The Fifth European Semantic Web Conference

1-5 June 2008, Tenerife, Spain

ESWC 2008

Workshop Proceedings

"SIEDL 2008: Semantic Interoperability in the European Digital Library"

Stefanos Kollias & Jill Cousins (eds.)

Semantic Interoperability in the European Digital Library

Proceedings of the First International Workshop, SIEDL 2008
Tenerife, June 2, 2008

Sponsors:

European Digital Library, http://europeana.eu

Preface

One of Europe's current most important initiatives is the construction of the European Digital Library (EDL), which aims at making Europe's diverse cultural and scientific heritage (books, films, maps, photographs, music, etc.) easy to access and use for work, leisure, or study. EDL will take advantage of existing rich European heritage, including multicultural and multilingual environments, of technological advances and of new business models. It will generate a common multilingual access point to Europe's distributed digital cultural heritage, including all types of multimedia cultural content, and all types of cultural institutions, including libraries, museums, archives.

The short-term objective of the European Digital Library is to create a prototype within 2008, providing multilingual access to over 2 million digital objects, focusing on libraries, while including as many museums and archives it is possible. The long-term objective, for 2010 is to increase the available digital content to over 6 million digital objects from all types of institutions, i.e., libraries, museums, audiovisual organisations, archives.

The EDLnet Thematic Network is the project approved by the European Commission's eContentPlus programme to prepare the ground for the realisation of the European Digital Library. Consistent with this vision about the European Digital Library, the project addresses particularly the area of improving cross-domain accessibility to cultural content- a pillar of the European Commission's i2010 Digital Library initiative. EDLnet tackles the fragmented European cultural heritage map, by bringing on board the key European stakeholders to build consensus on creating the European Digital Library. In this framework, interoperability has been defined as one of the most crucial issues, with a specific EDLnet workpackage being devoted to it. Semantic interoperability is one of the related key technological issues, and for this reason it constitutes the topic of the Workshop.

Several definitions of semantic interoperability have been proposed in the literature, covering different application domains. In this Workshop, we focus on how the late advances on Semantic Web technologies can facilitate the way that European digital libraries exchange information within the framework of the web. The key in the definition of semantic interoperability is the common automatic interpretation of the meaning of the exchanged information, i.e. the ability to automatically process the information in a machine-understandable manner.

June 2008

Vassilis Tzouvaras,
Program Chair, SIEDL 2008

The SIEDL Workshop

General Chairs

Stefanos Kollias - National Technical University of Athens (NTUA), Greece
Jill Cousins - Koninklijke Bibliotheek (KB), The Netherlands

Program Chair

Vassilis Tzouvaras - National Technical University of Athens (NTUA), Greece

Program Committee

Stefan Gradmann, University of Hamburg, Germany
Carlo Meghini, Consiglio Nazionale delle Ricerche (CNR), Itally
Guus Schreiber, Free University Amsterdam, the Netherlands
Jacco van Ossenbruggen, Centrum voor Wiskunde en Informatica (CWI), The Netherlands
Antoine Isaac, Free University Amsterdam, the Netherlands
Miles Alistair, e-Science Centre, UK
Carole Goble, University of Manchester, UK
Giorgos Stamou, National Technical University of Athens, Greece
Yannis Ioannidis, National and Kapodistrian University of Athens, Greece
Stavros Christodoulakis, Technical University of Crete, Greece
Lora Aroyo, Free University Amsterdam, The Netherlands
Eero Hyvonen, Helsinki University of Technology, Finaland
Johan Oomen, Sound and Vision, The Netherlands
Emmanuelle Bermes, Bibliotheque nationale de France, France
Jeff Z. Pan, University of Aberdeen, UK

Supporters

The European Digital Library (EDL), http://europeana.eu/
MinervaEC, http://www.minervaeurope.org/about/minervaec.htm
The Video Active Project, http://videoactive.wordpress.com/
Semantic Web, http://www.semanticweb.org
European Semantic Web Conference 2008, www.eswc08.org

Table of Contents

Invited Talks

Accepted Papers

It's the semantics, stupid! Remarks on the strategic potential of semantic foundations for Europeana

Prof. Stefan Gradmann[1]

Library and Information Science at Humboldt-University in Berlin
stefan.gradmann@RRZ.UNI-HAMBURG.DE,
WWW home page: http://www1.uni-hamburg.de/RRZ/S.Gradmann/

Short biography

Dr. Stefan Gradmann, Professor of Library and Information Science at Humboldt-University in Berlin with a focus on knowledge management and semantics based operations. He studied Greek, philosophy and German literature in Paris and Freiburg (Brsg.) and received a Ph.D in Freiburg in 1986. After additional training in scientific librarianship (Cologne, 1987-1988) he worked as scientific librarian at the State and University Library in Hamburg from 1988-199image2. From 1992-1997 he was the director of the GBV Library Network. 1997-2000 he was employed by Pica B.V. in Leiden as product manager and senior consultant. Form 2000 to March 2008, he was Deputy Director of the University of Hamburg Regional Computing Center. He was the Project Director of the GAP (German Academic Publishers) Project of the German Research Association and technical co-ordinator of the EC funded FIGARO project. He was an international advisor for the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences and as such has contributed to the report Our Cultural Commonwealth (http://www.acls.org/cyber infrastructure/OurCulturalCommonwealth.pdf) Stefan currently is heavily involved in building Europeana, the European Digital Library, and more specifically is leading WP2 on technical and semantic interoperability as part of the EDLnet project.

Abstract

The presentation will start from a short overview of Europeana's intended functional and technical architecture with articular attention given to its semantic foundation layer and then explore some of the opportunities to be expected from such an approach for European citizens and scientists with particular respect to the social sciences and humanities (SSH). Working with SSH scholars in turn could help computer science to add some hermeneutical foundations to the Semantic Web framework, as will be shown in a concluding presentation of the Scholarsource project proposal.

Improving Berrypicking Techniques Using Semantically Similar Information in a Historical Digital Library System

Ismail Fahmi[1], Henk Ellermann[1], Peter Scholing[1], and Junte Zhang[2]

[1] University of Groningen, University Library and Information Science Department
[2] University of Amsterdam, Archives and Information Studies

{i.fahmi,h.h.ellermann}@rug.nl,peter@scholing.net,j.zhang@uva.nl

Abstract. First, we describe the inadequacies of the classic information retrieval model and discuss the proposed model called the 'berrypicking model', which is closer to users' searching behavior. Second, considering the advantages of the Semantic Web standards, we propose a method to improve the berrypicking technique by adding semantically similar information into the berrypicking processes of a digital library system. We map historical metadata in the form of bibliographic finding aids into an ontology, and use the properties of an ontology's instances for measuring the similarity information. Third, to overcome the operation time and search capability problems faced by the implemented ontology storage and query system, we describe our strategy in indexing and searching the ontology using an open source information retrieval system. Finally, we provide the results of our evaluation.

Key words: berrypicking, semantic web, ontology, information retrieval, metadata, finding aids, digital library, history

1 Introduction

Metadata and other forms of annotated data are the foundation for both traditional and digital libraries. Because of the huge amount of such metadata that has been gathered over many years, especially digital libraries could be a successful primary adopter of Semantic Web technologies [10, 6]. Semantic Web technologies seem ideally suited to improve and widen the services that digital libraries offer to their users. Digital libraries rely heavily on information retrieval technology. Semantic web might be used to introduce meaningful and explicit relations between documents, based on their content, thereby allowing services that introduce forms of semantic browsing supplementing, or possibly replacing keyword-based searches. As we will discuss later, the element of browsing might be needed to better adapt search tools to the search behavior of information seekers.

The are several implicit limitations of relying on keywords as the basis vehicle for searching in (library) data: it can exclude works in different languages or from

different historical periods in which people use differently spelled words, or even different words to describe a certain phenomenon. Recall, to use the classical Information Retrieval terms, is there not optimized: not all relevant documents are found in such cases. The opposite can also be true: one and the same word can have different meanings in different contexts. Searching for an ambiguous keyword can return a lot of irrelevant items. In this case, precision is negatively influenced.

