



Proceedings of the 3rd International Workshop on **Semantic Digital Archives**

held in conjunction with the
17th Int. Conference on Theory and Practice of Digital Libraries (TPDL)
on September 26, 2013 in Valetta, Malta.

<http://sda2013.dke-research.de>

Edited by

Livia Predoiu, University of Oxford, UK

Annett Mitschick, University of Dresden, Germany

Andreas Nürnberger, University of Magdeburg, Germany

Thomas Risse, University of Hannover, Germany

Seamus Ross, University of Toronto, Canada

November, 2013

Copyright © 2013 for the individual papers by the papers' authors. Copying permitted only for private and academic purposes. This volume is published and copyrighted by its editors.

Acknowledgement: Original photo of the Siege Bell War Memorial in Valetta/Malta taken by Seán Russell, PhD Researcher, Clarity Centre for Sensor Web Technologies, School of Computer Science & Informatics, Science North, University College Dublin, Belfield, Dublin 4 Ireland, Sean.Russell@ucd.ie.

Preface

The 3rd Workshop on Semantic Digital Archives (SDA 2013) has built upon the success of the previous editions in 2011 and 2012, and has been held in conjunction with the International Conference on Theory and Practice of Digital Libraries, TPDL (formerly known as European Conference on Digital Libraries, ECDL). Organized as full-day workshop, SDA 2013 has aimed to advance and discuss appropriate knowledge representation and knowledge management solutions specifically designed for improving Archival Information Systems. The main objective was 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 was indeed 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.

The talks given at the workshop were joined by an invited keynote by *Giorgio Maria Di Nunzio* (University of Padua, Italy) on the subject of *Digital Geolinguistics: On the Use of Linked Open Data for Data-Level Interoperability between Geolinguistic Resources*, which gave an interesting overview on how linked data is used within the Open Language Archives community.

In the paper *Towards Preservation of semantically enriched Architectural Knowledge* Stefan Dietze et al. present an overview of early work within the DURAARK project for creating a semantic digital archive for the building and architecture domain.

Nina Tahmasebi and Thomas Risse (*The Role of Language Evolution in Digital Archives*) discuss problems that arise as a result of language evolution in digital archives and review methods to make digital archives semantically accessible and interpretable in the future.

With *BlogNEER: Applying Named Entity Evolution Recognition on the Blogosphere?* Helge Holzmann et al. propose an approach for the identification of named entity changes within blog content using a novel semantic filtering method and semantic resources like DBpedia.

In the paper *Elevating Natural History Museums' Cultural Collections to the Linked Data Cloud* Giannis Skevakis et al. introduce an infrastructure and methodology developed and applied for the transition and conversion of cultural heritage metadata to Linked Data within the Natural Europe project.

Alexandre Rademaker et al. showcase *A linked open data architecture for contemporary historical archives*, being proposed for the Center for Teaching and Research in the Social Sciences and Contemporary History of Brazil (CPDOC).

In *Pundit: Creating, Exploring and Consuming Semantic Annotations* Marco Grassi et al. present new results and aspects of their matured project (a semantic web annotation tool) focusing on demonstrative use cases for sharing, exploring and visualizing semantic annotations.

Stefan Haun and Andreas Nürnberger (*Towards Persistent Identification of Resources in Personal Information Management*) discuss the requirements and critical aspects of persistent identifiers for resources and entities in Personal Information Management, especially in personal file systems, which is also important in digital archives and digital libraries.

The paper *A Storage Ontology for Hierarchical Storage Management Systems* (Sandro Schmidt et al.) introduces an ontology-based concept to extensively describe files, their properties and interrelations as a basis to decide whether they should be stored on fast or cheap storage media.

In *Semantic Retrieval Interface for Statistical Research Data* Daniel Bahls and Klaus Tochtermann propose an intuitive multi-step retrieval prototype which supports researchers in finding, accessing and composing distributed statistical data sets.

Finally, the paper *Semantification of Query Interfaces to Improve Access to Deep Web Content* (Arne Martin Klemenz and Klaus Tochtermann) discusses the limitations of Deep Web Information Retrieval and the significant potential of additional semantic annotations and proposes the idea of 'Semantic Deep Web Services'.

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. During the workshop itself we had many fruitful and inspiring discussions which would not have been possible without the well done presentations and the interested audience. Many thanks to all workshop attendants for a great workshop!

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

November 2013

L. Predoiu, A. Mitschick, A. Nürnberger, T. Risse and S. Ross

Program Committee

Sören Auer	University of Leipzig, Germany
Kai Eckert	University of Mannheim, Germany
Marco Grassi	Università Politecnica delle Marche, Italy
Tudor Groza	University of Queensland, Australia
Armin Haller	CSIRO, Australia
Andreas Harth	KIT, Karlsruhe, Germany
Steffen Hennicke	Humboldt-Universität zu Berlin, Germany
Stijn Heymans	SRI International, USA
Pascal Hitzler	Wright State University, USA
Claus-Peter Klas	FernUniversität in Hagen, Germany
Thomas Lukasiewicz	University of Oxford, UK
Mathias Lux	Klagenfurt University, Austria
Knud Möller	Datalysator, Berlin, Germany
Kai Naumann	State Archive of Baden-Württemberg, Germany
Gillian Oliver	Victoria University of Wellington, New Zealand
Jacco van Ossenbruggen	VU University Amsterdam, Netherlands
Andreas Rauber	Vienna University of Technology, Austria
Sebastian Rudolph	University of Dresden, Germany
Heiko Schuldt	Universität Basel, Switzerland
Kunal Sengupta	Wright State University, USA
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

Invited Talk:

Digital Geolinguistics: On the Use of Linked Open Data for Data-Level Interoperability between Geolinguistic Resources	1
<i>Giorgio Maria Di Nunzio</i>	

Digital Preservation and Language Evolution

Towards Preservation of semantically enriched Architectural Knowledge	4
<i>Stefan Dietze, Jakob Beetz, Ujwal Gadiraju, Georgios Katsimpras, Raoul Wessel and René Berndt</i>	
The Role of Language Evolution in Digital Archives	16
<i>Nina Tahmasebi and Thomas Risse</i>	
BlogNEER: Applying Named Entity Evolution Recognition on the Blogosphere?	28
<i>Helge Holzmann, Nina Tahmasebi and Thomas Risse</i>	

Linked Open Data and Semantic Web Content

Elevating Natural History Museums' Cultural Collections to the Linked Data Cloud	40
<i>Giannis Skevakis, Konstantinos Makris, Polyxeni Arapi and Stavros Christodoulakis</i>	
A Linked Open Data Architecture for Contemporary Historical Archives	52
<i>Alexandre Rademaker, Suemi Higuchi and Dario Augusto Borges Oliveira</i>	
Pundit: Creating, Exploring and Consuming Semantic Annotations.....	65
<i>Marco Grassi, Christian Morbidoni, Michele Nucci, Simone Fonda and Francesca Di Donato</i>	

Aspects of Information Management and Access

Towards Persistent Identification of Resources in Personal Information Management	73
<i>Stefan Haun and Andreas Nürnberger</i>	
A Storage Ontology for Hierarchical Storage Management Systems.....	81
<i>Sandro Schmidt, Torsten Wauer, Ronny Fritzsche and Klaus Meißner</i>	
Semantic Retrieval Interface for Statistical Research Data	93
<i>Daniel Bahls and Klaus Tochtermann</i>	
Semantification of Query Interfaces to Improve Access to Deep Web Content	104
<i>Arne Martin Klemenz and Klaus Tochtermann</i>	

Digital Geolinguistics: On the use of Linked Open Data for Data-Level Interoperability Between Geolinguistic Resources

Giorgio Maria Di Nunzio

Dept. of Information Engineering – University of Padua
dinunzio@dei.unipd.it

Abstract. The Open Language Archives Community which recently celebrated its first 10 years of activity, is a worldwide network dedicated to collecting information on language resources and developing standard protocols for interoperability. In this context, Linked Open Data paradigm is very promising, because it eases interoperability between different systems by allowing the definition of data-driven models and applications.

In this talk, we give an overview of present geolinguistics projects and an approach which moves the focus from the systems handling the linguistic data to the data themselves. As a concrete example, we present a geolinguistic application build upon a real linguistic dataset which provides linguists with a system for investigating variations among closely related languages.

1 Introduction

The research field of linguistics studies all aspects of human language, including morphology (the formation and composition of words), syntax (the formation and composition of phrases and sentences from these words) and phonology (sound systems) [1]. Research in the variations in languages allows linguists to understand the fundamental principles that underlie language differences, language innovation and language variation in time and space.

Geolinguistics is an interdisciplinary field that incorporates language maps depicting spatial patterns of language location or the results of processes that lead to language change [2]. In this context, the linguistic atlas has proved to be a vital tool and product of geolinguistics since the earliest stages of the field, and it has provided a stage for the incorporation of modern GIS.

In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide. The Open Language Archives Community,¹ which recently celebrated its first 10 years of activity, is a worldwide network dedicated to collecting information on language resources (field

¹ <http://www.language-archives.org>

notes, grammars, audio/video recording, descriptive papers, and so on) and developing standard protocols for interoperability. GOLD² was the first ontology to be designed specifically for linguistic description on the Semantic Web [3]. It proposes a solution to the lack of interoperability between linguistic projects and projects designed specifically for NLP applications. It can act as a kind of lingua franca for the linguistic data community, provided that data providers are willing to map their data to GOLD or to some similar resource. In [4], the authors present a framework for producing multi-layer annotated corpora: a pivot format serving as “interlingua” between annotation tools, an ontology-based approach for mapping between tag sets, and an information system that integrates the various annotations and allows for querying the data either by posing simple queries or by using the ontology.

Language resources that have been made publicly available can vary in the richness of the information they contain: on the one hand, a corpus typically contains at least a sequence of words, sounds or tags; on the other hand, a corpus may contain a large amount of information about the syntactic structure, morphology, prosody and semantic content of every sentence, plus annotations of discourse relations or dialogue acts [5]. However, the quality of such corpora may have been reduced by the intense, and often poorly controlled, usage of automatic learning algorithms [6].

The heterogeneity of linguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections [7]. The rate of re-use for linguistic database technology together with related processing tools and environments is still too low. For example, the Edisyn search engine – the aim of which was to make different dialectal databases comparable – “in practice has proven to be unfeasible”.³ In order to find common ground where linguistic material can be shared and re-used, the methodological and technological boundaries existing in each research linguistic project needs to be overcome.

2 Linked Open Data for Geolinguistic Resources

The research direction we want to discuss in this talk is to move the focus from the systems handling the linguistic data to the data themselves. For this purpose the LOD paradigm [8] is very promising, because it eases interoperability between different systems by allowing the definition of data-driven models and applications. LOD is based on the definition of real-world objects, identified by means of a dereferenceable URI⁴. Objects are related to one another by means of typed links. Interoperability is achieved by a unifying data model (i.e. RDF⁵), a standardized data access mechanism (i.e. HTTP), hyperlink-based data discovery (i.e. URI), and self-descriptive data (based on shared open vocabularies from different namespace) [8].

² <http://linguistics-ontology.org/>

³ <http://www.dialectsyntax.org/>

⁴ <http://tools.ietf.org/html/rfc3986>

⁵ <http://www.w3.org/RDF/>

In this context, a relevant initiative is ISOcat,⁶ a linguistic concept database developed by ISO Technical Committee 37, Terminology and other language and content resources, to provide reference semantics for annotation schemata. The goal of the project is to create a universally available resource for language-related metadata that can be used in a variety of applications and environments. It also provides uniform naming and semantic principles to facilitate the interoperability of language resources across applications and approaches [9].

In this talk, we discuss the steps of a possible approach for exposing geolinguistic data into LOD [10–12] by presenting:

- the ASIt⁷ linguistic project which is based on micro-variations of Italo-Romance dialects;
- a geolinguistic Web application that provides functionalities for accessing, browsing, searching the linked open data by means of linguistic features, and visualizing the data on dynamically generated maps.

References

1. Akmajian, A., Demers, R.A., Farmer, A.K., Harnish, R.M.: *Linguistics, Sixth Edition - An Introduction to Language and Communication*. The MIT Press (2010)
2. Hoch, S., Hayes, J.J.: Geolinguistics: The Incorporation of Geographic Information Systems and Science. *The Geographical Bulletin* **51**(1) (2010) 23–36
3. Farrar, S., Langendoen, T.: A Linguistic Ontology for the Semantic Web. *Glott International* **7**(3) (March 2003) 97–100
4. Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., Stede, M.: A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *TAL* **49**(2) (2008) 217–246
5. Bird, S., Klein, E., Loper, E.: *Natural Language Processing with Python*. O'Reilly Media (2009)
6. Spärck Jones, K.: Computational linguistics: What about the linguistics? *Computational Linguistics* **33**(3) (2007) 437–441
7. Chiarcos, C.: Interoperability of corpora and annotations. In Chiarcos, C., Nordhoff, S., Hellmann, S., eds.: *Linked Data in Linguistics*. Springer Berlin Heidelberg (2012) 161–179
8. Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers (2011)
9. Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P., Wright, S.: ISOcat: Remodelling Metadata for Language Resources. *IJMSO* **4**(4) (2009) 261–276
10. Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A system for exposing linguistic linked open data. In Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F., eds.: *TPDL. Volume 7489 of Lecture Notes in Computer Science.*, Springer (2012) 173–178
11. Di Buccio, E., Di Nunzio, G.M., Silvello, G.: An open source system architecture for digital geolinguistic linked open data. In Aalberg, T., Papatheodorou, C., Dobrev, M., Tsakonas, G., Farrugia, C.J., eds.: *TPDL. Volume 8092 of Lecture Notes in Computer Science.*, Springer (2013) 438–441
12. Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A curated and evolving linguistic linked dataset. *Semantic Web* **4**(3) (2013) 265–270

⁶ <http://www.isocat.org/>

⁷ <http://svrims2.dei.unipd.it:8080/asit-enterprise/>

Towards Preservation of semantically enriched Architectural Knowledge

Stefan Dietze¹, Jakob Beetz², Ujwal Gadiraju¹, Georgios Katsimpras¹,
Raoul Wessel³, René Berndt⁴

¹ L3S Research Center, Leibniz University, Hannover, Germany
{dietze; gadiraju, katsimpras}@l3s.de

² Department of the Built Environment, Eindhoven University of Technology, The Netherlands
{j.beetz}@tue.nl

³ Computer Graphics Group, University of Bonn, Germany
wesselr@cs.uni-bonn.de

⁴ Fraunhofer Austria Research GmbH, Visual Computing, Graz, Austria
rene.berndt@vc.fraunhofer.at

Abstract. Preservation of architectural knowledge faces substantial challenges, most notably due the high level of data heterogeneity. On the one hand, low-level architectural models include 3D models and point cloud data up to richer building information models (BIM), often residing in isolated data stores with insufficient support for ensuring consistency and managing change. On the other hand, the Web contains vast amounts of information of potential relevance for stakeholders in the architectural field, such as urban planners, architects or building operators. This includes in particular Linked Data, offering structured data about, for instance, energy-efficiency policies, geodata or traffic and environmental information but also valuable knowledge which can be extracted from social media, for instance, about peoples' movements in and around buildings or their perception of certain structures. In this paper we provide an overview of our early work towards building a sustainable, semantic long-term archive in the architectural domain. In particular we highlight ongoing activities on semantic enrichment of low-level architectural models towards the curation of a semantic archive of architectural knowledge.

Keywords. Architecture, Semantic Web, Linked Data, Digital Preservation, Information Extraction, Building Information Model

1 Introduction

Long-term preservation of architectural knowledge - from 3D models to related Web data - faces a wide range of challenges in a number of use cases and scenarios, which are illustrated in Figure 1. In these diverse use-cases, preservation has to satisfy needs of a range of stakeholders, including architects, building operators, urban planners and archivists.

During the lifecycle of built structures, several engineering models are produced, updated and maintained, ranging from purely geometric 3D/CAD models and point clouds to higher level, semantically rich Building Information Models (BIM). Partial

domain models at different stages are highly interrelated and interdependent and include meronomic, spatial, temporal and taxonomic relationships. Apart from these BIM-internal explicit and implicit inter-relationships, a considerable number of references are also made to external information and data sets which imposes new challenges for digital long term preservation. For example, buildings to some degree can be considered as assemblies of various concrete building products which are specified by individual product manufacturers that have to be accessed in future maintenance, modification or liability scenarios.

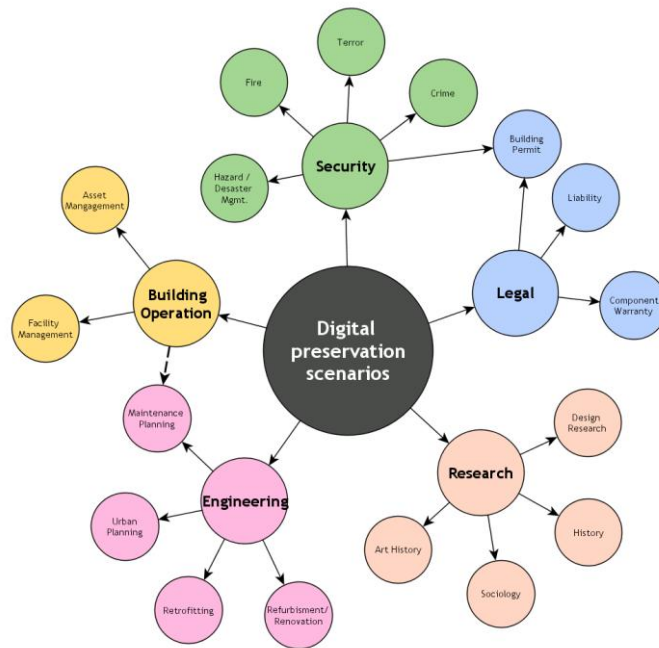


Fig. 1. Schematic overview of digital preservation scenarios & stakeholders

The individual building components and the building as a whole on the other hand have to comply with standards and local building regulations that are subject to constant evolvement and have to be preserved alongside the building model. Apart from such technical engineering information, the Web, in particular the Web of data and the social Web, contain an increasing amount of contextual information about buildings, their geo-location, history, legal context, the surrounding infrastructure or the usage and perception of structures by the general public. Examples include in particular the wide range of Linked Data [2] about geo-data¹, building-related policies² or traffic statistics³ as well as the wide range of information which can be extracted from the frequency and content of social media, such as tweets or Flickr images, for in-

¹ For instance, <http://www.geonames.org/> or <https://geodacenter.asu.edu/datalist/>

² For instance, energy efficiency guidelines at <http://www.gbpn.org/databases-tools/building-energy-rating-policies>

³ A wide range of traffic and transport-related datasets at <http://data.gov.uk>

stance about the perception and use of buildings by the general public. Such information is distributed across the Web, is evolving constantly and is available in a variety of forms, structured as well as unstructured ones. Integration and interlinking as well as preservation strategies are of crucial importance. Particularly with regards to preservation, i.e. the long-term archival of all forms of architecturally relevant knowledge, challenges arise with respect to:

- *Semantic enrichment* of low-level architectural models
- *Interlinking & archiving* of related models (across different abstraction levels and model types, across different datasets and repositories including open data and manufacturer-specific data, covering evolution at different points in time, covering parts or related contexts of particular models)
- *Preservation & temporal analysis*: capturing and supporting the evolution of models, buildings and related data
- *Maintaining consistency* across archived data over time

The Web of (Linked) Data is a relatively recent effort derived from research on the Semantic Web, whose main objective is to generate a Web exposing and interlinking data previously enclosed within silos. The Web of Data is based upon simple principles based on the use of dereferencable HTTP URIs, representation and query standards like RDF, OWL [1] and SPARQL⁴ and the extensive use of links across datasets. While Linked Data (LD) principles have emerged as de-facto standard for sharing data on the Web, our work is fundamentally aiming at (a) creating a semantic digital archive (SDA) for the architectural domain according to LD principles and (b) leveraging on the existing wealth of Web data, particularly Linked Data, to gradually enrich the archive. Given the distributed evolution of all considered knowledge and data types, dedicated archiving and preservation strategies are of crucial importance.

In this paper, we introduce our current vision and future work within the recently started project DURAARK ("Durable Architectural Knowledge")⁵, aimed at the long-term preservation of low-level architectural models gradually enriched with higher level semantics. The archived models are described as a part of a well-interlinked knowledge graph which in particular incorporates the temporal evolution of building structures and their contexts. We introduce an early draft of the overall architecture together with the semantic enrichment components. One of the requirements for preservation of structured Web data is dataset curation – i.e. profiling and classification of available datasets into their coverage (geographical, topics, knowledge types). We introduce our research activities on curation, aiming at generating catalogs (and archives) of available datasets useful to the architecture and construction sector and other interested parties.

⁴ <http://www.w3.org/TR/rdf-sparql-query/>

⁵ <http://www.duraark.eu>

2 Durable Architectural Knowledge - Approach & Overview

The novel approach of the DURAARK project in comparison to earlier efforts in the domain of digital preservation of building related information is the consideration of *open, self-documenting information standards* as well as the *enrichment and correlation of architectural models with related Web data*. This approach applies to both, the building models as well as interlinked data. While earlier efforts were focused on the preservation of proprietary, binary file formats such as Autodesk's DWG and DXF on a byte stream level [8][10], DURAARK makes distinct use of open, text-based formats from the family of ISO 10303 standards, referred to as STEP – Standard for the Exchange of Product data. In particular, the Industry Foundation Classes (IFC) [9] model along with its open specifications published and governed by the buildingSMART organization⁶, has been identified as the most suitable choice for sustainable long-term archival. This model features around 650 entity classes with approx. 2000 schema-level attributes and additional set of several hundred standardized properties that can be attached to individual entity instances and can conveniently be extended, providing a meta-modeling facility to end-users and software vendors alike.

Most, commonly IFC models are serialized as Part 21 – SPFF (STEP Physical File Formats) [14] and to a lesser extent as a content-equivalent XML representation following ISO 10303 part 28 [15]. These formats (albeit using different model schemas) have also been chosen in long-term preservation scenarios in other engineering domains [7][11], and have earlier been identified as most promising candidates for future research endeavors [8][10]. The self-documenting clear-text encoding of both instance files and schemas increase the likelihood of future reconstruction and make them less error-prone on physical levels of bit-rotting. Next to the aforementioned part 21 and 28 serializations, the DURAARK project will also provide an RDF representation of these models[12][13], which allows easier semantic enrichment and integration with other archival process chains. The architecture of the DURAARK system can be roughly divided into three layers:

1. *Processing tools* that help users to semantically and geometrically enrich and prepare architectural models for ingestion.
2. A *Semantic Digital Archive* that provides a common registry, access and preservation facility for enriched BIM and related Web data.
3. An *OAIS⁷ compliant archival system* that maintains AIPs consisting of IFC files, RDF graphs of the linked data used for semantic enrichment as well as compressed and uncompressed point cloud data sets to document as-build states of documented buildings. This will be implemented on top of the existing state-of-the-art archival products such as Rosetta.

⁶ <http://buildingsmart.org>

⁷ <http://public.ccsds.org/publications/archive/650x0m2.pdf>

One of the key features of the DURAARK approach is the gradual enrichment of low-level architectural models. Enrichment starts with *geometric enrichment*, which produces structured metadata (IFC, BIM) out of low-level architectural models and scans. Based on such structured metadata, *semantic enrichment* aims at retrieving higher-level semantic information about the described structure, for instance, about its geolocation, history or surrounding infrastructures. All data, low-level models as well as the enriched metadata will be archived in an OAIS-compliant preservation system.

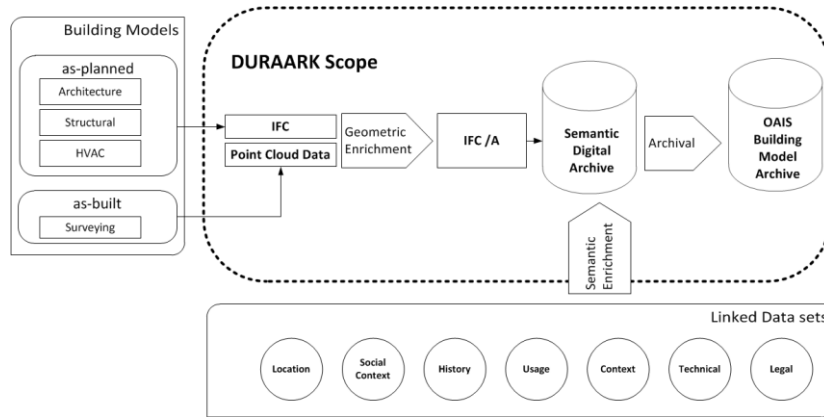


Fig. 2. DURAARK architecture overview

3 Semantic Enrichment & Preservation of Architectural Knowledge

As part of the preprocessing tools to be developed for the ingestion and preservation of building information models, an essential component facilitates the *semantic enrichment* (see [23]) and annotation as well as the extraction and compilation of relevant metadata. Semantic enrichment exploits both, *expert-curated domain models* and *heterogeneous Web data sources*, in particular Linked Data, for gradually enriching a BIM/IFC model with related information. The metadata enrichment aims to populate BIM according to the schema shown in Figure 3 and includes:

1. **During the creation** and modification of initial BIM/IFC models individual objects in the building assembly are enriched *by architects and engineers*. For example, general functional requirement specifications of a particular door set in early stages of the design (“door must be 1.01 m wide and have a fire resistance of 30 min according to the local building regulation”) are gradually refined with the product specification of an individual manufacturer that has been chosen as (“Product type A of Vendor B, catalogue number C, serial number D in configuration E3 with components X, Y, Z”). While a number of such common requirements and product parameters can be specified using entities and facets of standardized model schemas such as the IFCs, a great deal of information is currently

modeled in a formally weak and ad hoc manner. To address this, a number of structured vocabularies have been proposed in the past but have fallen short of wide adaption due to their limited exposure via standard interfaces. This includes, for instance, the buildingSMART Data Dictionary (bsDD) exposing several tens of thousands of concepts. While currently limited to custom SOAP and REST web services, the DURAARK project will expose this information as 5 star Linked Data preserved as part of the SDA.

2. ***Automated & manual interlinking and correlation with related Web data:*** as part of this step, architectural models (IFC/BIM) will be enriched with related information prevalent on the Web, for instance, about the geolocation (and its history), surrounding traffic, transport and infrastructure and the usage and perception by the general public. Building on previous work on entity linking [4], data consolidation and correlation for digital archives [6][5], dedicated algorithms for the architectural domain will be developed, for instance tailored to detect data relating to specific geospatial areas or to specifically architecturally relevant resource types. Additionally, *during the ingestion for archival* which will be carried out by *librarians and archivists* or members organizations such as municipalities, construction companies and architectural offices, other types of data sets need to be referenced.

The graph-based yet distributed nature of Linked Data has serious implications for enriching digital archives with references to external datasets. While distributed datasets (schemas, vocabularies and actual data) evolve continuously, these changes have to be reflected in the archival and preservation strategy. This joint and simultaneous consideration of *semantic enrichment and preservation* aspects is usually under-reflected in archival efforts and has to be tackled in an integrated fashion.

Generally, while within the LD graph, in theory all datasets (and RDF statements) are connected in a way, LD archiving strategies are increasingly complex and have to identify a suitable balance between correctness/completeness on the one hand and scalability on the other. These decisions are highly dependent on the domain and characteristics of each individual dataset, as each poses different requirements with regards to the preservation strategies. For instance, datasets, differ strongly with respect to the *dynamics* with which they evolve, that is, the frequency of changes to the dataset. For instance, there might be fairly static datasets where changes occur only under exceptional circumstances (for instance, *2008 Road Traffic Collisions in Northern Ireland* from data.gov.uk⁸) while on the other hand, other datasets are meant to change highly frequently (for instance, Twitter feeds or *Highways Agency Live Traffic Data*⁹). For the majority of datasets, changes occur moderately frequently (i.e. on a daily, weekly, monthly or annual basis) as is the case for datasets like BauDataWeb¹⁰

⁸ http://www.data.gov.uk/dataset/2008_injury_road_traffic_collisions_in_northern_ireland

⁹ <http://www.data.gov.uk/dataset/live-traffic-information-from-the-highways-agency-road-network>

¹⁰ <http://semantic.eurobau.com/>

or DBpedia¹¹. Depending on the specific requirements, nature and dynamics of individual datasets, we are exploring *Web data preservation strategies*, including (a) non-recurring capture of URI references to external entities as is common practice within the LD community, (b) non-recurring archival of subgraphs or the entire graph of the external dataset, (c) periodic crawling and archiving of external datasets.

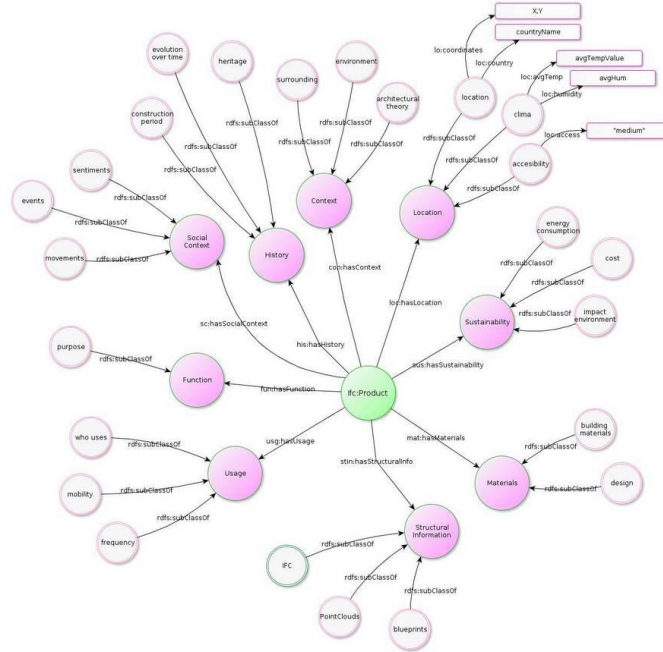


Fig. 3. Enrichment data schema

In order to facilitate informed decisions about suitable preservation strategies for individual datasets, additional structured information about the characteristics of each dataset is required, what is addressed through dedicated *data curation strategies*.

4 Curation and Preservation of Datasets and Vocabularies

In order to enable the discovery and retrieval of suitable datasets and to identify dedicated and most efficient preservation strategies for each relevant dataset, we need to provide structured metadata about available datasets, which includes in particular preservation-related information, for instance about the temporal and geographic coverage of a dataset, the estimated update frequency or the represented types and topics (for instance, whether the data contains building-related policy information or traffic or environmental data). For this purpose we are currently in the process of establish-

¹¹ <http://dbpedia.org>

ing dedicated *data curation and profiling strategies* for architecturally relevant Web data. Dataset curation and preservation follows a two-fold strategy:

- Semi-automated curation and preservation of distributed Web data
- Expert-based curation and preservation of core vocabularies

4.1 Towards semi-automated generation of a dataset observatory & archive for the architectural domain

While there exists a wealth of relevant Web datasets, particularly Linked Data, providing useful data of relevance to the architectural field (see Figure 4 for examples), metadata about available datasets is very sparse.

