposed method. Finally, [10] presents an ontology language, an example of formally defined mapping language and a query engine, all of which are based on the description logic *DL-Lite*.

To the best of our knowledge there are no proposals for methods pointing out the rationale and the steps one should follow in order to let a DB and an ontology-based store cooperate under a single system. The most similar work can be found in [2], where the authors present some use cases of integration of ontologies and relational databases. The main difference with our work is that in [2] there is the limitation of accessing data contained in the database read only, while our approach allows for the modification of data. Thus, the proposed method aims to the cooperation of data from repositories of different nature in order to provide the final user fully fledge access to the data, instead of a read-only RDF-based view of data stored in RDBMSs.

3 Cooperation method

The method we propose is aimed at guiding the *separation of data* between relational database (RDB) objects and ontology instances and *defining a suitable mapping* between the two repositories, in order to let them cooperate consistently. To achieve this, we have to address several issues, namely:

- the treatment of consistent references between the two schemas,
- the integration in an existing repository of an external data source,
- the identification of static and changing data,
- the decision on where to store schema instances.

The method defines a *mapping table*, which specifies, for each conceptual object (entity or relationship) in both the re-engineered repositories, where to store the respective instances and whether and how to refer to them. The mapping table should be sufficient to define a formal mapping between the sources, either by modifying the representation of conceptual objects in both sides or defining mapping in both directions e.g. by using mapping languages as D2RQ [5]. As we discuss in the guidelines (see Section 3.2), the choices about separation of instances should be guided by the cost and feasibility of modifications to each of the two knowledge bases. Note that the method does not assume the existence of one or both of the sources: if the ontology or the RDB already exists, its underlying conceptual model is extracted, otherwise the model has to be defined from the system specifications and requirements. The method mainly operates over conceptual representations of the two repositories: intuitively, entities correspond to ontology classes and DB tables, while relationships correspond to ontology properties and attributes in DB tables. Our method assumes that the conceptual models are to be defined in a formalism suitable for representing relevant properties of both sides: we assumed to use graphical models defined following the notations presented in [7,11]. Note that in the case of the RDB, defining such schema roughly corresponds to the extraction of its relational schema.

3.1 Method specification

In this section we present the tasks of our method using the following schema, derived from the definitions in [17,20]: we divide our method in *activities* composed

by *tasks*. In Fig. 2 we show the outline of the activities and tasks of our method: we shortly describe each task and its required *input* and *output* documents.

The method is composed by two distinct activities: the first one is a reverse engineering phase on the available information about DB and ontology, while the second is a forward engineering phase for the definition of the mapping. In the first activity **A1** the method analyzes the available description of the database and ontology (either the conceptual schema, the requirements or directly the sources structure) in order to extract a conceptual representation of the entire system. In **T1** and **T2**, thus, the conceptual schemas of the two sources is retrieved or generated from the descriptions. The two are combined in **T3** by recognizing and merging (possibly automatically [12]) the entities shared by both schemas. This represents the conceptual schema of the integrated system and it is the starting point to the following activity **A2**, in which the decision on the instance separation is taken and the related mappings are defined. In **T4** the entities which instances have to be shared are recognized by the knowledge engineer. In **T5** the decision about the distribution to such instances can take place, thus also defining the direction of mapping for their representation in the other schema. The same is done in **T6** for relationships. The last task **T7** consists in the logical modelling of the mapping, defining the actual objects in both knowledge stores to be mapped with the technical solutions of choice.

This structure is coherent with the one presented in [17,20] for the non-ontological resource re-engineering process: however, our method does not aim at the development of a new ontology but to a re-engineering of both sources in the context of the development of a semantic technology-based application. Moreover, we remark that our method can be applied either when one of the sources is available (by extracting its conceptual schema), when only its conceptual schema is available or when we just have the information (e.g. requirements) to derive the conceptual schema of each part. We also remark that some of the tasks described in the method specification (e.g. **T1** and **T2**) are easily mechanized, in particular when the knowledge bases are already present.

The outputs of our method are two *mapping tables*: the *entity table* and the *properties table*. The *entity table* should describe, for each conceptual entity:

- *Ontology class, DB table:* where its instances are stored,
- *Mapping:* logical mapping on classes and DB tables,
- *ID:* property chosen as identifier,
- *Source:* original source.

The *properties table* should describe, for each relation:

- *Ontology property, DB column:* where its instances are stored,
- *Mapping:* mapping on ontology properties and DB columns,
- *Domain and range:* conceptual entities linked by the property.

We present an execution of our method and an example of the resulting documents in the following sections.

A1. Conceptual Modeling Activity
Reverse engineering on sources to extract complete conceptual schema of the system.
Input: DB and ontology, their conceptual models or requirements
Output: Total conceptual model

T1. DB schema extraction
Extract conceptual schema from DB.
Input: DB, requirements or original DB schema
Output: DB conceptual schema

T2. Ontology schema extraction
Extract conceptual schema from ontology.
Input: Ontology, requirements or original ontology schema
Output: Ontology conceptual schema

T3. Total schema definition
Merge previous schemas to obtain a complete system schema.
Input: DB and ontology schemas
Output: Total conceptual model

A2. Mapping Definition Activity
Forward engineering on extracted model for the definition of mapping tables.
Input: Total conceptual model
Output: Complete mapping tables

T4. Shared schema extraction
Identify conceptual objects to be shared between schemas.
Input: Total conceptual model
Output: Shared conceptual schema

T5. Instances distribution
For every entity, decide where to store its instances.
Input: Shared conceptual schema
Output: Instances table

T6. Relationships distribution
For every relationship, decide where to store its instances.
Input: Shared schema, instances table
Output: Properties table

T7. Logical modeling
Define actual classes and tables to be mapped.
Input: Shared schema, mapping tables
Output: Complete mapping tables

Fig. 2. Method specification

3.2 Guidelines

In the following we suggest some guidelines useful for the application of our method and the definition of the mapping between DB and ontology. First of all, the following guidelines may drive the decision on which of the two schemas refer when storing instances of a conceptual object.

- *Ontology instances:* data can be stored as instance of ontology classes and properties mostly because it is necessary to draw inferences from this data. This can be useful when arranging data in complex taxonomies or meronimies or when it is needed to verify correctness of the data with respect to the ontology logical constraints. Another scenario where this choice is necessary is when one needs to comply with an external (possibly standard) ontology. In general, ontology instances should be treated as fixed and non-changing data, mostly representing "metadata" of the application to be developed.
- *RDB instances:* on the other hand, DB instances should represent the "working data" of the application, that is the data that one expect to be updated

and changed the most. Other reasons to leave out such data from the ontology include the fact that only simple queries (and no inferences) over this data are needed, or the fact that they represent only "administrative" data that is uninteresting to map and publish over the ontology.

Note that this means that the actual *data* would be stored in the DB, while the *metadata* would be stored as ontology instance. Note also that, in both cases, the choice can be affected by where the original instances were stored, the modifiability of the sources or the impact that these modification can cover. Moreover, as it is clear from the method, not every conceptual object in both parts takes part in the mapping: however, by re-engineering the conceptual schema we can decide to move an object from a schema to the other.

Once the choice on where to store instances has been done, the following guidelines suggest how to map and identify these instances in the two directions, so that they are visible in the other schema.

- *Instances in DB, class in ontology:* this partly corresponds to the case treated by mapping tools. From DB to ontology, entity instances can be mapped to individuals of a class naming them, e.g. as `ClassName_ID`. To identify in the DB mapped instances once they are retrieved from the ontology, one can map the ID or primary key of the instance as a `hasID` datatype property value referred to the ontology instance.
- *Instances in ontology, values in DB:* we can suggest different solutions to access or refer to external ontology instances into the DB records. A solution consists in directly using the URI of the referred individual in the DB tuples: however, this solution can be non satisfactory in that one can not check the validity and consistency of the references and can not add information to such individuals in the DB. Another solution is to keep in the DB a table relating the DB tuples to their counterpart individual in the ontology: additional data for the objects (not available in the ontology) can be stored in other columns of such table. A similar solution, but more demanding in terms of updates to the DB schema, consists in using the URI (or a transformation of it) of the ontology instance as the ID of the tuple of their counterpart in the DB. A relaxation of the previous solution consists in defining a transformation from the URI (or other property value) of ontology individuals to the value used as ID in the DB.

4 Example

In the following section we present a simple example for the application of the previously proposed method. More precisely, we present a simplification of the actual integration of the DB schema and ontology mentioned in our motivating scenario. Note that the operations described in the following of the section have been performed manually because of the dimension of the problem. In order to deal with more complex scenarios it will be required to develop tools supporting the tasks described in Section 3.