In this paper we present our work in improving the information retrieval technique of our digital library system SWHi [6]. Our primary users are historians with an interest in American history. The main data source are the bibliographic finding aids (metadata) of the Early American Imprints[3]. We use semantic web standards to extract semantically similar information on this large set of documents and formalize them in the SWHi ontology. This ontology was derived, not just automatically, from the original metadata (in the MARC format). For evaluating the ontology and the system, we have defined three similarity measures and have compared this measure to the judgement of four historians (representing the target users).

2 Related Work

The traditional focus of Information Retrieval has always been on keyword based searching or ad-hoc retrieval instead of on browsing. This traditional model of information retrieval is the basis for most (archival) information systems in use today [11]. This is also the primary means of locating information on the Web by using a search engine. It also serves as the basis of many theories and models: one of the most-frequently used and cited information-seeking paradigms, Ben Shneiderman's Visual Information-Seeking Mantra too makes use of the traditional IR model as its underlying foundation [14]. But this model, deeply rooted in Information Retrieval theory, has a number shortcomings. It does not seem to conform to real practice, to the way many users would like to seek information: 'real-life searches frequently do not work this way' [2].

Studies show that there is a clear division in the scholarly community between scholars from the 'hard' sciences, who prefer searching by query formulation, and scholars from the humanities and social sciences, who show a preference for more informal methods of information-seeking [4, 5, 3]. This model also does not take into account that many users dislike being confronted with a long disorganized list of retrieval results that do not directly address their information needs [8].

Enhancing information search in the Cultural Heritage domain is an active area for Semantic Web researchers (e.g. [15, 16, 1]), and related to our domain of historical research in digital libraries, because both involve scholars and others users in the Humanities. In [15] it is described how semantic interoperability between two collections of a museum can be achieved by aligning controlled vocabularies, which makes it possible to browse in the collections using facets.

[3] Early American Imprints, Series I: Evans, 1639-1800. Contains books, pamphlets, broadsides and other imprints listed in the renowned bibliography by Charles Evans.

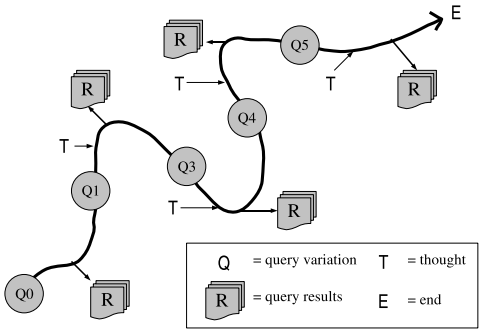


Fig. 1. Bates' Berrypicking Information Retrieval Model. Adapted from [2].

The work of [16] goes further with this idea by aligning more vocabularies and collections from multiple sources, where keyword-based search is extended with semantic clustering of RDF graphs to make use of the background knowledge from the underlying thesauri and taxonomies. The research of [1] builds on this framework as well. They demonstrate how to provide a personalized experience for the museum visitors both on the Web site and in the museum using recommendations and tour wizards driven by users profiles based on explicit user feedback, i.e. (manual) user assessments of art objects.

Although Semantic Web research in the Cultural Heritage domain related to information seeking deals with building an alternative for the traditional IR model, there is no specific implementation or evaluation yet that relates it to the theory of the Berrypicking Model. In 1989, Marcia Bates came up with the Berrypicking Model as shown in Figure 1. The intention of this model was to be 'much closer to the real behavior of information searchers than the traditional model of information retrieval is, and, consequently will guide our thinking better in the design of effective interfaces' [2].

Berrypicking is named after the process of picking blueberries or huckleberries from bushes, in the forest. The berries are scattered on the bushes; they do not grow in bunches, and can therefore only be picked one at a time. This one-at-a-time collecting of interesting pieces of information is the basis of the Berrypicking model. Evolving Search is a second key element in this model, meaning that the results found often co-determine the relevancy of new results: it plays an important role in the Berrypicking Model[2].

Our goal is to cater for search behavior that follows the Berrypicking Model more closely than what we have just named the traditional model of information retrieval. When a document is viewed by a user, we provide him or her with a context of other similar documents.

The selection of semantically similar objects based on ontology has been reported in [12], which extend the *tf-idf* method by involving essential features of the domain ontologies and RDF(S) languages. In the present paper, we use

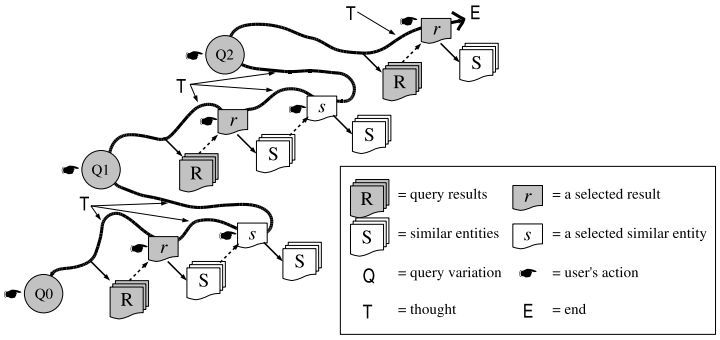


Fig. 2. The berrypicking technique improved with similar information.

a similarity measurement method described in [7], using the SWHi ontology as the source, and compare it with two simpler methods.

3 Method

Our approach to improve the berrypicking technique is based on the way a historian seeks information. We use an example in Cole's study [4], where a history student used the following steps to get required information: proper name compilation, reformulation of search queries, anomaly and pattern discovery, understanding the pattern, and research reformulation reflecting the new-found pattern. These steps demonstrate that the historian's information-seeking centers around two notions that are important to achieve their goals: name-collecting and pattern discovery.

3.1 Improving the Berrypicking Technique

The berrypicking process experienced by the user is expected to be improved as shown in Figure 2. During the evolving search, the system will recommend a set of objects S which are similar to the result r being viewed by the user. And by scanning these objects, which may consist of related documents, persons, organizations, locations, and events, the user will learn (T) the context or the related context of the viewed result r. If he views one of the recommended object s, he will also get other related objects. This iterative process may turn up new findings, which allow the user to keep browsing the suggested entities to get further context or to construct a new query based on the developed contextual knowledge.

The berrypicking model is a combination between searching and browsing activities. For a user who already knows what he wants to find, searching is the most suitable activity. With initial contextual knowledge, this will be the most efficient search strategy. For those who have less contextual knowledge, browsing will probably be a a better method because it shows existing and related information in the repository. However, the information might not be relevant to their information needs, especially if the pieces of information are returned

based on categories which are not of their interests. To overcome this problem, the browsing method in this model is aimed at recommending information related to only the viewed information. If the viewed information is of user's interest, then the user will get other information similar to what he needs. This in turn will help in developing his contextual knowledge. The recommended results may give benefits to both the experienced and the less experienced users.

3.2 Measuring Semantic Similarity

Finding similar information is process of pattern discovery, not unlike the search process in our example above. The pattern here will be formalized as a set of properties contained in one item. When more items are simultaneously considered, the pattern is a set of properties shared by all items. As an example of properties shared, if most of these objects are related to the following *subject* properties: *Pennsylvania, History, Revolution*, and *1775-1783*; then we can use these properties (i.e. the pattern) to extract similar items by looking at the *subject* properties of other documents which share the same set of properties.

Sources of Similarity Depending on the user's purposes, the pattern can be based on one of the following sources:

A piece of berry This refers to one item found after search. It has a pattern. This pattern can be used to find items with similar patterns that will function as a context for further search. For example, to a document being viewed the system may return persons, organizations, locations or other documents that share some of the properties. From these objects, the user will understand that, for example, there are some people related to that document in particular locations.