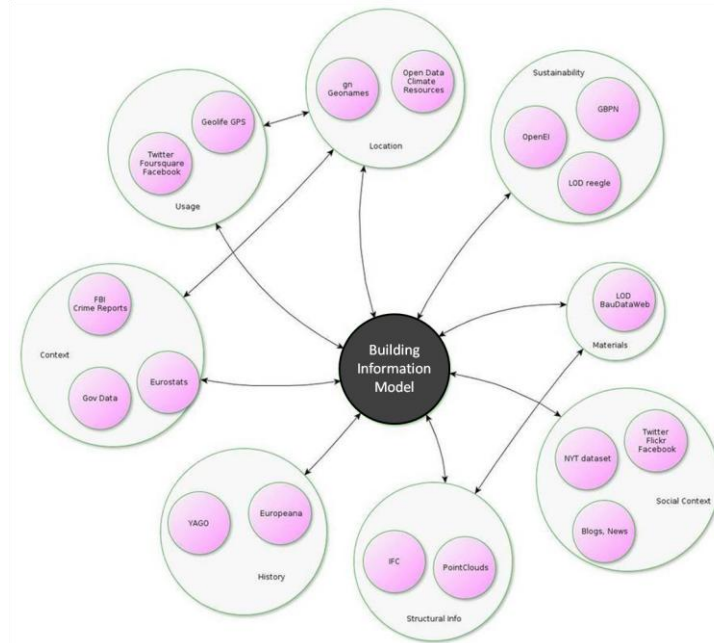


Fig. 4. Overview of building related knowledge types and datasets

Considering LD and Open Data in general, the main registry of available datasets is the DataHub¹², currently containing over 6000 open datasets and, as part of the Linked Open Data group¹³, over 337 datasets. However, while the range of data is broad, covering information about building-related policies and legislation, geodata or traffic statistics, finding and retrieving useful datasets is challenging and costly. This is due to the *lack of reliable and descriptive metadata* about content, provenance,

¹² <http://datahub.io>

¹³ <http://datahub.io/group/locloud>

availability or data types contained in distributed datasets. Thus previous knowledge of the data or costly investigations to judge the usefulness of external datasets are required. In addition, while distributed datasets evolve over time, capturing the *temporal evolution of distributed datasets* is crucial but not yet common practice. We currently conduct a number of *data curation* activities, aimed at assessing, cataloging, annotating and profiling all sorts of Web data of relevance to the architectural domain (independent of their original intention) where the overall vision entails the creation of (a) a well-described *structured catalog of datasets* and (b) an *architectural knowledge graph* which enables architects, urban planners or activists to explore all forms of suitable Web data and content captured in our SDA. This work covers several areas:

- *Data cataloging on the DataHub*: similar to the approach followed by the Linked Open Data community effort, a dedicated group ("linked-building-data"¹⁴) has been set up (though not yet populated) to collect datasets of relevance to the architectural field. While the DataHub is based on CKAN¹⁵, our group can be queried through the CKAN API, allowing further processing.
- *Automated data assessment, profiling and annotation*: while existing dataset annotations often do not facilitate a comprehensive understanding of the underlying data, we aim at creating a structured (RDF-based) catalog of architectural-related datasets, by
 - gaining new *insights and understanding* about the nature, coherence, quality, coverage and architectural relevance of existing datasets
 - *automatically obtaining annotations* and tags of existing datasets towards a more descriptive dataset catalog
 - *improving coherence and alignment* (syntactic and semantic) of existing datasets towards a unified *knowledge graph* (see [22][23])

As part of such activities, we are currently in the process of generating a *structured dataset catalog*, which adopts *VoID*¹⁶ for the description, cataloging and annotation of relevant datasets. Schema (type and property) mappings facilitate an easier exploration of data across dataset boundaries. This work builds on our efforts in [3] and follows similar aims as the work described in [24], yet we aim to not only provide metadata about the dynamics of datasets but also additional metadata about, for instance, topic, spatial or temporal coverage of the data itself. Automated data assessment exploits a range of techniques, such as Named Entity Recognition (NER) techniques together with reference graphs (such as DBpedia) as background knowledge for classifying and profiling datasets, for instance, to automatically detect the geographical and temporal coverage of a dataset or the nature of the content, for instance, whether it describes traffic statistics for the Greater London area or energy efficiency policies for Germany.

¹⁴ <http://datahub.io/group/linked-building-data> (recently founded group on the DataHub)

¹⁵ <http://ckan.org/>

¹⁶ <http://vocab.deri.ie/void>

As described in Section 3, different preservation strategies are considered for each dataset, depending on the dynamics and frequency and size of updates. While each strategy requires knowledge about the datasets to interact with, for instance, the URI of their SPARQL endpoints, our VoID-based "Linked Building Data" catalog will provide the basis for realising such individual preservation strategies and will be enriched with preservation-related metadata, for instance about the update procedures and evolution of each dataset.

4.2 Expert-based curation of domain vocabularies

In the past, a number of research efforts have aimed at providing manually curated, structured vocabularies of the various building-related engineering domains. Among them are the EU-projects eConstruct [16], IntelliGrid [18] and SWOP [17], as well as other national and international initiatives such as FUNSIEC [19]. The buildingSMART data dictionary (bsDD)¹⁷ has the ambition to be a central vocabulary repository that allows the parallel and integrated storage of different vocabularies such as the various classification systems (OMNICLASS Masterformat¹⁸, UNICLASS[20], or SfB(-NL)¹⁹) which are widely adopted in the respective countries to structure building data. The bsDD also serves as the central repository to store meta-model extensions of IFCs - referred to as PSets - which are not part of the core model schema but are recognized as typical properties of common building component. A number of commercial domain-specific building product catalogs and conceptual structures have been established that are captured in proprietary data structures that are not yet exposed as Open Data, yet have gained the status of de facto industry standards. These include the international ETIM²⁰ classification for the description of electronic equipment in buildings, the Dutch Bouwconnect²¹ platform, the German Heinze²² product database and the CROW library for infrastructural objects²³. Such structured vocabularies are often tightly integrated and oriented at local building regulation requirements and best practices and are often underlying structures for ordering higher-level data sets such as standardized texts for tendering documents (the German StLB²⁴, the Dutch STABU system²⁵, Finnish Haahtela²⁶ etc.)

Even though their use and application in the context of the Semantic Web and LD has been suggested time and again [21], the uptake of harmonized structures is still in its infancy although internationally anticipated by large end-user communities.

¹⁷ <http://www.buildingsmart.org/standards/ifd>

¹⁸ <http://www.csinet.org/Home-Page-Category/Formats/MasterFormat.aspx>

¹⁹ <http://nl-sfb.bk.tudelft.nl>

²⁰ <http://e5.working.etim-international.com>

²¹ <http://www.bouwconnect.nl>

²² <http://www.heinze.de/>

²³ <http://www.gww-ob.nl/>

²⁴ <http://www.stlb-bau-online.de/>

²⁵ <http://www.stabu.org>

²⁶ <https://www.haahtela.fi/en/>

5 Discussion and future works

In this paper we have presented an overview of the current and future work within the DURAARK project for creating a semantic digital archive for the building and architecture domain. While the project is in its early stages, currently focusing on gathering requirements and designing initial prototypes for the main components, our main contributions are the proposed architecture for digital preservation of architectural knowledge, the semantic enrichment approach and our currently ongoing work towards curation of architecturally relevant Web datasets, which builds the foundation for implementing tailored, specific and efficient strategies for preservation of continuously evolving Web datasets.

Our future work will be dedicated to fully realising our data curation approach by creating a structured dataset catalog containing meaningful metadata of architectural-related datasets. This will form the basis to implement (a) enrichment and interlinking algorithms which gradually enrich Building Information Models and (b) to fully realise preservation strategies which will enable to assess and analyse the temporal evolution of architectural models as well as correlated Web data.

Acknowledgments

This work is partly funded by the European Union under FP7 grant agreement 600908 (DURAARK).

References

- [1] Antoniou, G., van Harmelen, F. Web Ontology Language: OWL. in S. Staab, & R. Studer (eds.) Handbook on Ontologies, pp. 67-92, 2004,
- [2] 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.
- [3] D'Aquin, M., Adamou, A., Dietze, S., Assessing the Educational Linked Data Landscape, ACM Web Science 2013 (WebSci2013), Paris, France, May 2013.
- [4] Nunes, B. P., Dietze, S., Casanova, M.A., Kawase, R., Fetahu, B., Nejd, W., Combining a co-occurrence-based and a semantic measure for entity linking, ESWC 2013 – 10th Extended Semantic Web Conference, Montpellier, France, May (2013).
- [5] Risse, T., Dietze, S., Peters, W., Doka, K., Stavarakas, 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.
- [6] Dietze, S., Maynard, D., Demidova, E., Risse, T., Peters, W., Doka, K., Stavarakas, Y., Entity Extraction and Consolidation for Social Web Content Preservation, in Proceedings of 2nd International Workshop on Semantic Digital Archives (SDA), Pafos, Cyprus, September 2012.
- [7] prEN 9300-003:2005, 2005. Long Term Archiving and Retrieval of digital technical product documentation such as 3D, CAD and PDM data. PART 003: Fundamentals and Concepts.
- [8] Smith, M., 2009. Curating Architectural 3D CAD Models. International Journal of Digital Curation, 4(1), pp.98–106.

- [9] ISO 16739:2013 Industry Foundation Classes, Release 2x, Platform Specification (IFC2x Platform).
- [10] Berndt, R. et al., 2010. The PROBADO Project - Approach and Lessons Learned in Building a Digital Library System for Heterogeneous Non-textual Documents. In M. Lalmas et al., eds. *Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 376–383.
- [11] VDA, 2006. VDA 4958 Long term archiving (LTA) of digital product data which are not based on technical drawings.
- [12] Beetz, J., Van Leeuwen, J. & De Vries, B., 2009. IfcOWL: A case of transforming EXPRESS schemas into ontologies. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM*, 23(1), pp.89–101.
- [13] Pauwels, P. et al., 2011. Three-dimensional information exchange over the semantic web for the domain of architecture, engineering, and construction. *AI EDAM*, 25(Special Issue 04), pp.317–332.
- [14] ISO 10303-21:2002 Industrial automation systems and integration -- Product data representation and exchange -- Part 21: Implementation methods: Clear text encoding of the exchange structure.
- [15] ISO 10303-28:2007 Industrial automation systems and integration -- Product data representation and exchange -- Part 28: Implementation methods: XML representations of EXPRESS schemas and data, using XML schemas,
- [16] Tolman, F. et al., 2001. eConstruct: expectations, solutions and results. *Electronic Journal Of Information Technology In Construction (ITcon)*, 6, pp.175–197.
- [17] Böhm, M. et al., 2009. Semantic product modelling and configuration: challenges and opportunities. , 14, pp.507–525.
- [18] Dolenc, M. et al., 2007. The IntelliGrid platform for virtual organisations interoperability. , 12, pp.459–477.
- [19] Lima, C. et al., 2006. A framework to support interoperability among semantic resources. In *Interoperability of Enterprise Software and Applications*. Springer, pp. 87–98
- [20] Crawford, M., 1997. UNICLASS: Unified Classification for the Construction Industry, RIBA Publications.
- [21] Beetz, J. & de Vries, B., 2009. Building product catalogues on the semantic web. *Proc.CIB W78 "Managing IT for Tomorrow"*, pp.221–226.
- [22] Paes Leme, L. A. P., Lopes, G. R., Nunes, B. P., Casanova, M.A., Dietze, S., Identifying candidate datasets for data interlinking, in *Proceedings of the 13th International Conference on Web Engineering*, (2013)
- [23] Taibi, D., Fetahu, B., Dietze, S., Towards Integration of Web Data into a coherent Educational Data Graph, in Leslie Car, Alberto H. F. Laender, Bernadette F. Lóscio, Irwin King, Marcus Fontoura, Denny Vrandečić, Lora Aroyo, José Palazzo M. de Oliveira, Fernanda Lima, Erik Wilde (editors), *Companion Publication of the IW3C2 WWW 2013 Conference*, May 13–17, 2013, Rio de Janeiro, Brazil. IW3C2 2013, ISBN 978-1-4503-2038-2
- [24] Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A., Observing Linked Data Dynamics, in the *Proceedings of the 10th Extended Semantic Web Conference (ESWC2013)*, Montpellier, France, 26–30 May, 2013.

The Role of Language Evolution in Digital Archives^{*}

Nina Tahmasebi¹ ^{**} and Thomas Risse²

¹ Computer Science & Engineering Department,
Chalmers University of Technology,
412 96 Gothenburg, Sweden

² L3S Research Center,
Appelstr. 9, 30167 Hannover, Germany
`ninat@chalmers.se, risse@L3S.de`

Abstract. With advancements in technology and culture, our language changes. We invent new words, add or change meanings of existing words and change names of existing things. Left untackled, these changes in language create a gap between the language known by users and the language stored in our digital archives. In particular, they affect our possibility to firstly *find* content and secondly *interpret* that content. In this paper we discuss the limitations brought on by language evolution and existing methodology for automatically finding evolution. We discuss measures needed in the near future to ensure semantically accessible digital archives for long-term preservation.

Keywords: language evolution, finding and understanding content, digital archives

1 Introduction

With advancements in technology, culture and through high impact events, our language changes. We invent new words, add or change meanings of existing words and change names of existing things. This results in a dynamic language that keeps up with our needs and provides us the possibility to express ourselves and describe the world around us. The resulting phenomenon is called **language evolution** (or **language change** in linguistics).

For all contemporary use, language evolution is trivial as we are constantly made aware of the changes. At each point in time, we know the most current version of our language and, possibly, some older changes. However, our language does not carry a memory; words, expressions and meanings used in the past are forgotten

^{*} This work is partly funded by the European Commission under ARCOMEM (ICT 270239)

^{**} This work was done while the author was employed at L3S Research Center

over time. Thus, as users, we are limited when we want to find and interpret information about the past from content stored in digital archives.

In the past, published and preserved content were stored in repositories like national libraries and access was simplified with the help of librarians. These experts would read hundreds of books to help students, scholars or interested public to find relevant information expressed using any language, modern or old. Today, because of the easy access to digital content, we are no longer limited to physical hard copies stored in one library. Instead we can aggregate information and resources from any online repository stored at any location. The sheer volume of content prevents librarians to keep up and thus there are no experts to help us to find and interpret information. The same applies to the increasing number of national archives that are being created by libraries which crawl and preserve their national Web. Language in user generated content is more dynamic than language in traditional written media and, thus, is more likely to change over shorter periods of time [TGR12].

Much of our culture and history is documented in the form of written testimony. Today, more and more effort and resources are spent digitizing and making available historical resources that were previously available only as physical hard copies, as well as gathering modern content. However, making the resources available to the users has little value in itself; the broad public cannot fully understand or utilize the content because the language used in the resources has changed, or will change, over time. To fully utilize these efforts, this vast pool of content should be made semantically accessible and interpretable to the public. Modern words should be *translated* into their historical counterparts and words should be represented with their past meanings and senses.

In this paper we will discuss the role of language evolution in digital archives and the problems that arise as a result. We will review state-of-the-art in detecting language evolution and discuss future directions to make digital archives semantically accessible and interpretable, thus ensuring useful archives also for the future. The rest of the paper is organized as follows: In Sec. 2 we discuss different types of evolution and the corresponding problem caused. In Sec. 3 we discuss the differences between digitized, historical content and archives with new content, e.g., Web archives. In Sec. 4 we provide a review of current methods for detecting evolution and finally, in Sec. 5 we conclude and discuss future directions.

2 Evolution

There are two major problems that we face when searching for information in long-term archives; firstly *finding* content and secondly, *interpreting* that content. When things, locations and people have different names in the archives than those we are familiar with, we cannot find relevant documents by means of simple string matching techniques. The strings matching the modern name

will not correspond to the strings matching the names stored in the archive. The resulting phenomenon is called **named entity evolution** and can be illustrated with the following:

“The Germans are brought nearer to Stalingrad and the command of the lower Volga.”

The quote was published on July 18, 1942 in The Times [TT42] and refers to the Russian city that often figures in the context of World War II. In reference to World War II people speak of *the city of Stalingrad* or the *Battle of Stalingrad*, however, the city cannot be found on a modern map. In 1961, *Stalingrad* was renamed to *Volgograd* and has since been replaced on maps and in modern resources. Not knowing of this change leads to several problems; 1. knowing only about *Volgograd* means that the history of the city becomes inaccessible because documents that describe its history only contain the name *Stalingrad*. 2. knowing only about *Stalingrad* makes it difficult to find information about the current state and location of the city³.

The second problem that we face is related to interpretation of content; words and expressions reflect our culture and evolve over time. Without explicit knowledge about the changes we risk placing modern meanings on these expressions which lead to wrong interpretations. This phenomenon is called **word sense evolution** and can be illustrated with the following:

“Sestini’s benefit last night at the Opera-House was overflowing with the fashionable and gay.”

The quote was published in April 27, 1787 in The Times [The87]. When read today, the word *gay* will most likely be interpreted as *homosexual*. However, this sense of the word was not introduced until early 20th century and instead, in this context, the word should be interpreted with the sense of *happy*.

Language evolution also occurs in shorter time spans; modern examples of named entity evolution include company names (*Andersen Consulting* → *Accenture*) and Popes (*Jorge Mario Bergoglio* → *Pope Francis*). Modern examples of word sense evolution include words like *Windows* or *surfing* with new meanings in the past decades.

In addition, there are many words and concepts that appear and stay in our vocabulary for a short time period, like *smartphone face*, *cli-fi* and *catfishing*⁴ that are examples of words that have not made it into e.g., Oxford English Dictionary, and are unlikely to ever do so.

³ Similar problems arise due to spelling variations that are not covered here.

⁴ <http://www.wordspy.com/>

2.1 Formal problem definition

Formally, the problems caused by language evolution (illustrated in Figure 1) can be described with the following: Assume a digital archive where each document d_i in the archive is written at some time t_i prior to current time t_{now} . The larger the time gap is between t_i and t_{now} , the more likely it is that current language has experienced evolution compared to the language used in document d_i . For each word w and its intended sense s_w at time t_i in d_i there are two possibilities; 1. The word can still be in use at time t_{now} and 2. The word can be out of use (outdated) at time t_{now} .

Each of the above options opens up a range of possibilities that correspond to different types of language evolution that affect finding and interpreting in digital archives.

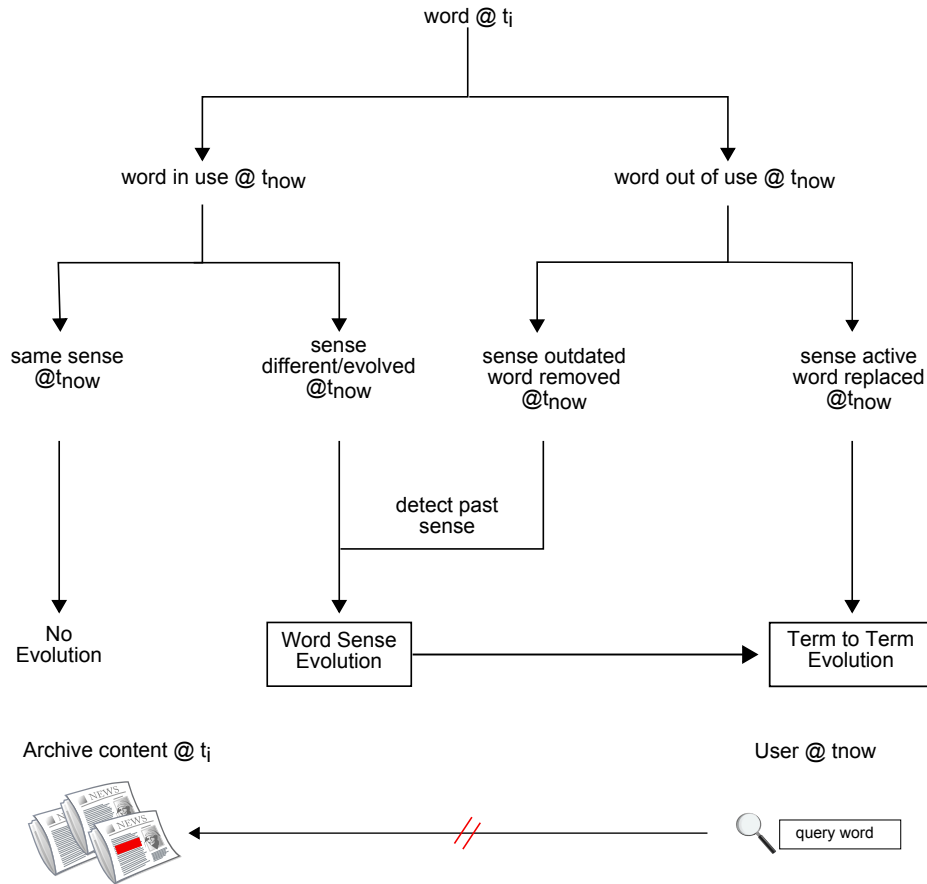


Fig. 1: Diagram of Word Evolution

Word w at time t_i in use at t_{now}

No Evolution: The word is in use at time t_{now} and has the *same sense* s_w and thus there has been no evolution for the word. The word and its sense are stable in the time interval $[t_i, t_{now}]$ and no action is necessary to understand the meaning of the word or to find content.

Word Sense Evolution: The word is still in use at time t_{now} but with a *different sense* s'_w . The meaning of the word has changed, either to a completely new sense or to a sense that can be seen as an evolution of the sense at time t_i . The change occurred at some point in the interval (t_i, t_{now}) . We consider this to be the manifestation of word sense evolution.

Word w from t_i out of use at t_{now}

Word Sense Evolution - Outdated Sense: The word is out of use because the word sense is outdated and the word is no longer needed in the language. This can follow as a consequence of, among others, technology, disease or occupations that are no longer present in our society. The word w as well as the associated word sense s_w have become outdated during the interval (t_i, t_{now}) . To be able to interpret the word in a document from time t_i it becomes necessary to detect the active sense s_w at time t_i . Because it is necessary to recover a word sense that is not available at time t_{now} we consider this to be a case of word sense evolution.

Term to Term Evolution: The word w is outdated but the sense s_w is still active. Therefore, there must be another word w' with the same sense s_w that has replaced the word w . That means, different words, in this case w and w' , are used as a representation for the sense s_w and the shift is made somewhere in the time interval (t_i, t_{now}) . We consider this to be term to term evolution where the same sense (or entity) is being represented by two different words. If the word w represents an entity, we consider it to be **named entity evolution**.

In addition to the above types of evolution, there are also *spelling variations* that can affect digital archives; historical variations with different spellings for the same word or modern variations in the form of e.g., abbreviations and symbols. Spelling variations are not considered in this paper.

3 Historical Data vs. Modern Data – Old Content vs. New Content

When working with language evolution from a computational point of view there are two main perspectives available. The first considers today as the point of reference and searches for all types of language evolution that has occurred until today. In this perspective the language that we have today is considered

as common knowledge and understanding past language and knowledge is the primary goal.

In the second perspective the goal is to prepare today's language and knowledge for interpretation in the future. We monitor the language for changes and incrementally map each change to what we know today. We can assume that knowledge banks and language resources are available and all new changes are added to the resources. In the next paragraphs we will discuss the differences between the two perspectives, and the affect on digital archives, in more detail.

3.1 Looking to the Past – The Backward Perspective

When looking to the past we assume that we have the following scenario. A user is accessing a long-term archive and wants to be able to find and interpret information from the past. There are several problems which the user must face. Firstly, there are few or no machine readable dictionaries or other resources like Wikipedia, which sufficiently cover language of the past. The user must rely on his or her own knowledge or search extensively in other resources like encyclopedias or the Web in order to find an appropriate reformulation for modern words. Once the resource is found the user must repeat the process to find the meanings of words, phrases and names in the document. Because of the low coverage of the past, the user can find only limited amount of help in this process.

In order to help users in their research of the past we need to automatically find and handle language evolution. This can be done by making use of existing algorithms and tools or by developing new ones. For both existing and new tools there are severe limitations caused by the lack of digital, high quality, long-term collections. Most existing tools have been designed and trained on modern collections and can have difficulty with problems caused by language evolution. For example, part-of-speech tagging, lemmatization and entity recognition can be affected by the age of the collection and thus limit the accuracy and coverage of language evolution detection which relies on the mentioned technologies.

There is much work being done currently to overcome this lack of resources by digitizing historical documents by means of optical character recognition (OCR). However, many older collections have been stored for a long time which leads to less than perfect quality of the resulting text. Degraded paper, wear or damage as well as old fonts cause errors in the OCR process. This leads to problems in the processing, for example to detect word boundaries or to recognize characters, as well as to verify the results. If words cannot be understood by humans then the correctness of the algorithms cannot be judged. Because of the historical nature of the language, it is also difficult to find people that are qualified to verify, improve or help detect language evolution on such collections.

3.2 Looking to the Future – The Forward Perspective

When looking to the future to find language evolution we have many advantages compared to when looking to the past. The largest advantage is that most resources are born digitally today and thus many of the problems with degraded paper quality and OCR errors are avoided. In addition, there is an abundance of available data. Most concepts, senses and entities are described and referenced over and over again which makes it easier to gather evidence for each one individually.

In addition to the higher amount and quality of the text, there are plenty of tools and resources available that can solve many smaller tasks automatically. Natural language processing tools, machine readable dictionaries, and encyclopedias form an army of resources which can be used to tackle current language. Changes in our world are captured in resources like Wikipedia and questions like *What is the new name of the city XYZ?* can be answered using machine readable resources like Yago [SKW07] or DBpedia [BLK⁺09]. To prevent information loss in the future, resources like Wikipedia, WordNet and natural language processing tools can be stored alongside the archives. This can significantly simplify finding and verifying language evolution in the future.

Table 1: Processing Comparison - Looking to the Past and Future

Aspect	Past	Future
Content	Digitized after creation, risk of decreased quality.	Increasingly born digital no need for digitization.
Resources	Limited availability	Increasing availability, WordNet, LinkedData etc.
Tools	Mostly modern tools few specialized NLP tools	Existing tools, will be continuously updated
Quality	OCR errors, outdated terms	Noise in user generated text, abbreviations, slang
Crowd sourcing	Limited possibility requires experts	Possible to make use of crowd sourcing

In the perspective of looking to the future we assume that current language is common knowledge and therefore we can employ humans to help detect language evolution. *Crowd sourcing* [How06] is collaborative work performed by large amounts of people and is the mechanism behind creating and maintaining Wikipedia. Such mechanisms could be used to monitor language and detect evolution. If models for representing and storing language evolution are provided,

crowd sourcing could be used to detect language evolution manually or to verify automatically detected language evolution. It is important to note that crowd sourcing is time sensitive and must be done together with the data harvesting to avoid that the crowd forgets.

There are however several limitations. The first limitation is noisy data being published on the Web. With increasing amounts of user generated text and lack of editorial control, there are increasing problems with grammars, misspellings, abbreviations, etc. To which level this can be considered as real noise like with OCR errors is debatable, however, it is clear that this noise reduces the efficiency of tools and algorithms available today. This in turn limits the quality of evolution detection as we depend on existing tools and their efficiency. The second limitation is the restricted nature of resources like Wikipedia. As with dictionaries, Wikipedia does not cover all entities, events and words that exist. Instead, much is left out or only mentioned briefly which limits to which extent we can depend exclusively on these resources.

In order to avoid that future generations face the same problems as we have to face, we need to start thinking about these problems already now. In particular for Web archives that are continuously created and updated, with ephemeral words, expressions and concepts. Otherwise we risk to render a large portion of our archives semantically inaccessible and cannot utilize the great power of crowd sourcing.

4 State-of-the-art

Word Sense Evolution Automatic detection of changes and variations in word senses over time is a topic that is increasingly gaining interest. During the past years researchers have evaluated and researched different parts of the problem mainly in the field of computational linguistics.

[SKC09] presented work on finding *narrowing* and *broadening* of senses over time by applying semantic density analysis. Their work provides indication of semantic change, unfortunately without clues to what has changed but can be used as an initial warning system.

The work presented by [LCM⁺12] aims to detect word senses that are novel in a later corpus compared to an earlier one and use LDA topics to represent word senses. Overall, the method shows promising results for detecting novel (or outdated) word senses by means of topic modeling. However, alignment of word senses over time or relations between senses is not covered in this work.

[WY11] report on automatic tracking of word senses over time by clustering topics. Change in meaning of a term is assumed to correspond to a change in cluster for the corresponding topic. A few different words are analyzed and there is indication that the method works and can find periods when words change

their primary meaning. In general, the work in this paper is preliminary but with promising indications.

Our previous work presented in [Tah13] was the first to automatically track individual word senses over time to determine changes in the meanings of terms. We found *narrowing* and *broadening* as well as slow shifts in meaning in individual senses and relations between senses over time like *splitting*, *merging*, *polysemy* and *homonymy*. For most of the evaluated terms, the automatically extracted results corresponded well to the expected evolution with regards to the main evolution. However, word senses were not assigned to individual word instances, which is necessary to help users understand individual documents.

In general, word sense disambiguation methods are not sufficient to solve the problem of word sense evolution because discrimination methods 1. often rely on an existing set of word sense; and 2. do not map word senses to each other over time.

Named Entity Evolution Previous work on automatic detection of named entity evolution has been very limited. The interest has largely been from an information retrieval (IR) point of view as named entity evolution makes finding relevant documents more challenging. Unfortunately, no effort has been put towards scalable methods and presentation of evolution to users.

Query reformulation is proposed in [BBSW09] where the degree of relatedness between two terms is measured by comparing co-occurring terms from different time periods. The approach requires recurrent computation for each query as it depends on a target time specified by the user and is not well suited for large datasets.

Semantically identical concepts (nouns) used at different time periods are discovered using association rule mining in [KVB⁺10]. Entities are associated to events (verbs) and linked across time via the event. The method could be used for shorter time spans but is less suited for longer time spans as verbs are more likely to change over time than nouns [Sag10].

Time-based synonyms (i.e., named entity evolution) are found in [KN10] by utilizing link anchor texts in Wikipedia articles. Unfortunately, link information, such as anchor text, is rarely available in historical archives but might be well suited for Web data.

In our previous work, [Tah13, TGK⁺12], we proposed NEER, an unsupervised method for named entity evolution recognition independent of external knowledge sources. Using burst detection we find *change periods*, i.e., periods with high likelihood of name change, and search exclusively in these periods for changes. We avoid comparing terms from arbitrary time periods and thus overcome a severe limitation of existing methods; the need to compare co-occurring terms or associated events from different time periods. The method needs to be targeted to Web data and streams of data to avoid re-computation.

In addition to detecting evolution, it is necessary to store evolution and to utilize it for finding and interpreting at query time. Though there is some work done in indexing and retrieval, e.g., [ABBS12, BMRV11], few target the particularities of language evolution.

5 Conclusions and Outlook

Language evolves over time. This leads to a gap between language known to the user and language stored in digital archives. To ensure that content can be found and semantically interpreted in our digital archive, we must consider **semantic preservation** and prepare our archives for future processing and long-term storage. Automatic detection of language evolution is a first step towards offering semantic access, however, several other measures need to be taken. Dictionaries, natural language processing tools and other resources must be stored alongside each archive to help processing in the future. Data structures and indexes that respect temporal evolution are needed to utilize language evolution for searching, browsing and understanding of content. To take full advantage of continuously updated archives that do not require expensive, full re-computation with each update, we must invest effort into transforming our digital archives into **living archives** that continuously learn changes in language.

There are methods for automatically finding language evolution, however, these are initial and have little focus on scalability. Effort needs to be invested into finding large scale methods that provide high quality evolution detection. In addition, the possibility to make use of **crowd sourcing** to improve detection of language evolution should be investigated. Studies are needed to establish where and in which format human input is most beneficial, in particular, when the input is in the form of the crowd without explicit domain expertise. If crowd sourcing solutions are to be employed, the processing must take place at the time of archiving to avoid the crowd forgetting up-to-date changes in the language.

To make the most out of our digital archives, language evolution must be given a **cultural dimension**. For example, the term *travel* has had the same overall meaning over time; *transporting from location A to location B*. However, this does not tell the full story of the word or the concept represented by the word. Today travel is mostly for business or as a happy occasion for holidays, without any substantial risks involved. In the past, traveling contained great dangers and was done at the risk of life. This inherent meaning of a word should be communicated to the user to allow for a full interpretation of language and to entail all dimensions of our language and culture. One possible solution is the **use of images** that can better capture and more easily convey culture.

In addition to viewing language as variant over time, language can be considered variant over demographics. When archiving the Web we have the possibilities to gather knowledge of many subcultures and parts of the world. By continuously

detecting language evolution, we can better determine what content to harvest and store for the future to ensure diverse archives.

Bibliography

- [ABBS12] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. Index maintenance for time-travel text search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 235–244, New York, NY, USA, 2012. ACM.
- [BBSW09] Klaus Berberich, Srikanta J. Bedathur, Mauro Sozio, and Gerhard Weikum. Bridging the Terminology Gap in Web Archive Search. In *12th Int. Workshop on the Web and Databases (WebDB'09)*, 2009.
- [BLK⁺09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A crystallization point for the Web of Data. *Journal of Web Semantics*, 7(3):154–165, September 2009.
- [BMRV11] Siarhei Bykau, John Mylopoulos, Flavio Rizzolo, and Yannis Velegrakis. Supporting queries spanning across phases of evolving artifacts using steiner forests. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1649–1658, New York, NY, USA, 2011. ACM.
- [How06] Jeff Howe. The Rise of Crowdsourcing. *Wired Magazine*, 14(6), 06 2006.
- [KN10] Nattiya Kanhabua and Kjetil Nørnvåg. Exploiting time-based synonyms in searching document archives. In *Joint Conference on Digital Libraries (JCDL'10)*, pages 79–88, Australia, 2010.
- [KVB⁺10] Amal Chaminda Kaluarachchi, Aparna S. Varde, Srikanta J. Bedathur, Gerhard Weikum, Jing Peng, and Anna Feldman. Incorporating terminology evolution for query translation in text retrieval with association rules. In *Proceedings of ACM Conf. on Information and Knowledge Management, (CIKM'10), Canada, October 26-30*, pages 1789–1792, 2010.
- [LCM⁺12] Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. Word Sense Induction for Novel Sense Detection. In Walter Daelemans, Mirella Lapata, and Lluís Màrquez, editors, *EACL*, pages 591–601. The Association for Computer Linguistics, 2012.
- [Sag10] Eyal Sagi. Nouns are more stable than Verbs: Patterns of semantic change in 19th century English. *The 32nd Annual Conference of the Cognitive Science Society*, 2010.