In the case of our example, we assume to already have the ontology conceptual schema (since basically we are adding a new ontology to an existing DB based system) and to be able to extract the conceptual schema from DB tables. Moreover, we assume (by relaxing the situation of our motivating scenario) that we can freely modify both parts. In our example, ideally, the DB mostly contains the data pertaining to the actual files representing medical images, while the ontology stores the relations between the concepts represented in the subject of images.

Given these premises, in the following we proceed trough the tasks of our method, providing examples for the most relevant produced documents. For simplicity, we only represent the properties and relations of the main entities of our system.

After the first two tasks **T1** and **T2**, given the previous assumptions, we obtain the conceptual schema of the DB and of the ontology, which are shown in Fig. 3 and Fig. 4. In particular, note that their structure is slightly different: e.g. the entity *Origin* (representing the method of acquisition of an image) appears only in the DB schema, while in the ontology schema *Image* is specialized in *MacroImage* and *MicroImage* (actual photographs versus microscopy images) which need to be treated differently in our system. Most notably, only the ontology contains the relations between the entities representing subject properties as in the case of *hasPosition* for *Lesion*.

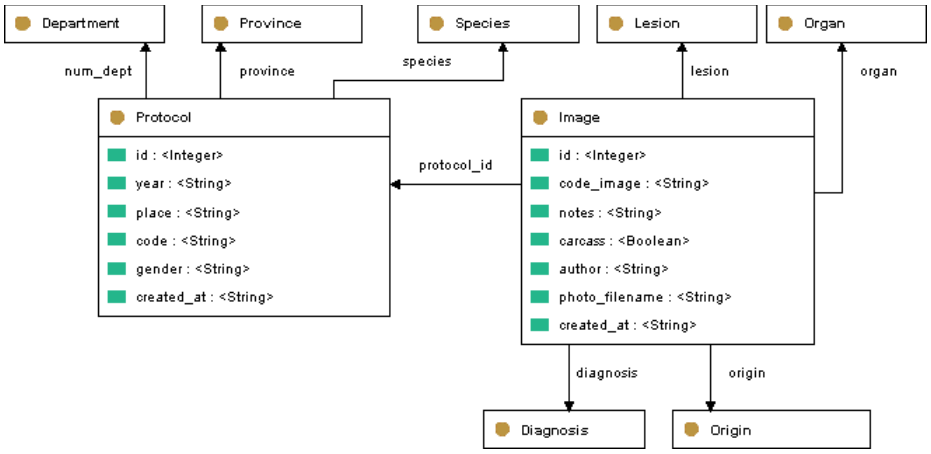


Fig. 3. DB Conceptual schema

In task **T3** we merge the two schemas in the total conceptual schema, considering the shared attributes: for example, note the case of the information about gender, in the DB represented as attribute and in the ontology as object property. We do not show this schema, for space and significance reasons. After obtaining the total schema, we can also begin to define the contents of the entity and property tables, mainly by filling in the names of entities and the properties with their specified domain and range.

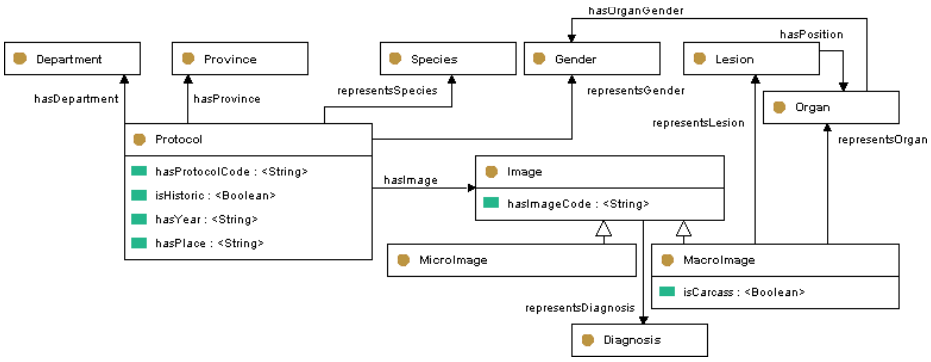


Fig. 4. Ontology Conceptual schema

We can now begin the forward engineering activity: the first task **T4** consists in identifying the shared schema in the total schema, which corresponds in picking out the instances that are not to be mapped, following the given guidelines. For example *Origin* and the attributes as *author* and *notes* only belongs to the DB while *MicroImage* is only to be contained in the ontology. The shared schema obtained in this task is shown in Fig. 5.

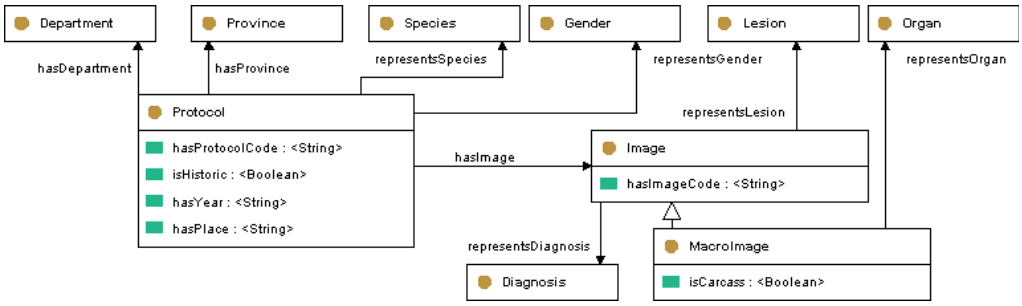


Fig. 5. Shared schema

The next two tasks **T5** and **T6** consist in separating the instances and relations of DB and ontology and thus defining the direction and the choices for the mapping, as suggested in our guidelines. For example, note that since *Protocol* and *Image* represent the data of our system, they are stored in the DB and mapped to their classes in the ontology. On the other hand, the objects actually detailed in the ontology have to be only referred in the DB instances.

In the last task **T7**, the mapping tables are completed with the actual DB tables and columns to be mapped. The final mapping tables for our example are shown in Table 1. Note that the proposed tables only contain relevant parts of the actual mapping tables for our schemas. We remark that the structure and notations used to present our method are simply suggestions for a manual execution of the method and can be replaced or hidden to the user in case of an implementation.

Entity	Mapping	Onto.Class	DB Table	ID	Source
Protocol	DB - O	Protocol	Protocol	Protocol.ID	O, DB
Image	DB - O	Image	Image	Image.ID	O, DB
MacroImage	DB - O	MacroImage	Image	Image.ID	O
MicroImage	O	MicroImage	-	URI	O
Species	O - DB	Species	-	URI	O, DB
Gender	O - DB	Gender	-	URI	O
Organ	O - DB	Organ	-	URI	O, DB
Origin	DB	-	Origin	Origin.ID	DB

Domain	Property	Range	Type	Mapping	DB column
Image (DB)	representsDiagnosis	Diagnosis (O)	object	DB - O	Image.diagnosis
	origin	Origin (DB)	object	DB	Image.origin
	hasImageCode	<string>	datatype	DB - O	Image.code_image
	notes	<string>	datatype	DB	Image.notes
	author	<string>	datatype	DB	Image.author
	...	...	...	...	...
MacroImage (DB)	representsOrgan	Organ (O)	object	DB - O	Image.organ
	representsLesion	Lesion (O)	object	DB - O	Image.lesion
	isCarcass	<boolean>	datatype	DB - O	Image.carcass
Lesion (O)	hasPosition	Organ (O)	object	O	-

Table 1. Entity and Property tables (excerpt)

5 Conclusions

In this paper we presented a method that allows a relational database and an ontology – as deployed in a Semantic Web application – to collaborate towards a fruitful distribution of data between them. We also provided guidelines in order to support the decisions to be taken in the deployment of our method. We defined the presented method motivated by the scenario of a system for the management of bio-medical images. In such project, semantic technologies are used to relate data from a relational database containing information about images to ontologies containing complex metadata classifying them.

The method we proposed in these pages represents a first step towards the definition of a generally applicable re-engineering process: for its further development, it is certainly necessary to refine and evaluate the proposal with experiences on several real-world applications scenarios. Moreover, the proposed guidelines are not thought to constitute a complete best practice, but they want to draw the attention to some relevant aspects of the cooperation and possibly promote discussion about these issues. Another interesting direction for further developments is the study of the automatization possibilities and the effective implementation for the tasks of our method.

Acknowledgments

We would like to thank the IZSLER institute for the support and collaboration.