A basket of berries It is a global similarity pattern that common to all of the objects in the basket (e.g. in a bookmark), which is useful in constructing user's contextual knowledge over their collected information.

Considering that our recommendation system in the present time is just focused on the local similarity, we leave the discussion of the global similarity method for the future work.

Definitions The similarity of two objects is defined in [7] as "the ratio of the shared information between the objects to the total information about both objects." This definition is based on an intuition that similar objects share more common properties. Based on this definition, the similarity of objects a (source object) and b (target similar object) can be measured.

Beforehand, we introduce the term *description set* of an object a, which is defined in [7] as,

$$desc(\mathbf{a}) = (\mathbf{a}, p, o) \in O \tag{1}$$

where a is an object in an ontology O, $desc(a)$ is a set of all RDF[4] triples (a, p, o) in O with a as the subject of the triple and with any predicate p and object o. A set of descriptions of a can be seen as a set of properties and their values of a.

[4] Resource Description Framework (RDF) - `http://www.w3.org/RDF/`

From two objects a and b, we can define a *shared description* set as follows,

$$sharedDesc(a, b) = desc(a) \cap desc(b) \tag{2}$$

which is a subset of triples in a whose (p, o) pairs are found in b.

And the total information contained in both objects a and b, called a *collected description* set, is defined as,

$$collectedDesc(a, b) = desc(a) \cup desc(b) \tag{3}$$

whose total number is computed from the set of unique (p, o) pairs collected from a and b.

Semantic Similarity Functions To measure the semantic similarity of a target object b to a source object a, we experiment with three methods. The first method, which is the baseline, is computed based on the number of shared triple's objects between a and b, regardless the predicates of the triples. We define this baseline as,

$$baseSim(a, b) = f_{object}(a, b) \tag{4}$$

The second method is calculated based on the number of the shared descriptions between the two objects, regardless the number of their total descriptions. This method is formally defined as,

$$sharedSim(a, b) = f_{shared}(a, b) \tag{5}$$

where f_{shared} is the shared function measuring the the number of information values in a's description set that are shared by b's description set. This function is computed using equation 2.

For the third method, we follow the similarity definition in [7], which can be formulated as,

$$ratioSim(a, b) = \frac{f_{shared}(a, b)}{f_{collection}(a, b)} \tag{6}$$

where $f_{collection}(a, b)$ is the collection function providing the number of total unique description contained in a and b. This function is computed using equation 3.

3.3 Retrieving Similar Object Candidates and Their Description Sets

Having a description set of a, we send a query to an ontology repository system to get a set of target object candidates. To accommodate the baseline function, we do not restrict the query to match with *predicate-object* pairs, instead to match with at least one of the *object*s of some discriminating *predicate*s, such as **type, subject, language, date, interest, knows, exists, involvedIn**, and **locatedIn**. This query results in a set of target object candidates along with their description sets.

4 Resources

4.1 Metadata

Our system is focussed on the Early American Imprints, Series I: Evans, 1639-1800. This source contains all published works of the 17th- and 18th-century America. Its bibliographic finding aids consist of 36,305 records, which are elaborately described (title, author, publication date, etc) with numerous values, and have been compiled by librarians in the format MARC21.

4.2 Metadata-to-Ontology Mapping

We do not convert the bibliographic finding aids encoded in MARC21 directly one-to-one to RDF, which would be trivial, but try to use extra knowledge, links and descriptives provided by different resources and align them together comprehensively. The key classes and relations of the SWHi ontology used for the "berrypicking" are outlined in [17]. A plethora of existing resources can be reused and merged to create a single historical ontology. We have decided to use an existing general ontology, namely PROTON[5], as the base, and modify and enrich it with existing historical resources and vocabularies like Dublin Core and FOAF. In addition to the PROTON's basic ontology modules, we get 152 new classes from this mapping. And in total, we get 112,300 ontology instances from the metadata. The motivation for reusing PROTON is that in the future we could also implement the ontology for other knowledge domains in our digital library. We found that this ontology is certainly sufficient for our data and specific purposes. However, in the future we will migrate to the CIDOC Conceptual Reference Model (CRM), which is an international (ISO) standard that provides an extensible ontology for concepts and information in archives, libraries, and museums. This makes it better possible not only to use our ontology consistently within our library, but we can achieve global interoperability with other institutions as well.

5 Implementation and Evaluation

5.1 Storing and Indexing the Ontology

We use the Sesame RDF Framework[6] for storing and querying the ontology. Besides the functionality of the system, we also aim at developing responsive system in term of the operation time since it is very crucial for a real-life application. In our experiments, getting information directly from an ontology through inference could cost an unacceptable processing time, especially if the inference queries are complex or repeated many times. For this purpose, we retrieve from that ontology all instances of some selected classes to be indexed using the Solr,[7] an open source information retrieval system.

[5] PROTON Ontology `http://proton.semanticweb.org/`

[6] Sesame `http://www.openrdf.org/`

[7] Solr is an enterprise search based on the Lucene Java. `http://lucene.apache.org/solr/`

There are 7 main classes in ontology, that are indexed, namely **protont: Document** (35,832 instances), **protont:Person** (23,688 instances), **swhi: Subject** (12,828 instances), **protont:TimeInterval** (4,042 instances), **protont: Location** (1,787 instances), **protont:Organization** (1,605 instances), and **protont:Event** (296 instances). The **protont:Document** class has the largest number of instances since it represents the set of records in the metadata. Among the extracted-named-entity classes, **protont:Person** is the largest one since it consists of authors and mentioned persons in the metadata records.

5.2 User Interfaces

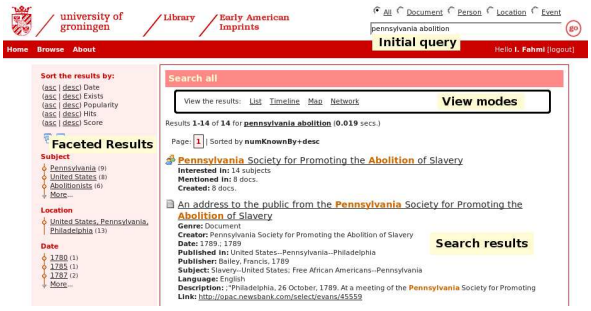
The method to measure similarity above has been implemented in the recommendation module of our SWHi system. Figure 3 illustrates the user interfaces of SWHi, that support the berrypicking process. In the sub-figure 3(a), a user starts with an initial query term *pennsylvania abolition* on any concept type. The results are presented in two ways: according to the results' facets (e.g. subject, location, date) and according to a preferred view mode (i.e. plain, timeline, map, or network graph). Then, the sub-figure 3(b) shows that the user has walked further in collecting berries from the viewed objects and from the suggested results into his 'berry basket'. As shown in that figure, the user has even not submitted a new query term to get related information. The similar document being viewed may not contain the query term (e.g. *abolition*), however it may contain related information (e.g. about *slavery*). Thus, in addition to the search, browsing the recommended documents may also turn up relevant information that could fulfill user's needs.

To help the user to understand a pattern that may be constructed from the collected information, the system shows a relationship network as presented in sub-figure 3(c). In this figure, 4 *persons* and 6 *documents* are linked to each other directly or via their shared properties, e.g. *subjects*. The user may quickly grab a pattern such as how close the relationship between *Thomas Jefferson* and *Benjamin Banneker* is when compared to with the other persons in the network.