- [SKC09] Eyal Sagi, Stefan Kaufmann, and Brady Clark. Semantic density analysis: comparing word meaning across time and phonetic space. In *Proc. of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 104–111. ACL, 2009.
- [SKW07] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.
- [Tah13] Nina Tahmasebi. *Models and Algorithms for Automatic Detection of Language Evolution. Towards Finding and Interpreting of Content in Long-Term Archives*. PhD thesis, Leibniz Universität Hannover, To be published 2013.
- [TGK⁺12] Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *Proceedings of COLING 2012*, pages 2553–2568, Mumbai, India, December 2012.
- [TGR12] Nina Tahmasebi, Gerhard Gossen, and Thomas Risse. Which Words Do You Remember? Temporal Properties of Language Use in Digital Archives. In *TPDL*, volume 7489, pages 32–37, 2012.
- [The87] The Times. Sestini’s benefit last night at the Opera-House was overflowing with the fashionable and gay. In *London, England, Apr 27, 1787; pg. 3; Issue 736*. Gale Doc. No.: CS50726043, 1787.
- [TT42] DIPLOMATIC CORRESPONDENT The Times. Menace To The Volga. In *London, England, Jul 17, 1942; pg. 3; Issue 49290*. Gale Doc. No.: CS52116209, 1942.
- [WY11] Derry Tanti Wijaya and Reyyan Yeniterzi. Understanding semantic change of words over centuries. In *Proceedings of the Int. workshop on DETecting and Exploiting Cultural diversiTy on the social web*, DETECT '11, pages 35–40. ACM, 2011.

BlogNEER: Applying Named Entity Evolution Recognition on the Blogosphere^{*}

Helge Holzmann¹, Nina Tahmasebi^{2**}, and Thomas Risse¹

¹ L3S Research Center,
Appelstr. 9, 30167 Hannover, Germany
{holzmann,risse}@L3S.de

² Computer Science & Engineering Department,
Chalmers University of Technology,
412 96 Gothenburg, Sweden
ninat@chalmers.se

Abstract. The introduction of Social Media allowed more people to publish texts by removing barriers that are technical but also social such as the editorial controls that exist in traditional media. The resulting language tends to be more like spoken language because people adapt their use to the medium. Since spoken language is more dynamic, more new and short lived terms are introduced also in written format on the Web. In [1] we presented an unsupervised method for Named Entity Evolution Recognition (NEER) to find name changes in newspaper collections. In this paper we present BlogNEER, an extension to apply NEER on blog data. The language used in blogs is often closer to spoken language than to language used in traditional media. BlogNEER introduces a novel semantic filtering method that makes use of Semantic Web resources (i.e., DBpedia) to gain more information about terms. We present the approach of BlogNEER and initial results that show the potentials of the approach.

Keywords: Named Entity Evolution, Blogs, Semantic Web, DBpedia

1 Introduction

The introduction of new technology changes the way we express ourselves [2]. In Social Media, like blogs, everyone can publish content, discuss, comment, rate, and re-use content from anywhere with minimal effort. The constant availability of computers and mobile devices allows communicating with little effort, few restrictions, and increasing frequency. As there are no requirements for formal or correct language, authors can change their language use dynamically. Under these circumstances we expect people to adapt their language to the means of communication by using more creative language and unconventional spellings.

^{*} This work is partly funded by the European Commission under ARCOMEM (ICT 270239)

^{**} This work was done while the author was employed at L3S Research Center

Also words which might otherwise have been reserved for use only in conversations between friends can be introduced in written text.

These changes lead to a more dynamic language where new and short lived terms are introduced also in written format. Local as well as global language trends can spread via forums on the Web to a larger audience. This shortened gap between written “Web Language” and spoken language coupled with the inherent dynamics of spoken language leads to the introduction of new terms and high dynamics also in written language.

With the increasing efforts in documenting and preserving the public view on certain events and topics like the Financial Crisis or the Olympic Games, there is also an increasing need to make use of this content. To turn user generated content into valuable information requires a better “understanding” of the content. A systems that is aware of this knowledge can support information retrieval by augmenting the query term. Awareness of language evolution is in particular important for searching tasks in archives due to the different ages of the involved texts.

Language evolution is a broad area and covers many sub-classes like word sense evolution, term to term evolution, named entity evolution and spelling variations. In [1] we presented our approach for Named Entity Evolution Recognition (NEER). NEER is an unsupervised method to find name changes without using external knowledge sources. As an example consider *Pope Benedict XVI*, formerly known as *Joseph Ratzinger*. NEER can detect those changes in a high quality newspaper dataset that reports this evolution by analyzing co-occurring terms.

In this paper we present a first extension of NEER towards “Web Language” by adapting and applying the method to blog content. The language used in blogs is often closer to spoken language than to language used in traditional media [3]. BlogNEER, an extension of NEER that introduces a novel semantic filtering method, makes use of semantic resources (here exemplarily DBpedia) to gain more information about terms.

In the following section we present the related work in the field of named entity evolution. In Section 3 we give an introduction to NEER and motivate BlogNEER. Section 4 explains our novel filtering method utilizing external resources from the Semantic Web. In Section 5 we describe our experiments and show an example. Section 6 concludes the work and gives an outlook on future work.

2 Related Work

Previous work on automatic detection of language evolution has mainly focused on named entity evolution. The interest has mainly been from an information retrieval point of view as search results can be affected by named entity evolution.

Berberich et al. [4] proposed a solution to this problem by reformulating a query into terms prevalent in the past. They measure the degree of relatedness between two terms when used at different times by comparing the contexts as

captured by co-occurrence statistics. This approach requires a recurrent computation each time a query is submitted as it requires a target time for the query reformulations which reduces efficiency and scalability. The results presented in this paper are “anecdotal” (to use the words of the authors) and thus do not provide a basis for comparison. However, because of the promising results we use the same method for defining a context.

Kaluarachchi et al. [5] propose to discover semantically identical concepts (or named entities) used at different times. They discover these changing entities using association rule mining by associating distinct entities to events. Sentences containing a subject, a verb, objects, and nouns are targeted and the verb is interpreted as an event. Two entities are considered semantically related if their associated event is the same and the event occurs multiple times in a document archive. The temporally related term of a given named entity is used for query translation (or reformulation) and results are retrieved appropriately w.r.t. specified time criteria. They present precision and recall for three queries and evaluate only indirectly on the basis of retrieved documents.

Kanhabua et al. [6] define a time-based synonym as a term semantically related to a named entity at a particular time period. They extract synonyms of named entities from link anchor texts in Wikipedia articles using the full history. The paper evaluates the precision and recall of the time-based synonyms by measuring increased precision and recall in search results rather than directly evaluating the quality of the found synonyms.

In more recent work, Mazeika et al. [7] consider semantically similar entities from different time periods. They extract named entities from the YAGO ontology and provide a visual analytics tool to analyze the evolution of named entities of the New York Times Annotated Corpus. No name changes are tracked but the tool offers a visualization of the evolution of an entity in the relation to other entities.

3 Named Entity Evolution Recognition

The NEER approach addresses the problem of automatically detecting named entity evolution. It works unsupervised and without incorporating external resources. This section gives an overview of NEER and its limitations on blog data.

3.1 Definitions

We consider a **term** w_i to be a single or multi-word lexical representation of an entity at time t_i . The **context** C_{w_i} is the set of all terms related to w_i at time t_i . Similar to Berberich et al. [4] we consider the most frequently co-occurring terms within a distance of k words as the context, however, other contexts can be used. We consider a **change period** to be a period of time in which one term evolves into another. We consider **temporal co-references** to be different lexical representations that have been used to reference the same concept or entity at the

different periods in time. **Direct temporal co-references** are temporal co-references that are variations of each other with some lexical overlap. **Indirect temporal co-references** are temporal co-references that lack lexical overlap on the token level. A **temporal co-reference class** contains all direct temporal co-references for a given named entity, denoted as $coref_r \{w_1, w_2, \dots\}$. Each temporal co-reference class is represented by a class representative r which is also a member of the class. For example, Joseph Ratzinger is the representative of the co-reference class containing the terms $\{Joseph\ Ratzinger, Cardinal\ Ratzinger, Cardinal\ Joseph\ Ratzinger, \dots\}$.

3.2 Overview of NEER

The major steps of the Named Entity Evolution Recognition (NEER) approach are depicted in Figure 1. NEER utilizes change period for finding named entity evolution. These periods are identified by detecting high frequency bursts of an entity. Those are considered to indicate a change period. Texts from the year around a burst are regarded for collecting the co-reference candidates by extracting the relevant terms. These are used to build up contexts represented as graphs. Based on the contexts four rules are being applied to find direct co-references among the extracted terms. These are merged to co-reference classes as follows:

1. *Prefix/suffix rule*: Terms with the same prefix/suffix are merged (e.g., Pope Benedict and Benedict).
2. *Sub-term rule*: Terms with all words of one term are contained in the other term are merged (e.g., Cardinal Joseph Ratzinger and Cardinal Ratzinger).
3. *Prolong rule*: Terms having an overlap are merged into a longer term (e.g., Pope John Paul and John Paul II are merged to Pope John Paul II).
4. *Soft sub-term rule*: Terms with similar frequency are merged as in rule 2, but regardless of the order of the words.

Ultimately, the graphs are being consolidated by means of the co-references classes. Afterwards filtering methods filter out false co-references that do not refer to the query term. For this purpose, statistical as well as machine learning (ML) based filters were introduced. A comparison of the methods revealed their strengths and weaknesses in increasing precision while keeping a high recall. The ML approach performed best with noticeable precision and recall of more than 90%. While it is possible to deliver a high accuracy with NEER + ML, training the needed ML classifier requires manual labelling.

3.3 Limitations of NEER Applied on Blog Data

Tahmasebi et al. [3] showed that language in blog texts behaves differently than traditional written language. Blog language is much more dynamic and closer to spoken language than written language in traditional media. Therefore, we treat blog texts differently than texts from newspapers.

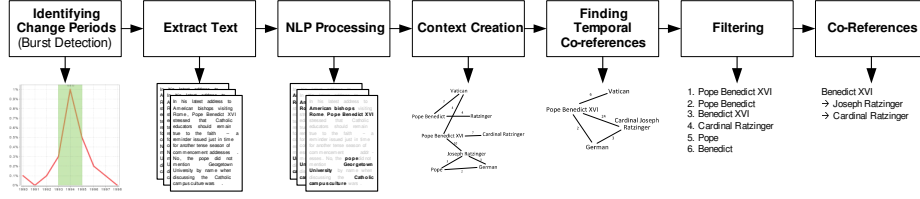


Fig. 1. Pipeline used to detect temporal co-references[1].

The machine learning filter, which delivered best results in NEER experiments, achieved a precision of more than 90% by filtering out false detected co-references. Applying this to blog data leads to a much wider contexts, containing many unrelated terms due to the large amount of relatively low quality texts. Therefore the NEER filtering methods would have a much lower effect.

NEER makes no use of external resources like DBpedia since the main development goal was to apply on historical document collections. Incorporating the Semantic Web allows us to filter out false detected names using semantic information, which is reasonable when working with data from the Web, like blogs.

4 Semantic Filtering Approach

Semantic Filtering is a novel a-posteriori filtering method for NEER incorporating the Semantic Web. With this approach we exemplarily use external data from DBpedia to augment a term with semantic information. Employing these, we are able to filter out names that do not refer to the same entity. Two terms referring to entities of different types or categories can not be evolutions of each other. A-posteriori means we apply this filter after applying NEER to our dataset given a query term and one or more change periods. At this step we have access to the NEER results which consist of a collection of indirect co-references and a co-reference class for the query term, composing the direct co-references. Using this filter, all co-references that could be identified as names for other entities will be filtered out.

The semantic filter incorporates semantic information from DBpedia which are structured as resources. A **resource** on DBpedia is the structured representation of a Wikipedia page, which is automatically extracted as described by Bizer et al. [8]. While an ambiguous name can refer to multiple resources, every resource has its own unique name and every name only points to one resource directly. This is realized by using **disambiguation resources**. E.g., *Apple_(disambiguation)* is the disambiguation resource of the resource *Apple* (the fruit) and *Apple_Inc.* Unlike this example, disambiguation resources do not always have the "disambiguation" suffix. However, every resource has **properties**, which either point to a textual or numeric value, or to another resource.

Disambiguation resources can be identified by the existence of **disambiguation properties** that point to their corresponding unambiguous resources.

Other properties which are important for our work are the **types** of resources as well as **subjects**, which can be conceived as categories. In addition to the property relations (resource \rightarrow property \rightarrow value), DBpedia also provides the inverse relations (value \rightarrow is-property-of \rightarrow resource). These can help to detect ambiguous resources where the corresponding disambiguation resource points to the ambiguous one (e.g., *Apple (disambiguation)* disambiguates *Apple*).

By mapping a query term as well as all of its co-references (direct and indirect) to DBpedia resources we can augment the terms with semantic properties. These properties can help to filter out false positive results derived by NEER as new names for the entity. It is important to mention that we only make use of descriptive properties and will not utilize already known name evolution information and co-references from DBpedia. In this paper we focus on a term's types and subjects, but also make use of redirects and disambiguations. Although, in some cases redirects represent a name change as well by redirecting an old name to its new name, we do not use this information explicitly. Hence, we treat all terms separately, even if they redirect to the same resource, like there is no redirection available (e.g., for *Czechoslovakia* and *Czech Republic* or *Slovakia*).

4.1 Disambiguation and Aggregation of Properties

To map a term to a DBpedia resource, we replace spaces with underscores and append it to the DBpedia resource URI (e.g., for "Project Natal" the resource URI becomes http://dbpedia.org/resource/Project_Natal). In case we are able to resolve a term to a resource we fetch all property relations as well as the inverse relations and save them in a lookup table. In this table, every property gets indexed twice, by the complete property URI (e.g., <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>, short *rdf:type*) and by the name extracted from the URI (e.g., *type*). In the lookup table, every property for a term points to a list of values, either URIs or strings for textual/numeric values. By indexing the property names in addition to the unique identifiers we are able to retrieve a list of all types independently from the used ontology. This is important since some resource have assigned same properties from different ontologies (e.g., <http://dbpedia.org/property/type> in addition to *rdf:type* from the example above). By indexing these using their name (i.e., *type*), we unify them to the same property.

After mapping the found terms to their corresponding resources, we follow four strategies to extend and disambiguate their semantic meanings. The first strategy is to follow DBpedia redirections if present. The second strategy is to explore disambiguation resources for ambiguous terms that do not redirect to a disambiguation resource. The remaining two strategies disambiguate ambiguous terms.

Redirection Strategy Redirections are realized on DBpedia by a redirection property (i.e., <http://DBpedia.org/ontology/wikiPageRedirects>, short *dbpedia-*

owl:wikiPageRedirects). This is assigned to the resource that is supposed to redirect to another. We leverage this by fetching the resource the property points to (s. Figure 2). Redirects are followed recursively. During this procedure we fetch and index all new found properties and aggregate them. The rationale behind this is that, in case there is a redirection pointing to another resource, this is supposed to give a better entity description. Therefore, it represents the same entity and its properties belong to the entity as well.

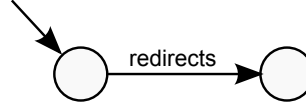


Fig. 2. Follow redirections.

Ambiguation Strategy If a resource has an ambiguous meaning, it mostly points to a disambiguation resource using the *dbpedia-owl:wikiPageRedirects* property. In this case, we apply the first redirection strategy. However, there are ambiguous resources that do not redirect. For instance, the resource *Apple* (i.e., <http://dbpedia.org/resource/Apple>) represents the fruit, even though *Apple* is an ambiguous term. The disambiguation resource for *Apple* is *Apple_(disambiguation)*, but there is no redirection between these two. Therefore, *Apple_(disambiguation)* uses the *dbpedia-owl:wikiPageDisambiguates* property to point to its non-ambiguous resources, like *Apple* (the fruit).

To discover ambiguous terms, we analyze all inverse disambiguation relations of a resources and follow backwards if there is a relation originating in a resource with the exact same name as the original term, but with the suffix "(disambiguation)" appended (s. Figure 3). Unlike for the redirection, we do not collect all properties. Instead, we only keep the properties of the disambiguation resource, because the original term might not be the one we are interested in (e.g., Apple fruit).

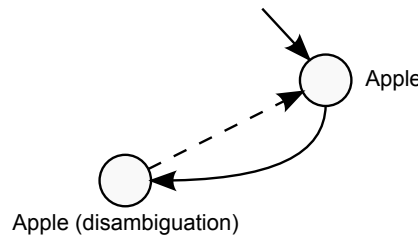


Fig. 3. Redirect to disambiguation resource, in case it exist with the same name.

Direct Disambiguation Strategy If a disambiguation resource has been identified we need to decide for one of the suggested resources as a representation

for the entity name under consideration. In case one of the candidates proposed by DBpedia is also a direct co-reference of the term we take this one as shown in the example in Figure 4. The term we try to resolve in the example is *Pope Benedict*. The corresponding disambiguation resource proposes all popes with name Benedict up to XVI. Since *Pope Benedict XVI* is a direct co-reference in the co-reference class of *Pope Benedict* derived by NEER we follow this resource as described for our redirection strategy and aggregate its properties with the properties that have been fetched so far.

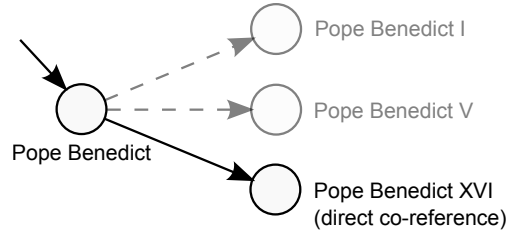


Fig. 4. Disambiguate entity by following resource with the same name as a direct co-reference.

Indirect Disambiguation Strategy For the disambiguation of terms for which we do not have a direct co-reference as disambiguation candidate, we make use of indirect co-references derived by NEER for that term. Using these indirect co-references ind_1, ind_2, \dots we form a term vector. Additionally, a term vector is formed for each disambiguation candidate based on the property values of the corresponding resource. These vectors consist of the frequencies of every indirect co-reference occurring in the property values: $(freq(ind_1), freq(ind_2), \dots)$. Similar to [9] we calculate the cosine similarity between two vectors to measure which resource fits the term in our context best. That resource will be selected as the semantic representation for the ambiguous term. This procedure is illustrated in Figure 5.

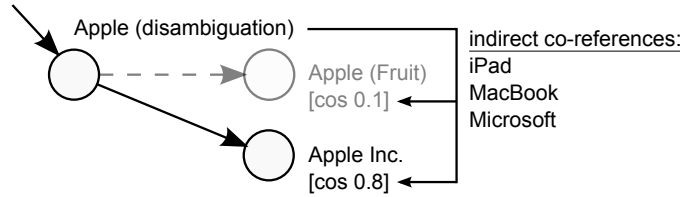


Fig. 5. Disambiguate entity by following resource that is most similar to the indirect co-references.

4.2 Filtering

After the disambiguation and aggregation of properties from DBpedia we proceed with the filtering. We consider the properties *type* and *subject* despite their ontology or namespace (i.e., URI), as described in Section 4.1. We treat DBpedia under the open world assumption. That means the fact a resource does not have a certain property does not mean that the corresponding entity does not have the property either. The resource has perhaps just not been annotated with the property. However, if a resource has a certain property, we consider this to be complete. For instance, if a resource is annotated with types, we assume these are all types it has and there is no type missing.

Similarity Filtering The first filter we apply to the result set of co-references derived by NEER compares the similarity of the query term with its co-reference candidates based on the their types and subjects from DBpedia. We compare the set of types and subjects of the query term with sets of each co-reference, direct and indirect. This only works if the query term or its corresponding DBpedia resource respectively has been annotated with types or subjects at all. Otherwise, this filtering method is not applicable. The same holds for the co-references. It would be wrong to consider two term referring to different entities just because one of them has not been annotated with types or subjects while the other one has (open world assumption, s. above). In this case we treat them as correct co-references for the query term and keep them in our result set. In case the query term's resource and the resource of the co-reference under consideration have both been annotated with types or subjects we require them to have at least one type and/or subject in common. To check this requirement, we compute the intersections of their type sets as well as their subject sets. In case one of the set intersections is empty, we consider the two terms as different and filter out those co-references. Otherwise, we keep them in our result set and pass them to the type filter.

Type Filtering Other than the similarity filter, the type filter considers hierarchies of types in addition to the types a resource is directly annotated with. For instance, both *Pope Benedict XVI* and *Barack Obama* are *persons* (resources of type *dbpedia-owl:Person*). Therefore, the similarity filter would not have filtered out one of them as co-reference of the other. However, *Pope Benedict XVI* is of type *dbpedia-owl:Cleric* while *Barack Obama* is annotated with *dbpedia-owl:OfficeHolder*. Both types are sub-types of *Person*. Thus, the two terms refer to different kinds of persons on DBpedia and do most likely not correspond to the same entity.

To achieve this filtering we need to analyze the sub-class relations of all types assigned to a resource. Each type on DBpedia is represented as an URI that points to a resource of that type. To obtain the hierarchy of a type, we leverage the *rdfs:subClassOf* property (i.e., <http://www.w3.org/2000/01/rdf-schema#subClassOf>) of the resource. This points to its super-type and allows us to

perform this procedure recursively until there is no `rdfs:subClassOf` property available or no resource corresponding to a type's URI exists.

After we have fetched the hierarchies for all types top-down, starting by a type and fetching the super-types, we analyze them bottom-up. For all types that the query term and its potential co-reference have in common we compare all of their sub-types. For instance, for *Pope Benedict XVI* and *Barack Obama*, having type *Person* in common, we compare their sub-types of type *Person: Cleric* and *OfficeHolder*. As these are different we consider the two terms not to be the same or referring to the same entity respectively and do not keep the co-reference candidate in our result set. In case they are equivalent we proceed with the next sub-type. This will be done recursively as long as both terms have sub-types in common or until they are not annotated with further sub-types.

The open world assumption holds again if the terms under consideration have a type in common, only one of them has been annotated with a further sub-type though. As we cannot tell whether the sub-type is missing on the other DBpedia resource or the entity is actually not an instance of that type, we do not filter out that co-reference and keep it in the final result set.

5 Experiments

For our experiments we created a Ruby implementation of NEER and added the introduced extensions for BlogNEER. For the entity extraction we used a Ruby implementation of the Lingua English Tagger by Coburn [10].

For the evaluation we created two datasets. The techblog dataset consists of five popular tech blogs covering five years from 2008 to 2013, fetched from Google Reader: TechCrunch, Gizmodo, SlashGear, Ubergizmo and GottaBeMobile. For the general blog dataset we fetched the top 100 blogs from nine different categories (sports, autos, science, business, politics, entertainment, technology, living, green), based on the ranking of Technorati [11], also from Google Reader. In addition, we used the Blogs08 TREC dataset, described by Ounis et al. [12].

Prior to creating contexts with NEER we applied a frequency filtering to avoid feeding NEER with too many noisy terms. Those terms often do not have a corresponding DBpedia resource and thus they cannot be filtered out by using the semantic filter with similarity or type filtering and remain as noise in the end result. Applying the frequency filter lead to much better results by keeping the contexts smaller.

To demonstrate the results of BlogNEER we use the term “Kinect” as an example. “Kinect” is the name of a gaming accessory from Microsoft. During its development it was known under the name “Project Natal” until the announcement of Kinect in June 2010. We used that month as the change period and applied the frequency as well as the semantic filter to our results. The following set of terms is a result containing both, direct and indirect co-reference without semantic filtering:

Apple, Engadget, GameStop, Project Natal, Kotaku, Nintendo, Redmond, USA Today, Microsoft Kinect, Microsoft

After applying the semantic filter (s. Section 4.2) we get an improved result set:

Project Natal, Microsoft Kinect

Due to the preliminary stage of our research, we are unable to compare precision and recall. However, in recent experiments we already reached a recall similar to the recall we achieved with our baseline, NEER on the New York Times dataset [1]. Even though the precision was still lower due to noise, the semantic filter helped with filtering out false positives as shown in the example above. In case the noise consists of misspelled, informal or rarely used terms, which are not known in DBpedia, we are not able to filter them out using semantic filtering. In future work we will tackle this problem by using advanced frequency filtering methods.

Our results also indicated how differently NEER behaves on blog data. Although both datasets consist of blogs, we observed much less noise with the general blog dataset specialized in certain categories than in arbitrary, unspecialized and partly private blogs from the TREC blogs. Our experiments, even if not yet final, already indicate the impact of frequency and semantic filtering. We were already able to reduce the noise and achieve a constantly high recall.

6 Conclusions and Future Work

For applying the NEER method on the Blogosphere we proposed BlogNEER, an extension to the original approach. BlogNEER uses a novel a-posteriori filtering method incorporating the Semantic Web. The semantic filter applied to the results of NEER increased the precision by making use of data from DBpedia. Using properties like types and subjects (i.e., categories) we are able to keep apart terms that refer to different entities. Therefore, we can filter out names that refer to another entity than a query term and thus, can not be an new name.

We presented a first evaluation and a simple example showed the potential of BlogNEER. However, to further reduce the noise we will need to filter terms a-priori before they are processed by BlogNEER. We are also planning on incorporating additional web resources in BlogNEER as well as making use of other web specific feature, for instance tags.

References

- [1] Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. Neer: An unsupervised method for named entity evo-

12 Holzmann, Tahmasebi, Risse

- lution recognition. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, Mumbai, India, December 2012.
- [2] Y.H. Segerstad. *Use and adaptation of written language to the conditions of computer-mediated communication*. PhD thesis, Göteborg University, 2002.
- [3] Nina Tahmasebi, Gerhard Gossen, and Thomas Risse. Which words do you remember? temporal properties of language use in digital archives. In *Theory and Practice of Digital Libraries*, volume 7489, pages 32–37. Springer, 2012.
- [4] Klaus Berberich, Srikanta J. Bedathur, Mauro Sozio, and Gerhard Weikum. Bridging the terminology gap in web archive search. In *WebDB*, 2009.
- [5] Amal Chaminda Kaluarachchi, Aparna S. Varde, Srikanta J. Bedathur, Gerhard Weikum, Jing Peng, and Anna Feldman. Incorporating terminology evolution for query translation in text retrieval with association rules. In *CIKM*, pages 1789–1792. ACM, 2010.
- [6] Nattiya Kanhabua and Kjetil Nørvg. Exploiting time-based synonyms in searching document archives. In *Proceedings of the 10th annual joint conference on Digital libraries, JCDL '10*, pages 79–88, New York, NY, USA, 2010. ACM.
- [7] Arturas Mazeika, Tomasz Tylenda, and Gerhard Weikum. Entity timelines: visual analytics and named entity evolution. In *CIKM*, pages 2585–2588, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2064026. URL <http://doi.acm.org/10.1145/2063576.2064026>.
- [8] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165, 2009.
- [9] A. Garcá-Silva, M. Szomszor, H. Alani, and O. Corcho. Preliminary results in tag disambiguation using dbpedia. In *Knowledge Capture (K-Cap 2009)-Workshop on Collective Knowledge Capturing and Representation-CKCaR*, 2009.
- [10] Aaron Coburn. Lingua::EN::Tagger - search.cpan.org. (accessed October 27, 2009), 2008. URL <http://search.cpan.org/perldoc?Lingua::EN::Tagger>.
- [11] Technorati Inc. accessed June 05, 2013, 2013. URL <http://www.technorati.com>.
- [12] Iadh Ounis, Craig Macdonald, and Ian Soboroff. Overview of the trec-2008 blog track. In *In Proceedings of TREC-2008*, 2009.

Elevating Natural History Museums' Cultural Collections to the Linked Data Cloud

Giannis Skevakis, Konstantinos Makris, Polyxeni Arapi, and Stavros Christodoulakis

Laboratory of Distributed Multimedia Information Systems and Applications,
Technical University of Crete (TUC/MUSIC), 73100 Chania, Greece
{skevakis, makris, xenia, stavros}@ced.tuc.gr

Abstract. An impressive abundance of high quality scientific content about Natural History and Biodiversity is produced in a distributed, open fashion by Natural History Museums (NHMs) using their own established standards and best practices. Managing publication of such richness and variety of content on the Web, and also supporting distributed, interoperable content creation processes, poses challenges that traditional publication approaches are not adequate to meet. The Natural Europe project offers a coordinated solution to those challenges at European level that aims to improve the availability, discoverability and relevance of environmental cultural content for education and life-long learning use, in a multilingual and multicultural context. Cultural heritage content is collected from six Natural History Museums around Europe into a federation of European Natural History Digital Libraries that is directly connected with Europeana. In this paper we present the architecture of the semantic infrastructure developed for the transition of the Natural Europe federation of NHMs' cultural repositories to the Semantic Web, as well as the methodology followed for ingesting and converting the NHMs' cultural heritage metadata into Linked Data.

Keywords: digital curation, preservation metadata, Europeana, Linked Data

1 Introduction

Cultural heritage and biodiversity data are syntactically and semantically heterogeneous, multilingual, semantically rich, and highly interlinked. They are produced in a distributed, open fashion by organizations like museums, libraries, and archives, using their own established standards and best practices [4]. As a result, an impressive abundance of high quality scientific content available around the world remains largely unexploited. Managing the publication of rich content on the Web and supporting distributed, interoperable content creation processes, poses challenges that traditional publication approaches are not adequate to meet.

The Semantic Web and Linked Data is a promising approach to address these problems. The Semantic Web standards and best practices provide a basis on which interoperable Web systems can be built in a well defined manner. W3C recommendations like RDF(S), SKOS, SPARQL, and OWL are considered as corner-stones for

cross-domain and domain-independent interoperability. Moreover, the exploitation of common ontologies, taxonomies and published datasets make the reusability of existing data possible. The exploitation of the aforementioned standards and practices in the Cultural Heritage domain and their adaptation by collaborative tools allowing open content publishing on the Semantic Web leads to: (a) semantically richer content, (b) creation of large national and international Cultural Heritage portals, such as Europeana, (c) large open data repositories, such as the Linked Open Data Cloud, and (d) massive publications of linked library data [4].

The Natural Europe project [11] offers a coordinated solution at European level that aims to improve the availability and relevance of environmental cultural content for education and life-long learning use, in a multilingual and multicultural context. Cultural heritage content related to natural history, natural sciences, and nature/environment preservation, is collected from six Natural History Museums (NHMs) around Europe into a federation of European Natural History Digital Libraries, directly connected with Europeana. The Natural Europe project adopts and integrates the strong requirements for metadata management and interoperability with cultural heritage, biodiversity, and learning repositories. It offers appropriate tools and services that allow the participating NHMs to: (a) uniformly describe and semantically annotate their content according to international standards and specifications, as well as (b) interconnect their digital libraries and expose their Cultural Heritage Object (CHO) metadata records to Europeana.eu.

In this paper we present the architecture of the semantic infrastructure developed for the transition of the Natural Europe Cultural Digital Libraries Federation to the Semantic Web, as well as the methodology followed for ingesting and converting the NHMs' cultural heritage metadata to Linked Data, supporting the Europeana Data Model (EDM) [2].