References

1. Alexe, B., Chiticariu, L., Miller, R.J., Tan, W.C.: Muse: Mapping understanding and design by example. In: ICDE. pp. 10–19. IEEE (2008)

2. Auer, S., Feigenbaum, L., Miranker, D., Fogarolli, A., Sequeda, J.: Use Cases and Requirements for Mapping Relational Databases to RDF. RDB2RDF XG Working Draft, W3C (Jun 2010), http://www.w3.org/TR/rdb2rdf-ucr/
3. Aumueller, D., Do, H.H., Massmann, S., Rahm, E.: Schema and ontology matching with coma++. In: Özcan, F. (ed.) SIGMOD Conference. pp. 906–908. ACM (2005)
4. Barrasa, J., Corcho, O., Gómez-Pérez, A.: R2O, an Extensible and Semantically based Database-to-Ontology Mapping Language. In: Proceedings of SWDB2004. pp. 1069–1070. Springer (2004)
5. Bizer, C.: D2R MAP - A Database to RDF Mapping Language. In: Proceedings of WWW03 (Posters) (2003)
6. Bizer, C., Cyganiak, R.: D2RQ - Lessons Learned. W3C Workshop on RDF Access to Relational Databases (Oct 2007), http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/w3c-d2rq-positionpaper/
7. Brockmans, S., Haase, P.: A Metamodel and UML Profile for Networked Ontologies - A Complete Reference. Tech. rep., Institute AIFB, Universität Karlsruhe, Germany (2006)
8. Handschuh, S., Staab, S., Volz, R.: On deep annotation. In: Proceedings of WWW03. pp. 431–438 (2003)
9. He, B., Patel, M., Zhang, Z., Chang, K.C.C.: Accessing the deep web. Commun. ACM 50(5), 94–101 (2007)
10. Poggi, A., Lembo, D., Calvanese, D., Giacomo, G.D., Lenzerini, M., Rosati, R.: Linking data to ontologies. J. Data Semantics 10, 133–173 (2008)
11. Presutti, V.: D2.5.1. A Library of Ontology Design Patterns: reusable solutions for collaborative design of networked ontologies. NeOn Project Deliverable D2.5.1/v1.2, NeOn (Feb 2008)
12. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB J. 10(4), 334–350 (2001)
13. Raunich, S., Rahm, E.: Atom: Automatic target-driven ontology merging. In: Abiteboul, S., Böhm, K., Koch, C., Tan, K.L. (eds.) ICDE. pp. 1276–1279. IEEE Computer Society (2011)
14. Sahoo, S., Halb, W., Hellmann, S., Idehen, K., Thibodeau, T., Auer, S., Sequeda, J., Ezzat, A.: A Survey of Current Approaches for Mapping of Relational Databases to RDF. RDB2RDF XG Report, W3C (Jan 2009), http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf
15. Stojanovic, L., Stojanovic, N., Volz, R.: Migrating data-intensive web sites into the semantic web. In: Proceedings of SAC. pp. 1100–1107. ACM (2002)
16. Stojanovic, N., Stojanovic, L., Volz, R.: A reverse engineering approach for migrating data-intensive web sites to the semantic web. In: Proceedings of IFIP. pp. 141–154 (2002)
17. Suárez-Figueroa, M.C.: D5.4.1. NeOn Methodology for Building Contextualized Ontology Networks. NeOn Project Deliverable D5.4.1/v1.0, NeOn (Feb 2008)
18. Suchanek, F.M., Abiteboul, S., Senellart, P.: Paris: Probabilistic alignment of relations, instances, and schema. CoRR abs/1111.7164 (2011)
19. Sugumaran, V., Storey, V.C.: The role of domain ontologies in database design: An ontology management and conceptual modeling environment. ACM Trans. Database Syst. 31(3), 1064–1094 (2006)
20. Villazón-Terrazas, B.: D2.2.2. Methods and Tools Supporting Re-engineering. NeOn Project Deliverable D2.2.2/v2.0, NeOn (Feb 2009)

Yet Another Triple Store Benchmark? Practical Experiences with Real-World Data

Martin Voigt, Annett Mitschick, and Jonas Schulz

Dresden University of Technology, Institute for Software and Multimedia Technology,
01062 Dresden, Germany
{martin.voigt, annett.mitschick, jonas.schulz}@tu-dresden.de

Abstract. Although quite a number of RDF triple store benchmarks have already been conducted and published, it appears to be not that easy to find the right storage solution for your particular Semantic Web project. A basic reason is the lack of comprehensive performance tests with real-world data. Confronted with this problem, we setup and ran our own tests with a selection of four up-to-date triple store implementations – and came to interesting findings. In this paper, we briefly present the benchmark setup including the store configuration, the datasets, and the test queries. Based on a set of metrics, our results demonstrate the importance of real-world datasets in identifying anomalies or differences in reasoning. Finally, we must state that it is indeed difficult to give a general recommendation as no store wins in every field.

Keywords: RDF triple stores, benchmark, real-world datasets, reasoning, multi-user

1 Introduction

The last months inevitably reveal the advance of Semantic Web technologies in organizing and finding information, e. g., through the advent of *schema.org* or *Google Knowledge Graph* [1]. This is especially fostered by the widespread W3C standards like RDF(S), OWL, and SPARQL to allow for publishing and consuming Linked (Open) Data. As their benefits become more and more clear, also vendors in the publishing sector, e. g., *moresophy* [2], are applying these technologies to facilitate the semantic tagging and searching of media assets within their archives. An important and critical issue when developing large-scale Semantic Web applications is the right choice of an appropriate storage solution for RDF-based data. Comprehensive benchmarking results may help to estimate the applicability of state-of-the-art triple stores for ones own project.

Although, a number of performance reports already exist, we soon discovered that available results are of limited significance for our particular purposes. The goal of our research project *Topic/S* [3] is to provide a topic-based ranking and search of texts, images, and videos delivered by press media agencies. Therefore, we rely on the information automatically extracted from the media assets using NLP algorithms, but also consume other datasets to broaden the knowledge

graph of our archive, e. g., with information about people or organizations from New York Times [4] or YAGO2 [5], to improve the search. Thus, we heavily depend on a high-performance RDF storage solution – on the one hand for the extracted semantic data, and on the other hand for simulating public SPARQL endpoints for the required third party datasets (for the reason of continuous availability and serviceability).

A review of the existing work in the area of RDF store benchmarking [6] exposes that the results are interesting but not quite helpful for our use case due to varied reasons. A prominent reason is the lack of comprehensive tests on real-world datasets (non-synthetic). According to [7] automatically generated datasets used for benchmarking differ from real datasets, like *DBpedia* or *WordNet*, regarding "structuredness", and inevitably lead to different benchmarking results. BSBM [8, 9] is the most advanced benchmark available with regard to data size, parameters, or number of RDF stores. Unfortunately, the results are building on a generated dataset apart from our media archive domain. Further, the last test does not address SPARQL 1.1 [10]. Another project to be mentioned is *SP²Bench* [11] which dealt with the use of a broad range of possible SPARQL constructs and their combination – without reasoning. As the benchmark was carried out in 2008 on a generated dataset the results were also not beneficial for us. In contrast to that, the *FedBench* suite [12] focused on query federation on real-world datasets. However, the benchmark does not address RDFS reasoning nor SPARQL 1.1. Further, current RDFS stores like Virtuoso are not tested. A very comprehensive and most up-to-date survey on existing RDF stores is given in [13]. The results of this survey, carried out in the context the *Europeana* project [13], were build upon previous studies which the authors extended by their own benchmark using the (real-world) Europeana dataset (bibliographic metadata, approx. 380 million triples). Even though the results are the most up-to-date available (March 2011), the performance of the stores are maybe improved meanwhile. Moreover, the tests also did not consider RDFS reasoning, SPARQL 1.1, and heavy load (multiple queries in parallel).

In this paper we present the latest results of *"yet another"* benchmark of current RDF stores – but with the following *unique features*: loading and querying real-world datasets, testing RDFS reasoning and SPARQL 1.1 queries, conducting multiple queries in parallel, and recording the memory requirements.

Thus, the paper is organized as follows: In the next section, we briefly summarize the benchmark setting, including the selected RDF stores, the datasets and queries. In Section 3, we shortly introduce the metrics, present and discuss the results of our benchmark. A conclusion and outlook on future work is given in Section 4.

2 Benchmark Setup

In this section, we briefly introduce which RDF stores we have selected and how we set them up in our benchmark. In the second part, we present the four real-world datasets and the queries we utilized for our evaluation.

2.1 RDF Triple Stores

Although, there are more RDF stores available, we focused on Apache Jena [14], BigData RDF Database [15], OWLIM Lite [16], and OpenLink Virtuoso [17] within our benchmark due to the restrictions of our project setup: freely available, allows to handle up to 100 million triples, supports RDFS reasoning as well as SPARQL 1.1, and is build for the Java runtime environment.