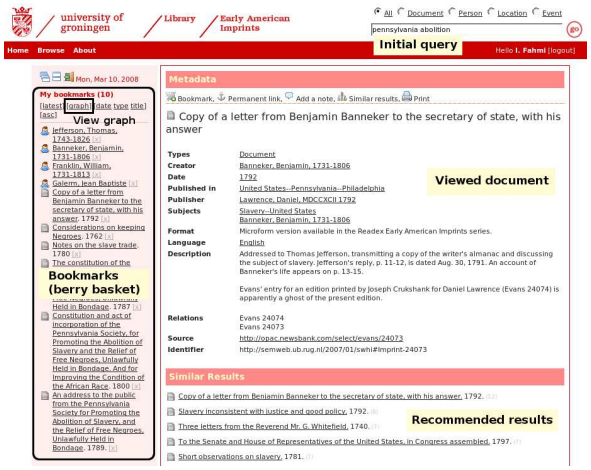
5.3 Evaluation Method

Our evaluation is aimed at answering the research question: which semantic similarity ranking method is the most correspond to the end-user's judgment? In the evaluation, the order of the documents is actually very important since normally a user will look at the suggested documents starting from the highest rank. However, a preliminary study in [12] indicated that the ordering evaluation was difficult for the user, while on the other hand, a binary classification of a document was rather easy. Therefore, we ask the user to judge if a document in a subset of recommended documents is semantically similar to the source document.

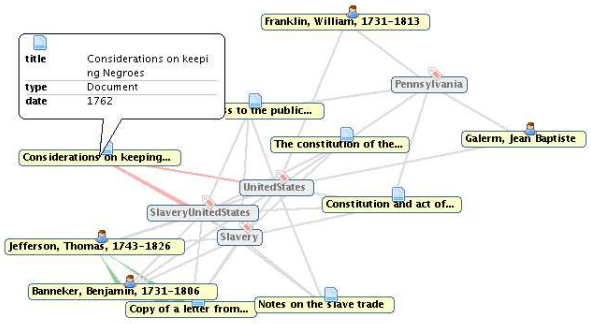
The precision of each suggestion set is calculated using the precision (P) and the uninterpolated average precision (UAP) methods [13]. We also use the UAP method, because, besides considering the precision, it also considers the order of the documents in the ranking set. The UAP of a set of K relevant documents is defined as,



(a) Faceted and multi-view search results



(b) Bookmarks and suggestions



(c) Network graph

Fig. 3. The user interfaces of the SWHi system showing (a) search results, (b) a metadata (viewed document), its similar results (recommended results), and user's bookmarks, and (c) the pattern (network graph) of the relationship between objects in the bookmarks. URL: http://evans.ub.rug.nl.

$$UAP(K) = \frac{1}{K} \sum_{i=1}^{K} P_i \tag{7}$$

where P_i (precision at a rank position i) equals i/H_i, and H_i is the number of hypothesized documents required to find the i^{th} relevant document.

5.4 Dataset

Four test users with historical knowledge background were involved in the evaluation. In the first step, a test user was asked to submit 6 search queries of his interests to the SWHi system, using the user interface as shown in Figure 3(b). For each search query, the test user selected from the results the most relevant source document according to his judgement. For each of the viewed source document, the system suggested 20 similar results according to each of the three similarity measurement methods. Thus, in total, there are 60 suggestions for each viewed source document, or 360 suggestions for all search queries. In the second step, we asked the four test users to judge whether each suggestion is relevant to the source document or not. For the general evaluation, a suggested document is considered as relevant if all of the test users judged it as relevant. The inter-rater reliability was measured using the Fleiss' kappa measure [9] resulted in $\kappa = 0.361$, which is regarded as reliable.

5.5 Results

Table 1 presents the precision (P) and the uninterpolated average precision (UAP) values of the three similarity measures given each search query. A value of P is calculated based on the first 20 similar results, while a value of UAP is calculated based on the first K true similar results for each viewed source document. We grouped the queries into two groups according to the coverage of documents matching the query strings. The second group containing two queries ('west india' and 'florida') has a small article coverage in the collection.

No	**Query**	**P**			**UAP**			
		BS	SS	RS	K	BS	SS	RS
1	The Coercive Acts	0.53	0.63	0.74	10	0.73	0.75	0.96
2	Articles of Confederation [...]	0.79	0.79	0.84	15	0.91	0.91	0.79
3	Saratoga	0.68	0.68	0.95	13	0.77	0.77	0.99
4	Presbyterian	0.53	0.58	0.68	10	0.60	0.59	0.65
	Average	0.63	0.67	**0.80**		0.75	0.76	**0.85**
5	west indies	0.47	0.47	0.37	7	0.57	0.62	0.75
6	florida	0.21	0.21	-	4	0.95	0.95	-
	Average	0.34	0.34	-		0.76	**0.79**	-

Table 1. The precision (P) and the uninterpolated average precision (UAP) of the similarity methods (*BS, SS, RS*) on each viewed source document matched with the query.

In the first group, where the topics of the queries are widely covered in the collection, all of the method show similar performance with at least 10 of the 20 suggested results are judged as relevant by the test users. In both of the evaluation methods, SS was the second best method with 67% of P and 76% of UAP, which were slight improvements to the BS performance, and RS shows the best performance with 80% of P and 85% of UAP. These results show that the shared *predicate-object* pairs are good indicators for measuring the similarity of two documents, and when the collected *predicate-object* pairs are involved in the measurement, the performance will be significantly increased. The higher values of UAP by RS show that most of relevant documents are ranked higher.

A similar result is shown in the second group, where SS demonstrates a better UAP compared to BS. However, RS only returns 1 true similar-result for the query 'florida', which results in an empty UAP value for the $K = 4$. This low of recall is probably caused by the denominator of the RS formula which requires the total number of descriptions of both the source document and the similar document to be as small as possible in order to get a better ratio value. Therefore, similar documents with small numbers of descriptions will appear in the top of the ranks, while those with higher numbers of shared descriptions but have significantly more descriptions will be descended.

At the end of the evaluation, there are two interesting remarks from the test users. First, when picking a rather specific document from the search results, even when the search query itself is not very specific, the test users found the suggested similar results to contain a real interesting set of related documents. Second, a test user was "impressed with the results and thought that such a system would be an interesting tool for historians when digging in large primary-source corpora."

6 Conclusion and Future Work

We have presented our approach in improving the Berrypicking technique by adding semantically similar information to the viewed document. We have compared three similarity measurement methods and evaluated their performance using precision (P) and uninterpolated average precision (UAP) methods. It turned out that the ratio score (RS) methos significantly outperforms the baseline. In some cases, assigning a denominator with the number of collected descriptions, as in the ratio score (RS), may harm the performance. The evaluation also shows that, if the ranking of the results is matter, UAP can assign a better precision score to a method which better ranks the results.

As for the future work, we would investigate whether selecting shared descriptions based on a set of selected *predictes* will improve the performance. We also would like to extend the similarity measurement to cover the whole 'picked berries in basket.'

References

1. Lora Aroyo, Natalia Stash, Yiwen Wang, Peter Gorgels, and Lloyd Rutledge. Chip demonstrator: Semantics-driven recommendations and museum tour generation.