2 The Natural Europe Cultural Digital Libraries Federation and the Transition to the Semantic Web

In the context of Natural Europe, the participating NHMs provide metadata descriptions about a large number of Natural History related CHOs. These descriptions are semantically enriched with Natural Europe shared knowledge (vocabularies, taxonomies, etc.) using project provided annotation tools and services. The enhanced metadata are aggregated by the project, harvested by Europeana and exploited for educational purposes. The architecture of the Natural Europe Cultural Federation (**Fig. 1**) consists of the following components:

- The *Natural Europe Cultural Environment (NECE)* [6], which facilitates the complete metadata management lifecycle (i.e., ingestion, maintenance, curation, and dissemination) of CHO metadata and specifies how legacy metadata are migrated into Natural Europe. NECE provides (among others) the following tools and services for each participating NHM:

- The *MultiMedia Authoring Tool (MMAT)*¹ is a multilingual web-based management system for museums, archives and digital collections, which facilitates the authoring and metadata enrichment of CHOs. It employs modules for CHO/multimedia manipulation, persistency and vocabulary management.
- The *CHO Repository* is the underlying repository of MMAT, responsible for the ingestion, maintenance and dissemination of both content and metadata. It is backed up by an eXist XML database and exposes an OAI-PMH interface, able to disseminate metadata records complying with the Natural Europe CHO Application Profile.
- The *Vocabulary Management* module enables the access to taxonomic terms, vocabularies, and authority files (persons, places, etc.).
- The *Natural Europe Cultural History Infrastructure (NECHI)* [7] interconnects the NHM digital libraries and exposes their metadata records to Europeana.eu. It provides (among others) the following tools and services:
 - The *Natural Europe Harvester* manages the harvesting of metadata records provided by Natural Europe content providers. It employs modules for persistent identification, metadata transformation and metadata validation.
 - The *Metadata Repository* is the underlying repository of the Natural Europe Harvester, responsible for the maintenance of the harvested metadata records. It is backed up by an RDBMS and exposes an OAI-PMH service interface, able to disseminate metadata records complying with Europeana Semantic Elements (ESE) [3].

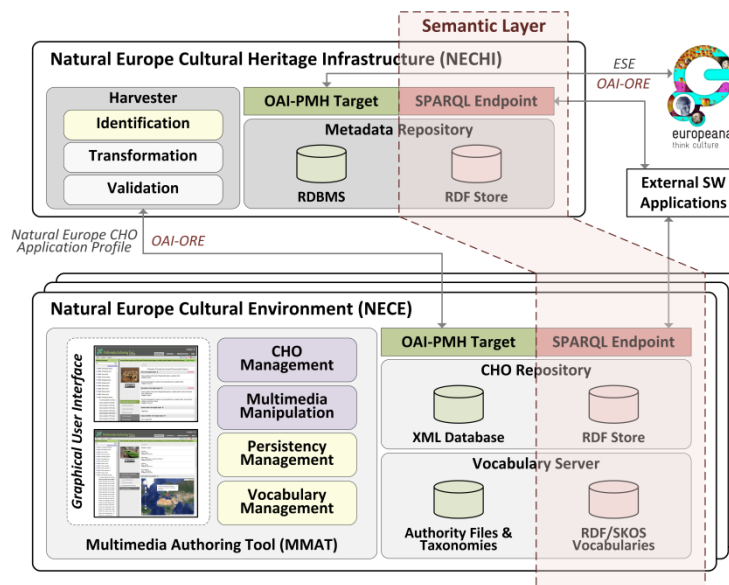


Fig. 1. Enhanced Natural Europe Cultural Digital Libraries Federation, supporting the transition to the Semantic Web and Linked Data Cloud

¹ A demo version of MMAT is available at: <http://natural-europe.tuc.gr/music/mmat>

Our aim was to develop a semantically rich cultural heritage infrastructure for NHMs, providing a Semantic Web perspective to the Natural Europe cultural content in terms of: (a) creating the Natural Europe Ontology in order to introduce semantics to the current Natural Europe Schema for inferring new knowledge, (b) using the RDF data model to publish the Natural Europe data on the Web, (c) linking the Natural Europe's cultural content to external commonly used vocabularies, thesaurus and published datasets, (d) enabling data retrieval through SPARQL, and (e) supporting interoperability with the Europeana Semantic Layer by offering the appropriate Europeana Data Model (EDM) [2] dissemination mechanisms.

In order to achieve the above objectives, the modules of the federated instances (NECE) and the federal node (NECHI) of the Natural Europe Cultural Federation have been enhanced with software components supporting the Semantic Web technologies. The Natural Europe RDF data are aggregated to the federal node in order to allow the inference of new knowledge from all NHM federated nodes. This data format allows the execution of domain specific queries. Each federated node provides an RDF store which allows the retrieval of a single museum's data through SPARQL, and enables the future connection with MMAT which will be modified to provide up-to-date triples. Specifically, NECE has been enhanced with the following modules/functionality:

- The *Vocabulary Server* has been extended with published taxonomies expressed in RDF/SKOS format.
- The *RDF Store* is managed by the *CHO Repository* and keeps the triples generated from the conversion of the XML metadata to the RDF format.
- The *SPARQL Endpoint*, exposed by *CHO Repository*, enables semantic queries on top of the triples stored in the RDF Store.
- The *OAI-PMH Target* has been refactored to support the harvesting of OAI-ORE packages by NECHI, which in turn allows the further exploitation of the data.

NECHI has been enhanced with the following modules/functionality:

- The *RDF Store*, managed by the *Metadata Repository*, keeps the harvested triples of NECE instances, along with knowledge inferred from the aggregated datasets.
- The *SPARQL Endpoint*, exposed by *Metadata Repository*, allows external systems to query the aggregated data from the entire Natural Europe infrastructure.
- The *OAI-PMH Target* has been refactored to support the dissemination of the aggregated linked data to Europeana, when the respective Europeana services become available.

2.1 From the Natural Europe Schema to the Natural Europe Ontology

The Natural Europe data complies with the Natural Europe CHO Application Profile [10], which is a superset of the Europeana Semantic Elements (ESE) [3] metadata format. The Natural Europe CHO Application Profile describes the cultural heritage objects as records and consists of the following parts:

- The *Cultural Heritage Object (CHO) information* that provides metadata information about the analog resource or born digital object (specimen, exhibit, cast, painting, documentary, etc.). It is composed of the following sub-categories:
 - The *Basic information* holds general descriptive information (mostly scientific) about a Cultural Heritage Object.
 - The *Species information* holds information related to the species of a described specimen (animals, plants, minerals, etc.) in the context of Natural Europe.
 - The *Geographical information* contains metadata for the location in which the specimen has been collected.
- The *Digital Object information* that provides metadata information about a digital (photo, video, etc.) or digitized resource (scanned image, photo, etc.) in the context of Natural Europe. It is composed of the following sub-categories:
 - The *Basic information* deals with general descriptive information about a digital or digitized resource.
 - The *Content information* is related to the physical characteristics and technical information exclusive to a digital or digitized resource (URL, Format, etc.).
 - The *Rights information* describes the intellectual property rights and the accessibility to a digital or digitized resource.
- The *Meta-metadata information* that provides metadata information for a CHO record. These include the creator of the record, the languages that appear in the metadata, the history of the record during its evolution in the MMAT, etc.
- The *Collection information* that provides metadata information for logical groupings of contributed CHOs within a museum.

When creating a rich cultural heritage infrastructure that aims to provide a Semantic Web perspective to the Natural Europe cultural content, it is not sufficient to use a flat Schema or a Schema providing weak semantics. With this objective in mind we described the Natural Europe Schema as an OWL Ontology, exploiting the use of class and property axioms in order to enable the inference of new knowledge out of the existing data. Notions such as CHO, CHO collection, specimen, observation, multimedia object, person, and organization have been described as OWL classes, while the underlying attributes have been described using object and datatype properties. As a result, the contributed flat Natural Europe records can be organized in aggregations of different kinds of objects, e.g., a specimen may be described by multiple observations and an observation may contain multiple multimedia objects.

The Natural Europe Ontology references other well-known Ontologies/Schemas (e.g., SKOS) and has been aligned with EDM, allowing any system supporting the Natural Europe Ontology to work seamlessly with other systems/organizations supporting EDM.

3 Vocabularies

Exposing data to the Linked Data cloud is not only about creating an Ontology with sufficient semantics and converting them to RDF. The quality of the exposed Linked Data is measured by their linkage with already published, external data. To this end,

we tried to find external sources that provide data which we can be linked to our datasets. Most of the vocabularies that we came across already provided RDF data and ways to access/query them. Nevertheless, the Catalogue of Life (CoL) which is used extensively in the biodiversity context did not expose any data in RDF format. To overcome this issue, we supported the publishing of its database to RDF (Section 3.1). In the context of the Natural Europe infrastructure, we chose the following services/datasets:

- **GeoNames:** Geographical database containing over 10 million geographical names and consisting of over 8 million unique features whereof 2.8 million populated places and 5.5 million alternate names. The GeoNames Ontology is described using OWL, exploiting its inferencing capabilities for extracting new knowledge. In addition, it supports interoperability by being mapped to several well-known ontologies, including schema.org, linkedgeodata.org, dbpedia.org, and INSEE. The GeoNames website offers numerous web services² for searching any kind of information available in the system. The response data is available in multiple formats, like XML and JSON.
- **DBpedia:** A knowledge base describing more than 3.64 million things, including 764,000 persons, 573,000 places, and 202,000 species. The dataset has been mainly created by extracting structured data from Wikipedia and has been classified in a consistent cross-domain Ontology. The data can be accessed through web services using the provided SPARQL Endpoint³, or the XML based search api⁴.
- **Catalogue of Life (CoL):** A comprehensive catalogue of all known species of organisms on Earth. It is compiled by 99 taxonomic databases from around the world providing critical species information on: (a) synonymy, enabling the effective referral of alternative species names to an accepted name, (b) higher taxa, within which a species is clustered, and (c) distribution, identifying the global regions from which a species is known. It is being used by several Global Biodiversity Programmes including GBIF⁵ and EoL⁶. The CoL website offers web services for searching the latest taxonomy⁷ and the available response formats are: JSON/XML/PHP-based.
- **Uniprot:** A comprehensive, high-quality and freely accessible database of protein sequence and functional information including among others, a taxonomic classification, literature citations and keywords. The dataset is also available in the RDF format, conforming to the highly structured Uniprot OWL Ontology, while the Uniprot taxonomic classification has been described with SKOS. The Uniprot database can be downloaded⁸ or queried through the provided RESTful services⁹.

² <http://www.geonames.org/export/ws-overview.html>

³ <http://wiki.dbpedia.org/OnlineAccess>

⁴ <http://wiki.dbpedia.org/lookup/>

⁵ <http://www.gbif.org/>

⁶ <http://eol.org/>

⁷ <http://webservice.catalogueoflife.org/>

⁸ <http://www.uniprot.org/downloads>

⁹ <http://www.uniprot.org/faq/28>

- **GEMET:** A general multilingual thesaurus aimed to define a common language and core terminology for the environment. GEMET's data is available in SKOS (RDF/XML) format and can be either downloaded¹⁰ or accessed through the provided RESTful and XML-RPC interfaces¹¹.

3.1 SKOSification of Catalogue of Life

The extended use of taxonomies in the biodiversity domain dictates for a formal way of describing complex vocabularies and taxonomies, in compliance to the Semantic Web standards. The most popular standard for describing these types of controlled vocabularies is SKOS (Simple Knowledge Organization System) [8]. SKOS is formally described as an OWL Full Ontology, providing the basic notions and semantics needed for describing knowledge in knowledge organization systems. Its use facilitates the semantic linkage of museum objects to well-established KOS, including GEMET and Uniprot which have already been expressed in SKOS. Another system that is widely used in biological classification but is not available in SKOS format is the Catalogue of Life, which has been described above.

The current implementation of CoL provides a web-based system for browsing the taxonomy of the species, as well as services for searching, but lacks support for persistent URIs able to be referenced by external applications, and RDF representation of its data. Towards this end, we have worked on a method of exposing the taxonomy of CoL to RDF, and more specifically SKOS, using the annual checklist, which is a downloadable package containing the relational database of the CoL.

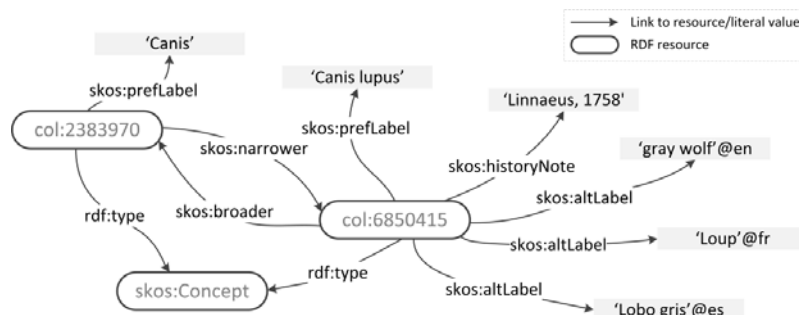


Fig. 2. An example of the Catalogue of Life SKOSified data in the form of a graph

For the conversion of the CoL dataset to SKOS we used the D2R Server [1], which allows the publishing of relational databases in RDF format. The features of the SKOS model that we employed are: (a) the class *Concept*, and (b) the properties *broader*, *narrower*, *prefLabel* and *altLabel*. The first step was the representation of all the taxonomy nodes as *Concepts*. The scientific name of each node was transformed into a *prefLabel*, and the common names into *altLabels*. Finally, the hierarchy of the

¹⁰ <http://www.eionet.europa.eu/gemet/rdf>

¹¹ <http://taskman.eionet.europa.eu/projects/zope/wiki/GEMETWebServiceAPI>

taxonomy was retained by connecting the parent and children nodes with the properties *broader* and *narrower*. An example of the CoL SKOSified data in the form of a graph is shown in **Fig. 2**.

4 Methodology

The methodology for the transition of the Natural Europe Cultural Federation and cultural data to the Semantic Web and the Linked Data Cloud includes the following stages: (1) enrichment of Natural Europe metadata records with knowledge from well-known vocabularies and thesaurus (e.g., Geonames, DBpedia, GEMET and CoL/Uniprot), (2) conversion of metadata from XML to RDF, (3) connection of Natural Europe LOD node to the Linked Data cloud (RDF store, SPARQL endpoint), and (4) transition to EDM. These stages are described in the following sections.

4.1 Metadata Enrichment

The metadata enrichment is a very crucial step in the production of rich Linked Data, especially in the case where data already exist in other legacy formats (Relational Databases, XML Databases, etc.). Existing data in these systems are rarely connected to external data because of the structure of the information storage and the fact that most of these have been created long before the introduction of the Open Data.

The Natural Europe datasets have been linked to the above vocabularies/thesaurus by executing customized batch operations that exploit the services exposed by the datasets. More specifically, the spatial information of a Natural Europe CHO record that generally describes places is matched to place names in Geonames. This provides unique references for places and enables spatial information enhancement with: (a) multiple multilingual versions of place names, (b) geographic coordinates and (c) broader geographic areas associated with the places. Unique references for places have also been retrieved from Geonames by exploiting any available geographic coordinates associated with the CHO.

The CHO scientific names of the species information are matched to the accepted scientific names of CoL/Uniprot. By doing so, unique references to well-known taxonomic databases are established and scientific information regarding species common names, species distribution and literature citations is added to the CHO records.

The scientific information of CHO records is further enriched with knowledge retrieved from the DBpedia database. To this end, the scientific names appearing in the Natural Europe CHO records are matched to DBpedia resources, providing links to external bibliographic references, as well as additional information such as the abstract description and conservation status of the CHO's referred species.

The keywords describing CHOs and CHO collections are matched to terms in the Gemet thesaurus. This provides unique references for keywords, and enhances the CHO information with terms in multiple languages, as well as labels or references of broader terms.


```

<record xmlns="http://www.natural-europe.eu/nhm/aip/">
  <objectUri>http://nhmc.natural-europe.eu/12dda2d5</objectUri>
  <contextUri>http://www.nhmc.uoc.gr/museum/40319</contextUri>
  <contentType>http://purl.org/dc/dcmitype/Image</contentType>
  <licenseUri>http://creativecommons.org/licenses/by-nc-nd/3.0</licenseUri>
  <scientificName xml:lang="el">http://www.catalogueoflife.org/col/6850415</scientificName>
  <commonName xml:lang="el">Λύκος</commonName>
  <commonName xml:lang="en">Wolf</commonName>
  <title xml:lang="en">Wolf, Canis lupus</title>
  <title xml:lang="el">Λύκος, Canis lupus</title>
  <creator>http://nhmc.natural-europe.eu/persons/158ggse7</creator>
  <subject>http://www.eionet.europa.eu/gemet/concept/4982</subject>
  <description xml:lang="en">Photo of wolves in forest diorama in the Paranefti NHM.</description>
  <description xml:lang="el">Φωτογραφία λύκων σε διόραμα δάσους στο ΜΦΙ στο Παράνεφτι.</description>
  <contributor>http://nhmc.natural-europe.eu/persons/1dg5hhd7</contributor>
  <type xml:lang="en">Preserved specimen</type>
  <type xml:lang="el">Συντηρημένο δείγμα</type>
  <format>image/jpeg</format>
  <identifier>nhmc.image.40319</identifier>
  <rights annotation="Rights Reserved - Free Access">Natural History Museum of Crete ©</rights>
  <alternative xml:lang="en">Photo of Canis lupus</alternative>
  <alternative xml:lang="el">Φωτογραφία ενός Canis lupus</alternative>
  <extent>500 x 323 pixels</extent>
  <spatial>http://www.geonames.org/390903</spatial>
  <geolocation latitude="35.296227084320144" longitude="23.91901402771254"/>
  <relation>http://live.dbpedia.org/page/Gray_wolf</relation>
  <relation>http://live.dbpedia.org/page/Carl_Linnaeus</relation>
</record>

```

Fig. 3. An example of an enriched Natural Europe XML record

Apart from the use of external vocabularies and thesauri, person authority files have been created using information about CHO record creators and contributors. This way, information about persons in the CHO records has been replaced with references to the created authority files. It is worth to note that the authority file metadata are available in RDF, becoming resolvable and linkable by other external applications. In addition, information regarding CHO relations within each federated node's repository is enriched by matching existing CHO records based on their scientific name. An example of a semantically enriched Natural Europe record is presented in **Fig. 3**. Although the CHO metadata enrichment process has been performed automatically, we plan to support the inspection of the results using MMAT.

4.2 Conversion of Metadata from XML to RDF

Generally, the basic operations that have to take place in order to convert XML data to RDF include: (a) mapping of every complex XML element to a resource (often a blank node) and of every atomic attribute to an attribute of this resource, and (b) assignment of a namespace prefix to each XML name to create fully qualified URIs.

In the case of Natural Europe, the XML to RDF data conversion has been performed through automatic transformation processes, taking into account the Natural Europe Ontology. The *Identification* module of the Natural Europe's federal node has a central role in this process by providing unique identifiers for previously anonymous objects. The generated RDF data have been persisted in an RDF store and can be queried through SPARQL. The use of the Natural Europe Ontology allows the inference of new RDF statements by applying well known reasoning techniques that exploit OWL axioms. An example of the Natural Europe RDF data in the form of a graph is shown in **Fig. 4**.

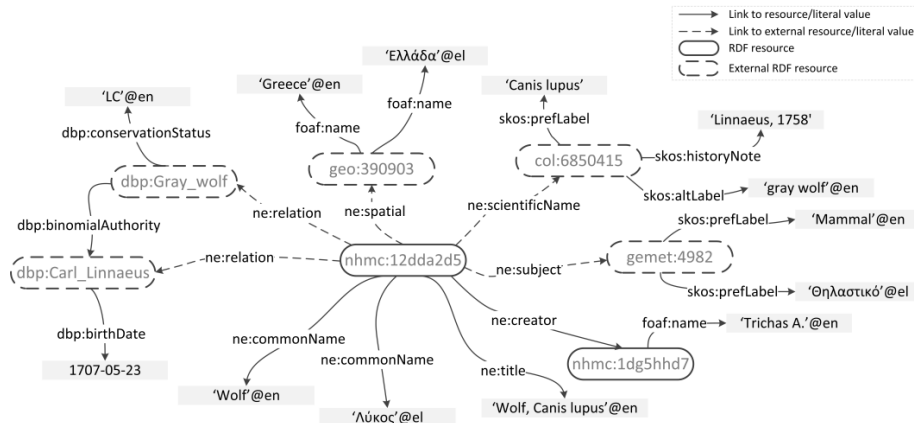


Fig. 4. An example of the Natural Europe RDF data in the form of a graph

4.3 Connection of Natural Europe LOD Node to the Linked Data Cloud

The exposure of the Natural Europe data in RDF format as well as the availability of the semantic services (SPARQL Endpoint, Resolvable URIs) allows all the museums' specimens to be available on the Linked Data cloud. This way anyone can reference any node of the knowledge graph, based on the Linked Data paradigm.

All the data in the Natural Europe environment (even those coming from different institutes) have been automatically interconnected using the aforementioned vocabularies/taxonomies. As an example, consider two museums that have described a specimen of a gray wolf (*canis lupus*). During the enrichment step the CHOs are connected to the SKOS *Concept* describing "*Canis lupus*", and as soon as the data are available in RDF triples, there will be at least two resources of the class *Specimen* that are linked to "*Canis lupus*". Using the SPARQL endpoint in the federated node, or the feature of the federated query of SPARQL 1.1 specification, we can utilize the relation between these two specimens.

4.4 Transition to EDM

From a technical point of view, EDM adheres to the modeling principles that underpin the approach of the Web of Data ("Semantic Web"). In this approach, there is no such thing as a fixed schema that dictates just one way to represent the data. A common model like EDM can be seen instead as an anchor to which various finer-grained models can be attached, making them at least partly interoperable at the semantic level, while the data retain their original expressivity and richness. It does not require changes in the local approaches, although any changes that increase the cross-domain usefulness of the data are encouraged (e.g., the usage of publicly accessible vocabularies for persons, places, subjects etc.).

Nevertheless, an ingestion mechanism is yet to be provided by Europeana. Until such an option is available, the only way to expose external data to the system is

through the ingestion of XML records in ESE format. Our approach ensures that the generated data complies with the EDM specification, thus allowing the immediate dissemination to the Europeana infrastructure. To this end, we plan to support the ingestion of EDM (OAI-ORE) packages through the OAI-PMH protocol on the federated node. This will be implemented very closely to the way that the data are aggregated from the federated to the federal node.

5 Related Work

The *STERNA* project [12] focuses on the enrichment of existing content in the natural history domain. It has developed a methodology on how to integrate one's content into the *STERNA* information space. Its Reference Network Architecture (RNA) is a web-based information architecture that allows connecting various knowledge resources and provides an accessible and unambiguous way of retrieving the heterogeneous content within those resources. RNA's architecture is based on RDF and SKOS. In an RNA environment, content items can be stored in several different RDF stores that can be located on different servers and on various locations. However, they can still be approached as one integrated environment when using the RNA Toolset or when searching the RNA environment.

The *MultimediaN E-Culture* project [9] developed a search portal and engine served as a joint prototype Semantic Web application for subsets of digital collections and thesauri from a number of heritage institutions. Several datasets from Dutch art and ethnographic collections have been ported to the Semantic Web. The core Getty vocabularies (AAT, TGN and ULAN) have been converted from the Getty XML files into RDF, and together with the SKOSified thesauri and other controlled vocabularies form the RDF graph underlying the E-Culture semantic search portal demonstrator. The project has developed a generic Java-based framework for converting collection metadata and controlled vocabularies into RDF/SKOS (AnnoCultor).

The *STITCH* project [13] examined the extent to which current Semantic Web techniques can solve issues presented by the heterogeneity of cultural heritage collection databases and controlled vocabularies. To this purpose, *STITCH* developed methods for aligning and browsing reference structures such as SKOSified thesauri and classification systems. SKOS representations of Iconclass and Aria thesaurus aligned these representations using state-of-the-art mapping tools, and implemented a faceted Web browsing environment to visualize and examine the results.

6 Conclusion

We presented a semantic infrastructure and a methodology making possible the transition of the Natural Europe Cultural Digital Libraries Federation, providing cultural and biodiversity content, to the Semantic Web and the Linked Data Cloud. The methodology includes the following stages: (a) enrichment of Natural Europe metadata, (b) conversion of metadata from XML to RDF, (c) connection of Natural Europe LOD node to the Linked Data cloud (RDF store, SPARQL endpoint), and (d) transition to

EDM. This methodology can be applied in other domains as well, exploiting their schemes, and related with the domain vocabularies/taxonomies.

Our current research focuses on investigating the integration of the Natural Europe NHM federated nodes with cultural heritage and biodiversity RDF data providers, utilizing different metadata schemas (e.g., ABCD), in an ontology-based mediator system. Such an infrastructure is extremely important for Semantic Web applications and end users, since it will enable the retrieval of up-to-date triples, unlike the data warehousing approaches applied by data aggregators. To this end, the SPARQL-RW Framework [5], developed by TUC/MUSIC Lab, is considered as a corner-stone component for transparently accessing federated RDF data sources complying to different Ontology Schemas.

Acknowledgements. This work has been carried out in the scope of the Natural Europe Project (Grant Agreement 250579) funded by EU ICT Policy Support Programme.

References

1. Bizer C., Cyganiak R.: D2r server-publishing relational databases on the semantic web. In: Proceedings of the 5th International Semantic Web Conference (ISWC), (2006).
2. Europeana Data Model Definition V.5.2.3, <http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22>
3. Europeana Semantic Elements Specification V.3.4.1, <http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57>
4. Hendler J., Ding Y.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on Semantic Web: Theory and Technology. Morgan & Claypool Publishers series (2012).
5. Makris K., Bikakis N., Gioldasis N., Christodoulakis S.: SPARQL-RW: Transparent Query Access over Mapped RDF Data Sources. In: Proceedings of the 15th International Conference on Extending Database Technology (EDBT), Berlin (2012).
6. Makris K., Skevakis G., Kalokyri V., Arapi P., Christodoulakis S.: Metadata Management and Interoperability Support for Natural History Museums. In: Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL), Malta (2013).
7. Makris K., Skevakis G., Kalokyri V., Gioldasis N., Kazasis F., Christodoulakis S.: Bringing Environmental Culture Content into the Europeana.eu Portal: The Natural Europe Digital Libraries Federation Infrastructure. In: Proceedings of the 5th Metadata and Semantics Research Conference (MTSR), Izmir (2011).
8. Miles A., Matthews B., Wilson M., Brickley D.: SKOS Core: Simple knowledge organization for the Web, (2005).
9. MultimediaN E-Culture project, <http://e-culture.multimedien.nl>
10. Natural Europe Cultural Heritage Object Application Profile, http://wiki.natural-europe.eu/index.php?title=Natural_Europe_Cultural_Heritage_Object_Application_Profile
11. Natural Europe Project, <http://www.natural-europe.eu>
12. STERNA project, <http://www.sterna-net.eu/>
13. STITCH project, <http://www.cs.vu.nl/STITCH/>

A linked open data architecture for contemporary historical archives

Alexandre Rademaker¹, Suemi Higuchi², and Dário Augusto Borges Oliveira²

¹ IBM Research and FGV/EMAp

² FGV/CPDOC

Abstract. This paper presents an architecture for historical archives maintenance based on Open Linked Data technologies and open source distributed development model and tools. The proposed architecture is being implemented for the archives of the Center for Teaching and Research in the Social Sciences and Contemporary History of Brazil (CPDOC) from Getulio Vargas Foundation (FGV).

1 Introduction

The Center for Teaching and Research in the Social Sciences and Contemporary History of Brazil (CPDOC) was created in 1973 and became an important historical research institute, housing a major collection of personal archives, oral histories and audiovisual sources that document the country memory.

CPDOC is a vibrant and diverse intellectual community of scholars, technicians and students, and has placed increasing emphasis on applied research in recent years, working in collaborative projects with other schools and institutes, aiming at extending the availability and scope of the valuable records it holds. It is part of Getulio Vargas Foundation (FGV), a prestigious Brazilian research and higher education institution founded in 1944, considered by Foreign Policy Magazine to be a top-5 “policymaker think-tank” worldwide [25].

Thanks to the donation of personal archives of prominent Brazilian figures from the 1930s onward, such President Getulio Vargas himself, CPDOC started to develop its own methodology for organizing and indexing documents. By the end of the 1970s, it was already recognized as a reference among research and historical documentation centers. In 1975, the institute launched its Oral History Program (PHO), which involved the execution and recording of interviews with people who participated in major events in Brazilian history. In 1984, CPDOC published the Brazilian Historical-Biographical Dictionary (DHBB) [1], a regularly updated reference resource that documents the contemporary history of the country. In the late 1990s, CPDOC was recognized as center of excellence by the Support Program for Centers of Excellence (Pronex) of the Brazilian Ministry of Science and Technology.

This year, celebrating 40 years of existence, CPDOC received the support of the Brazilian Ministry of Culture (MinC), which provided a fund of R\$ 2.7 million to finance the project “Dissemination and Preservation of Historical Documents”. This project has the following main goals: (1) digitizing a significant

amount of textual, iconographic and audiovisual documents; (2) updating the dictionary DHBB; and (3) prospecting innovative technologies that enable new uses for CPDOC's collections.

The advances in technology offer new modes of dealing with digital contents and CPDOC is working to make all data available in a more intelligent/semantic way in the near future, offering swift access to its archives. In collaboration with the FGV School of Applied Mathematics (EMAp), CPDOC is working on a project that aims to enhance access to documents and historical records by means of data-mining tools, semantic technologies and signal processing. At the moment, two applications are being explored: (1) face detection and identification in photographs, and (2) voice recognition in the sound and audiovisual archives of oral history interviews. Soon it will be easier to identify people in the historical images, and link them to the entries in CPDOC archives. Additionally, voice recognition will help locate specific words and phrases in audiovisual sources based on their alignment with transcription – a tool that is well-developed for English recordings but not for Portuguese. Both processes are based on machine learning and natural language processing, since the computer must be taught to recognize and identify faces and words.

CPDOC also wants its data to constitute a large knowledge base, accessible using the standards of semantic computing. Despite having become a reference in the field of organization of collections, CPDOC currently do not adopt any metadata standards nor any open data model for them. Trends for data sharing and interoperability of digital collections pose a challenge to the institution to remain innovative in its mission of efficiently providing historical data. It is time to adjust CPDOC's methodology to new paradigms.

In Brazilian scenario many public data is available for free, but very few are in open format following the semantic web accepted standards. Examples in this direction are the Governo Aberto SP [13], the LeXML [23] and the SNIIC project ³.

In this sense, we present hereby a research project that reflects a change in the way CPDOC deals with archives maintenance and diffusion. The project is an ongoing initiative to build a model of data organization and storage that ensures easy access, interoperability and reuse by service providers. The project proposal is inspired by: (1) Open Linked Data Initiative principles [21]; (2) distributed open source development model and tools for easy and collaborative data maintenance; (3) a growing importance of data curating concepts and practices for online digital archives management and long-term preservation.

The project started with an initiative of creating a linked open data version of CPDOC's archives and a prototype with a simple and intuitive web interface for browsing and searching the archives was developed. The uses of Linked Open Data concept are conformed to the three laws first published by David Eaves [14] and now widely accepted: (1) If it can't be spidered or indexed, it doesn't exist;

³ Sistema Nacional de Informaes e Indicadores Sociais, <http://culturadigital.br/sniic/>.

- (2) If it isn't available in open and machine readable format, it can't engage; and
- (3) If a legal framework doesn't allow it to be repurposed, it doesn't empower.

Among the project objectives we emphasize the construction of a RDF [24] data from data originally stored in a relational database and the construction of an OWL [18] ontology to properly represent the CPDOC domain. The project also aims to make the whole RDF data available for download similarly to what DBpedia does [6].

This paper reflects a real effort grounded in research experience to keep CPDOC as a reference institution in the field of historic preservation and documentation in Brazil.

2 CPDOC information systems

Figure 1 presents the current CPDOC database architecture. The archives are maintained in three different information systems that share a common relational database. Each of the systems has independent management and adopts idiosyncratic criteria concerning the organization and indexing of the information, which vary depending on the specifications of the content they host: personal archives documents, oral history interviews and the Brazilian Historical – Biographic Dictionary entries. CPDOC's web portal provides a query interface to archives data. In the following subsections we briefly describe each of the systems.

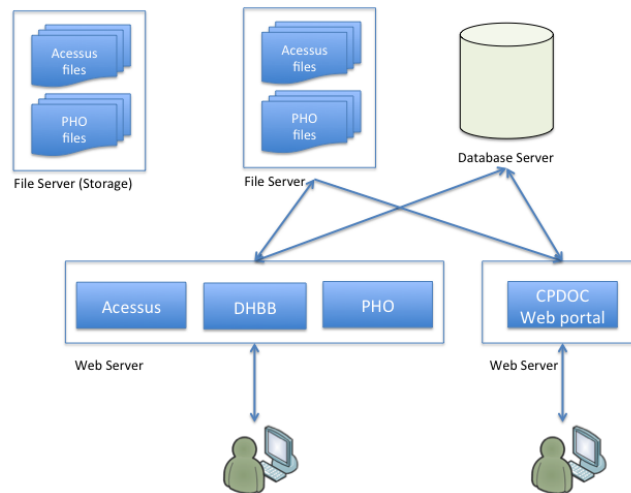


Fig. 1. CPDOC's current architecture

2.1 Personal Archives (Acessus)

This system is composed by personal files from people who influenced the political and social scenario of our country. These historical documents, in textual or audiovisual form, in form of handwritten and printed texts, diaries, letters, photographs, speeches or memos, represent much more than private memories: they are the registry of a whole collective memory.