The Apache Jena projects comes up with a bunch of sub-systems. For our purpose we needed a fast triple store as well as a SPARQL endpoint, thus, we relied on the Fuseki server with the version 0.2.3 which includes TDB 0.9.0 – a high performance RDF store. We used Fuseki with the default configuration. The RDFS reasoning is applied using an assembler description file.

BigData®is a high-performance RDF database which includes the NanoSparqlServer as SPARQL endpoint. We used the version 1.2.0 which comprises SPARQL UPDATE functionality. We employed the default setting except for setting up the RDFS reasoning within the RWStore.properties file.

Ontotext distributes three OWLIM editions. We deployed OWLIM-Lite 5.0.5 which builds on top of Sesame 2.6.5 and is freely available. Further, it is designed for datasets up to 100 million triples which fits our requirements.

Virtuoso provides an open source edition of their RDF store which includes a SPARQL endpoint. We installed version 6.1.5 und used the default setting. Inference is done only on runtime while querying.

Our benchmarks were conducted within a Ubuntu Linux 12.04 64bit virtual machine on a Intel Xeon CPU X5660 2.80GHz with 4 cores, 16GB RAM and 120GB virtual hard drive. The stores ran within Java 1.7.0 64bit runtime environment. If an application server was required, e. g., for OWLIM-Lite, we used Apache Tomcat 7.0.28.

2.2 Datasets and Queries

Especially as we want to reuse existing semantic datasets within the Topic/S project [3], e. g., to link named entities to DBpedia or YAGO2 [5], we chose to test the stores with real-world data. Furthermore, we could check if some of the stores had problems in loading or querying the data. To allow for a better comparability we transformed all sets to the N-Triple format what means that every row contains a single RDF triple. Therefore, for all but YAGO2 Core we used the current version of TopBraid composer. For YAGO2 we used the RDF2RDF tool [18]. The New York Times dataset [4] is originally distributed over four files which were merged. Table 1 gives an overview of the used sets which illustrates their difference of the general size but also of the schema and the number of instances.

We defined 15 queries per dataset using the SPARQL 1.1 query language. To compare the performance between stores and datasets at once, we created six general queries, e. g., counting the number of distinct subjects or the occurrence of the properties. The further nine queries are dataset-dependent and designed for real-world scenarios in Topics, for instance to retrieve the names of persons

	NY Times	Jamendo	Movie DB	YAGO2 Core
Size (in MByte)	56,2	151,0	891,6	5427,2
Triples (in Mio)	0,35	1,05	6,15	35,43
Instances (in k)	13,2	290,4	665,4	2648,4
Classes	19	21	53	292861
Properties	69	47	222	93

Table 1. Overview of the benchmark datasets

living in a dedicated time. Four of them are SELECTs with rising complexity, e. g., by using UNION, regex filters, or subqueries. Query 11 is used to investigate the DESCRIBE performance. In query 12 and 13 we especially considered the performance of RDFS inference. The last two UPDATE queries delete and insert triples. The complete list of queries can be found at [19].

3 Benchmark Results

In the following, we give a brief introduction to the metrics and the execution plan of our benchmark tests. The actual results of the tests are presented in Section 3.2.

3.1 Benchmark Metrics and Execution

In our benchmark we propose several metrics to capture different evaluation aspects for the stores. They are tested with all four datasets which are different in size and structure (c.f. Sect. 2.2). Furthermore, all the metrics are evaluated with and without RDFS reasoning (except the multi-client performance tests which were conducted exclusively with reasoning).

1. *Loading time*: At first, we measured the loading time of each dataset with the stores three times and calculated the average.
2. *Memory requirement*: Further, we measured the memory consumption of each store after the datasets were loaded.
3. *Per-query type performance*: Our main focus was to compare the query performance of the stores. We mainly distinguish between the generic, the dataset-specific, and the UPDATE queries. For our report we calculate the average but also extract the min and max values.
4. *Success rate*: We also investigate the success rate of each query. We define a query to be successful if it delivers the expected results without any error or timeout.
5. *Multi-client performance*: For the multi-client scenario we measured the average query performance as well as how many queries could be executed within a 10 minutes time slot.

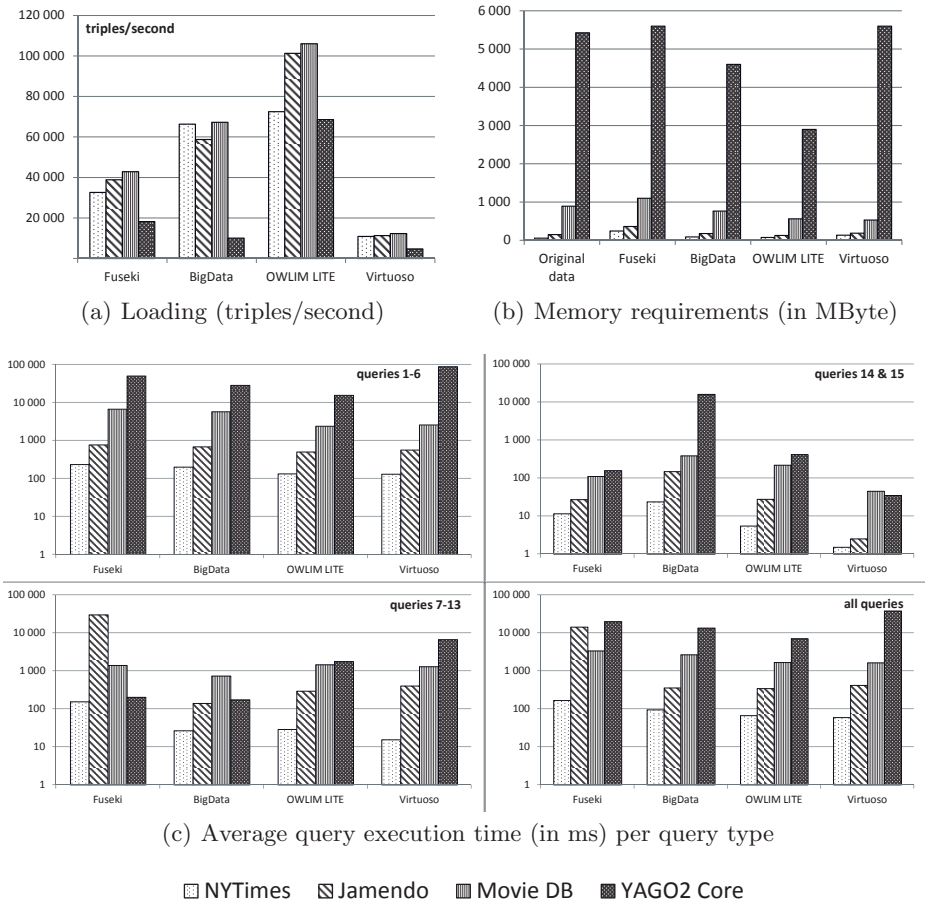
For the execution of the query benchmark we loaded the same dataset into all installed and pre-configured stores. Further, we wrote a simple Java client (test driver), which is available at our website [19], to rise automation and comparability. It requests a store with all 15 queries 20 times in a round robin manner. Having all values, we compute the average for each query. After all four sets were evaluated, we enabled RDFS reasoning for the stores, loaded the data, and did the same request using the test driver. Besides theses metrics, we evaluate how the stores scale in a multi-client scenario. Therefore, we loaded the NY Times dataset in every store with enabled RDFS reasoning. We selected four queries – two generic and to dataset-specific – which had approximately the same average execution time and called them using our test driver in a randomized round robin mechanism.

3.2 Experimental Results and Discussion

From our benchmark we gain several insights we will discuss in the following paragraphs. The selected Figures 1, 2 and 3 give a short summary of our results, whereas our website [19] provides more detailed information. Please mind, that Figures 1(c), 2(c) and 3 use logarithmic scaling.

Fig. 1 showcase the benchmark results with RFDS inference turned *off*. The first interesting finding is that OWLIM Lite performs best and Virtuoso worst to load all four datasets. BigData was fast but we identified a strange behaviour with the regex filters. The worked well for all datasets exempt from Jamendo. Here, all queries with a regex deliver no result. Further, query 3 and 13 on YAGO2 Core caused timeouts but BigDatas log didn't provide any insight to solve the problem. Next, Virtuoso is the store of your choice if you had many UPDATE transactions (query 14 and 15) to handle. If we compare all INSERT and DELETE queries of all datasets, it is approximately 4 times faster then second-placed Fuseki. With regard to all queries made, the performance of OWLIM Lite and Virtuoso is nearly the same. Only with the 35 million dataset Virtuoso fell back. Finally, we identified that Fuseki scales really bad with subquery requests on NY Times and Jamendo. Unfortunately, we could not identify a particular reason as the query complexity is quite the same as for the other two datasets.