In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007), LNCS 4825*, pages 879–886. Springer, 2007.
2. M. J. Bates. The design for browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–431, 1989.
3. M. J. Bates. Berrypicking. *American Society for Information Science and Technology, Medford, NJ, USA*, pages 58–62, 2005.
4. C. Cole. Inducing expertise in history doctoral students via enabling information retrieval design. *Library Quarterly*, 70(1):444–455, 2000.
5. W. M. Duff and C. A. Johson. Accidentally found on purpose: Information-seeking behaviour of historians in archives. *Library Quarterly*, 72(4):472–496, 2002.
6. Ismail Fahmi, Junte Zhang, Henk Ellermann, and Gosse Bouma. SWHi system description: A case study in information retrieval, inference, and visualization in the semantic web. In *Proceedings of European Semantic Web Conference (ESWC) 2007, LNCS 4519*, pages 769–778. Springer, 2007.
7. J. Hau, W. Lee, and J. Darlington. A semantic similarity measure for semantic web services. In *Workshop of WWW2005, Web Service Semantics: Towards Dynamic Business Integration*, 2005.
8. M. A. Hearst. *User interfaces and visualization*, pages 257–324. Modern Information Retrieval. ACM Press, New York, NY, USA, 1999.
9. Jason E. King. Software solutions for obtaining a kappa-type statistic. In *Annual Meeting of the Southwest Educational Research Association, Dallas, Texas*, 2004.
10. E Miller. Digital libraries and the semantic web. http://www.w3.org/2001/09/06-ecdl/slide1-0.html, 2001. Accessed 11 December 2006.
11. E. M. Rasmussen. *Libraries and bibliographical systems*, pages 397–413. Modern Information Retrieval. ACM Press, New York, NY, USA, 1999.
12. Tuukka Ruotsalo and Eero Hyvönen. A method for determining ontology-based semantic relevance. In *DEXA*, pages 680–688, 2007.
13. Patrick Schone and Daniel Jurafsky. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In Lillian Lee and Donna Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108, 2001.
14. Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages*, pages 336–343, College Park, Maryland 20742, U.S.A., 1996.
15. Marjolein van Gendt, Antoine Isaac, Lourens van der Meij, and Stefan Schlobach. Semantic Web Techniques for Multiple Views on Heterogeneous Collections: A Case Study. In *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), LNCS 4172*, pages 426–437. Springer, 2006.
16. Jacco van Ossenbruggen, Alia Amin, Lynda Hardman, Michiel Hildebrand, Mark van Assem, Borys Omelayenko, Guus Schreiber, Anna Tordai, Victor de Boer, Bob Wielinga, Jan Wielemaker, Marco de Niet, Jos Taekema, Marie-France van Orsouw, and Annemiek Teesing. Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques. In J. Trant and D. Bearman, editors, *Proceedings of Museums and the Web 2007*, San Francisco and California, March 2007. Archives & Museum Informatics.
17. Junte Zhang, Ismail Fahmi, Henk Ellermann, and Gosse Bouma. Mapping metadata for SWHi: Aligning schemas with library metadata for a historical ontology. In *Proceedings of the International Workshop on Collaborative Knowledge Management for Web Information Systems (WEKnow'07), LNCS 4832*, pages 103–114. Springer, 2007.

Porting Cultural Repositories to the Semantic Web

Borys Omelayenko

Vrije Universiteit Amsterdam, the Netherlands
b.omelayenko@cs.vu.nl,
WWW: http://borys.name

Abstract. In the eCulture project we ported a dozen datasets of different cultural institutions to the Semantic Web. In this paper we share our experiences: we sketch the technical data conversion problem, describe the conversion rules that were needed, and the methods that we used to align terms to vocabularies. We wrap it up with the statistics that give some insight on practical conversion cost and its success rate.

1 Problem

Let us look at the eCulture [Schreiber et al., 2006] demonstrator,[1] a prototypical Semantic Web application that won the Semantic Web challenge award at ISWC 2006. eCulture is a semantic search engine that allows simultaneously searching collections of multiple cultural heritage institutions. This is done by porting these collections to RDF[2] and linking collection instance objects via shared vocabularies, thus, building a large RDF graph. Then, during the search, this graph is traced and some subgraphs are extracted and returned as a result [Wielemaker et al., 2007]. In this paper we will focus on the problem of porting collections to RDF referring the reader to the demonstrator to see what can be done with it.

In eCulture we developed a methodology for porting cultural repositories to Semantic Web and RDF [Tordai et al., 2007]. This methodology is based on the fact that we typically can expect two kinds of data from a cultural heritage institution:

- meta-data describing cultural objects and their photos,
- local vocabularies that are used in some of these meta-data descriptions.

The institutions tend to use different formats and different data models to represent their collections and vocabularies, where databases, XML files, or even tab-separated plain text files are common choices.

Following the methodology, we first transform the schema of both meta-data and vocabularies to standardized RDF schemas, where Dublin Core[3] is used to describe the meta-data and SKOS[4] is used to represent the vocabularies. Then,

[1] e-culture.multimedian.nl
[2] http://www.w3.org/RDF/
[3] http://dublincore.org/
[4] http://www.w3.org/2004/02/skos/

we align local vocabulary terms with their counterparts from some standard vocabularies. Finally, we process the meta-data values to extract possible terms eventually used there, and align these terms need to the standard terms as well.

Similar process has been used to MuseumFinland project where collections of Finnish museums were ported to the Semantic Web [Hyvyonen et al., 2005]. In that project additional effort has been put on creating new terminologies aiming at the creation of a national cultural heritage ontology. In eCulture we do not develop any terminology focusing on the integration of the existing ones.

There exist a number of large standard vocabularies known and watched in the cultural heritage domain. These are the Getty[5] vocabularies: AAT (Art and Architecture Thesaurus of 34,000 concepts and 131,000 terms), ULAN (the Union List of Artist Names with 120,000 records and nearly 300,000 names), TGN (the Thesaurus of Geographic Names describing more than a million places). In addition, museums working in a specific niche tend to create shared vocabularies, such as the Dutch SVCN[6] ethnographic thesauri of more than 10,000 terms, or the Dutch artist list RKD holding around 400,000 artist records.

Technically, in the repository conversion task we receive the following inputs:

- Data sources describing meta-data and vocabularies. They may be described according to various data models stored as databases, XML dumps, or formatted text files.
- The target data models to which we need to convert the data (Dublin core and SKOS).
- The standard vocabularies to which we need to align the eventual terms used in the meta-data and local thesauri.

Given this input we need to develop a system that converts the source data to the target schema, and extracts and aligns possible terms to the standard terms. This is the `AnnoCultor` system,[7] developed in Java and freely available under GPL. In the rest of this paper we present the technology that we built to construct this system.

2 State of art

When converting something in the XML world, we need to start with XSLT,[8] the XML transformation language that allows creation of rules to translate on XML document to another. It has high-quality tool support and forms the standard solution for XML transformation. Our target format, RDF, is a specific kind of XML, and RDF documents can be easily constructed at the XML (syntactical) level with XSLT. In this way XSLT is widely used to perform syntactical conversions [Papotti and Torlone, 2004,Butler et al., 2004]. However XSLT is not

[5] `http://www.getty.edu/research/conducting_research/vocabularies/`
[6] `http://svcn.nl`
[7] `http://annocultor.sourceforge.net`
[8] `http://www.w3.org/TR/xslt`

suitable for semantic conversion and enrichment, or the porting problem as we face it:

- The data is often provided by cultural institutions in the form of databases or text files. XSLT requires the source data to be presented in XML.
- The terms need to be looked up in separate large vocabularies with some reasoning. XSLT does not provide any means to work with them.
- As we found out, nearly every dataset requires some dataset-specific code to be written and integrated. XSLT is not really meant for being integrated with custom code.
- Filtering broken images is nearly impossible to implement in XSLT.

Some approaches, such as [Sperberg-McQueen and Miller, 2004], perform automatic conversion of XML documents to RDF. However, they typically assume the target RDF model to follow the source XML model, that is really not the case in the cultural heritage conversion, as we will see later.

Practical state of art consists of custom program code written in general-purpose programming languages to convert a specific dataset. For example, manual creation of mappings is used in the cultural heritage domain as described in [Kondylakis et al., 2006]. While in [Kondylakis et al., 2006] an attempt to come up with a declarative language for conversion is made, in eCulture we quickly abandoned this path as, to solve practical conversion problems, it requires a language as expressive as a programming language.

The main goal of this work was to perform the conversions needed by the eCulture project. Accordingly, we started with programming specific converters, separating conversion rules for reuse. When programming we were constantly unifying the rules, maximizing their reuse. These reusable rules form some kind of a declarative rule language for conversion that is easy to be augmented with custom code.

3 Data models

We would illustrate our story with the sample source record, a simplified version of the description of the Rembrandts 'Night watch' from Rijksmuseum of Amsterdam.

3.1 Source models

A fragment of an XML dump produced out of the Rijksmuseum's information system is presented in Figure 1.