Currently, more than 200 personal archives from presidents, ministers, military personal and other Brazil's important public figures compose the CPDOC's collections. Together, they comprise nearly 1.8 million documents or 5.2 millions pages. From this, nearly 700 thousands pages are in digital format and the expectation is to digitize all collections in the next few years. The collection entries metadata are stored in an information system called Acessus. It can be accessed through the institution's intranet for data maintenance or by internet for simple data query. Currently, allowed queries are essentially syntactic, i.e., restricted to keywords searches linked to specific database fields defined in an *ad hoc* manner. For those documents that are already digitized, two digital file versions were generated: one in high resolution aiming long-term preservation and another in low resolution for web delivery. High resolution files are stored in a storage system with disk redundancy and restricted access, while low resolution files are stored in a file server ⁴ (Figure 1).

2.2 Oral History Interviews (PHO)

The CPDOC collection of Oral History hosts currently more than 6.000 hours of recording, corresponding to nearly 2.000 interviews. More than 90% of it, video or audio, are in digital format. For the time being, two kinds of queries are available in the database: query by subject and query by interviewed. Each interview record holds a brief technical information and a summary with descriptions of the interview themes in the order they appear in the record. Almost 80% of the interviews are transcribed, and to access the audio/video content the user is requested to come personally to CPDOC.

Currently, CPDOC is analyzing better ways of making this data available online, considering different aspects such as the best format, use policies, access control and copyrights.

As in the case of Acessus, the database actually stores only the metadata about the interviews, while digitized recorded audios and videos are stored as digital files in the file servers.

2.3 Brazilian Historical-Biographic Dictionary (DHBB)

The Brazilian Historical-Biographic Dictionary (DHBB) is certainly one of the main research sources for contemporary Brazilian politicians and themes. It contains more than 7.500 entries of biographic and thematic nature, i.e., people, institutions, organizations and events records carefully selected using criteria that

⁴ https://en.wikipedia.org/wiki/File_server.

measure the relevance of those to the political history for the given period. The entries are written evenly, avoiding ideological or personal judgments. CPDOC researchers carefully revise all entries to ensure the accuracy of the information and a common style criteria.

The DHBB's database stores few metadata concerning each entry, and the query is limited to keywords within the title or text.

3 Current Status

In this section we summarize the main problems identified in CPDOC's current infrastructure and daily working environment.

As described in Section 2, CPDOC's archives are maintained by three different information systems based on traditional relational data models. This infrastructure is hard to maintain, improve and refine, and the information is not found or accessed by standard search engines for two reasons mainly: (1) an entry page does not exist until it is created dynamically by a specific query; (2) users are required to login in order to make queries or access the digital files. Service providers do not access data directly and therefore cannot provide specialized services using it. Users themselves are not able to expand the queries over the collections, being limited to the available user interfaces in the website. Thereupon, data of CPDOC's collections is currently limited to what is called "Deep Web" [3].

The maintenance of current different information systems is very problematic. It is expensive, time demanding and ineffective. Improvements are hard to implement and therefore innovative initiatives are usually postponed. A relational database system is not easily modified, because relational data models must be defined *a priori*, i.e., before the data acquisition's stage. Moreover, changes in the database usually require changes in system interfaces and reports. The whole workflow is expensive, time consuming and demands different professionals with different skills from interface developers to database administrators.

Concerning terminology, CPDOC's collections do not follow any metadata standards, which hinders considerably the interoperability with other digital sources. Besides, the available queries usually face idiosyncratic indexing problems with low rates of recall and precision. These problems are basically linked to the *ad hoc* indexing strategy adopted earlier to define database tables and fields.

Finally, data storage is also an issue. Digitized Acessus's documents and Oral History's interviews are not stored in a single place, but scattered in different file servers. The CPDOC database only stores the metadata and file paths to the file servers, making it very difficult to ensure consistency between files, metadata information and access control policies.

4 The proposal

As discussed in Section 3, relational databases are often hard to maintain and share. Also, the idea of having in-house developed and closed source information systems is being increasingly replaced by the concept of open source systems. In such systems the responsibility of updating and creating new features is not sustained by a single institution but usually by a whole community that share knowledge and interests with associates. In this way the system is kept up-to-date, accessible and improving much faster due to the increased number of contributors. Such systems are usually compatible with standards so as to ensure they can be widely used.

Our objective is to propose the use of modern tools so CPDOC can improve the way they maintain, store and share their rich historical data. The proposal focuses on open source systems and a lightweight, shared way of dealing with data. Concretely, we propose the substitution of the three CPDOC systems by the technologies described as follows.

The ACESSUS data model comprises personal archives that contains one or more series (which can contain also other series in a stratified hierarchy) of digitalized documents or photos. The PHO system data model is basically a set of interviews grouped according to some defined criteria within the context given by funded projects. For instance, a political event could originate a project which involve interviewing many important people taking part on the event.

Therefore, ACESSUS and PHO systems can be basically understood as systems responsible for maintaining collections of documents organized in a hierarchical structure. In this way, one can assume that any digital repository management system (DRMS) have all the required functionalities. Besides, DRMS usually have desirable features that are not present in ACESSUS or PHO, such as: (1) data model based on standard vocabularies like Dublin Core [20] and SKOS [26]; (2) long-term data preservation functionalities (tracking and notifications of changes in files); (3) fine-grained access control policies; (4) flexible user interface for basic and advanced queries; (5) compliance with standard protocols for repositories synchronization and interoperability (e.g., OAI-PMH [22]); (6) import and export functionalities using standard file formats and protocols; and more.

In our proposal the data and files from ACESSUS and PHO systems are planned to be stored in an open source institutional repository software such as Dspace⁵ or Fedora Commons Framework⁶. In this article we assume the adoption of Dspace with no prejudice of theoretical modeling.

As to DHBB, its relational model can be summarized to a couple of tables that store metadata about the dictionary entries (stored in a single text field of a given table). The actual dictionary entries are created and edited in text editors outside the system and imported to it only after being created and revised.

The nature of its data suggests that DHBB entries could be easily maintained as text files using a lightweight human-readable markup syntax. The files would

⁵ <http://www.dspace.org/>

⁶ <http://www.fedora-commons.org>

be organized in an intuitive directory structure and kept under version control for coordinated and distributed maintenance. The use of text files ⁷ is justified by a couple of reasons. They are: easy to maintain using any text editor allowing the user to adopt the preferred text editor (tool independent); conform to long-term standards by being software and platform independent; easy to be kept under version control by any modern version control system ⁸ since they are comparable (line by line); and efficient to store information ⁹.

The use of a version control system will improve the current workflow of DHBB reviewers and coordinators, since presently there is no aid system for this task, basically performed using Microsoft Word text files and emails. The adoption of such tool will allow file exchanges to be recorded and the process controlled without the need of sophisticated workflow systems, following the methodology developed by open sources communities for open source software maintenance. For instance, Git ¹⁰ is specially suited to ensure data consistency and keeps track of changes, authorship and provenance.

Many of the ideas here proposed were already implemented as a proof of concept to evaluate the viability of such environment in CPDOC. Figure 2 illustrates the necessary steps to fully implement our proposal. In the following text we briefly describe each step.

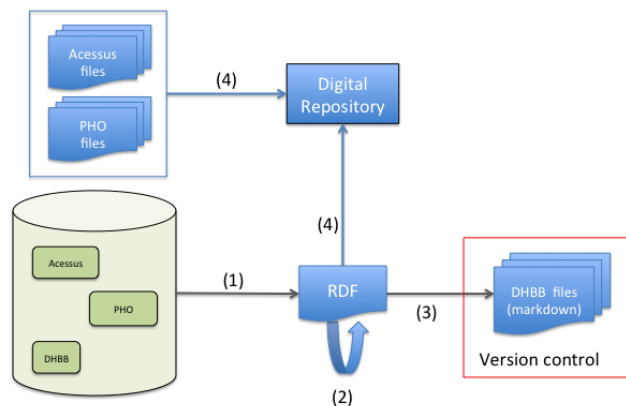


Fig. 2. Migrating from relational databases to the proposed model

Step (1) is implemented and the relational database was exported to RDF [24] using the open source D2RQ [5] tool. The D2RQ mapping language [11] allows the definition of a detailed mapping from the current relational model to a graph

⁷ http://en.wikipedia.org/wiki/Text_file

⁸ https://en.wikipedia.org/wiki/Revision_control.

⁹ A text file of a DHBB entry has usually 20% the size of a file DOCX (Microsoft Word) for the same entry.

¹⁰ <http://git-scm.com>.

model based on RDF. The mapping from the relational model to RDF model was already defined using the standard translation from relational to RDF model sketched out in [4]. The mapping created so far defers any model improvement to step (2) described below.

Step (2) is planned and represents a refinement of the graph data model produced in step (1). The idea is to produce a data model based on standard vocabularies like Dublin Core [20], SKOS [26], PROV [17] and FOAF [7] and well-known conceptual models like [10]. The use of standard vocabularies will make the data interchangeable with other models and facilitate its adoption by service providers and users. It will also help us to better understand the database model and its semantics. In Section 5 we describe the refinement proposal in detail.

Step (3) is already implemented and deploys a text file for each DHBB entry. Each text file holds the entry text and metadata ¹¹. The files use YAML [2] and Markdown [19] markup languages to describe the metadata and entry content. YAML and Markdown were adopted mainly because both languages are human-readable markups for text files and are supported by almost all static site generators ¹². The use of a static site generator allows DHBB maintainers to have full control over the deployment of a DHBB browsable version.

Note that step (3) was actually implemented to use the initial version of the RDF produced in step (1). The code can be easily adapted to use the final RDF model produced by step (2).

In the planned step (4) the digital files and their metadata will be imported into a DRMS. This step is much more easily implemented using the RDF produced in step (2) than having to access the original database. It is only necessary to decide which repository management system will be adopted.

The proposed workflow architecture is presented in Figure 3. Recall that one of the main goals is to make CPDOC archive collections available as open linked data. This can be accomplished by providing data as RDF/OWL files for download and a SPARQL Endpoint [9] for queries. Since data evolve constantly, CPDOC teams would deliver periodical data releases. Besides the RDF/OWL files and the SPARQL Endpoint, we believe that it is also important to provide a lightweight and flexible web interface for final users to browse and query data. This can be easily done using a static website generator and Apache Solr ¹³ for advanced queries. As a modern index solution, Solr can provide much powerful and fast queries support when compared to traditional relational database systems. Note that the produced website, the RDF/OWL files and SPARQL Endpoints are complementary outputs and serve to different purpose and users.

Finally, it is vital to stress the main contrast between the new architecture and the current one. In the current CPDOC architecture the data is stored in relational databases and maintained by information systems. This means that

¹¹ It is out of this article scope to present the final format of these files.

¹² In this application we used Jekyll, <http://jekyllrb.com>, but any other static site generator could be used.

¹³ <http://lucene.apache.org/solr/>

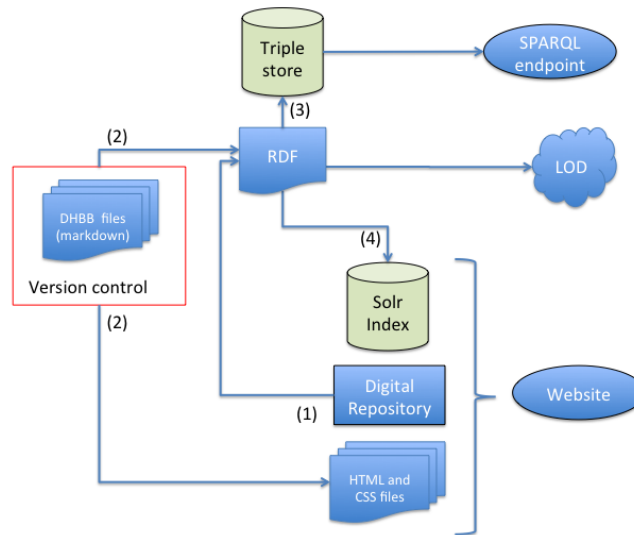


Fig. 3. The final architecture

any data modification or insertion is available in real time for CPDOC website users. However, this architecture has a lot of drawbacks as mentioned in Section 3, and also the nature of CPDOC data does not require continuous updates, which means that the cost of this synchronous *modus operandi* is not needed. Usually, CPDOC teams work on projects basis and therefore new collections, documents and metadata revisions are not very often released.

The results obtained so far encouraged us to propose a complete data model aligned with open linked data vocabularies, presented in detail in next section.

5 Improving Semantics

More than improving the current infrastructure for storing and accessing the CPDOC's data, we would like to exploit the semantic possibilities of such rich source of knowledge. One of the ways to do that is to embed knowledge from other sources by creating links within the available data. Since much of the data is related to people and resources with historical relevance, or historical events, some available ontologies and vocabularies can be used in this task.

The personal nature of the data allows us to use projects that are already well developed for describing relationships and bonds between people, such as FOAF [7] (Friend of a Friend) – a vocabulary which uses RDF to describe relationships between people and other people or things. FOAF permits intelligent agents to make sense of the thousands of connections people have with each other, their belongings and historical positions during life. This improves accessibility and generates more knowledge from the available data.

The analysis of structured data can automatically extract connections and, ultimately, knowledge. A good example is the use of PROV [17], which provides a vocabulary to interchange provenance information. This is interesting to gather information of data that can be structurally hidden in tables or tuples.

The RDF graph model enables also the merging of data content naturally. The DBpedia project, for instance, allows users to query relationships and properties associated with Wikipedia resources, and users can link other datasets to the DBpedia dataset in order to create a big and linked knowledge knowledge base. CPDOC could link their data to DBpedia and then make their own data available for a bigger audience.

In the same direction, the use of lexical databases, such as the WordNet [15] and its Brazilian version OpenWordnet-PT [12], will allow us to make natural language processing of DHBB entries. Named entities recognition and other NLP tasks can automatically create connections that improve dramatically the usability of the content. Other resources like YAGO [29] and BabelNet [27] links Wikipedia to WordNet. The result is an “encyclopedic dictionary” that provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations. Finally, the SUMO Ontology [28] could also be used to provide a complete formal definition of terms linked to WordNet. All of these lexical resources and ontologies will be further explored when we start the natural language processing of DHBB entries.

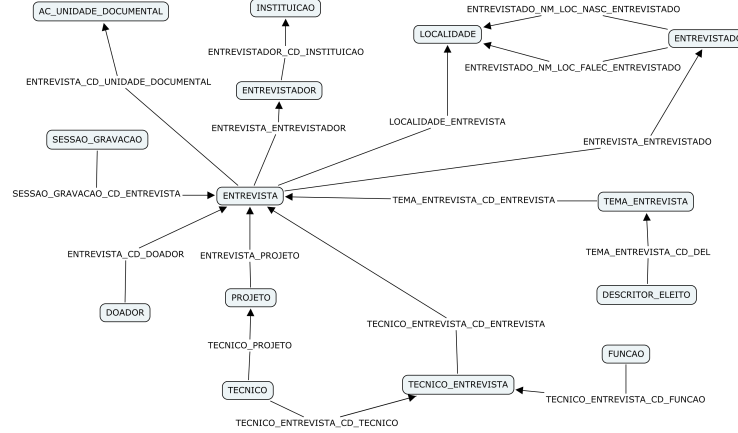


Fig. 4. PHO first RDF model

Figure 4 shows a fragment of the current RDF model produced by D2RQ (in step (1) of Figure 2) using the original CPDOC database relational model. This fragment shows only some PHO classes (derived from the tables) and some properties (derived from the foreign keys). Classes are written inside the boxes and properties are represented by the names in arrows that connect boxes.

The model presented in Figure 4 depicts that D2RQ was not able to automatically improve much further the model. D2RQ was able to correctly translate relations N:M in the relational model, such as `entrevista_entrevistador` (originally a table in the relational model) to a property that connect directly instances of `entrevista` (interview) with instances of `entrevistador` (interviewer). Nevertheless, the N:M relation between `entrevista` and `tecnico` (technician) was kept as an intermediary class called `tecnico_entrevista` due to the existence of an additional information in this N:M relation, the role (`funcao` class) of the interview technician. The relational model also seems to have some inconsistencies. Although the connection of technician and interview is parameterized by different roles, the donator, interviewer and interviewed of an interview are modeled each one in a specific table. In this case interviewed, interviewer, donator and technician are all people that share a lot of common properties like name, address, etc, and could be modeled as people. These problems are all result of a “ad hoc” modeling process. The model defined this way only makes sense for CPDOC team and it could hardly be useful outside CPDOC.

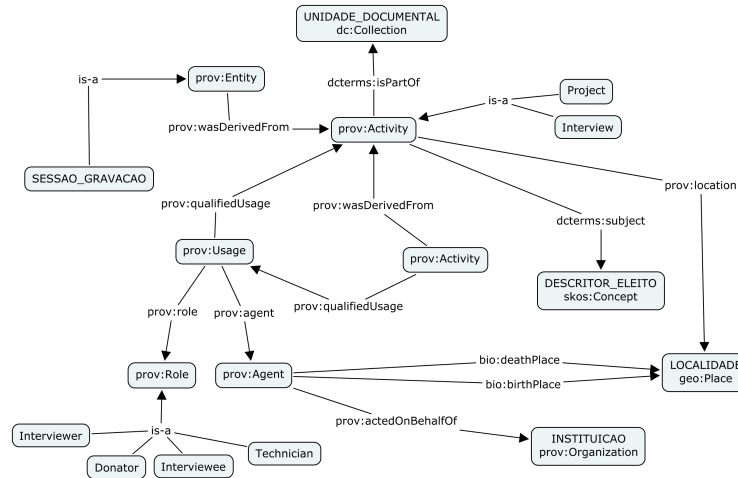


Fig. 5. PHO revised RDF model

Figure 5 shows how PHO model can be refined. The new model uses standard vocabularies and ontologies, making the whole model much more understandable and interoperable. In the Figure 5, `prov:Activity` was duplicated only for a better representation. The prefixes in the names indicate the vocabularies and ontologies used: `prov`, `skos`, `dcterms`, `dc`, `geo`, and `bio`. We also defined a CPDOC ontology that declares its own classes and specific ontology links, such as the one that states that a `foaf:Agent` is a `prov:Agent`. In this model, we see that some classes can be subclasses of standard classes (e.g. `Interview`), while some classes can be replaced by standard classes (e.g. `LOCALIDADE`).

6 Conclusion

In this paper we presented a new architecture for CPDOC archives creation and maintenance. It is based on open linked data concepts and open source methodologies and tools. We believe that even though CPDOC users would need to be trained to use the proposed tools such as text editors, version control software and command line scripts; this architecture would give more control and easiness for data maintenance. Moreover, the architecture allows knowledge to be easily merged to collections data without the dependency of database refactoring. This means that CPDOC team will be much less dependent from FGV's Information Management and Software Development Staff.

Many proposals of research concerning the use of lexical resources for reasoning in Portuguese using the data available in CPDOC are being carried out so as to improve the structure and quality of the DHBB entries. Moreover, the automatic extension of the mapping proposed in Section 5 can be defined following ideas of [8]. Due the lack of space, we do not present them in this paper.

Finally, we aim to engage a wider community and an open-source development process in order to make the project sustainable. As suggested by one of the reviewers, we must also learn from experiences of projects like Europeana¹⁴ and German National Digital Library [16].

References

1. Alzira Alves Abreu, Fernando Lattman-Weltman, and Christiane Jalles de Paula. *Dicionário Histórico-Biográfico Brasileiro pós-1930*. CPDOC/FGV, 3 edition, February 2010.
2. Oren Ben-Kiki, Clark Evans, and Ingy dot Net. Yaml: Yaml ain't markup language. <http://www.yaml.org/spec/1.2/spec.html>.
3. Michael K Bergman. White paper: the deep web: surfacing hidden value. *journal of electronic publishing*, 7(1), 2001.
4. Tim Berners-Lee. Relational databases on the semantic web. Technical report, W3C, 1998. <http://www.w3.org/DesignIssues/RDB-RDF.html>.
5. Christian Bizer and Richard Cyganiak. D2R server-publishing relational databases on the semantic web. In *5th international Semantic Web conference*, page 26, 2006. <http://d2rq.org>.
6. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics*, 7(3):154–165, 2009.
7. Dan Brickley and Libby Miller. Foaf vocabulary specification. <http://xmlns.com/foaf/spec/>, 2010.
8. Isabel Cafezeiro, Edward Hermann Haeusler, and Alexandre Rademaker. Ontology and context. In *IEEE International Conference on Pervasive Computing and Communications*, Los Alamitos, CA, USA, 2008. IEEE Computer Society.
9. Kendall Grant Clark, Lee Feigenbaum, and Elias Torres. SPARQL protocol for RDF. Technical report, W3C, 2008.

¹⁴ <http://www.europeana.eu>.

10. Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff. Definition of the CIDOC conceptual reference model. Technical Report 5.0.4, CIDOC CRM Special Interest Group (SIG), December 2011. <http://www.cidoc-crm.org/index.html>.
11. Richard Cyganiak, Chris Bizer, Jorg Garbers, Oliver Maresch, and Christian Becker. The D2RQ mapping language. <http://d2rq.org/d2rq-language>.
12. Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. Openwordnet-pt: An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics*, 2012.
13. Governo do Estado de São Paulo. Governo aberto sp. <http://www.governoaberto.sp.gov.br>, 2013.
14. David Eaves. The three law of open government data. <http://eaves.ca/2009/09/30/three-law-of-open-government-data/>.
15. Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.
16. Natalja Friesen, Hermann Josef Hill, Dennis Wegener, Martin Doerr, and Kai Stalman. Semantic-based retrieval of cultural heritage multimedia objects. *International Journal of Semantic Computing*, 06(03):315–327, 2012. <http://www.worldscientific.com/doi/abs/10.1142/S1793351X12400107>.
17. Yolanda Gil and Simon Miles. Prov model primer. <http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/>, 2013.
18. W3C OWL Working Group, editor. *OWL 2 Web Ontology Language Document Overview*. W3C Recommendation. World Wide Web Consortium, 2 edition, 2012.
19. John Gruber. Markdown language. <http://daringfireball.net/projects/markdown/>.
20. Dublin Core Initiative. Dublin core metadata element set. <http://dublincore.org/documents/dces/>, 2012.
21. Open Data Initiative. Open data initiative. <http://www.opendatainitiative.org>, 2013.
22. Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting. <http://www.openarchives.org/OAI/openarchivesprotocol.html>, 2008.
23. LexML. Rede de informação informativa e jurídica. <http://www.lexml.gov.br>, 2013.
24. Frank Manola and Eric Miller, editors. *RDF Primer*. W3C Recommendation. World Wide Web Consortium, February 2004.
25. James McGann. *The Think Tank Index*. Foreign Policy, February 2009.
26. Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. <http://www.w3.org/2004/02/skos/>, 2009.
27. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
28. Ian Niles and Adam Pease. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM, 2001.
29. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA, 2007. ACM Press.

Pundit: Creating, Exploring and Consuming Semantic Annotations

Marco Grassi^{a,1}, Christian Morbidoni^{b,1}, Michele Nucci^{c,1}, Simone Fonda^{d,2},
and Francesca Di Donato^{e,3}

¹ Samedia Group, Università Politecnica delle Marche, Italy
^am.grassi@univpm.it, ^bchristian.morbidoni@gmail.com, ^cm.nucci@univpm.it
<http://www.samedia.dibet.univpm.it/>

² NET7, Italy
^dfonda@netseven.it - <http://www.netseven.it>

³ Scuola Normale Superiore, Italy
^efrancesca.didonato@sns.it - <http://www.sns.it/>

Abstract. This paper presents Pundit, a novel semantic web annotation tool, and demonstrates its use in producing structured data out of users annotations. Pundit allows communities of scholars to produce machine-readable annotations that can be made public and thus consumable as web data via SPARQL and ad-hoc REST APIs. Pundit is highly configurable and can be deployed in custom instances to include well-defined and agreed annotation vocabularies. Such instances can be distributed as bookmarklets to community users so they can create uniformly structured data in a certain application scenario. Basing on the provided APIs, some demonstrative applications have been developed, exploring different use scenarios, ranging from philosophy to journalism and cultural heritage. The main aim of this paper is to demonstrate how such uniformly structured annotations can be quickly re-used on the web to make information discoverable or to visualize it in interesting ways.

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

1 Introduction

Annotation is a primary activity for scholars and professionals. It consists in enriching a content with some new information, which possibly helps in understanding or searching the content itself. While until few decades ago annotations were sketched by hand on the side of a book, today web technologies have the potential to make them infinitely replicable, remotely accessible and easy to share. Web annotations systems and bookmarking/clipping tools are popular nowadays both among generic users (e.g. social tagging) and among scholarly communities (e.g. Zotero⁴, Mendeley⁵ users). However, existing annotation systems are generally limited to textual comments, tags or predefined metadata

⁴ <http://www.zotero.org/>

⁵ <http://www.mendeley.com/>

templates (e.g. bibliographic records). Furthermore, annotations are often isolated into closed systems and very rarely are connected to the Web of Data. The simple idea behind our work is that of making of annotations a vehicle to create new semantic web data, actually adding links and, ultimately, knowledge to the so called Global Data Space [1]. Once annotations become available in a standard and highly expressive form, a variety of applications can be built to visualize the resulting knowledge in specific domains. Pundit is a novel annotation system that aims at implementing this vision, by enabling annotators (e.g. scholars) to use semantically specified relations and link to web of data entities, producing in fact accessible RDF graphs out of their work. Such RDF graphs are collections of annotations that we call “notebooks”. Notebooks can be consumed via REST APIs or standard SPARQL endpoints. In this paper we first overview Pundit at a high level, then we focus on the issue of effectively re-using the annotations produced in Pundit to drive demonstrative use cases and address end-user needs such as sharing, exploring and visualizing annotations. Two main directions are currently being targeted. In Ask⁶, we attempt at creating a portal to manage annotations, share them and explore public notebooks. We then explore, by means of some demonstrative developments, the possibility of basing on the Pundit “framework” to build vertical, specialized applications. In the latter case, the basic pattern we follow is that of configuring and deploying custom instances of Pundit, which can be distributed among users. Such instances generate annotations that conform to pre-defined data schemas and can be quickly fed into existing open-source tools to produce more interesting visualizations. Nevertheless they maintain the generality and flexibility of RDF, thus being compatible with Ask or other “general purpose” usages of data.

2 Related Works

An exhaustive state of the art in semantic annotation goes beyond the purpose of this paper and can be found in the literature [3] and this section focus only on tools related to our work. The semantic tagging paradigm, which exploits publicly available Linked Data sources to retrieve unambiguous tags, has been implemented in Faviki⁷ and Europeana Connect Media Annotation Prototype (ECMAP)[4]. Other tools such as One click annotation [5], CWRC-Writer [6] and LORE (Literature Object Reuse and Exchange) [7] also allow the use of restricted vocabularies or ontologies in the annotations. Some annotations tools, as LORE and CWRC-Writer enable also the editing of more expressive annotations in the form of subject-predicate-object statements. Although not based on Semantic technologies and not supporting semantic annotations, Open Knowledge Foundation (OKFN) Annotator⁸ has been conceived as a JavaScript library that can be added to any Web page, both adding it into HTML and injecting it using a bookmarklet, to make it annotatable, similarly to Pundit.

⁶ <http://ask.as.thepund.it>

⁷ <http://www.faviki.com>

⁸ <http://okfnlabs.org/annotator/>



Fig. 1: Creating annotations with Pundit

3 Pundit overview

Annotations in Pundit are essentially triples that connects different kinds of items together. A triple has the form [subject - predicate - object], where the subject and object can be segments of text and images (e.g. [text - describes - image]_i) or entities from the web of data (e.g. [text - has author - Dante(from Freebase.com)] or [image - depicts - Florence(from DBPedia.org)]). The most expressive annotation interface provided by the Pundit client is the “triple composer”. It allows users to drag and drop items into triples, or select them from the web page (e.g. by selecting a text or an image), as well as searching into available vocabularies and data sources. However, other annotation wizards support specific kind of annotations, as putting two segments of text in relation, or attaching tags and comments to a text segment. Image annotation of a segments of images is supported by a dedicated module as shown in 1. The Pundit client is a JavaScript application that can be deployed as a library, to then be easily included in existing web sites to make the content “annotable”⁹, as well as delivered as a bookmarklet. A bookmarklet is a simple link (bookmark) that, once added to a web browser allows loading Pundit on every web page and annotating its content.

In Pundit, an annotation contains information at a twofold level. The first one is the “annotation metadata” and deals with the act of annotating, including information on the author, the time of creation and the involved web resources. Pundit bases on the Open Annotation data model (OA)¹⁰ [8] for representing this dimension. The second, the “annotation graph”, is an RDF graph resulting from metadata and relations among web resources that a user has created by annotating. In other words, it captures the semantics of the annotation representing the user’s contribution in terms of “domain knowledge”. For example, an annotation graph could contain Wikipedia pages corresponding to Italian writers and relevant text segments from their works on wikisource.org or other open web archives, perhaps linking each text to a number of other texts from relevant contemporary writers. We call “items” the nodes of such a graph, which represent

⁹ this has been done in wittgensteinsource.org

¹⁰ Open Annotation core specification: <http://www.openannotation.org/spec/core/>

the annotated web resources, being them web pages segments or other kind of entities (places, persons, etc.) While no restrictions are applied and no assumptions are made by Pundit regarding the ontologies used in annotation graphs, a certain knowledge of the structure of the single annotations into a notebook has to be owned by a developer to implement a meaningful visualization based on such “free shaped” data.

So far, one of the most successful approaches to foster the reuse of data on the web is to create a consensus around vocabularies and ontologies within a certain community. In Pundit we try to follow this pattern by make it possible to deploy customized annotation clients, in the form of JavaScript libraries or bookmarklets, which can be distributed to users by “community leaders”. A custom client possibly includes a precise set of a well-defined set of “relations” to be used in annotations to create typed links among items or taxonomies where relevant web entities are collected and ready to be annotated. Both taxonomies and relations are represented in JSON and can be easily extracted from existing vocabularies (e.g. SKOS) or ontologies, as we did in the Wittgenstein’s brown book pilot¹¹. Aggregating items in collection, for sharing and publishing, is a common pattern in social clipping and bookmarking tools. In Pundit, annotations are collected in notebooks that users can optionally make publicly accessible. When a notebook is public, the annotations contained in it are not only shown in the Pundit client (e.g. when a user loads the Pundit bookmarklet on one of the annotated web pages) but, more interestingly, a notebook can be consumed by means of open REST APIs and accessed by a variety of web applications. Each notebook provides a SPARQL endpoint to query its content. In other words, a notebook is an independent RDF graph created by a given user in time and connecting a variety of web resources.

4 Consuming annotations

Regarding how to use annotations, and the semantic data they enclose, to drive end-user applications, there are mainly two “dimensions” that can be explored:

- Annotation centric approach. This is commonly used in clipping systems where each clip is the result of a single annotation and is shown as a “box” containing some information (e.g. pictures, links, tags) about the annotated items. We mainly based on this approach in designing Ask, a prototype web application to search over public notebooks and manage personal ones.
- Item centric approach. As annotations graphs in a notebook can be in fact consumed as a unique and bigger RDF graph, a possible way of looking at the data is that of focusing the visualization on the annotated items and their relations with other items. This approach clearly benefits from an a-priori knowledge on ontologies and custom vocabularies used in annotations, as it needs to take into account the nature of the information and deals with the “meaning” of annotations.

¹¹ DM2E blog, Wittgenstein Brown Book experiment, <http://dm2e.eu/dm2e-to-start-work-on-wittgensteins-brown-book/>

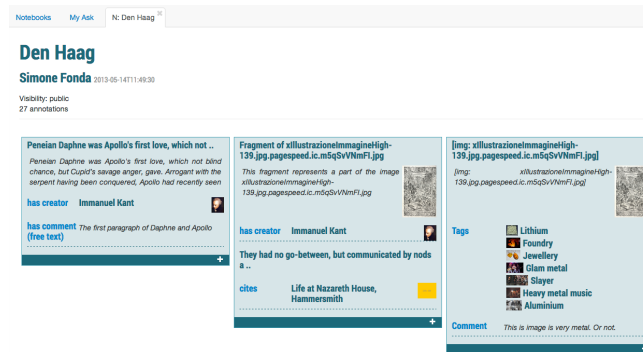


Fig. 2: Exploring notebooks with Ask.