Our findings regarding the abilities of the stores to support RDFS reasoning are displayed in Fig. 2. Here, Fig. 2(a) confirms that OWLIM Lite is approximately 80% faster than Virtuoso regarding the load performance. As Fuseki is almost as fast like without reasoning, we run into performance issues with BigData. For our technical setup, it was not possible to load YAGO2 Core into the store because the created temporary file exceeded the disk space. But all in all, the memory consumption of the store changes only slightly (Fig. 2(b)). The average execution times of the queries (c.f. Fig. 2(c)) show that Virtuoso is still the fastest on UPDATEs by far – approximately 8 times faster than the second-placed Fuseki. The performance of OWLIM and BigData decreases dramatically with this query type. Similar to the subquery problem of Fuseki we found that Virtuoso performs really bad with two generic queries (query 4 and 5) on the

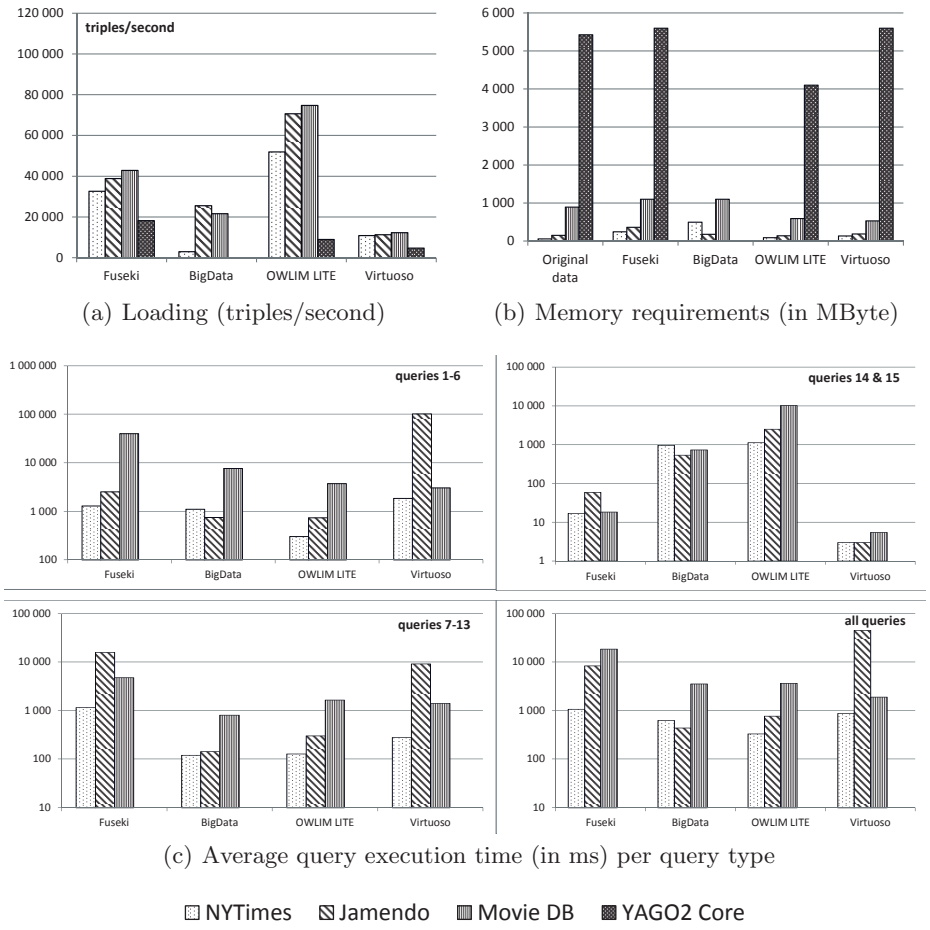


(a) Loading (triples/second)

(b) Memory requirements (in MByte)

(c) Average query execution time (in ms) per query type

Fig. 1. Overview of the results **without** RDFS reasoning: (a) describes the loading time of the stores and (b) shows the final memory consumption; (c) allows for comparing the performance per-query type



(a) Loading (triples/second)

(b) Memory requirements (in MByte)

(c) Average query execution time (in ms) per query type

Fig. 2. Overview of the results **with** RDFS reasoning: (a) describes the loading time of the stores and (b) shows the final memory consumption; (c) allows for comparing the performance per-query type. Here, YAGO2 is left out because of some shortcomings.

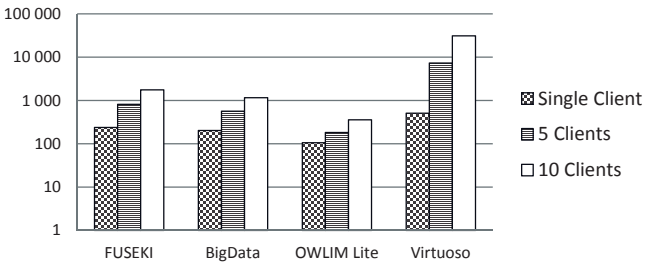


Fig. 3. Comparison on average query execution time (in ms) in our multi-client scenario using the NY Times dataset

Jamendo dataset. As the reason is not obvious it strengthens our finding that benchmarks on real-world datasets is important. In the end, we must state that the query performance decreases in general.

The bar chart in Fig. 3 illustrates the average query execution time in our multi-client setup. OWLIM Lite scales best with factor 2 from single to five as well as 5 to 10 clients. Fuseki and BigData are running shoulder on shoulder as both scale quite linear to the number of clients. For Virtuoso, the performance of query 2 does not depend on the number of clients. But it unfolds issues for some queries, e. g., query 9 is around 533 time slower with 10 clients compared to the single client scenario.

In the end, we want to review the error rate in a qualitative way. First, we need to state that we faced issues regarding YAGO2 Core with RDFS inference turned *on* so that we did not measure any query performance. For instance, BigData produces a temporary file which exceeded the disk space of our virtual machine or Fuseki timed out for some of the queries. Second, our generic queries comprise the SPARQL *COUNT* statement so that we could easily compare the results. Within the RDFS inference scenario the benchmark was very surprising as for the queries 1, 2, 3, and 6 *every* triple store returned a different result. Further, OWLIM Lite was the only store which counts more triples for query 4 and 5. Thus, our urgent advice for your own project is to cross-check the results at random if you need to use RDFS reasoning. Third, another anomaly we faced was that BigData had problems with *regex* filters but only on Jamendo dataset. This underlines again the need to benchmark an RDF store with different real-world datasets.

4 Conclusion

In this paper we present *yet another triple store benchmark* as we did not find anyone with evaluation criteria like they are required for our research project: loading and querying real-world datasets, testing RDFS reasoning and SPARQL 1.1 queries, as well as conducting multiple queries in parallel. In the following, we discuss our four findings.

As first result on our work with four up-to-date store we need to state, that comprehensive tests with real-world data are necessary. Otherwise it is not possible to detect anomalies like we identified. Second, every tested store allows for RDFS inference. But be careful as the result set may differ from store to store. Third, SPARQL 1.1 is well implemented nowadays. But the performance on UPDATE queries is varying. Here, Virtuoso stands out. Finally, we could not recommend any triple store in general as no store could win on all fields. Thus, the selection strongly depends on your specific project requirements. For our work, we will rely on OWLIM Lite because we need one which is fast in reading the datasets and multi-client query processing.

As we had problems with YAGO2 Core and inference, especially with regard to our technical setup, we are evaluating the requirements for bigger datasets. Besides the core version we like to benchmark the stores with YAGO2 Full as well. Another future work is to evaluate the performance of OWL reasoning for some common constructs, e. g., cardinalities. Therefore, we need to identify suitable real-world datasets.

5 Acknowledgments

Work on this paper is partly funded within the Topic/S project by the European Social Fund / Free State of Saxony, contract no. 99457/2677.