The key property of the datasets is that they describe objects that are separatable at the schema level, i.e. there is an XML path that wraps the description of each object. In the example from Figure 1 objects are wrapped with tag

```
<recordList>
 <record>
  <maker>
   <maker>
    <name>Rembrandt Harmensz. van Rijn</name>
    <name>Rembrandt</name>
    <name>Rembrandt van Rijn</name>
    <birth_on>1606-07-15</birth_on>
    <birth_place>Leiden</birth_place>
   </maker>
   <maker.role>
    <term>schilder</term>
   </maker.role>
  </maker>
  <title>
   <title>Het korporaalschap ... bekend als de 'Nachtwacht'</title>
   <title.gb>The company ... known as 'The nightwatch'</title.gb>
   <title.type>display</title.type>
  </title>
  <title>
   <title>De Nachtwacht</title>
   <title.type>secondary</title.type>
  </title>
  <object.number>SK-C-5</object.number>
  <date>1642</date>
  <reproduct>
    <reproduction_reference>M-SK-C-5-00</reproduction_reference>
  </reproduct>
 </record>
</recordList>
```

Fig. 1. Sample XML description of 'The Nightwatch' by Rembrandt

`recordList/record`. As we found them, these objects always have a tag for a unique identifier, such as tag `object.number` in our example. [9]

The other key property of the source objects is that they may include (repeating) parts. These parts would be representing other objects with their properties, that are treated as an integral part of the whole object. Similar to the wholes, the part objects are always wrapped up into a known XML tag. The parts occur quite frequent in the datasets. In our example we have several of them: multiple makers (wrapped with tag `maker`), titles, and reproductions (tag `reproduct`).

[9] In an SQL-accessible repository it is always possible to construct a query that would return one or more rows for each object, with one identifying field, when all the rows describing a single object would have the same value of the identifying field.

While the objects are always identifiable, the parts typically have no identifying property, e.g. tag `title` representing unidentified parts. [10]

3.2 Target models

For the meta-data we use Dublin Core, the leading ontology for meta-data descriptions of resources. It specifies resources with properties like `title`, `creator`, `location`, etc. For the specific purpose of describing visual resources the Visual Resources Association VRA[11], a leading organization in the area of image management, has proposed a specialization of Dublin Core where nearly each Dublin Core property is elaborated into several properties specific for this domain. In eCulture we developed an RDF representation of the resources based on the VRA model.

We represent each work with an RDF object, an instance of VRA class `vra:Work`, and each image of the work with an instance of VRA class `vra:Image`, linked to the work with property `vra:relation.depicts` of the image. Accordingly, the example from Figure 1 should be represented in RDF as depicted in Figure 2.

```
<vra:Work rdf:about="#SK-C-5">
 <rma:painter rdf:resource="#Rembrandt_Harmensz._van_Rijn"/>
 <dc:date>1642</dc:date>
 <dc:title xml:lang="nl">De Nachtwacht</dc:title>
 <dc:title xml:lang="nl">Het korporaalschap...</dc:title>
 <vra:title.alt xml:lang="en">The company ...</vra:title.alt>
</vra:Work>

<vra:Image rdf:about="#id=M-SK-C-5-00">
 <vra:relation.depicts rdf:resource="#SK-C-5"/>
</vra:Image>

<vra:Image rdf:about="#id=M-SK-C-5-01">
 <vra:relation.depicts rdf:resource="#SK-C-5"/>
</vra:Image>
```

Fig. 2. Sample RDF description of 'The Nightwatch' by Rembrandt

For vocabularies we use the Simple Knowledge Organization System (SKOS), a W3C model for expressing structure of thesauri and other vocabularies in RDF.

[10] In SQL results the parts are represented with multiple rows. In XML subtags have no identifiers, while in a database each row is likely to have (an internal) one. Accordingly, a part would be represented with multiple rows sharing the same value of this internal identifier of the part.

[11] `http://www.vraweb.org/`

It defines the major properties needed to represent vocabulary terms, such as `skos:prefLabel` (preferred label), `skos:broader` (link to a broader term), etc.

The following example illustrates how term 'Leiden', the birth town of Rembrandt, may be represented in SKOS:

```
<skos:Concept rdf:about="#Leiden">
  <skos:prefLabel xml:lang="nl">Leiden</skos:prefLabel>
  <skos:broader rdf:resource="#Netherlands"/>
</skos:Concept>
```

4 Conversion

From Figure 1 and Figure 2 one can see a number of modeling differences:

- tags `record` are converted to RDF objects that are identified according to XML tag `object.number`,
- (potentially multiple) makers are converted to new objects and their roles are converted to RDF properties (`rma:painter`),
- (potentially multiple) reproductions (tag `reproduct`) are converted to new objects of type `vra:Image` and linked to works,
- etc.

These differences are quite common in practical conversion.

4.1 Object conversion

First of all, we need to find the source objects: the tag that wraps the objects up, and the tag that holds object identifiers (`record` and `object.number` in our example). In vocabularies the terms are often identified with their labels. We use an existing methodology [van Assem et al., 2004] to represent them as RDF objects.

To do this we developed a special conversion rule, the `ConvertObject`, that is responsible for creating target RDF objects and apply (property- and value-level) rules to flesh them up. In this rule we also select the objects to be converted:

- based on field values of the record in question;
- based on external lists of records that have online images.

It is common that some part of a dataset, up to as large as 80% may be filtered out because it is either not yet meant to be published, or not yet approved for publishing, or has too few annotations, or, finally, has no image of the cultural object being made and available online.

Source objects may contain repeating parts to be converted according to one of the following options:

- a part may need to be converted into a separate object linked to the main object. For example, tag `reproduct` is converted to an instance of class `vra:Image` linked to the whole with property `vra:relation.depicts`.

– a part may be converted to a specific RDF property. For example, tag `title` is converted to properties `dc:title` and `vra:title.alt`.
– a part may be converted to an RDF property, which name depends on another tag, such as the `maker` converted to property `rma:painter` according to tag `maker/maker.role/term`.

These operations are performed by the `ConvertObject` as well.

A number of other conversion rules were created and used to convert eCulture repositories. These rules and their usage statistics are presented in Table 1. There the rules are grouped into: vocabulary lookup rules, value translation rules, sequential ordering of other rules, property renaming rules, branching of other rules, and custom rules.

Table 1 lists a number of datasets, where each dataset may consist of works meta-data, a local vocabulary of terms, and a local directory of artists. For each conversion rule the table shows the number of times this rule was used in each dataset-specific converter. These counts are summarized in three ways: the totals per rule and per datasets, the number of datasets depending on each rule, and the number different kinds of rules needed to convert a dataset that called 'diversity'.

4.2 Object schema conversion

We will now briefly sketch the conversion rules that were reused.

Property rules. As can be seen from Table 1, property transformation rules (group `Property`) account for more than half of all rules written. Here the rule `RenameLiteralProperty` creates a literal RDF property and stores there the source value, leaving in intact. This rule is used most in the converters (107 times out of 283), and all the 13 datasets depend on it. Its sibling, rule `RenameResourceProperty` creates a resource RDF property, adding a namespace prefix to the value. It is used just 16 times.

The other rules in this group create literal and resource constants (rules `CreateLiteralProperty` and `CreateResourceProperty`), and extract values with a regular expression pattern (rules `ExtractLiteralValueByPattern` and `ExtractResourceValueByPattern`). They are used less frequently.

Rule ordering. Two types of ordering rules: sequential (rule `Sequence`) and branching (rules from group `Branch`) were used in the converters. The sequential rule simply executes other rules in a sequence, and the branching rules evaluate a condition to choose one rule to execute, out of a pool of options.

Rule `BranchOnPattern` evaluates if the value of a specified source property matches a regular expression pattern provided with the rule to choose one of two options to continue.