Ask¹² is a web application where public notebooks stored in Pundit can be searched and explored. At the time of writing a new version of the tool is being released. Personal notebooks are accessible to their owner and can be made public or kept private. By default, Ask provides a general purpose visualization of notebooks where single annotations are shown as “metadata boxes”. However, alternative visualizations (such as the one described in the following sections) can be easily plugged by providing a compliant REST API. Ask is currently subject to intense development, and one of the most interesting recent features is the prototypal faceted browser available in alpha version¹³.

4.1 Edgemaps Visualization: A Demonstrative Use Case.

Edgemaps [10] is an open-source web tool that drives an interesting visualization demonstration in the field of philosophy¹⁴. The graph is generated by Freebase.com data, which includes “influences” slot in the description of authors. While for a “generic” user such a visualization is enough, we cant probably say the same for scholars that consider such relations as a matter of study and might probably ask: “Why exactly do you say that Marx influences Gramsci?”, “What is the evidence of that in the actual primary sources?”, “Who said that?”. Structured annotations in conjunction with online open content as the one provided by Wikisource¹⁵ make it relatively easy to bring the philosophers demo a little further: generating the graph from scholars annotations made on primary sources (thus including the evidence of the connections), rather than from centralized data. We did so proving an opportunely tuned instance of Pundit and extending, with little programming effort, the Edgempas code. The demonstration is documented on the web site¹⁶. The Pundit instance uses relations picked from

¹² <http://ask.as.thepund.it>

¹³ <http://demo.ask.thepund.it>

¹⁴ <http://mariandoerk.de/edgemaps/>

¹⁵ <http://wikisource.org>

¹⁶ <http://www.thepund.it/visualization-demos/philosophers-demo-howto/>

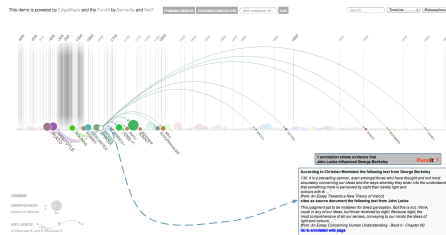


Fig. 3: Showing evidences of philosopher influence with a Edgemap Visualization

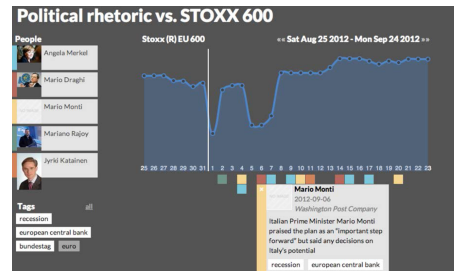


Fig. 4: Political Rhetoric vs. STOXX 600

the CiTO ontology¹⁷ and includes predicates like “cites” and “quotes”, as well as other more specific ones like “discusses”, “cites as sources”, “agrees with”, etc. Each time two philosophers are connected by an “influenced by” relation, the corresponding annotations are shown so that the scholar can immediately get an evidence of “why the relation is there”. It is also possible to load multiple notebooks from different scholars, thus in fact enabling a collaborative scenario, where annotation authorship is always tracked back and each user can decide what notebook to see or trust.

4.2 Data Journalism

The same pattern can be applied to several contexts and to address very diverse use cases, as economics or journalism. The data journalism demonstrative application shows the use of annotations, in this case quotations from politicians and public persons taken from online news papers, to produce dynamic graphics. A Pundit bookmarklet has been deployed containing a small set of relations (or properties) to tag, describe and date in time politicians’ declarations. The associated visual tool has been developed in JavaScript and provided as a web API, which gets a notebook (id) as argument and builds a timeline where annotated declarations are shown along with the trend of a financial indicator. The idea is that of creating a tool for journalists to demonstrate and reveal possible existing connections among what important persons says and how the market behaves. (Fig. 4). A live demo can be found online¹⁸.

4.3 Tracking Annotated Resources Over Time

Timeline visualization has become a common practice for showing data containing time-related information and several tools already exist that allows creating such type of visualization. Instead of developing another one, Pundit reuses TimelineJS¹⁹. This is an example of the advantages of decoupling annotation

¹⁷ CiTO Ontology: <http://purl.org/spar/cito/>

¹⁸ Journalism demo, <http://ask.thepund.it/?#/timeline/31951d93/20120927>

¹⁹ TimelineJS: www.timeline.verite.co

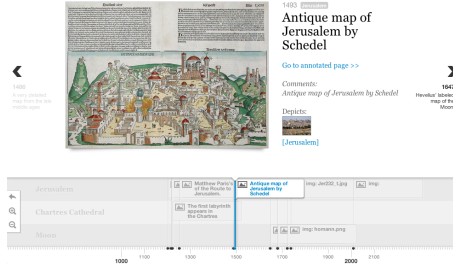


Fig. 5: Timeline visualization

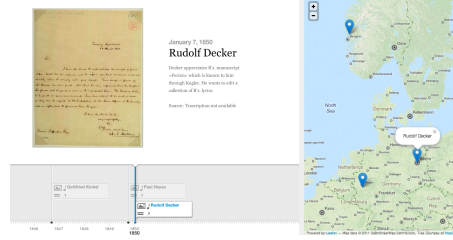


Fig. 6: Tracking annotated resources position over time and space.

creation and consumption. A pundit API has been created that allows to extract time-related information from a notebook given its id and to convert them in a TimelineJS compliant JSON to feed the timeline, as shown in Fig. 5. Annotations in the notebook simply need to have date information, i.e. they have to contain triples having as subject a text fragment or an image, as predicate “dates to” (for a date) or “start date” and “end date” (for a period in time) and as object a date.

A similar approach has been used in another experiment conducted in the context of the Burckhardtsource.org platform that aims at mapping and producing a critical edition of the extensive correspondence of 400 European intellectuals with Jacob Burckhardt over a period of more than half a century from 1842 to 1897.²⁰ [11]. In this case study, the resources of interest were of three types: Persons, Places and Works of art. Freebase has been used data source for such resources as it contained already several of them. In line with the principle of contributing to the Web of Data, rather than only consuming it, missing resources have been added to Freebase. As a result, at the time of writing this paper, scholars have added several hundreds of new entries to Freebase.org, providing basic metadata and descriptions. Pundit has been configured to use a simple set of properties, to cover the different relations that can occur among resources. These relations allow explicitly relating dates, places and persons with text in the letters. The Timeliner open-source tool²¹ has been used to show dynamic visualizations built from the corpus of annotations that scholars created so far, mainly about places and persons, see Fig. 6. The visualization shows the letters in a timeline, based on their sending date. It also graphs all of the mentioned places and persons on a map, where person location is determined by their birthplace and their movements can be tracked over time.

5 Conclusions

In this paper we presented preliminary results in leveraging on structured semantic annotations to create interactions and visualization of collaboratively

²⁰ www.burckhardtsource.org

²¹ <http://timeliner.okfnlabs.org/>

created data. In our examples we used Pundit: a customizable and flexible semantic web annotation tool. In deploying the tool for different use scenarios, we highlighted a simple pattern consisting of developing custom vocabularies, perhaps aggregating existing data, distributing a simple tool to annotate web resources of interest and, finally building on third party applications to consume the generated information and address specific data visualization needs.

6 Acknowledgments

The research activity underlying this work is being partially funded by the European Union's Seventh Framework Programme managed by REA-Research Executive Agency²² ([FP7/2007-2013][FP7/2007-2011]) under grant agreement n. 262301, and by the GramsciSource project of the MIUR, FIRB 2012, p. RBFR12MZ8R.003. Pundit was originally developed in the Semlib project²³.

References

1. C. Bizer, T. Heath, "Linked Data: Evolving the Web into a Global Data Space", <http://linkeddatabook.com/editions/1.0/>
2. 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.
3. 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>
4. 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.
5. M. L. Ralf Heese, "One Click Annotation" in 6th Workshop on Scripting and Development for the Semantic Web, 2010.
6. 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
7. A. Gerber and J. Hunter, "Authoring, Editing and Visualizing Compound Objects for Literary Scholarship", Journal of Digital Information, vol. 11, 2010.
8. "Open Annotation: Alpha3 Data Model Guide" 15 October 2010 Eds. R. Sanderson and H. Van de Sompel. <http://www.openannotation.org/spec/alpha3/>
9. M. Grassi, C. Morbidoni, M. Nucci, S. Fonda, G. Ledda. "Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries". Proceedings of the Second International Workshop on Semantic Digital Archives (SDA 2012).
10. M. Dörk, S. Carpendale, C. Williamson, "EdgeMaps: Visualizing Explicit and Implicit Relations". Proceedings of VDA 2011: Conference on Visualization and Data Analysis, IS&T/SPIE. 2011
11. F. Di Donato. "Working on scholarly contents: A semantic vision". In Proceedings of Open Platforms for Digital Humanities, 17-18 January 2013, 2013.

²² <http://ec.europa.eu/research/rea>, DM2E Project: <http://dm2e.eu/>

²³ <http://www.semllibproject.eu/>, SemLib EU project

Towards Persistent Identification of Resources in Personal Information Management

Stefan Haun and Andreas Nürnberger

Data and Knowledge Engineering Group,
Faculty of Computer Science,
Otto-von-Guericke-University Magdeburg, Germany
<http://www.findke.ovgu.de>

Abstract. Persistent identification is necessary for recognition, dissemination and (external) cross-references to digital objects. *Uniform Resource Identifiers (URIs)* provide an established scheme for this task, but do not guarantee stable and persistent identification. In the context of (personal) archives, stability is needed when references are stored on a medium where later changes to identifiers cannot be corrected at all or only with a very large overhead, such as WORM media or tape archives. Additionally, resources like contacts or appointments do not have a URI, while other URIs, such as file system paths or the IMAP URI, are unstable by design and cannot represent the dynamic aspects of Personal Information Management (PIM). This paper discusses problems of archiving that arise with entity identification in PIM, especially on the example of the personal file system.

1 Introduction

During the past decades, our everyday work-life has rapidly moved from the real world to the digital domain. Then we used to put our hands on paper files, books, business cards, calendars or photographs, nowadays everything is a digital object stored on the computer, in the smart phone or the Cloud. Still we need ways for virtually grabbing and referencing those objects.

The *Uniform Resource Identifier* (URI) serves exactly this need: to identify digital resources for later look-up and to reference objects. However, there is only a very generic specification for how a URI must look like. This results in high flexibility and viability for many applications, but URIs may not match the requirements posed by the task. For example, the proposed URI for a file path is unstable in the way that those URIs become invalid whenever the file is moved. Therefore external references in objects no longer point at that file and have become stale and thus useless.

Even worse, if these objects are archived, changes to the stored data are often prohibited either by policy or due to the storage medium. For example, A CD-ROM is read-only and re-writing a tape storage takes much effort. Therefore identifiers need to be stable so that existing archives are not broken.

Having reliable identifiers for digital objects is a requirement to mine and store relationships between those entities, especially in dynamic environments such as Personal Information Management. With an integrated view novel interaction concepts, such as graph based interaction, are possible. These approaches enable the user to explore large graphs—such as the graph emerging from connected objects in Personal Information Management—in order to find specific elements without the need for keyword-based search or to get an overview on structures in the personal information space (e.g. [4]).

The paper starts with a short analysis of related work and a definition followed by an elaboration of requirements and problems with stable identification systems. The findings are further discussed on the example of a personal file system to point out some problems that arise with current systems.

2 Related Work

Entity identification has uses in several areas: In geometry, entity identification is a known problem when objects change, but reference points must be recognized [16]. In the context of the *World Wide Web*, identification refers to digital resources. The W3C¹ creates and maintains standards related to the Internet and especially the WWW, which are highly relevant to this paper. In the *Persisting Identifier Linking Infrastructure (PILIN) project* options for identification of public entities and necessary infrastructures are researched [9]. Semantic annotation of data, for example in linked-data sets, induces the problem of finding feasible identifiers in a specific domain, such as the files [12] or biological data [10]. Analysis of stability and reliability of digital identifiers can be found in the field of *digital forensics*, e.g. a survey of Message-IDs in e-mails [11].

No related work could be found towards building stable identifiers in Personal Information Management and regarding personal archives, hence this paper aims at starting a discussion in this direction by identifying some general problems based on examples of stable URIs and personal file systems.

3 Persistent Identifiers

3.1 Identifiers

A general definition for the term *identifier* can be found in [9]:

Any association of a name with a thing – by anyone – establishes an identifier. A name is not an identifier unless it identifies something. (E.g. an unassigned phone number is a name, but not an identifier.)

Digital identifiers exist in contexts which are seldom made explicit or included when citing an identifier.

¹ World Wide Web Consortium (W3C) <http://www.w3.org/>

RFC 3986 [1] specifies the *Uniform Resource Identifier (URI)* as “a compact sequence of characters that identifies an abstract or physical resource”. URIs can be divided into *Uniform Resource Names (URN)*, denoting the name of a specific resource, and *Uniform Resource Locators (URL)*, referring to the location of a resource. However, a formal distinction is often neglected based on the findings from RFC 3305 [5]. Throughout this paper, the term *URI* will be used whenever a further distinction is not necessary. Yet it is important to notice that an identifier in terms of a URI, while referencing a resource, not necessarily has to point at this resource in the sense of a URL.

When looking for an identifier, the URI makes a good choice. Sharing the syntax with URLs, any entity—real, abstract or on the Worldwide Web—can be identified without introducing a new standard.

3.2 Requirements

The URI specification defines requirements regarding syntax and semantics of a URI. However, it is agnostic regarding requirements for a system using the URI. This leads to high flexibility, but demands further analysis in the context of an application using the URI. In [7, 13] the following functional requirements for resource identification systems are defined:

Global scope Identifiers are location-independent and have the same meaning everywhere.

Global uniqueness Identifiers are unique, i.e. one identifier does not reference information associated with multiple resources.

Persistence Identifiers uniquely reference resources beyond their lifetime. This specifically means that an identifier must not decay when the resource is no longer available and, in a broader interpretation, that an identifier must not be re-used for a resource once it has been used.

Scalability Identifiers are scalable and can be assigned to any resource.

Legacy support Identifiers can support legacy identification schemes to the extent that these satisfy minimum requirements.

Extensibility Identification schemes can accommodate future extensions.

Independence Responsible authorities maintain and assign resource identifiers within a given system.

Resolution Identifiers are supported by services that enable their translation.

The W3C demands **opacity** for all identifiers [3]:

Axiom: Opacity of URIs The only thing you can use an identifier for is to refer to an object. When you are not dereferencing, you should not look at the contents of the URI string to gain other information.

URI opacity has been a source of many debates, especially since technical solutions like the GET method for submitting HTML form content explicitly violates this rule.

A more ontological approach towards identifier requirements can be found in [9]. However, the discussion goes beyond the scope of this paper.

3.3 Resolution and Retrieval

In order to obtain the resource denoted by an identifier, the identifier must be *resolved* to a locator which allows to *retrieve* the resource [9]. According to [7] there are basically two ways of implementing persistent identifier management systems:

First, identifier schemes based on *Uniform Resource Names (URN)* as defined in RFC 2141 [6]. Each URN contains a globally unique part, denoting the namespace, and a namespace specific string, containing the actual reference. The reference data is a complete set of all attributes necessary to identify the referenced object. For example, the URN `urn:ISSN:0259-000X` references the journal with ISSN *0259-000X*. With this information, a copy can be obtained from the local library or a respective web service. As the URN does not contain any information about the storage of the referenced entity, changes to the storage paradigm do not render the identifier invalid.

A second option are handle-based systems: Instead of encoding the information needed to find an entity into the identifier, a handle to a database entry in a resolver is created. In order to retrieve the entity, i.e. *resolve* the identifier, a look-up call to the resolver is necessary. The identifier is then transformed in either the resource itself or a different identifier which can be used to retrieve the entity. Among many, a well-known handle-based example is the *DOI[®] System*, maintained by the *International Digital Object Identifier Foundation (IDF)*². The IDF hands out unique prefixed for publishers, who create unique identifiers in their own namespace. Thus each digital document, such as publications in conference proceedings, can be referenced without knowing its actual storage location. The DOI resolver maps a certain DOI to a URI in the database, which can be used to retrieve the document. In this case, the resolution consists of multiple steps, as the IDF does not store the documents themselves but identifier for the storage locations. Other known handle systems are *Persistent Uniform Location Locators (PURL)*³ and the *Archival Resource Key*⁴. Please refer to [15] for further information on those systems.

Many systems are handle based and when the task of resource identification arises, handles and central registries seem to be the preferred choice. In the context of Internet connectivity and centralized systems this makes sense: Having a control entity avoids naming clashes and makes it simple to decouple identifier and resource location. However, a handle based system also requires the availability of a handle database. In the case of personal archives, it may be necessary to store the resolution database alongside with the archive. This may be possible for a complete snapshot with stable references, but if the references may still change, the archived database will be outdated and if only parts of the personal information is archived, the resolution database must either be split or very likely results in a large storage overhead. For an offline scenario or when data is only available locally, e.g. on the personal computer, a handle

² <http://www.doi.org>

³ <http://purl.oclc.org/>

⁴ <https://confluence.ucop.edu/display/Curation/ARK>

database may not be feasible at all. Here the information needed for resolution must become a part of the identifier itself. There are URI specifications for many identifiers in the PIM context, such as file paths or e-mails, but these often do not take stability into account.

The following case study shows problems that may arise when identifying resources in the personal file system.

4 Case Study: Personal File System

Hierarchical file systems are still the main way of storing information on a personal computer. PIM systems have to provide means of identifying those files in order to create stable references.

Several systems have been devised towards file identifications, all with their own strengths and caveats. A small selection shall be discussed to give an overview on the problem field.

File Path A naive approach is given by the file path, i.e. the location of the file in the local file system. RFC 1738 [2] specifies in Section 3.10 the `file:` URI scheme “to designate files accessible on a particular host computer.” The solution has several caveats: First, path names are only *unique* in the scope of the host. If a reference to the file is transferred to another host, the context is lost and it becomes invalid. The uniqueness problem can be solved by adding user and host to the file path. While the URI scheme supports this full qualification of file paths, most computers do not have a globally unique host name, i.e. a name in the scheme `hostname.domain` anymore. The second caveat is much more pressing: Changing the location of a file—which is denoted by the file path—is a common and intended interaction with file objects, i.e. everyday interaction with files will result in broken links. The file URI scheme is not stable in the sense that it becomes invalid as soon as a file is moved or renamed, even though the file still exists. Using the file path to reference a file is only feasible if the file cannot be moved, which is the case for certain system files.⁵ Other means of identification are needed to reference files on a computer, than the paths built-in into modern operating systems.

Magnet Links A widespread solution for referencing files based on their content is provided by *magnet links*⁶. Although not listed as a standard, the `magnet:` prefix can be found in the list of URI schemes⁷. Instead of a file path, the URIs are generated from a hash code computed from the binary content of the file, such as SHA-1 [8]. To resolve the URIs, a mapper must keep track of file paths matching the given hash. When a file is moved or renamed, i.e. the file path changes, the mapper database must be updated accordingly. Magnet-link based

⁵ For example, the *Filesystem Hierarchy Standard* defines a quite rigid structure for *Unix*-based systems. (<http://www.pathname.com/fhs/pub/fhs-2.3.html>)

⁶ <http://magnet-uri.sourceforge.net/>

⁷ <http://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml>

URIs are not very common on local file systems for two reasons: First, they have a rather negative reputation, as these identifiers are mostly used in P2P file sharing. Here they serve the purpose very well, as they identify a certain content without knowledge of its location. Second, magnet links become invalid if the file content changes—another intended use of files in Personal Information Management.

Heuristics When both file path and file content change, heuristics must be used to identify a file. The *GIT version control system*⁸ tracks changes in the file path by comparing the content of edited files, i.e. if a file is removed and a new file with sufficiently similar content appears, the change is logged as “renamed” instead of “new”. Similar, changes in the file path could be detected by monitoring the file system operations, which is supported by common operating systems. Methods from *duplicate detection* can be employed to restore links by providing candidates to replace a broken file reference. Spinellis [14] presents an approach that augments file URIs with data from index vectors, so that broken links can be restored based on previously indexed content, i.e. a search query that will most likely recover the file is attached to the file path. Those approaches cannot guarantee stable references, as the changes cannot be reflected in the generated URI. However, they provide means of recovering a broken URI and represent a step towards stable file links.

Version Control Systems In an assumed system where content cannot be deleted but only updated, such as found in a version control system, the problem boils down to two distinct cases:

1. The generated identifier references a certain version of a file, i.e. a stable content.
2. The generated identifier references a certain path of a file, i.e. *move* or *rename* operations will not be applied.

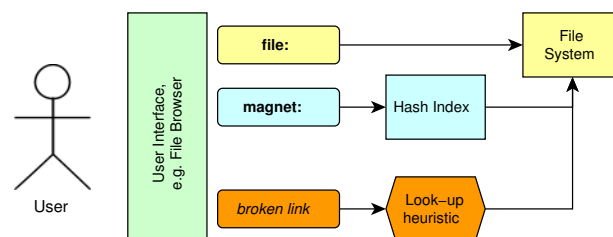


Fig. 1. Schematics of a file resolver combining several approaches

⁸ <http://git-scm.com/>

To create a stable identifier the intended use of a file must be known to select the right approach. Figure 1 shows how the described components could work together to resolve file links. In current operating system this leads to a semantic gap between the available data and the user's intention. It is left to future research to determine if and how this gap can be closed, e.g. based on the file type, by observation of user interaction with the file or by taking the file's provenance into account.

5 Conclusion

In this paper we have discussed some fundamental problems that arise when persistent identifiers for objects from the PIM domain are needed. Requirements and modeling of identifiers cover a growing research field, however, PIM solutions still rely on locally generated identifiers instead of stable, global URIs. In a case study some problems with persistent identifiers have been shown: Stable file links require more attention and may even lead to a semantic gap where the system cannot decide how to generate the correct link.

Future work includes research on remaining PIM entities, specifically e-mails, contacts and appointments as well as further tests on implementation of these identifiers and tests towards reliability in different data sets. The file case study has shown that reliable persistent identification will need methods from machine learning to repair broken links and determine the intended use of an object. Semantic archives may provide the necessary information for re-assigning an object to its previous link if a connection has been lost, e.g. by comparing meta-information. Finally, when persistent identification is possible, user interfaces and semantic desktop applications can make use of those identifiers.

Acknowledgments

Parts of this paper have been researched within the scope of the SENSE project which is funded by the Federal Ministry of Education and Research, German Aerospace Center. It is part of the *KMU-Innovativ: IKT* campaign and goes by the funding numbers FKZ 01IS11025E.

References

1. Berners-Lee, T., Fielding, R., Masinter, L.: Uniform Resource Identifier (URI): Generic Syntax. RFC 3986 (January 2005), <http://tools.ietf.org/html/rfc3986>
2. Berners-Lee, T., Masinter, L., McCahill, M.: Uniform Resource Locators (URL). RFC 1738 (December 1994), <http://tools.ietf.org/html/rfc1738>
3. Berners-Lee, T.: Design Issues, chap. Universal Resource Identifiers – Axioms of Web Architecture (1996), <http://www.w3.org/DesignIssues/Axioms.html>
4. Haun, S., Gossen, T., Nürnberger, A., Kötter, T., Thiel, K., Berthold, M.: On the Integration of Graph Exploration and Data Analysis: The Creative Exploration Toolkit, Lecture Notes in Computer Science, vol. 7250, pp. 301–312. Springer Berlin Heidelberg (2012)

5. Mealling, M., Denenberg, R.: Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations. RFC 3305 (August 2002), <http://tools.ietf.org/html/rfc3305>
6. Moats, R.: URN Syntax. RFC 2141 (May 1997), <http://tools.ietf.org/html/rfc2141>
7. Morgan, H.: Persistent Identification of Digital Resources – Environmental Scan. Tech. rep. (2008)
8. National Institute of Standards and Technology, USA: FIPS PUB 180-4: Secure Hash Standard. Federal Information Processing Standards Publication (March 2012), <http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf>
9. Nicholas, N., Ward, N., Blinco, K.: Abstract Modelling of Digital Identifiers. Ariadne issue 62 (January 2010), <http://www.ariadne.ac.uk/issue62/nicholas-et-al>
10. Pasquier, C.: Biological data integration using Semantic Web technologies. *Biochimie* 90(4), 584–594 (April 2008)
11. Pasupatheeswaran, S.: Email 'Message-IDs' helpful for forensic analysis? In: Proceedings of the 6th Australian Digital Forensics Conference. School of Computer and Information Science, Edith Cowan University, Perth, Western Australia (2008)
12. Schandl, B., Popitsch, N.: Lifting File Systems into the Linked Data Cloud with TripFS. In: Bizer, C., Heath, T., Berners-Lee, T., Hausenblas, M. (eds.) LDOW. CEUR Workshop Proceedings, vol. 628. CEUR-WS.org (2010), <http://dblp.uni-trier.de/db/conf/www/ldow2010.html#SchandlP10>
13. Sollins, K., Masinter, L.: Functional Requirements for Uniform Resource Names. RFC 1737 (December 1994), <http://tools.ietf.org/html/rfc1737>
14. Spinellis, D.: Index-based persistent document identifiers. *Inf. Retr.* 8(1), 5–24 (Jan 2005)
15. Tonkin, E.: Persistent Identifiers: Considering the Options. Ariadne issue 56 (July 2008), <http://www.ariadne.ac.uk/issue56/tonkin>
16. Wang, Y., Nnaji, B.O.: Geometry-based semantic ID for persistent and interoperable reference in feature-based parametric modeling. *Comput. Aided Des.* 37(10), 1081–1093 (Sep 2005)

A Storage Ontology for Hierarchical Storage Management Systems

Sandro Schmidt, Torsten Wauer, Ronny Fritzsche, and Klaus Meißner

University of Dresden, 01062 Dresden, Germany
{sandro.schmidt, torsten.wauer, ronny.fritzsche,
klaus.meissner}@tu-dresden.de

Abstract. The increasing capacity of storage media could store a huge amount of data while the price per Gigabyte is decreasing. On the other hand users and companies produce much more information that still overruns the available storage capacities. As a consequence, the management and storage of this data is very complex and expensive. Users and apps want to access their files directly and without any time delay, resulting in using fast but very expensive storage media. A deeper look on the usage of data, especially in companies, shows that only a very small amount of it is used every day. Summarizing these facts, a concept is needed to find out which data is important to provide them on fastest storage media, and less important one on cheapest storage media. Concepts derived from the Semantic Desktop can be a solution. We introduce a concept to describe files by an ontology. This allows for the description of their relations to each other, as well as their attributes. The resulting ontology offers an extensive volume of data that helps to find the adequate storage conditions for every single file. Another great advantage that showed up, is the independence from file system accesses to gather information about the stored files.

Keywords: Semantic Storage, Ontology, Hierarchical Storage Management

1 Introduction

As Gantz described in his IDC White Paper [4], there is a huge amount of information that overruns the available amount of storage capacity. That does not mean that all of this information represented as files is important during the whole lifetime. Studies showed that only 1 % of all files stored in a file system with 200,000 files are modified daily [15]. Furthermore, some data are more important, oriented on their content than others, or some should only be stored on special storage media. Some files are duplicates or redundant. As a consequence, only

Acknowledgments - Parts of this paper have been researched within the scope of the SENSE project which is funded by the Federal Ministry of Education and Research, German Aerospace Center. It is part of the *KMU-Innovativ: IKT* campaign and goes by the funding number FKZ 01IS11025D.

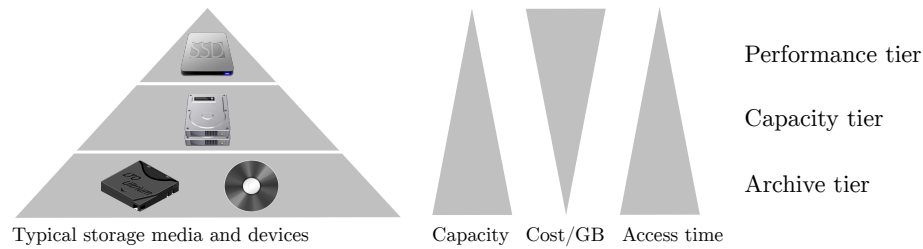


Fig. 1. Comparison of the three typical tiers of a HSM

a fraction of the stored files in a system needs to be on fast and, therefore, expensive storage media like the newest solid state drives to provide them in a performant way to users or programs.

Hierarchical Storage Management (HSM) Systems solve this problem by storing files on the most efficient storage media respective to the importance of these files. Efficient means the balance between the access frequency, the time it needs to load files from slower storage media, and how expensive the storage is. Thus, files that have a high importance because, e. g. , they were accessed a lot, should be accessed very fast by users or programs. Therefore, it is necessary to store these files on media with low access time and high data rate. Conversely files that are used rarely and, therefore, seems to be needed less often in future, should be stored on cheap media.

As shown in Fig. 1, a classical HSM System is divided into three tiers: Performance Tier (PT), Capacity Tier (CT) and Archive Tier (AT) [12, 6]. The PT has the fastest but most expensive storage media (e. g. , solid state drives), while the AT utilizes the cheapest but slowest storage media like Tape Libraries (Fig. 1). The CT acts as a compromise between cost and performance and uses storage media like SATA RAID systems, since they are faster than tape storage media but a lot cheaper than solid state drives. Due to economical reasons, the PT has the lowest storage capacity and the AT the greatest one.

Before the initialization phase no user, program or other data is stored on the HSM System and all new files will be placed down on the PT. If there is no more capacity available on the PT, files were migrated to the next lower tier and the same happens when there is no more capacity available on the CT. In practice, this migration is done in cyclic jobs, e. g. , once a day. The difficulty now, is to find the right files to migrate, since the migration is expensive due to read and write actions on the storage media and one does not want to migrate files that one need often in the future. A trivial method is to find all files that had no access for a specific time. Furthermore, every file would be analysed without context and relationship to other files. This reveals the main problems of classical HSM Systems. There is a lack of semantic and the very restricted amount of values to decide if a file has to be migrated between tiers. The lack of semantic means, that the system can not analyze files and consider files that are related by the same topic as a group of files and should be migrated together. A

second problem is the very restricted amount of file attributes such as file size and last access time.

The next section introduces our Storage Ontology (SO), that is the main part to overcome the lack of missing relations between files (Section 2). In Section 3, we present our experience and results we have made with a practical realization of our approach. Closing with related work (Section 4), we summarize our results and give an outlook in Section 5.

2 Storage Ontology

The purpose of the Storage Ontology (SO) is to offer an optimal information repository to find files that should be migrated from one tier to another. In this paper, we want to restrict to HSM Systems as described in Section 1. We also limit the files to be stored on the HSM to documents created by programs and users. This excludes files belonging to the operating system. The domain of the SO deals with files on the one side and HSM Systems on the other side and, of course, the interaction between both. This involves every information that is necessary to determine the best place of storage for each file. The following subsections provides a deeper look into the domain of the SO, starting with the domain file and going on with HSM.

2.1 Analyzing the Domain

Imagine a typical HSM System described in Section 1. Each of the files stored on it, are going to be analysed, and the extracted information is stored in a repository. Our purpose is not to collect each single possible information, since this ends up in a repository too big, where querying is not much faster than on file system level. Using too little semantic information leads to lack of knowledge and one can not determine the correct files for migration. To find a trade-off for this conflict, we took some typical scenarios from the real world to analyze which information is important. Starting with analysing files, we have a closer look on what is special about files that come from the same raw material, have the same project, or files that are grouped by the same topic.

Same raw material: At first, we observed that a lot of files are generated from the *same raw material*. E. g., professional cameras are taking photos in a very rich raw format, that needs a lot of storage capacity. After importing them, a graphic designer edits the photos and exports them maybe as JPEG files, which need less storage capacity. Later, these photos are presented on a website and small teasers from these photos are created. In this case, we don't need the raw photos from the camera any more, since we have more performant photos and the teaser ones.

This small scenario shows the following. First, there is a *group* of files that comes from a specific *source* (camera) at a specific *time*. Second, there are three *different characteristics* resulting from the *same raw material*, having different

amount of information. Third, there is a *purpose of use* for each of the characteristics. Among others, one could decide about the storage medium from this information, e.g. , the preview photos should always be stored in the PT.