References

1. Google Knowledge Graph: `http://www.google.com/insidesearch/features/search/knowledge.html`.
2. Moresophy: L4 suite `http://www.moresophy.com/l4_suite` (only in German).
3. Topic/S: Project website `http://www.topic-s.de/` (only in German).
4. New York Times Dataset: `http://data.nytimes.com/`.
5. YAGO2 Dataset: `http://www.mpi-inf.mpg.de/yago-naga/yago/index.html`.
6. W3C Wiki: `http://www.w3.org/wiki/RdfStoreBenchmarking`.
7. Duan, S., Kementsietsidis, A., Srinivas, K., Udrea, O.: Apples and oranges: a comparison of rdf benchmarks and real rdf datasets. In: Procs. of the Intern. Conf. on Management of Data, ACM (2011) 145–156
8. Bizer, C., Schultz, A.: The berlin sparql benchmark. In: International Journal on Semantic Web & Information Systems. Volume 5. (2009)
9. Berlin SPARQL Benchmark: `http://www4.wiwiss.fu-berlin.de/bizer/BerlinSPARQLBenchmark/`.
10. Harris, S., Seaborne, A.: SPARQL 1.1 Query Language (October 2010)
11. Schmidt, M., Hornung, T., Lausen, G., Pinkel, C.: Sp2bench: A sparql performance benchmark. CoRR **abs/0806.4627** (2008)
12. Schmidt, M., Görlitz, O., Haase, P., Ladwig, G., Schwarte, A., Tran, T.: Fedbench: A benchmark suite for federated semantic data query processing. In: The Semantic Web ISWC 2011. Volume 7031 of LNCS. Springer (2011) 585–600
13. Haslhofer, B., Roochi, E.M., Schandl, B., Zander, S.: Europeana rdf store report. Technical report, University of Vienna, Vienna (March 2011)

14. Apache Jena: http://jena.apache.org.
15. BigData RDF Database: http://www.systap.com/bigdata.htm.
16. OWLIM: http://www.ontotext.com/owlim.
17. OpenLink Virtuoso: http://virtuoso.openlinksw.com/.
18. RDF2RDF: http://www.l3s.de/~minack/rdf2rdf/.
19. Voigt, M., Mitschick, A., Schulz, J.: Yet another triple store benchmark? practical experiences with real-world data (website) http://mt.inf.tu-dresden.de/topics/bench.

Implementing CIDOC CRM Search Based on Fundamental Relations and OWLIM Rules*

Vladimir Alexiev, Ontotext Corp

vladimir.alexiev@ontotext.com

Abstract. The CIDOC CRM provides an ontology for describing entities, properties and relationships appearing in cultural heritage (CH) documentation, history and archeology. CRM promotes shared understanding by providing an extensible semantic framework that any CH information can be mapped to. CRM data is usually represented in semantic web format (RDF) and comprises complex graphs of nodes and properties.
An important question is how a user can search through such complex graphs, since the number of possible combinations is staggering. One approach "compresses" the semantic network by mapping many CRM entity classes to a few "Fundamental Concepts" (FC), and mapping whole networks of CRM properties to fewer "Fundamental Relations" (FR). These FC and FRs serve as a "search index" over the CRM semantic web and allow the user to use a simpler query vocabulary.
We describe an implementation of CRM FR Search based on OWLIM Rules, done as part of the ResearchSpace (RS) project. We describe the technical details, problems and difficulties encountered, benefits and disadvantages of using OWLIM rules, and preliminary performance results. We provide implementation experience that can be valuable for further implementation, definition and maintenance of CRM FRs.

Keywords: CIDOC CRM, cultural heritage, semantic search, Fundamental Concepts, Fundamental Relations

1 Introduction

The CIDOC Conceptual Reference Model (CRM) [1], ISO Standard 21127:2006, provides an ontology for describing entities, properties and relationships appearing in cultural heritage (CH) documentation, history and archeology. CRM promotes shared understanding by providing an extensible semantic framework that any CH information can be mapped to. CRM data is usually represented in semantic web format (RDF) and comprises complex graphs of nodes and properties.

* This work is partially supported by the Andrew W. Mellon Foundation under the ResearchSpace project of the British Museum. The author thanks the anonymous referees for their feedback

An important question is how a user can search through such complex graphs, since the number of possible combinations is staggering. [2] presents one approach that "compresses" the semantic network by mapping many CRM entity classes to a few "Fundamental Concepts" (FC), and mapping whole networks of CRM properties to fewer "Fundamental Relations" (FR). These FC and FRs serve as a "search index" over the CRM semantic web and allow the user to use a simpler query vocabulary.

We describe an implementation of CRM FR Search based on OWLIM Rules [6], done as part of the ResearchSpace project [6] funded by the Andrew W. Mellon foundation and run by the British Museum. We describe the technical details of our approach, problems and difficulties encountered, benefits and disadvantages of using OWLIM rules, and preliminary performance results. We provide implementation experience that can be a valuable guide for the further implementation, definition and maintenance of CRM FRs.

The FP7 project 3D COFORM [7] is also implementing FR search, and we have established a collaboration.

2 Example: Thing from Place

As an example, let's consider the FR "Thing from Place". It is intended to capture all alternatives through which a Thing's **origin** can be related to Place, and is defined in [8] as:

```
FC70_Thing --(P46i_forms_part_of* | P106i_forms_part_of* | P148i_is_component_of*)-> FC70_Thing:
 {FC70_Thing --(P53_has_former_or_current_location | P54_has_current_permanent_location)-> E53_Place:
  {E53_Place --P89_falls_within*-> E53_Place}
 OR FC70_Thing --P92i_was_brought_into_existence_by-> E63_Beginning_of_Existence:
  {E63_Beginning_of_Existence --P9i_forms_part_of*-> E5_Event:
   {E5_Event --P7_took_place_at-> E53_Place:
    {E53_Place --P89_falls_within*-> E53_Place}
   OR E7_Activity --P14_carried_out_by-> E39_Actor:
    {E39_Actor --P107i_is_current_or_former_member_of* -> E39_Actor:
     {E39_Actor --P74_has_current_or_former_residence  -> E53_Place:
      {E53_Place --P89_falls_within*-> E53_Place}
     OR E39_Actor --P92i_was_brought_into_existence_by-> E63_Beginning_of_Existence:
      {E63_Beginning_of_Existence --P9i_forms_part_of*-> E5_Event:
       {E5_Event --P7_took_place_at-> E53_Place:
        {E53_Place --P89_falls_within* -> E53_Place}}}}}}}
 OR E19_Physical_Thing  --P25i_moved_by-> E9_Move:
  {E9_Move --(P26_moved_to | P27_moved_from)-> E53_Place:
   {E53_Place  --P89_falls_within*-> E53_Place}}
 OR E19_Physical_Object --P24i_changed_ownership_through-> E8_Acquisition:
  {E8_Acquisition --P9i_forms_part_of*-> E5_Event:
   {E5_Event --P7_took_place_at-> E53_Place:
    {E53_Place --P89_falls_within*-> E53_Place}}}}
```

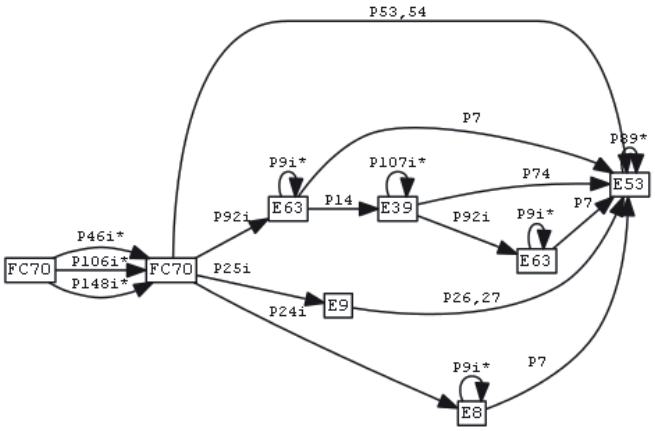
Note: we've made the following (mostly notational) simplifications:

- removed the construct "--P2F.has_type-> E55.Type" (allowing to search by event type) from a number of places
- removed "C2.Finding" which is a Find event of interest to archeology, defined in 3D COFORM but not part of CRM proper
- renamed "C1.Object" to "FC70_Thing" (which stands for Fundamental Concept "Thing")
- used Erlangen CRM [9] notation for entities (e.g. E5_Event) and properties (e.g. P89_falls_within, P24i_changed_ownership_through)
- used "property*" instead of "(property)(0,n)" to denote reflexive-transitive closure, and later use "property+" to indicate transitive closure
- used SPARQL Property Paths notation [10]: "(prop1 | prop2)" instead of "{prop1 OR prop2}" to indicate alternative (disjunction)

2.1 Interpretation and Graphical Represenatation

This FR can be interpreted as follows, where "(...)*" means "optionally and recursively" i.e. reflexive-transitive closure:

- a Thing (part of another Thing)* is considered to be "from" Place if it:
- is formerly or currently located at Place (that falls within another)*
- or was brought into existence (produced/created) by an Event (part of another)*
  - that happened at Place (that falls within another)*
  - or was carried out by an Actor (who is member of a Group)*
    - who formerly or currently has residence at Place (that falls within another)*
    - or was brought into existence (born/formed) by an Event (part of another)* that happened at Place (that falls within another)*
- or was Moved to/from a Place (that falls within another)*
- or changed ownership through an Acquisition (part of another)*
  - that happened at Place (that falls within another)*