Rule `FacetRenameProperty` is designed for the frequent situation where the target property name should be chosen based on the value of some other property of the work. In our example the role of the maker (tag `maker.role/term`,

Rule groups	Conversion rules	Datasets: aat terms	biblio terms	biblio works	geo terms	icn works	rkd artists	rma artists	rma terms	rma works	rmv works	svcn terms	tropen terms	tropen works	Total	Dependent
Vocabulary	LookupTerm		1	5	1	1	1		4	7	2	3	2		27	10
Value	AffixValue			1		1				1	2			2	7	5
Value	UseOtherPropertyValue											2	3		5	2
Value	ReplaceValues			1			1	1							3	3
Sequence	Sequence		1	4	1	1	4	2	2	4	1	2	2	1	25	13
Property	RenameLiteralProperty	3	9	13	3	4	14	9	4	29	7	1	1	10	107	13
Property	CreateResourceProperty	2	1	4	1	1	5	1	1	3	2	3	3	2	29	13
Property	RenameResourceProperty		3	1	4		2	1	1	4					16	7
Property	CreateLiteralProperty			1		1		1	1	2					6	6
Property	ExtractLiteralValueByPattern										1			1	2	2
Property	ExtractResourceValueByPatt											1	1		2	2
Branch	BranchOnPattern	1									2	6	5	2	16	5
Branch	FacetRenameProperty						1			1	4	1	1	3	11	7
Object	CreateNewObject			2			4								6	2
Custom	MakeEcultureImage			1		1				1	1			1	5	5
Custom	MakeEculturePerson						1	2		2			1		6	4
Vocabulary	LookupTermsFromText									2					2	1
Property	AAT$2	1													1	1
Branch	BranchOnTermInVocabulary												1		1	1
Custom	AAT$MakeAatSubject	3													3	1
Custom	RKDArtists$3						1								1	1
Custom	Biblio$ExtractRoleAndName			1											1	1
Object	RKDArtists$4						1								1	1
	Totals	10	15	34	10	10	35	17	13	56	22	19	20	22	283	
	Diversity	5	5	11	5	7	11	7	6	11	9	8	10	8		

Repositories: aat - AAT, biblio - Bibliopolis (`bibliopolis.nl`), geo - Geonames (`geonames.org`), icn - ICN (`icn.nl`), rma - Rijksmuseum, svcn - SVCN vocabulary (`svcn.nl`), tropen - Tropenmuseum (`tropenmuseum.nl`).

Table 1. The usage of conversion rules.

value `schilder`, Dutch for 'painter') should be converted to property `painter`. Another role, represented with another value of tag `maker.role/term` would need to be converted to a different property.

Value replacement. Several value replacement rules proved their usefulness, as shown in rule group `Value`. These rules deal with prefixing and affixing values (rules `ValueDePrefixer` and `AffixValue`), perform a value replacing based on fixed table (rule `ReplaceValues`), and may take value of another property of the same object (rule `UseOtherPropertyValue`).

Custom code. There are several rules shown in Table 1 below the double line that are used in a single dataset each. These are specific rules that cannot be reused further.

4.3 Term alignment

Rule `LookupTerm` is used in most datasets. While it counts for just 10% of the total rule base, this rule generates all cross-collection links, providing all the bridges between collections. Let us sketch the basic principles that we use to do vocabulary lookup and assess its practical utility.

Term extraction. The terms are used as values of some properties, where they occur in two situations in the meta-data:

- the property may explicitly contain terms, where property name suggests the vocabulary they are coming from, e.g. `creator='Rembrandt'`;
- the property may be a title or description and have terms mentioned in it as text, e.g. `title='Portret of Rembrandt'`.

In the first case no natural language processing is needed, as typically each property value corresponds to a single term (multiple semicolon-separated terms was the most complex case here). However, they need to be cleaned up: trimmed, brackets removed, etc. It is also quite common to find some clarifications included in term labels in brackets. For example, paper in AAT is referred to as `paper (fiber product)` to differentiate it from other papers, such as the research paper you are reading now. We hard-coded the cleaning rules.

In the second case we extract words from the textual descriptions and try to match them to vocabulary terms.

Term alignment. Interpreting context of the terms is the key to our alignment strategy. Here we use the fact that in the cultural heritage domain the terms are defined with labels augmented with a context. This context consists of the following:

- term label and its language,
- property name that uses this term,
- other properties of the resource (work or thesaurus entry) using the term,
- analysis of the other terms used in this property,
- implicit context of the dataset.

We find a counterpart of a term we, first, need to select the vocabulary to lookup. It is typically known from the name of the property and a brief look at its values. Properties that refer to places are likely to occur in a geographical vocabulary (TGN or Geonames), people names in a directory (ULAN or RKD), and other terms, such as styles or materials – in a generic thesaurus like AAT. For example, property `material` is likely to correspond to the materials subtree of AAT, or `place.origin` to a geographic place. Implicit context of the dataset may allow refining that further, e.g. to restrict the search for the origin places to Egypt, in the case of an Egyptian collection.

Within the vocabulary we rely on exact string match of the labels to find possible counterparts. While typically regarded as too strict, it works in the cultural domain where vocabularies often provide many alternative labels for

each term. For example, ULAN lists more than 40 ways of writing the name of Rembrandt, and it is quite likely that it will contain the name used by a specific museum.

An analysis of other terms helps in narrowing the part of a vocabulary that we need to search. For example, property `birth_place` with values like `Leiden` or `Amsterdam` suggests to store terms-cities. It should be aligned to cities from a geographical vocabulary TGN, while other geographic objects presented in the vocabulary, such as rivers, should be left out. In another example a property may contain values such as `Frankrijk` that suggests that (i) these terms correspond to countries rather than cities, and (ii) their labels are given in Dutch.

Using term context for disambiguating AAT terms seems to be more difficult than disambiguation of places and people.

4.4 Assessment

Schema conversion is easy to assess: it is either done or not. The number of properties is limited and can be assessed manually.

Our eCulture experience allows estimating the cost of conversion. It is realistic to estimate a single rule being written in an hour. To convert the Rijksmuseum repository of works, terms, and artists 86 rules are needed as shown in Table 1 (and other museums come at similar count). These 86 rules count to more than two weeks of working time, plus some time for the few custom rules. Accordingly, we can estimate one month of a conversion specialist being needed to convert an elaborated dataset.

Vocabulary alignment is more difficult because in each of its 30 usages the `LookupTerm` was applied to a specific property of a specific dataset that uses a specific part of a specific vocabulary. We do not make any statistical analysis of the mapping success as each of these 30 mapping cases is very specific, and any results aggregation would be as informative as the average patient temperature in a hospital. However, let us look at a few mapping cases presented in Table 2.

Table 2. Cases of thesauri alignment, '000

Repository - Vocabulary (concept count)	AAT (34)	TGN (1100)	ULAN (131)
Rijksmuseum	7 of 43	7.8 of 29 records	13 of 56
Ethnographic SVCN	4 of 5	3 of 5.5	
Artists list RKD			41 of 410
Bibliopolis	0.4 of 1.1		0.25 of 1

Rijksmuseum Amsterdam has developed an in-house thesaurus of 43,000 terms that we tried to map to AAT. At the moment, the mapping technique described above gave us 7,000 mappings, as shown in Table 2. Note, that the Rijksmuseum thesaurus is larger that AAT term-wise. From a random manual

check it seems that most of the terms that are not mapped simply do not occur in the AAT, or are phrased differently. Other vocabularies, such as SVCN fit AAT closer, and here we mapped 80% of the terms.

Aligning directories of people is also shown in Table 2, where 25% the Rijksmuseum people and 10% of the RKD[12] people were mapped to ULAN, that amounts to one third of ULAN records. In practical terms this is a good result as well.

Among the three kinds of vocabularies (terms, places, and people) places are the easiest as they have specific context and are organized. We map 3,000 out of 5,500 places in SVCN to TGN. In addition, mapping TGN to a comparable Geonames vocabulary gives us, for example, 50% success in mapping Dutch places. Again, the real overlap is not known, so this success is difficult to judge.

Here we are talking about semantic mappings, e.g. based not only on exact match of the labels of the terms, but also on the interpretation of some contextual properties of these terms.