Same project: As aforementioned, files often appear in groups. Groups are very important for the migration, because they offer the possibility to migrate more than just one file with only one query. As an example, imagine a video project. It consists of lots of files such as material from a video camera or audio files that should be underlayed into the video. There are also different chapters that occurs in different files. Also, the finished film, the trailer or specific still frames e.g. , for the web, could be contained in this project. As a consequence, if lots of files from one project are accessed permanently, one could infer that none of the files should be migrated to lower tiers. Only if the project is finished and very few accesses are made on the project files, these could be migrated in one step.

Same topic: Imagine there are lots of Christmas photos stored on the HSM System. Assume minimal accesses are made on these photos in the middle of the year, these photos are stored on the archive tier. But when Christmas time is coming, photos related to this topic are accessed more than in the middle of the year. Related to the migration of the files belonging the same topic, one could easily archive photos from this group in the archive tier or put them back on the higher tiers when a special topic becomes more important. Such topics could be inferred from more accesses on files related to this topic.

In every scenario, information related to time and place of a file is also needed, if the important files for the migration should be found. Therefore, it's necessary to know which storage media could offer the storage conditions that are needed for a special file. As introduced in Section 1, a typical HSM System is build up into three *Storage Tiers*, the *Performance Tier*, *Capacity Tier* and *Archive Tier*. Some of these tiers have directly mounted *Storage Media*, like solid state drives, while others have mounted whole *storage devices*, like tape libraries, that consists of several storage media (tapes). As defined, devices do not directly store files, they are more of an overlaying construct that is necessary to read, write and manage their containing media. Only physical storage media can directly store files and we do not want to lose the possibility of knowing where a file is stored concretely. Furthermore, an HSM System could consist of several *Storage Vaults*. Each of them manages one tiered hierarchy. Fig. 2 shows the structure of our understanding of an HSM System with the help of Extended Backus–Naur Form (EBNF). Since we map these structure into an ontology, EBNF is a good, simple and formal tool. Additionally, every Storage Medium should be characterized by its typical values like read/write access, average throughput or access time.

Furthermore, it is important to know about typical properties of every storage device and medium such as storage capacity, access time or data rate, while the latter could be distinguished into read and write values. These values are important to decide on which storage tier the devices or media should be placed, even in future when faster media is available and the current fast ones are not sufficient for the performance tier.

```

HSM                = { Storage Vault }
Storage Vault      = Performance Tier, Capacity Tier, Archive Tier
Performance Tier   = Storage Tier (* also for Capacity and Archive Tier *)
Storage Tier       = { Storage Device }
Storage Device     = { Storage Media }
Storage Media      = { File }

```

Fig. 2. Domain of HSM described in Extended Backus–Naur Form (EBNF)

2.2 Modeling of the SO

Ontologies are formal and their purpose is that every program and human using them will understand it exactly in the same way. To gain this goal, a modeling language like OWL [14], which we are using, is needed.

Since the XML-Syntax is not easy to read, we will use the *Manchester Syntax* [8], that uses natural language, to explain the structure of our SO. Thereby we can describe the set of all audio files (all things that are files and have an audio file format) as Defined Class by writing `File and (hasFileFormat exactly 1 AudioFileFormat)`. Furthermore, the term contains a (normal) *class* `File`. The differences between a Defined Class and a Class is the necessity of inference. While we must explicitly set classes to an *individual*, Defined Classes are inferred by reasoner and we do not have to assign them. There is also an *object property* in the upper term: `hasFileFormat`. Object properties describe the connection between two individuals. Additionally, we will also use *datatype properties* that describes the connection to data values, like string, int or self defined types. An example is `hasSize long`. In this case, the reasoner looks which of all individuals, who are member of the class `File`, has exactly one property `hasFileType` that has the range of another individual as an instance of `AudioFileFormat`, that is also a Defined Class. At last, we have two keywords from the Manchester Syntax: `and` and `exactly`. The keyword `and` is an intersection (\cap) and `exactly x (=)` means that every instance of such a class must have x times a given property. Another important concept of our ontology are *individuals*. These are concrete instances that belongs to concrete classes. While `File` is a class describing an individual belonging to its class, `File123` could be a concrete instance of the class `File` and could optionally have a file name or a size.

2.3 Structure of the SO

Following, we present the schema of our ontology by explaining the main ideas with the help of the Manchester Syntax. At first, we describe the classes and properties belonging to files, and later we want to end up with HSM Systems and the connection between both.

File Concluding from Section 2.1 we modelled the class `File` as the central class (see Fig. 3). There are some datatype properties, among others, that describes

attributes such as path, name, id, size or other important once such as last access or the date of creation. There is also a property `isPreview` that indicates if a file is a preview or not. Furthermore, a collection of files could be added to a `Group`. The connection between `File` and `FileFormat` is more complex. To explain the `FileFormat`, we have to go further to the two classes `MediaType` and `QualityType`. These classes are modeled as closed enumerations. `MediaType` has only five predefined instances based on the MIME Media Types by IANA [9]: `Audio`, `Application`, `Image`, `Video` and `Text`. All the same, we defined the class `QualityType`, that has two instances: `Performance` and `Raw`, considering that a lot of files exists in several characteristics (see Section 2.1). While the `MediaType` could be extracted through several tools, the `QualityType` must be set manually, except of file formats for preview files, because, if a system creates a preview from another file, one could observe this and set the property `isPreview` on true. As an example, PNG as instance of `FileFormat` could be created. The two properties `MediaType` and `QualityType` are set to `Image` and `Performance`. A reasoner could infer from this information, that the individual is also a member

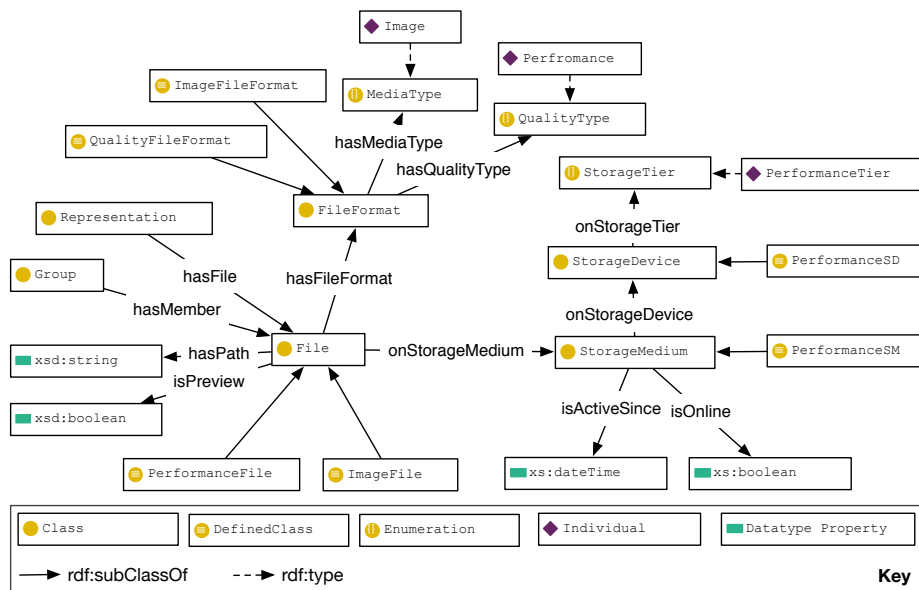


Fig. 3. Partial graph from the SO, that describes raw and image files on the performance tier

Furthermore, an individual, e.g. , `holiday.png` as member of the class `File` would be created. One could use an analyzing tool to get the MIME Media Type from this file and set the property `hasMediaType` to the individual, in this case with the range `PNG`. From this connection, a reasoner could infer that this file also belongs to the classes `ImageFile` and `PerformanceFile`, since the class description is `File and (hasFileFormat exactly 1 ImageFileFormat)` for the Defined Class `ImageFile`. As one could envision, there are also analogous class descriptions for every subclass from `File` that is build from every subclass of `FileFormat`. An exception is the subclass `PreviewFile` as described earlier. This class description is `File and (isPreview true)`.

We also defined the class `Representation` in the SO. This is an abstraction of `Files`, that have at least one similar characteristic. As an example, there could be three different characteristics from one photo as raw photo from a camera, a performance file that was created from the first one and a preview. If an individual of `Representation` would be created and add connections to the three `Files`, one could find these three ones with only one query. As an example, imagine an access on the preview file of a photo, and later the user wants to retrieve the raw file of this photo which is archived. In this case, it is easy to find this photo.

HSM Going on with the classes and properties belonging to the domain of the HSM System, we have made a meaningful basis in Section 2.1 by describing the HSM with the help of EBNF and it is easy to map this into our SO. Therefore, we add a class for every type from the EBNF description, but we omit the class `Storage Vault` in Fig. 1, since it is not necessary if only one is used in the HSM System. According to the classes `MediaType` and `QualityType`, we modelled the class `StorageTier` as an enumeration with the three members `PerformanceTier`, `CapacityTier` and also `ArchiveTier`. With the property `onStorageTier`, an instance of the class `StorageDevice` could be connected to one of the tier, and also with the property `onStorageDevice`, an instance of `StorageMedium` could be placed on a given storage device. Depending on to which kind of storage tier a storage device and on to which storage device a storage medium is connected, a further type could be added to these individuals. Taken the example from Fig. 1, a storage device that is connected to the performance tier would be additionally inferred as a `PerformanceDevice`, because the class description of the defined class `PerformanceDevice` is `StorageDevice and (onStorageTier) PerformanceTier`. We modelled the class description from the defined classes `CapacityDevice` and `ArchiveDevice` and also the different storage mediums, accordingly. Among others, a Storage Device has properties for read/write, access time or throughput.

3 Evaluation

As we focus on the meaning and context of documents we decided to use semantic data technologies instead of a relational database to store extracted information

[13]. This section introduces our test environment, gives an example to compare our approach with the classical one and shows the advantages.

3.1 Test setup and environment

In 2005, we started the K-IMM² project [10], which aim is to include the content and context of documents in personal document management. Although the architecture was designed to enable a simple replacement of the used ontology, the prototype supported it in a cumbersome way. Another problem was the support of only one ontology at runtime. Thus, we redesigned the architecture and reimplemented the core module to support multiple ontologies, exchange them at runtime and extended the approach to support small and medium-sized enterprises. We call it KIMM+. In the SENSE³ project⁴, KIMM+ is used as a semantic middleware. Web applications are used to upload documents such as videos and pictures to the SENSE application. These files are analysed by KIMM+ and extracted information is stored in the ontologies, especially the SO. The files are stored on a file system under control of an HSM System. In this infrastructure both, HSM and web applications are able to query the SO without gathering information or access to the file system and the files themselves. Therefore, the HSM System was extended to execute and analyse SPARQL queries. The following section will give two short examples of how to use the SO and highlights the advantages.

3.2 Use cases

In Section 1, we focussed on two aspects: (1) the lack of semantic (e. g., the unknown relations between documents represented by their content) and (2) that classical HSM Systems just use a restricted set of attributes to store files on a specific tier. For the following examples, we assume to have three folders: `music`, `documents` and `pictures`. Each folder contains files related to the topics: `flowers`, `birds` and `research paper`.

(1) In Section 2, we introduced the concept `Group`. We see the SO as a minimal subset of relations necessary to store a file in the best place. Therefore, we need a concept to group files. By keeping it abstract, a group can represent the same topic (e. g., `flowers`) or persons in documents. A file can be part of several groups. If a file named `national_fauna.pdf` should be migrated from the PT to the CT, the following query can be used to get the associated groups.

```
SELECT ?group WHERE {
    ?file :hasName "national_fauna.pdf". ?group :hasMember ?file . }
```

The result contains all groups the document is related to. The other documents of these groups can be found by executing for each group:

² Knowledge through intelligent media management

³ Intelligent Storage and Exploration of large Document Sets

⁴ <http://www.sense-projekt.de>

```
SELECT ?file WHERE { ?group :hasMember ?file .}.
```

To prevent the HSM System from migrating files that should not be migrated, other attributes have to be considered to get the final result. This is done by the logic of the extended HSM System in SENSE. The example shows that in this way the relations and semantic of documents will be involved in the decision making for migration. We chose the group concept to keep the semantic within the SO as simple as possible, because the HSM does not need to know what a topic is but which files are related by topic. Each concept of grouped entities can be broken down to a SO group.

(2) The second aspect we focussed on is the restricted set of attributes. They are limited to classical attributes to filter files for a policy definition. The folder **pictures** introduced above contains, for example, PNG, TIFF, NEF and JPEG files. If RAW formats can be archived shortly after they were accessed, this is difficult to model with standard policies. First, the attributes to match a RAW format have to be defined. Let's assume TIFF and NEF are RAW formats and need to have a file size greater than 1024 KByte. A policy would like this:

```
(File name matches pattern *.TIFF or *.NEF) and
  (File is larger than 1204KByte) .
```

If the definition of our RAW format changes or a new one needs to be regarded, the policy has to be modified. Using the new approach, the HSM System gets a list of files by executing the following query:

```
SELECT ?file WHERE
  { ?file a RawFile. ?file :hasFileFormat ?f. ?f a :ImageFileFormat. }.
```

In this case, changed requirements can be applied directly to the class definition for **RawFile**. Another benefit is the possibility that files can be excluded in the Defined Class if they match specific criteria. This improves the flexibility of the HSM System.

First tests within the SENSE framework showed another advantage of the Storage Ontology. Listing the size of files, no matter if a group of files or all files are selected, is much faster if the ontology is used. Linux as well as Windows and OS X showed problems, when the HSM filesystem was loaded, containing more than 500.000 files, although they were uniformly distributed in the filesystem. This section showed two small examples how the introduced SO can be used to improve existing HSM Systems. Currently, we are testing and refactoring the SO within the prototype of the SENSE application under real conditions to verify the improvement. The next sections give an overview on existing approaches on using semantic technologies to store, search and access files.

4 Related Work

As mentioned in Section 1, we focus on two main problems: the lack of relations regarding to file content and the usage of restricted values such as file size.

Services and platforms like Google Docs, Facebook, Flickr, Youtube and Twitter are used to store personal documents and to share them with others. Even for enterprises this becomes more and more important. Regarding this situation, Semantic Web Technologies can yield to a more flexible way to connect and use these services. Focussing on the requirements of small and medium sized enterprises, we concentrated on the storage of files within a centralized infrastructure. Reasons for such a infrastructure can be required by law. In this context, the Memex Vision published by Vannevar Bush [2] is of great interest. He described a system to handle a large amount of personalized heterogeneous data and defined four requirements, cited among, others by Gemmell et al.[5]: (1) collections and search must replace hierarchy for organization, (2) many visualizations should be supported, (3) annotations are critical to non-text media and must be made easy, and (4) authoring should be via transclusion⁵.

We understand (1) and (3) as necessary precondition to fulfill (2), and (4) given by the concept of Semantic Web Technologies. Describing the relations of files using an ontology supports the first requirement. While (semi)-automatic creation from meta data and information retrieval fulfills the third requirement. Having these information in form of an ontology, different visualization can be realized (3). And as shown in Section 2, using Semantic Web Technologies, relations between documents change when rules or schema change (4). A first published approach to store relations between documents and their contents in a ubiquitous way was WinFS [7]. Through an API, the documents and settings folder of Windows Vista was planned to handle files in transactions supported by relational database technologies. The aim was to handle all files in the same way and to get a homogeneous knowledge base on documents on a PC. Although WinFS never was released, parts of the concepts were included in following versions of NTFS on Windows 7 and Windows 8.

An interesting approach to fulfill the first requirement of Vannevar Bush [2] was published by Bloehdorn et al.[1] in 2006. Using WebDAV as an abstraction layer between user and file system enables Bloehdorn et al. to break with the classical hierarchical approach of managing files in folders. To the user, the file system hierarchy is presented, matching results of search queries. The files are sorted and organized in virtual folders depending on their tags. Each Folder represents a tag and a concatenation of multiple tags represents a location. In this way files are tagged by their location with multiple tags. On the one hand, this approach enables an orthogonal search, but on the other hand the subset of tags is limited to the folder names, as they are the tags. That means, that no real search for files ordered by persons is possible.

In [3] Crenze et al. the challenges of information management in enterprise environments are presented. They identified the: (1) amount of data, (2) the quality and performance of extraction tools, and (3) security and authentication of data and data access. Semantic Web Technologies are seen as possibility to replace a classical full-text search by a semantic search on ontologies. According to Crenze et al. a combination of full-text search with semantic-aware filtering

⁵ Including documents into each other by using a reference.

and proposal functions is the best combination to retrieve good search results. Because of weak performance, the authors intend to use Apache Lucene⁶ to index properties instead of Apache Jena⁷.

Regarding the description of files through properties and meta data, PREMIS [11] gives a good system to describe them in a standardized way. Especially, since there also exist RDFS and OWL schemas. Beside a subset of properties to describe data for archiving, PREMIS also covers security and authentication. Although it is designed to achieve a flexible technology-independent way to archive physical objects it is also useful to identify storage related properties for an HSM.

In 2011, we came up with the idea of using a ontology to optimize the placement of files in an HSM System. For the first prototype, we did not use all capabilities of ontologies and used a combination of semantic descriptions and relational rules to generate storage solutions [12]. The great disadvantage of this approach was the modeling of rules that depended directly on information stored in the ontology. Another disadvantage was the minimalistic schema, which was not able to represent storage-related properties. For example, it did not support the description of different file types like raw files that can be stored on cheap but slow memory⁸. In consequence, we focussed on developing the Storage Ontology. The next section gives a conclusion on our results and overview on upcoming work.

5 Conclusion

In Section 2 we introduced a schema to describe files stored on an arbitrary file system controlled by an HSM System. We focused on describing files and their relations among themselves and within an HSM System. The quality of the ontology depends on the used extraction tools to gather information. Using OWL allows a more flexible classification of files such as image files or files which should be moved to the archive. This leads to a more flexible configuration for HSM System policies as the underlying system uses SPARQL queries to get related files. In consequence, the policy engine of existing systems can be used with minimal modifications. The located files are related based on the two mentioned aspect. They are chosen either because of matching attributes or because of relations in content or type. In this way, we introduced a source for decision-making in HSM systems and enabled applications in the front-end to get file specific information as well as knowledge about the placement in the storage hierarchy. We integrated our schema in ontologies developed in the SENSE project. Currently, we evaluate and adapt the Storage Ontology by defined scenarios. The recorded data, like file access-times on the Performance Tier and migration tasks, are compared with the results of the classical approach using the policy engine without semantic. Furthermore, we improve the SENSE framework.

⁶ <http://lucene.apache.org/core/>

⁷ <http://jena.apache.org/>

⁸ Depending on the application domain.

References

- [1] Stephan Bloehdorn, Olaf Görlitz, Simon Schenk, et al. “TagFS - Tag Semantics for Hierarchical File Systems”. In: *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06)*, Graz, Austria, September 6-8, 2006. Sept. 2006.
- [2] Vannevar Bush. *As We May Think*. The Atlantic Monthly. 1945.
- [3] Uwe Crenze, Stefan Köhler, Kristian Hermsdorf, et al. “Semantic Descriptions in an Enterprise Search Solution”. In: *Reasoning Web*. Edited by Grigoris Antoniou, Uwe Aßmann, Cristina Baroglio, et al. Volume 4636. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pages 334–337.
- [4] J Gantz. “The Diverse and Exploding Digital Universe”. In: *IDC white paper*. White Paper 2 (2008), pages 1–16.
- [5] Jim Gemmell, Gordon Bell, Roger Lueder, et al. “MyLifeBits: fulfilling the Memex vision”. In: *Proceedings of the tenth ACM international conference on Multimedia*. MULTIMEDIA '02. Juan-les-Pins, France: ACM, 2002, pages 235–238.
- [6] PoINT Software & Systems GmbH. *Automated Storage Tiering: Optimizing the storage infrastructure concerning Cost, Efficiency and Compliance*. http://www.point.de/fileadmin/Documents/White_Paper_Automated_Storage_Tiering_by_PoINT_Storage_Manager.pdf. 2012.
- [7] Richard Grimes. *Code Name WinFS*. <http://msdn.microsoft.com/de-de/magazine/cc164028%28en-us%29.aspx>. Aug. 2004.
- [8] Matthew Horridge, Nick Drummond, John Goodwin, et al. “The manchester owl syntax”. In: *In Proc. of the 2006 OWL Experiences and Directions Workshop (OWL-ED2006)*. 2006.
- [9] IANA. *MIME Media Types*. <http://www.iana.org/assignments/media-types>. 2013.
- [10] Annett Mitschick. “Ontologiebasierte Indexierung und Kontextualisierung multimedialer Dokumente für das persönliche Wissensmanagement”. PhD thesis. Technischen Universität Dresden, 2009.
- [11] PREMIS Editorial Committee. *PREMIS Data Dictionary for Preservation Metadata*. Edited by PREMIS Editorial Committee. www.loc.gov/standards/premis/v2/premis-2-0.pdf. 2008.
- [12] Axel Schröder, Ronny Fritzsche, Sandro Schmidt, et al. “A Semantic Extension of a Hierarchical Storage Management System for Small and Medium-sized Enterprises”. In: *SDA*. 2011, pages 23–36.
- [13] M. Uschold. *Ontologies and Database Schema: What’s the Difference*. <http://semtech2011.semanticweb.com/programDetails.cfm?ptype=K&optionID=284&pgid=4>. 2011.
- [14] W3C OWL Working Group. *OWL 2 Web Ontology Language Document Overview (Second Edition)*. <http://www.w3.org/TR/owl2-overview/>. 2012.
- [15] Muzhou Xiong, Hai Jin, and Song Wu. “FDSSS: An Efficient Metadata Management Scheme in Large Scale Data Environment”. In: *Grid and Cooperative Computing* 90412010 (2006).

Semantic Retrieval Interface for Statistical Research Data

Daniel Bahls, Klaus Tochtermann

Leibniz Information Centre for Economics (ZBW), Kiel, Germany

Abstract. Statistical research data is the foundation for empirical studies. Researchers in economics or social sciences often obtain such data from external sources through specially designed retrieval interfaces from statistical offices, commercial data providers as well as from data agencies and other archives. With the advancements in data cataloguing and acquisition of long tail research data sets from individual scientists and institutes, the opportunity is there to install central services for a more holistic data search. In view of a rapid increase in amount of data available and by association an emerging retrieval problem, retrieval interfaces must make effective use of provided metadata in order to help find relevant data sets efficiently.

This paper presents a multi-step retrieval interface that aims to support the researchers' natural approach to data search and composition. Starting with an idea of the concepts that are to be compared, users kick off their search with thesauri terms and successively specify requirements according to their priorities until suitable data can be selected easily from a manageable number of matching data sets. The prototype presented in this paper also provides means for convenient data harmonization, which is an essential aspect especially when combining statistical data from different sources.

Keywords: Research Data Management, Semantic Digital Data Library, Linked Data, Statistics, Data Retrieval

1 Introduction

A significant number of scientific results are based on research data, since research has become increasingly data-driven over the years [1]. Therefore, to understand such scientific publications in depth, documentation on underlying data is a necessary means. To further provide transparency and enable replicability in the end, respective data sets must be available as such, for which a reliable infrastructure is required. Scientific data needs to be maintained and organized in archives.

With the advancement of computer technology, scientific analyses are more and more carried out with the aid of machines, as it allows for large amounts of data being processed in short amount of time which has never been possible before. While this certainly is one reason why science has become significantly

data-driven, it also leads to the fact that most scientific data is maintained in digital form already. This circumstance and the rise of the Web opens up possibilities for a powerful information infrastructure for supporting these aforementioned goals. Information resources nowadays can be delivered to any place in the world within seconds, laying the ground for delivering the right information to the right place at the right time, the precept of knowledge management.

The Web together with its well-established Web 2.0 technologies has already been recognized as a powerful media for promoting efficient exchange and advancement in the scientific domain. In this regard, the Leibniz Association has recently started the research alliance Science 2.0¹ with a growing number of 30 associated institutes to jointly venture into a well-organized and integrated environment of Web-based tools and services for the scientific community to support rapid exchange and good scientific practice.

The vision of a thought-out research data infrastructure fits well into this theme, and many initiatives have formed in the last years, a whole movement to effectively enable exchange, citation and preservation of research data. However, this task has proven non-trivial, as it opened up exhaustive discussions on meta-data schemes², organized preservation and curation [2], responsibilities [3], data publication policies [4] as well as solutions to overcome issues of data protection and usage rights, only to mention a few. Yet, these efforts have already lead to significant advancements (TheDataHub³, DataCite⁴, and other).

At present, efforts are being made to pick up research data as bibliographic artifacts for re-use, transparency and citation[5]. In view of a rapid increase in amount of data available and by association an emerging retrieval problem, retrieval interfaces must make effective use of provided metadata in order to help find relevant data sets efficiently.

In this paper, we investigate how to make use of Semantic Web technologies for providing an efficient and novel approach for the retrieval of statistical data sets that follows a natural approach for data retrieval in the domain of statistics, particularly in the context of economics or the social sciences. Section 2 elaborates on the practice of data acquisition in empirical research to gain a clear picture on the purpose of our system. Related work is discussed in the subsequent section, and Section 4 explains fundamental design decisions and outlines a system architecture. Section 5 describes the user interface itself and how the declared goals have been implemented into features. The paper eventually closes with conclusions and outlook.

¹ <http://www.leibniz-science20.de>

² particularly important, as in contrast to textual publications, data cannot be understood without documentation

³ <http://datahub.io/>

⁴ <http://www.datacite.org/>

2 Retrieving Statistical Data

In many cases, empirical researchers in economics and the social sciences are to put together statistical indicators in large data tables. Typically, each column represents one indicator while the rows represent respective data per year, country or other so-called dimension. The data itself may be self-produced in terms of studies and surveys or acquired from external sources such as statistical offices, affiliated institutes or purchased from commercial data providers. However, common practice is to combine several sources, since some indicators may be obtained from one source while the data for other indicators may be obtained from another one. In this regard, researchers have to be extra careful to make sure respective data represents the same or sufficiently similar statistical population.

To gain a clear picture of the goals of this research, we need to clearly understand the purpose of the system. We have conducted interviews with economic scientists which helped us gain insights in their work with research data. Empirical researchers typically start out with an idea of concepts relevant in their research (e.g. living standards, work conditions, economic growth, etc.). In addition, they have further details in mind, for instance on reference periods, regions to be included and distinguished or frequency of data acquisition in case of time series data. As a result, the data set should be as consistent as possible with respect to acquisition method, statistical universe and adjustments. To achieve user acceptance, the system has to be practical in research settings [6], and therefore we aim to support this data harmonization procedure in a light-weight manner.

As a result, user communication should follow the below steps:

1. Prompt for a list of concepts that are to be compared
2. Let user specify additional requirements on the data
3. Let explore and select matching data sets, allow for revisiting Step 2
4. Offer selected data for download

After finishing Step 1, data sets associated with the concepts named should be presented to the user. Specification of additional requirements should be based on the metadata available for the data sets found. As soon as all relevant requirements are given, the user may inspect and decide on these satisfying data sets and proceed to download at last.

3 Related Work

There are many repositories on the Web that provide statistical data. Some of them are provided by statistical offices and data agencies (e.g. Federal Statistical

Office of Germany⁵, EuroStat⁶, World Bank⁷), some are associated with commercial providers (e.g. Thomson Reuters Datastream⁸, Statista⁹) and yet others are maintained by journals, archives, libraries or independent organizations (e.g. GESIS¹⁰, The Data Hub¹¹, Dataverse repository of Economists Online¹²). All of these portals are as heterogeneous as the kind and spectrum of data they provide. Some of them provide interfaces for composition of customized data tables where users pick and choose indicators and data records according to their needs. Such features are also provided by the Nesstar system¹³, one of the most prominent systems for data publishing and online analysis that is being used by a large number of institutes. The Social Science Variables Database at ICPSR¹⁴ allows for direct comparison of indicators with respect to a variety of metadata, giving intuitive means to understand differences in universe, acquisition method and other between data sets. However, users of these systems are to run keyword-based queries and browse through category trees in order to find relevant data sets individually, and therefore our approach follows a different paradigm as presented in Section 2.

Technical challenges in dealing with distributed sources and applying the OLAP paradigm for retrieval of statistical data from the Linked Data cloud have been addressed in [7]. We view this work as a major contribution for building a scalable backend, whereas our work aims to provide a user interface and communication design for data search and retrieval within the specific setting research data sharing.

Other approaches are based on semantic links between data sets and research articles [8] which give textual context for otherwise sparsely described data content and therefore improve data search by established Information Retrieval techniques. These data links, typically given by persistent identifiers, however, point to entire data bundles as a whole, whereas our approach aims to make single indicators and values available for retrieval.

4 System Architecture

Following the steps presented in Section 2, we elaborate on the system architecture of our data retrieval system. To support Step 1, a thesaurus should be used, so that data sets associated with a particular concept can be found easily. To enable the specification of requirements, metadata must be given in detail

⁵ <https://www.destatis.de>

⁶ <http://epp.eurostat.ec.europa.eu>

⁷ <http://data.worldbank.org>

⁸ <http://online.thomsonreuters.com/datastream/>

⁹ <http://de.statista.com>

¹⁰ <http://www.gesis.org/en/>

¹¹ <http://thedatahub.org>

¹² <http://dvn.iq.harvard.edu/dvn/dv/NEEO>

¹³ <http://www.nesstar.com>

¹⁴ <http://www.icpsr.umich.edu>

and in association with individual indicators and records rather than a separate metadata block for a zipped data bundle. This enables the system to make sense of the data in depth and allow for requirement specification as explained later in Section 5.

The research on a data retrieval interface is part of our overall research activities on an infrastructure for scientific data for the field of economics. For several reasons we regard Semantic Web technologies most suitable for this purpose, among which is strength in dealing with distributed data and extensibility, which is required whenever highly specific long tail data from individual researchers needs additional vocabulary for description [9]. However, the data format should provide for typical data types, such as floats, strings, dates and other. It must provide metadata on fine-grained level as to open up possibilities for retrieval and composition. As a consequence, the retrieval system operates on statistical data in the format of the RDF Data Cube Vocabulary¹⁵ [10].

The prototype was implemented in Java and JavaScript under the use of the Play Framework¹⁶. The live system was tested on an Apache Tomcat¹⁷ and a Sesame Triple Store¹⁸, as the system operates on statistical data provided as RDF using the RDF Data Cube Vocabulary¹⁹ [10].

5 User Interface Design

The system implements a multi-step retrieval interface as described in Section 2. In the following, we are going to refer to the screenshots given in Figure 1 to 8 in parantheses. Since the expected result is a data table after all, the main screen starts with an empty spreadsheet (1). For Step 1, the user successively enters the names of the concepts that are to be compared in the empty column headers as shown in (2). This task is supported by autocompletion on the basis of concept terms contained in a thesaurus, STW²⁰ in our case. With the selection of a concept, the system displays the number of associated data sets beneath the concept label entered before. A click on this number lists all of them in alphanumerical order (3), and another click reveals a detailed description and further information on the particular data set (7). Yet, at this point, the number of data sets might be huge, and the user may decide to formulate requirements for the data first as per Step 2. With the selection of a single column header, the panel on the left lists down the *union* over all properties and property values available in the metadata of all the data sets associated with the concept of the column (4). Hovering over a property or property value produces an info box with documentation on the vocabulary. Selecting a particular property value specifies a requirement and tells the system that only those data sets are relevant for

¹⁵ <http://www.w3.org/TR/vocab-data-cube/>

¹⁶ <http://www.playframework.org>

¹⁷ <http://tomcat.apache.org>

¹⁸ <http://www.aduna-software.com/technology/sesame>

¹⁹ <http://www.w3.org/TR/vocab-data-cube/>

²⁰ STW Thesaurus for Economics, <http://zbw.eu/stw/>

this column that provide this respective property and property value, and the number of relevant data sets drops. With the selection of two or more column headers, the panel on the left shows the *intersection* between the properties and values of the single columns (5). This feature facilitates harmonization of data, as it reveals which data characteristics can be unified among the columns. To specify the contents of the rows, one must specify the *Dimension* property. A click on the respective header highlights all column headers of the entire table as to indicate that the property of choice must be available in the data sets of all columns. The user selects (multiple) values from the properties listed on the left and the Dimension column fills accordingly (6). This again sets requirements for the data sets, as it filters all data sets that do not provide respective records. Eventually, when all requirements are set, the user examines and selects from the remaining list of data sets for each column (7). If all remaining properties with multiple options are bound to a value, the table fills with actual data content (8). As a last step, the table is offered for download.