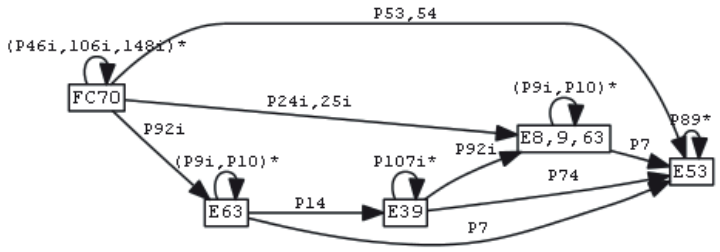
Although defined as a tree of property paths, the FR is better depicted as a network through a simple merge of leaf-level nodes

2.2 Corrections and Rationalization

We reviewed each FR, made some corrections, and rationalized the network. This FR:

- Allowed paths of mixed properties (e.g. P46i,P106i) at the beginning
- Allowed a loop P9i* at E9 (Move forms part of a bigger event) by merging the nodes E8, E9, and the second E63
  - We could even merge the first E63, but then we'd have a back-link, so before traversing P14 must check that the event is E12, E65, or E81 (i.e. the production/creation of a Thing), so that won't lead to simplification
- Allowed P10_falls_within in addition to P9i_forms_part_of (after consultation with the original authors)
- Skipped P26,P27 since these are subproperties of (infer) P7, so it's enough to check for P7

The result is this network:



3 Inverses, Transitive, Parallel-Serial Networks

The example above suggests several implementation considerations:

- Most CRM properties have an inverse and [9] declares them as owl:inverseOf (symmetric properties are their own inverse). FRs use CRM properties in both directions: forward (e.g. P53_has_former_or_current_location) and inverse (P24i_changed_ownership_through), so it's useful to rely on owl:inverseOf inferencing
- FRs use transitive closure (denoted +) to traverse the various "part" hierarchies of CRM (physical object parts, conceptual object parts, sub-places, sub-events), so it's useful to rely on owl:TransitiveProperty inferencing. CRM scope notes suggest that 14 properties (and their inverses) should be transitive: P9 P10 P46 P86 P88 P89 P106 P114 P115 P116 P117 P120 P127 P148. [9] declares them as owl:TransitiveProperty (except P9 P46 that were forgotten, so we declared them). In addition to these "atomic" properties, disjunctions of properties often also need to be declared as transitive.

- FRs often use reflexive-transitive closure (denoted *). However, we have opted not to use reflexive closure in the implementation, since it would generate a lot of trivial facts (self-loops). We use disjunction instead: the iterated property is applied 0 times in the first disjunct, and *n* times in the second
- FRs are defined *mostly* as parallel-serial networks of properties, using SPARQL Property Paths constructs [10]

3.1 Decomposing Thing from Place Into sub-FRs

The example network in section 2.2 can be decomposed into "sub-FRs" as follows. We use prefix FRT for a transitive sub-FR, FRX for a non-transitive sub-FR, and FR for the final result: FR7 "thing from place". A major challenge has been coming up with names for these sub-FRs, so we used numbering from CRM properties

```
    # self-loops and simple disjunctions
FRT_46i_106i_148i := (P46i|P106i|P148i)+
FRT_9i_10 := (P9|P10)+
FRT_107i := P107i+
FRT_89 := P89+
FRX_53_54 := (P53|P54)
FRX_24i_25i := (P24i|P25i)
    # growing fragments
FRX_92i := P92i | P92i/FRT_9i_10
FRX_92i_14 := FRX_92i/P14 | FRX_92i/P14/FRT_107i
FRX_FC70_E8_9_63 := FRX_92i_14/P92i | FRX_24i_25i
FRX_FC70_E8_9_63_P7 := FRX_FC70_E8_9_63/P7 | FRX_FC70_E8_9_63/FRT_9i_10/P7
FRX7 := FRX_53_54 | FRX_FC70_E8_9_63_P7 | FRX_92i_14/P74 | FRX_92i/P7
FRX7_P89 := FRX7 | FRX7/FRT_89
FR7 := FRX7_P89 | FRT_46i_106i_148i/FRX7_P89
```

3.2 Implementing Networks with RDFS and OWL

Parallel-serial networks can be implemented wholly within the RDFS and OWL vocabularies (we express the implementation fragments in RDF Turtle):

Pattern	Construct	Implementation
inverse	prop := ^prop1	prop1 owl:inverseOf prop2.
parallel	prop := prop1|prop2	prop1 rdfs:subPropertyOf prop. prop2 rdfs:subPropertyOf prop.
serial	prop := prop1/prop2	prop owl:PropertyChainAxiom (prop1 prop2).
transitive	prop := prop1+	prop1 rdfs:subPropertyOf prop. prop owl:TransitiveProperty
reflexive-transitive	prop := prop1 prop2*	Converted to the following: prop := prop1 | (prop1/prop2+)

3.3 Type Checking and Conjunctive Properties

The original definition [8] supposes type checks for every node (FC70, E63, etc). So for example the final definition of the target FR should be:

```
x FR7_from_place y := x a FC70_Thing;  x FR7 y; y a E53_Place.
```

Here x,y are variables, "a" stands for rdf:type as usual, and ";" separates triple patterns and stands for conjunction.

In many cases the type checks can be skipped since they are implied by the appropriate property ranges. E.g. all of P53 P54 P7 P47 P89 have range E53, so there is no need to check the type of the final node.

But in some cases type checks are required, e.g. for the "about" FR family that applies to various FCs and is segmented into several FRs: Thing about Thing, Thing about Place, Thing about Actor, etc. If the type check is at the first or last node, it can be added in SPARQL. But if the type check is needed in the middle of a network, we need a conjunctive property.

Unfortunately properties cannot be defined by conjunction in OWL 2 [11]. While the same answer suggests that adding role conjunctions in DLs increases computational complexity significantly, [12] shows conditions under which role conjunction can be added without increase in complexity. In particular, OWL RL can be extended with role conjunctions without any restrictions or increase in complexity, and [13] proposes extending OWL with such capabilities. Such extensions may become available in a future OWL version (OWL 3)

4 OWLIM Rules

Because of the difficulty described in 3.3, we chose to implement FRs using OWLIM Rules [6]. OWLIM [4,5] is a semantic repository by Ontotext Corp that provides high-performance and scalability, comprehensive OWL RL and QL reasoning, custom rules, incremental assert/retract, clustering and other enterprise features.

OWLIM Rules use simple unification: a set of premise triple patterns is checked against the repository, and if it matches, a set of consequence triples is inferred and stored in the repository. The rules are translated to Java bytecode for speed.

The OWLIM Rules syntax is verbose (one line per premise/conclusion). Since we had to define a lot of rules, we defined a simpler syntax (one line per rule, see examples below) that we translate using a simple script. The syntax is similar to N3 Rules, but simpler.

RDFS and OWL2 are implemented in OWLIM using OWLIM Rules. The user loads different rule sets (PIE files) depending on the required reasoning capabilities. We started from RDFS that implements sub-class and sub-property reasoning, and added a bit of OWL that implements inverse and transitive reasoning:

```
p <rdf:type> <owl:TransitiveProperty>; x p y; y p z => x p z
p1 <owl:inverseOf> p2; x p1 y => y p2 x
p1 <owl:inverseOf> p2; x p2 y => y p1 x
```

The implementation of owl:propertyChainAxiom is more complex (using the full rules syntax), mostly because it deals with RDF list iteration. We don't use it for the current implementation:

```
Id: prp_spo2_1
   p <owl:propertyChainAxiom> pc
   start pc last              [Context <onto:_checkChain>]
   ---------------------------
   start p last
Id: prp_spo2_2
   pc <rdf:first> p
   pc <rdf:rest> t            [Constraint t != <rdf:nil>]
   start p next
   next t last                [Context <onto:_checkChain>]
   ---------------------------
   start pc last              [Context <onto:_checkChain>]
Id: prp_spo2_3
   pc <rdf:first> p
   pc <rdf:rest> <rdf:nil>
   start p last
   ---------------------------
   start pc last              [Context <onto:_checkChain>]
```

Then we added specific rules for the FRs. We used a Literate Programming style to intersperse FR definitions and discussions with FR implementation in our wiki, then weaved the final FR rules using a simple script.