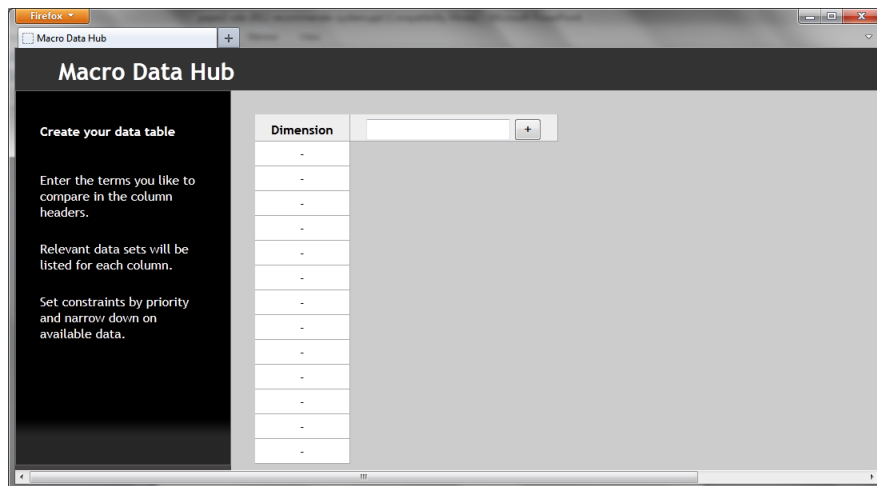


Fig. 1.

6 Conclusions and Outlook

Following the call for a research data infrastructure, we have addressed the issue of data retrieval for the domain of economics and social sciences where large amounts of scientific results are based on statistical data. With the prospect of a rapidly growing amount of data from individual researchers and institutes filed in the future, overviewing all relevant data sets efficiently becomes a problem. For this purpose, we have designed an innovative retrieval interface that aims

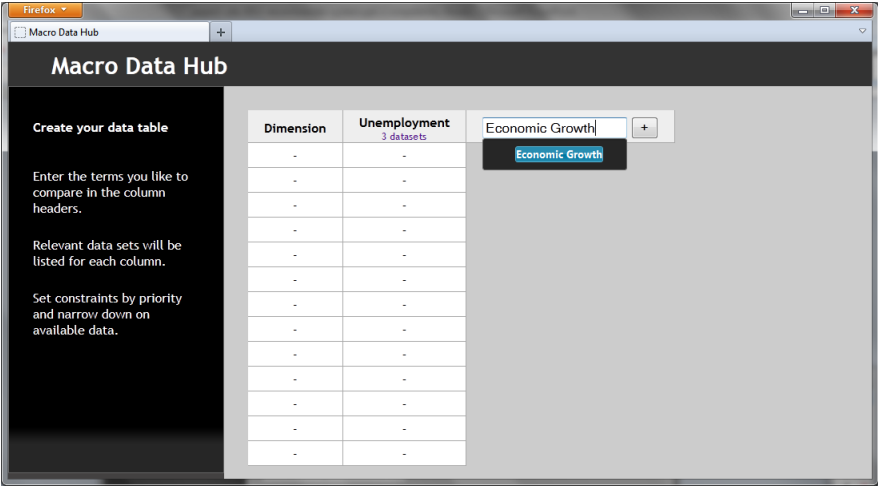


Fig. 2.

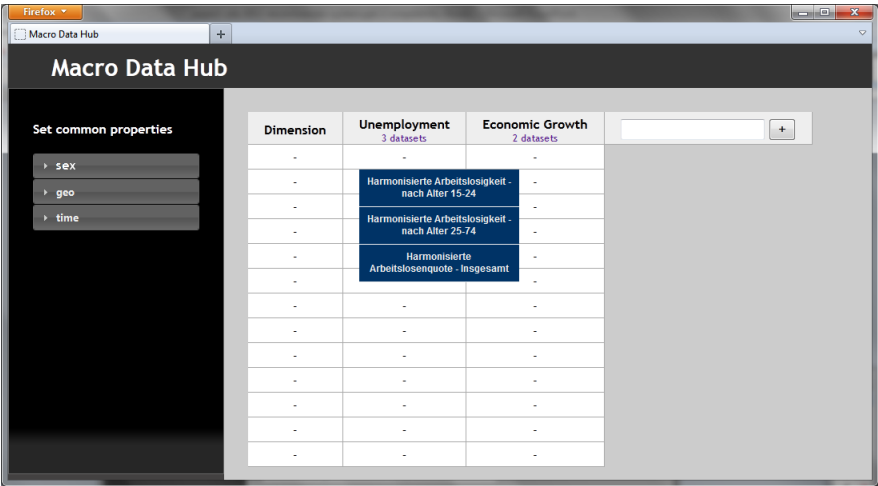


Fig. 3.

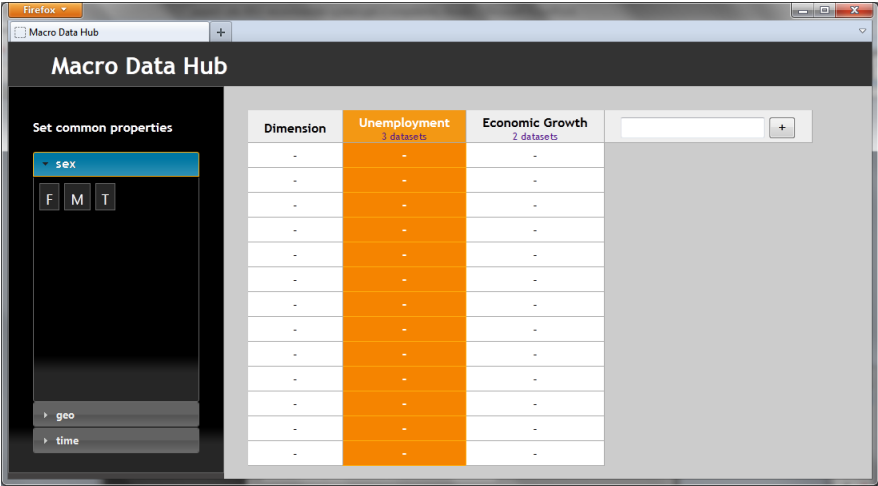


Fig. 4.



Fig. 5.

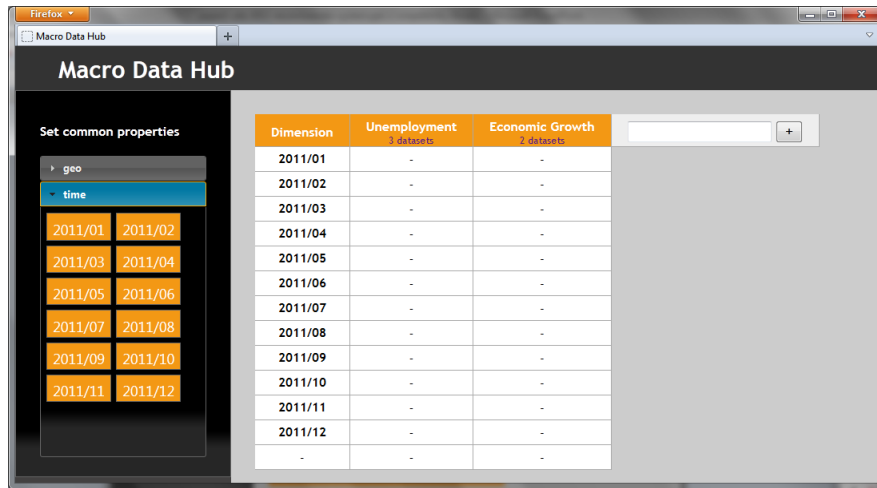


Fig. 6.

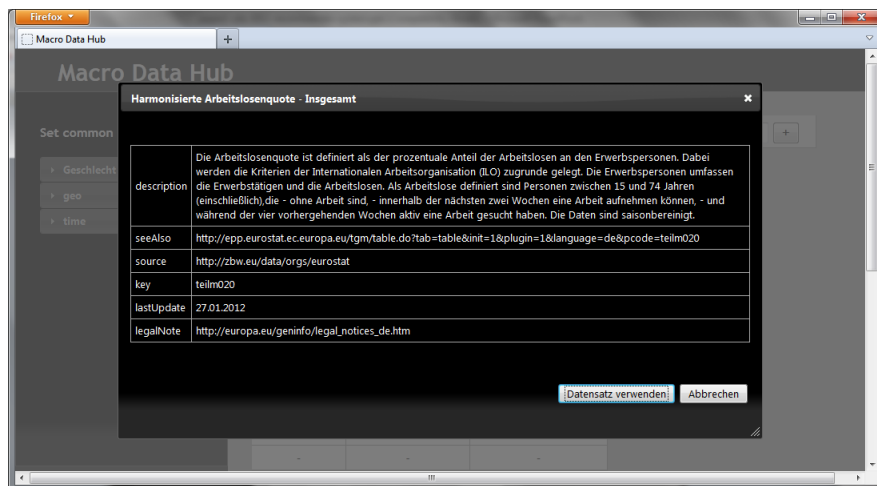
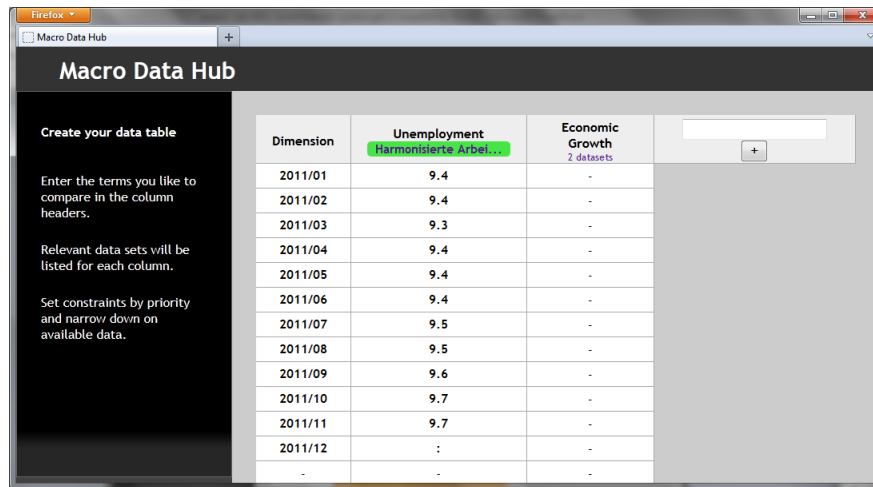


Fig. 7.



Dimension	Unemployment Harmonisierte Arbel...	Economic Growth 2 datasets
2011/01	9.4	-
2011/02	9.4	-
2011/03	9.3	-
2011/04	9.4	-
2011/05	9.4	-
2011/06	9.4	-
2011/07	9.5	-
2011/08	9.5	-
2011/09	9.6	-
2011/10	9.7	-
2011/11	9.7	-
2011/12	:	-
-	-	-

Fig. 8.

to support researchers in finding and composing data sets according to their natural way of approaching a research question. The prototype presented in this paper provides simple means for data harmonization to enable consistency within statistical population in intuitive ways. Under the use of these features, we expect a significant decrease of time needed for data search and composition in comparison to the current practice, although this is yet to be evaluated.

Future improvements of the system should include retrieval from distributed sources, as this version operates on a single triple store endpoint only. Moreover, the advantages of using subproperty relations should be investigated and made available to the user. Many other valuable ideas for improvements can be found with regard to user assistance, e.g. warning notifications when selected time series data include breaks, errors or changes in acquisition method which can be derived from well-maintained metadata.

Finally, this approach needs to be tested on a large archive of various kinds of statistical data and evaluated with end users from the target group of empirical researchers.

References

1. Gray, J.: Jim Gray on eScience: A Transformed Scientific Method (January 2007)
2. Treloar, A., Harboe-Ree, C.: Data management and the curation continuum: how the Monash experience is informing repository relationships. Proceedings of VALA 2008 (2007)
3. Rümpel, S.: Data Librarianship : Anforderungen an Bibliothekare im Forschungsdatenmanagement (2010)
4. Vlaeminck, S., Siegert, O.: Welche rolle spielen forschungsdaten eigentlich für fachzeitschriften? eine analyse mit fokus auf die wirtschaftswissenschaften. Technical report, German Council for Social and Economic Data (RatSWD) (2012)

5. 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., Giarretta, 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.
6. Feijen, M.: What researchers want - a literature study of researchers' requirements with respect to storage and access to research data (February 2011)
7. Kämpgen, B., Harth, A.: Transforming statistical linked data for use in olap systems. In: Proceedings of the 7th international conference on Semantic systems, ACM (2011) 33–40
8. Boland, K., Ritze, D., Eckert, K., Mathiak, B.: Identifying references to datasets in publications. In Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F., eds.: Theory and Practice of Digital Libraries. Volume 7489 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 150–161
9. Bahls, D., Tochtermann, K.: Addressing the long tail in empirical research data management. In: Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies. i-KNOW '12, New York, NY, USA, ACM (2012) 19:1–19:8
10. 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)

Semantification of Query Interfaces to Improve Access to Deep Web Content

Arne Martin Klemenz, Klaus Tochtermann

ZBW – German National Library of Economics
Leibniz Information Centre for Economics,
Düsternbrooker Weg 120, 24105 Kiel, Germany
{a.klemenz,k.tochtermann}@zbw.eu
<http://www.zbw.eu/>

Abstract. This position paper as part of a PhD thesis is a contribution to an automatic retrieval of information from the *Deep Web*. Addressing current limitations of the *Deep Web Information Retrieval* leads to the prevailing lack of semantics regarding the retrieval process. Focusing this problem from the information providing services perspective, indicates the significant potential of additional semantic annotations provided by websites. *Web query interfaces*, the interfaces to the majority of available information on the Deep Web, are interpreted as *Semantic Deep Web Services* (SDWS). The introduction of a SDWS annotation leads to great potential for *Information Retrieval* services based on the large variety of information available on the Deep Web.

Keywords: Deep Web, Semantic Deep Web Service, web query interface, semantic annotation

1 Introduction

A continuously increasing amount of content on the web is not directly accessible and indexable by search engines. The content might, for example, be hidden in non-public, inaccessible areas or might be stored in background databases and therefore only accessible through *web query interfaces*. This part of the web is known as the *Deep Web* (or *Hidden Web*) in contrast to the *Surface Web* which can be easily accessed and indexed by common search engines [1].

The Surface Web consists of mostly static content, which is directly inter-linked with static hyperlinks. "Search engines rely on hyperlinks to discover new webpages [...]" [11], but static websites are outnumbered by dynamic websites on an extremely large scale and the web has been rapidly deepened [6]. The content as part of dynamic websites is mostly not accessible through static hyperlinks, as this content is dynamically enwrapped into web pages as the response to a query submitted through a web query interface. These are intended to be used by human users to retrieve content from a background database often containing highly relevant content of a specific domain. Common search engines do not

reach this part of the web. This is caused by the fact, that search engines "[...] typically lack the ability to perform form submissions" [11].

Considering current arising services on the web like the Google Knowledge Graph, "we can use [...] [these services] to answer questions you never thought to ask and help you discover more"¹. These services are related to *Knowledge Discovery*, but in general the benefit from the automatic discovery of new knowledge from existing information on the web is depending on an excellent *Information Retrieval*. As the retrieval of information from the Deep Web is still limited, Knowledge Discovery services are also still limited in their potential. Therefore, more efficient and targeted retrieval mechanisms for the Deep Web are needed to achieve full potential of Knowledge Discovery services.

The usage of semantic annotations for information on the web play a crucial role "to assimilate information from multiple knowledge sources" [13]. This challenge has been addressed, resulting in standards like *Resource Description Framework in attributes* (RDFa) and Microdata markups like *schema.org* initiated by the search engine big players Bing, Google, Yahoo! and Yandex. Therefore, this paper addresses the improvement of accessing this semantically annotated content on the Deep Web.

2 Related Work

The retrieval and indexing of Deep Web content have been addressed from different perspectives in the past. The effort has mostly focused specific applications to discover, retrieve and index structured data from the Deep Web. This includes special emphasis on the automatic web query interface interpretation.

Common approaches focusing on exposing Deep Web content can be classified to *surfacing* and *virtual integration* approaches. The surfacing approach focuses a search engine initiated process to index the search result pages for pre-computed (randomized) queries to discover Deep Web content on large scale [10]. The virtual integration approach follows the data integration paradigm, using a mediator system to map queries to relevant web query interfaces [10]. The content, that is retrieved, is brought to the user by the virtual integration to the search result page. Both of these approaches have been approved as useful in some cases. But in general the virtual integration approach is related to a lot of manual effort setting up query mapping rules for each Deep Web query interface in the mediator system. Furthermore, the surfacing approach is too imprecise or ineffective and therefore not scalable regarding the pre-computation of queries for domain independent sets of Deep Web websites.

Regarding the discovery and cataloging of Deep Web sources Hicks et al. [8] highlight the challenges and demonstrate via prototype implementation, that their Deep Web discovery framework can achieve high precision using domain dependent knowledge for probing web query interfaces. Wenye et al. [15] focus "Manufacturing Deep Web Service Management [...] [by] Exploring Semantic

¹ <http://www.google.com/insidesearch/features/search/knowledge.html?hl=en>

Web Technologies” by semantically annotating the Deep Web Services to reflect their hidden, dynamic, and heterogeneous contents while the relevance of semantic annotations for the Deep Web has already been identified in 2003 by Handschuh et al. [5]. Whereas these publications as well as Chun et al. [3] discuss these challenges from the information retrieving services perspective this paper will set the focus to the information providing services perspective.

Furche et al. [4] introduced a promising automated form understanding ontology based approach, which is far beyond heuristics to fill out search forms [12], combining “[...] signals from the text, structure, and visual rendering of a web page”. But according to Li, Xian et al. in “Truth Finding on the Deep Web: Is the Problem Solved?” [9] the challenges arising from the Deep Web are regarded as not yet solved. In general, current approaches are still limited either in being domain specific or limited in their efficiency.

Until today, there still exists no general domain independent solution for the Deep Web Information Retrieval problem. Just a fraction of total available data in background databases may be covered by common state of the art approaches. This is particularly due to the fact, that for large data sets there exist nearly endless possible permutations of search results. This especially applies to the retrieval of dynamic content. Therefore, it seems to be improbable to improve retrieval and indexing mechanisms towards reaching a 100% coverage of all available Deep Web content. Consequently, this is not the focus of our current research. Currently still limited mechanisms have already “[...] succeeded largely by targeting narrow domains where a search application can be fine-tuned to query a relatively small number of databases and return highly targeted results” [14]. For that reason, we focus e.g. on the reduction of manual effort regarding the query mapping on the one hand and more precise query generation or pre-computation for the targeted retrieval from broader domains on the other hand. Therefore, this paper is intended to improve access to Deep Web content by providing great potential for new Information Retrieval mechanisms and for the significant improvement of previously existing mechanisms.

3 Approach

3.1 Research Focus

To step forward towards a *Semantic Deep Web*, which is the superordinated long-term objective, it is necessary to focus on additional research questions resulting from previously identified limitations. For the targeted retrieval especially of dynamic Deep Web content, the need of an efficient and in an ideal case fully automatic approach is essential. Therefore, the focus needs to be set to these challenges: content providing service *Discovery*, *Invocation & Execution* and *Composition*. Addressing these challenges will ensure the discovery of appropriate web query interfaces providing access to relevant content (\rightarrow *Discovery*), the appropriate query mapping and query submission (\rightarrow *Invocation & Execution*) and the service interoperability (\rightarrow *Composition*). Common approaches for Deep Web Information Retrieval focus these challenges from the information

retrieving services perspective. The conceptual idea being introduced in this section focuses these challenges from the information providing services perspective.

Common semantic annotation standards like *RDFa* and *schema.org* microdata address particularly the annotation of web content and do not have means for the prevailing *lack of semantics* at the crucial point of the Deep Web Information Retrieval process. This crucial point is regarding the web query interfaces. To improve common crawling, indexing and content retrieval mechanisms and to ensure new mechanisms, a semantic annotation for query interfaces is suggested. This will reuse the query interfaces originally intended for human users in a combined computer and human readable format. The abstract concept, to describe query interfaces in a computer readable format, is derived from the semantic annotation of *Web Services*. Standards like *Semantic Annotations for WSDL and XML Schema* (SAWSDL) provide a machine readable Web Service annotation describing the functionality and retrievable data. A semantic annotation for query interfaces will provide machine readable information for henceforth called *Semantic Deep Web Services* (SDWS).

3.2 Semantic Deep Web Service annotation

An implementation of the SDWS annotation should meet the following fundamental criteria: SDWS interface semantics, providing a generalization of SDWS interfaces (\rightarrow *abstract*) with the ability to include own vocabularies as for example thesauri (\rightarrow *extendable*). Going more into details, a SDWS annotation prototype should provide information about general properties regarding the content that is provided by the SDWS (\rightarrow *content* properties) and concrete interface field properties to describe the semantic structure and internal structural dependencies of the SDWS interfaces (\rightarrow *field* properties).

The prototype SDWS *content properties* describe the content *domain* of the retrievable information, the content *language*, as well as the content *type*. The content type attribute may be described based on *schema.org* microdata and the supplementary usage of other vocabularies. An additional content property might provide information about the amount of available data (property: *count*). These content properties are just the extendable basis for this prototype providing general information about the retrievable content. Further ideas for the extension of SDWS content properties will be discussed in the following section.

A simple example for the SDWS content properties is provided in Fig. 1, describing the basic SDWS field properties for the interface of the subject portal *EconBiz*². EconBiz provides highly relevant content for the domain of *economics* and *business studies* and access to specific content types (various types of *CreativeWork* and information about *Events*).

The prototype SDWS *field properties* describe the field *type* (e.g. *selectField*, *inputField*), as well as the input *domain* and output *range* of each particular SDWS interface field. The input *domain* attribute describes valid input values of a specified SDWS field. Furthermore, it is a trigger for the output *range* attribute,

² <http://www.econbiz.de/en/>

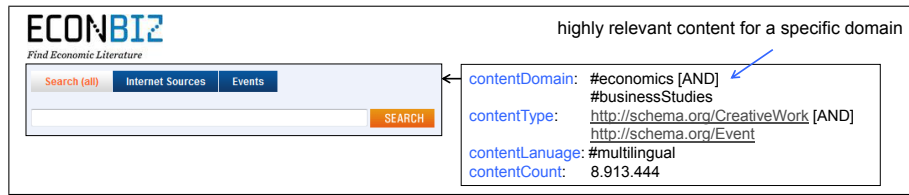


Fig. 1. SDWS content properties (example)

as its input value defines the restriction set for the retrieval process at time of form submission (examples, Fig. 2-4). Additionally, a *vocabulary* attribute may reference for instance a thesaurus that can be used as suggest-value vocabulary for the particular domain to ensure a targeted retrieval.

The basic SDWS field properties example in Fig. 2 refers to a standardized vocabulary, the *STW Thesaurus for Economics*. The STW provides "vocabulary on any economic subject" containing "[...] more than 6,000 standardized subject headings and about 19,000 entry terms to support individual keywords"³. This thesaurus is the basis for the annotation on metadata level in EconBiz and will therefore ensure a targeted retrieval. The benefit of the vocabulary property will especially apply to digital libraries but also to other domains. Therefore, regarding simple SDWS interfaces, this might be one of the most appropriate use cases for SDWS field properties as there is no complex interface structure.

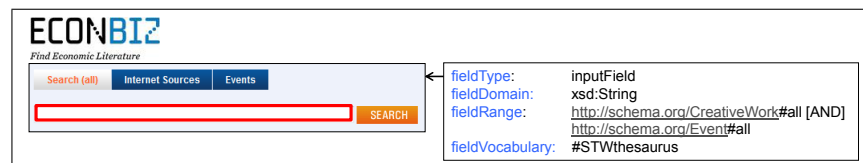


Fig. 2. SDWS field properties (basic example)

Focusing on more complex SDWS interfaces, the example in Fig. 3 contains *chunks* of related fields that affect each other. The first *selectField* as part of the marked chunk defines the relation to the other chunks. The second *selectField* as part of this chunk defines the input field domain and restricts the input field range of the *inputField* that is part of the focused chunk. Furthermore, Fig. 4 considers some exemplary effects triggered by the selection of different select values of the second *selectField* within the focused chunk in Fig. 3.

More complex examples as illustrated in Fig. 3 and 4 demonstrate that the semantic meaning behind a SDWS interface might be quite complex and automated form understanding approaches will quickly reach their limits. Especially

³ <http://zbw.eu/stw/versions/latest/about.en.html>

6 Arne Martin Klemenz, Klaus Tochtermann

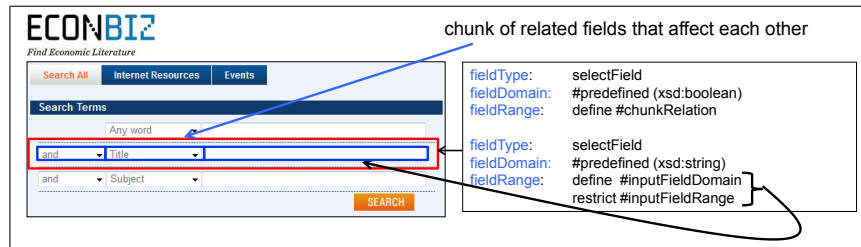


Fig. 3. SDWS field properties (chunk relation)

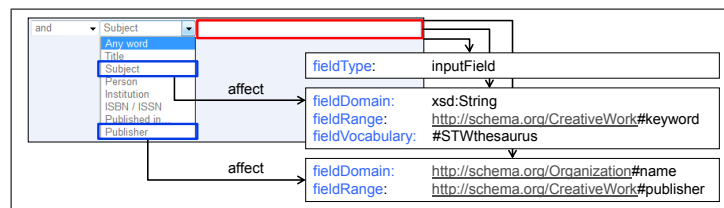


Fig. 4. SDWS field properties (triggered effects)

the automated detection of related fields and the detection of complex relations within chunks might be the most difficult part, where common approaches fail.

In addition to the SDWS interface annotation, it is advisable to link every SDWS interface from the websites root index. This will ensure a targeted SDWS discovery and can be realized by using XML Sitemaps to define a SDWS *retrieval index*. The SDWS annotation itself is suggested to be embedded directly to each particular SDWS interfaces.

4 Benefit and further Use Cases

The introduced SDWS annotation will lead to great potential for new information retrieval mechanisms and plays a significant role for the improvement of current mechanisms. Queries might for example be automatically mapped to various SDWS at the same time based on the SDWS annotation (\rightarrow *abstract semantic querying*). In accordance with Heath et al. the vision of "users [...] interacting with the Web as a data space" [7] will therefore also benefit from the introduced SDWS annotation. In general, ensuring new user oriented services especially new Knowledge Discovery services based on the large variety of available information on the Deep Web is one of the major purposes. Furthermore, the SDWS annotation might also be used for purposes, not directly focusing on the retrieval itself, as for example reasoning processes for client side query interface input validations. Another use case focuses content licensing issues as these are a problematic topic for digital libraries. These may be addressed by adding licensing information directly to the SDWS interfaces by extending the introduced annotation prototype. This will be an appropriate possibility to pro-

vide licensing information exactly at that point where the information itself is being provided e.g. based on the *Creative Commons* licensing model.

Overall, this approach will make webmasters aware of their responsibility to add SDWS annotations to SDWS interfaces in addition to current semantic content annotations. This process requires additional effort on the one hand, but on the other hand it also enables the webmasters to control the information content that may be retrieved by various retrieving services like search engines. For now webmasters may only use common HTML attributes like *nofollow* or *noindex* and the *Robots Exclusion Standard* to control the crawling behavior on their websites. The SDWS annotation ensures the targeted influence of the webmaster. Furthermore, only the webmaster knows the exact semantic statement intended by the implemented SDWS interface. Regarding web content, search engines rely on semantic content markups as it is more reliable than current automatic content interpretation approaches. Therefore, it is obvious, that this will also apply to the annotation of SDWS interfaces.

5 Conclusion

This paper addressed the lack of semantic information regarding web query interfaces in the process of Information Retrieval from the Deep Web. Transferring the concepts of semantic web content annotations on the one hand and Semantic Web Service Descriptions on the other hand, leads to the great potential of semantic annotations for SDWS interfaces. Equivalent to semantic web content annotations, the SDWS annotation provides an unambiguous semantic interpretation of the SDWS interface. A variety of current information retrieval mechanisms and form understanding systems try to analyze SDWS interfaces automatically by focusing the Deep Web Information Retrieval challenge from the retrieving services perspective. Instead of relying on these, the introduced SDWS interface annotation is focusing this challenge from the information providing services perspective.

In general, this approach follows the open knowledge sharing paradigm as part of the *Semantic Web* vision from Berners-Lee et al. [2]. This is based on the assumption, that the information provided on websites is intended to be retrieved by various services. Any additional licensing issues restricting the retrieval and further usage of the retrievable information have also been addressed.

This approach will contribute to domain independent and automatic Information Retrieval mechanisms based on the introduced SDWS annotation. Manual effort for currently still limited Deep Web Information Retrieval mechanisms will be reduced or even eliminated. Furthermore, these retrieval mechanisms will benefit regarding their efficiency and can be adapted targeting broader domains.

6 Future Work

Future work will especially focus on the critical evaluation based on further research studies. The definition of a concrete SDWS annotation syntax based

on the usage of existing annotation standards will concern the challenge how retrieving services will learn to understand the SDWS annotation. Reduction of manual effort for the annotation process also requires further effort. A semi-automatic generation process may provide support for the definition of SDWS annotations. This process may be based on sampling and probing the background database utilizing promising automated form understanding approaches. This may lead to semi-automatic generation approaches for the SDWS annotation.

References

1. BERGMAN, M. K. White paper: The deep web: Surfacing hidden value. *the journal of electronic publishing* 7, 1 (2001).
2. BERNERS-LEE, T., HENDLER, J., LASSILA, O., ET AL. The semantic web. *Scientific American* 284, 5 (2001), 28–37.
3. CHUN, S. A., AND WARNER, J. Semantic annotation and search for deep web services. In *E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008 10th IEEE Conference on* (2008), IEEE, pp. 389–395.
4. FURCHE, T., GOTTLÖB, G., GRASSO, G., GUO, X., ORSI, G., AND SCHALLHART, C. Opal: Automated form understanding for the deep web. In *Proceedings of the 21st international conference on World Wide Web* (2012), ACM, pp. 829–838.
5. HANDSCHUH, S., AND STAAB, S. *Annotation for the semantic web*, vol. 96. IOS Press, 2003.
6. HE, B., PATEL, M., ZHANG, Z., AND CHANG, K. C.-C. Accessing the deep web. *Communications of the ACM* 50, 5 (2007), 94–101.
7. HEATH, T., AND BIZER, C. Semantic annotation and retrieval: Web of data. *Handbook of Semantic Web Technologies* (2011).
8. HICKS, C., SCHEFFER, M., NGU, A. H., AND SHENG, Q. Z. Discovery and cataloging of deep web sources. In *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on* (2012), IEEE, pp. 224–230.
9. LI, X., DONG, X. L., LYONS, K., MENG, W., AND SRIVASTAVA, D. Truth finding on the deep web: Is the problem solved? In *Proceedings of the 39th international conference on Very Large Data Bases* (2012), VLDB Endowment, pp. 97–108.
10. MADHAVAN, J., AFANASIEV, L., ANTOVA, L., AND HALEVY, A. Harnessing the deep web: Present and future. *4th Biennial Conference on Innovative Data Systems Research (CIDR)* (Jan. 2009).
11. MADHAVAN, J., KO, D., KOT, L., GANAPATHY, V., RASMUSSEN, A., AND HALEVY, A. Google’s deep web crawl. *Proceedings of the VLDB Endowment* 1, 2 (2008), 1241–1252.
12. MASANÉS, J. Archiving the hidden web. In *Web Archiving*. Springer, 2006, pp. 115–129.
13. MUKHERJEA, S. Information retrieval and knowledge discovery utilising a biomedical semantic web. *Briefings in Bioinformatics* 6, 3 (2005), 252–262.
14. OGRAPH, T., AMANCA, Y., AND MAAHS, Y. Searching the deep web. *Communications of the ACM* 51, 10 (2008).
15. WENYU, Z., JIANWEI, Y., MING, C., JIAN, W., AND LANFEN, L. Manufacturing deep web service management: Exploring semantic web technologies. *Industrial Electronics Magazine, IEEE* 6, 2 (2012), 38–51.