4.1 Benefits of OWLIM Rules

The important benefits of OWLIM Rules used in our implementation are:

- Speed: OWLIM uses forward-chaining materializing inference, so consequences are stored in the repository and query answering is very fast. Custom rules are treated just like system rules.
- Rules are "reversible": when a triple is retracted, all relevant rules are checked. If an inferred triple matches the consequences and there are no other triples that support it, the triple is retracted as well. This suppors incremental retract and is extremely important for high-update use cases such as BBC World Cup, BBC Olympics, and ResearchSpace.
- Support conjunctive checks, i.e. overcome the problem described in section 3.3

4.2 Disadvantages of OWLIM Rules

The main disadvantages of OWLIM rules are:

- They are not flexible: every time the rule set is changed, the repository needs to be reloaded from scratch. In contrast, once the RDFS/OWL vocabularies are implemented as rules (see section 4), adding a meta-property (e.g. owl:TransitiveProperty or owl:inverseOf) recomputes all relevant consequences dynamically.
- They are proprietary to OWLIM. Ontotext is considering the implementation of proposed standard rule languages in future OWLIM versions.
- They don't support negation in a real sense (e.g. one can check that a rule variable is not bound to a specific class, but cannot check that a variable does not have a specific type or one of its sub-classes). Implications of this are discussed in sections 5.1 and 6.

5 Results and Performance

Once each FR is depicted as a diagram similar to the one in 2.2, the implementation as OWLIM rules is straightforward if tedious. E.g. the first line in the decomposition shown in 3.1 is implemented as these 3 rules ("rso" stands for "ResearchSpace Ontology"):

```
x <crm:P46i_forms_part_of> y => x <rso:FRT_46i_106i_148i> y
x <crm:P106i_forms_part_of> y => x <rso:FRT_46i_106i_148i> y
x <crm:P148i_is_component_of> y => x <rso:FRT_46i_106i_148i> y
```

We have implemented 11 FRs of Thing:

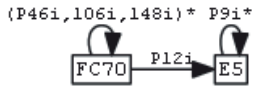
- refers to or is about Place: FR67_refers_to_or_is_about
- from Place: FR7_from_place
- is/was located in Place: FR53_is_was_located_in
- has met Actor: FR12_has_met
- by Actor: FR14_by
- refers to or is about Event: FR67_about_event
- has met Event: FR12_was_present_at
- is made of Material: FR45_is_made_of
- is/has Type: FR2_has_type
- used technique: FR32_used_technique
- identified by Identifier: FR1_identified_by

This took 86 OWLIM rules and 10 axioms. They use 44 source properties (from CRM0 and define and use 26 intermediate properties (sub-FRs, see 3.1).

5.1 Bug in Thing has met Event

We found a "bug" in the definition of Thing has met Event (FR12_was_present_at) that causes quadratic growth and exponential slowdown of data loading. The rule is defined benignly enough:

```
FC70_Thing --FR12_was_present_at-> E5_Event :=
  FC70_Thing --(P46i_forms_part_of | P106i_forms_part_of | P148i_is_component_of)* ->
    FC70_Thing --P12i_was_present_at-> E5_Event:
      E5_Event --P9i_forms_part_of*-> E5_Event
```



ResearchSpace currently deals with RKD and British Museum data, and we model an acquisition as an event having several of these types:

- E8_Acquisition: changes the current owner
- E10_Transfer_of_Custody: changes the current keeper
- E80_Part_Removal: removes the object from the old collection
- E79_Part_Addition: adds the object to the new collection

The acquisition is an event at which meet the object, buyer, seller, old collection and new collection. The object is part of the old collection (before the acquisition) and part of the new collection (after the acquisition). Because P46i_forms_part_of is included in the definition, this causes all objects in a collection to have met (witnessed) the acquisition of all other objects in the collection. This is logically undesirable:

- If Thing2 was added to Collection after Thing1, it's causally impossible for Thing2 to be present at the acquisition of Thing1
- If Thing2 was added to Collection before Thing1, one **could** say Thing2 quietly observed the addition of Thing1, but that is not really useful

More importantly for us, this is computationally very expensive for a large collection such as the British Museum that has over 1.5M objects.

We considered fixing the problem by adding a clause that the target of P46i_forms_part_of is not E78_Collection. However, OWLIM rules don't support true negation, so for the time being we've simply removed P46i from the definition, since our data does not deal with object parts.

5.2 Performance

Concerns were expressed that materializing sub-FR triples may increase the repository size too much and slow it down. We have preliminary performance results that are very promising and dispel these fears:

- A small repository of 11 Rembrandt paintings had 1.5M triples, including about 0.5M object triples (complex data about each painting, researches, documents, etc) and 1M thesaurus triples (people, places, etc). The FRs added only 25.8k triples, which is 1.7% of the total data or 5.1% of the object data.
- A large repository of over 1.5M British Museum objects and about 200M triples performs FR searches with no noticeable slow-down.

5.3 Benefits Compared to Straight SPARQL

To appreciate the query simplification that FRs afford, compare this simple query using the FR "Thing from Place" defined in sec. 3:

```
select * {?t FR7_from_place ?y}
```

To a "straight SPARQL" query:

```
select ?t ?p2 {
?t a FC70_Thing. ?t (P46i_forms_part_of* | P106i_forms_part_of* | P148i_is_component_of*) ?t1.
 {?t1 (P53_has_former_or_current_location | P54_has_current_permanent_location) ?p1}
 UNION
 {?t1 P92i_was_brought_into_existence_by ?e1. ?e1 P9i_forms_part_of* ?e2.
    {?e2 P7_took_place_at ?p1}
    UNION
    {?e2 P14_carried_out_by ?a1.
     ?a1 P107i_is_current_or_former_member_of* ?a2.
      {?a2 P74_has_current_or_former_residence ?p1}
      UNION
      {?a2 P92i_was_brought_into_existence_by ?e3. ?e3 P9i_forms_part_of* ?e4.
       ?e4 P7_took_place_at ?p1}}}
 UNION
 {?t2 P25i_moved_by ?e5. ?e5 (P26_moved_to | P27_moved_from) ?p1}
 UNION
 {?t2 P24i_changed_ownership_through ?e6.
  ?e6 P9i_forms_part_of ?e7. ?e7 P7_took_place_at ?p1}.
?p1 P89_falls_within* ?p2}
```

Even though it uses SPARQL 1.1 shortcut notation (Property Paths), the query is complex. It is also expensive, since it considers many alternative paths. When you consider that FRs are usually used in combination, the resulting queries become too complex. K.Tzompanaki reports that an FR implementation approach using straight SPARQL queries quickly becomes hard to manage (personal communication).

6 Summary and Future Work

We presented an implementation of CRM Search based on the "Fundamental Concepts" and "Fundamental Relations" approach [2]. FC and FRs serve as a "search index" over complex CRM semantic networks and allow the user to use a simpler query vocabulary.

Our implementation uses OWLIM Rules and was done over large repositories of Cultural Heritage objects. We describe the technical details, problems and difficulties encountered, benefits and disadvantages of using OWLIM rules, and preliminary performance results. We provided implementation experience that can be valuable for further implementation, definition and maintenance of CRM FRs.

Future work in this direction can include:

- Implement more FRs in collaboration with the 3D COFORM project. This includes more FRs of Thing, as well as FRs of other Fundamental Concepts (Person, Event, etc) that are not yet defined.
- Automate the discovery of shared sub-FRs to facilitate the implementation
- Take care of complications related to negation

7 References

1. CIDOC CRM website, http://www.cidoc-crm.org
2. Katerina Tzompanaki, Martin Doerr: A New Framework for Querying Semantic Networks. FORTH technical report TR419, May 2011
3. ResearchSpace project, http://www.researchspace.org
4. OWLIM website, http://www.ontotext.com/owlim
5. Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, Ruslan Velkov, OWLIM: A family of scalable semantic repositories, Semantic Web Journal, Volume 2, Number 1, 2011.
6. Barry Bishop, Spas Bojanov. Implementing OWL 2 RL and OWL 2 QL rule-sets for OWLIM. Proc. of 8th International Workshop on OWL: Experiences and Directions (OWLED 2011), San Francisco, USA, June 5-6, 2011, CEUR-WS.org, ISSN 1613-0073
7. FP7 project 3D COFORM, http://www.3d-coform.org
8. Katerina Tzompanaki, Martin Doerr: Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories, Technical Report ICS-FORTH/TR-429, April 2012, http://www.cidoc-crm.org/docs/TechnicalReport429_April2012.pdf
9. Erlangen CRM mapping of CRM to OWL, http://erlangen-crm.org
10. SPARQL Property Paths, http://www.w3.org/TR/sparql11-property-paths
11. Answer on SemanticWeb.com, http://answers.semanticweb.com/questions/11602/-property-intersection-impossible-in-owl-2-full
12. Sebastian Rudolph, Markus Krötzsch, Pascal Hitzler: Cheap Boolean Role Constructors for Description Logics. Proceedings of 11th European Conference on Logics in Artificial Intelligence (JELIA 2008), p362-374, LNAI 5293, Sep 2008
13. Pascal Hitzler, Suggestions for OWL 3, Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2009), Oct 2009. CEUR 529, http://ceur-ws.org/Vol-529/owled2009_submission_6.pdf