5 System

To implement the converters we developed `AnnoCultor`,[13] a Java-based system that provides the conversion infrastructure and implements a number of reusable conversion rules, shown in Table 1 and discussed in this paper. It is open for the inclusion of custom rules, modifying existing rules, and integrating with external systems.

To construct a converter for a specific dataset, one has to create a simple Java program, where the existing rules need to be put together. This does not require complicated Java programming: converters are created given a template by composing reusable rules.

6 Conclusions

The task of porting a cultural dataset to the Semantic Web consists of two subtasks: converting database schemas and aligning vocabulary data.

Programming required. As we practically proved in the eCulture project, the problem of integrating various database schemas in the cultural heritage area can be solved by converting the schemas to open web standards SKOS and specializations of Dublin Core. This conversion cannot de defined automatically. Moreover, the complexity of the conversion task requires writing conversion programs in a programming language.

One man-month per dataset. Given our experience we can estimate around 4 weeks needed for a skillful professional to convert a major museum database (assuming that the `AnnoCultor` conversion system will be used). For this (s)he will have to create a converter of 50-100 rules plus some custom code.

[12] The national directory of Dutch artists, `http://rkd.nl`

[13] `http://annocultor.sourceforge.net`

Reasonable alignment success. Existing technologies, including the `AnnoCultor` tool discussed in this paper allow finding semantic alignments between works, vocabularies and artists. Success of the alignments depends on the specific vocabulary, collections and the terms that happen to be used, as these vary a lot. However, it is possible to achieve reasonable 50-80% of terms being mapped to the standard vocabularies without any advanced natural language processing, but with the use of the term context.

Acknowledgements. The author would like to thank the Dutch NWO-funded eCulture MultimediaN project for supporting this work.

References

[Butler et al., 2004] Butler, M. H., Gilbert, J., Seaborne, A., and Smathers, K. (2004). Data conversion, extraction and record linkage using xml and rdf tools in project simile. Technical report, Digital Media Systems Lab, HP Labs Bristol.

[Hyvyonen et al., 2005] Hyvyonen, E., Myakelya, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., and Kettula, S. (2005). Museumfinland - finnish museums on the semanticweb. *Journal of Web Semantics*, 3(2).

[Kondylakis et al., 2006] Kondylakis, H., Doerr, M., and Plexousakis, D. (2006). Mapping language for information integration. Technical report, ICS-FORTH, `http://www.ics.forth.gr/isl/publications/paperlink/Mapping_TR385_December06.pdf`.

[Papotti and Torlone, 2004] Papotti, P. and Torlone, R. (2004). An approach to heterogeneous data translation based on xml conversion. In *Proceedings of the Int. Workshop on Web Information Systems Modeling at CaiSE-2004*, Riga.

[Schreiber et al., 2006] Schreiber, G., Amin, A., Assem, M., Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., Kersen, J., Niet, M., Omelayenko, B., Ossenbruggen, J., Siebes, R., Takema, J., Tordai, A., Wielemaker, J., and Wielinga, B. (2006). Multimedian e-culture demonstrator. In *International Semantic Web Conference*, volume 4273 of *LNCS*, pages 951–958, Athens, GA.

[Sperberg-McQueen and Miller, 2004] Sperberg-McQueen, C. M. and Miller, E. (2004). On mapping from colloquial xml to rdf using xslt. In *Extreme Markup Languages 2004*.

[Tordai et al., 2007] Tordai, A., Omelayenko, B., and Schreiber, G. (2007). Thesaurus and metadata alignment for a semantic e-culture application. In *Proceedings of the 4th international conference on Knowledge capture (KCAP-2007)*, pages 199–200, Whistler, British Columbia, Canada.

[van Assem et al., 2004] van Assem, M., Menken, M. R., Schreiber, G., Wielemaker, J., and Wielinga, B. (2004). A Method for Converting Thesauri to RDF/OWL. In McIlraith, S. A., Plexousakis, D., and van Harmelen, F., editors, *Proceedings of the Third International Semantic Web Conference (ISWC'04)*, number 3298 in Lecture Notes in Computer Science, pages 17–31, Hiroshima, Japan. Springer-Verlag.

[Wielemaker et al., 2007] Wielemaker, J., Hildebrand, M., and van Ossenbruggen, J. (2007). Using Prolog as the fundament for applications on the semantic web. In S.Heymans, A. Polleres, E. R. D. P. and Gupta, G., editors, *Proceedings of the 2nd Workshop on Applications of Logic Programming and to the Web, Semantic Web and Semantic Web Services*, volume 287 of *CEUR Workshop Proceedings*, pages 84–98. CEUR-WS.org.

Solutions for a Semantic Web-Based Digital Library Application

Andrew Russell Green and José Antonio Villarreal Martínez

Instituto Mora (National Council for Science and Technology, Mexico)
and Instituto de Investigaciones Estéticas (National Autonomous University of Mexico)
ahg@servidor.unam.mx,quetzal1910@gmail.com

Abstract. Digital libraries and archives stand to benefit greatly from the Semantic Web, which may provide a basis for novel end-user functions targeted at research and teaching. The project "Image Preservation, Information Systems, Access and Research" seeks to develop an adaptable digital library application based on a back-end of semantically modeled data. By "adaptable" we mean able to adapt to diverse library and archive scenarios, especially involving the integration of different types of material (photographic prints, negatives, drawings, periodicals, books, etc.) in a single system. This requires "mixing and matching" standard and non-standard catalogue record formats and ontologies to model them. A central problem we have encountered is: how to structure application logic in this context in a way that follows best-practice principles. In this paper we discuss the problem and propose, as a tentative solution, a Semantic Component Architecture, which would provide an integrated, encapsulating way of selecting vocabularies and establishing inference rules, recurrent path patterns, graph constraints, catalogue record templates and arbitrary logic. We also consider other related issues, including the encapsulation of low-level graph structures and possible features of record display templates.

Key words: Semantic Web application architecture, Semantic Web display systems, metadata management, digital libraries

1 Introduction

Digital libraries and archives stand to benefit greatly from the Semantic Web (SW). Semantically modeled catalogues should provide a basis for new functions to help users sift through large and diverse repositories, discover patterns, explore associations among objects, find relevant information, and create and share descriptions of objects in a structured, flexible manner. This is the promise the SW holds for knowledge repositories, and one can hardly underestimate its potential impact in History and other Social Sciences: archives are primary sources—essential deposits of partially processed information, used for research in these disciplines—and despite the high degree of interrelation among data in different

archives, catalogues are often isolated and employ divergent record formats that are hard to align using standard information technology.[1]

This issue is one of the reasons the project Image Preservation, Information Systems, Access and Research (IPISAR) set out to build a SW-based digital library application. The project investigates the dissemination, study and management of heritage resources, and attempts to provide solutions to common problems in these areas.

The application being built, called "Pescador", will store catalogue data in a persistent triple store (whose function will be similar to that of a relational database in traditional systems). The requirements for the application include the ability to integrate data in various catalogue formats and adapt to the cataloguing needs of diverse archives. In this paper, the terms "catalogue format" and "record format" refer to the selection, organization and meaning of fields used to describe objects in an archive or library catalogue, as well as other conventions related to catalogue creation. Since Pescador will use the SW to model catalogues, each record format will correspond to a distinct kind of graph structure, often requiring specialized vocabulary and rules, and related to specialized application logic.

The application will have three main types of user: (1) regular users (or "patrons") who will consult the material provided by the digital library, (2) cataloguers, who will provide and manage the library's materials and metadata, and (3) catalogue designers/modelers/programmers, who will select or create the catalogue record formats and corresponding ontologies, and adapt the system to the needs of a given scenario. Pescador will provide a Web interface for the first two kinds of users; on this level, numerous functions targeted at research, teaching and cataloguing are planned [7]. When these users view data from the catalogue, they will see a user-friendly organization of information extracted from the SW graph; similarly, when cataloguers modify elements in the catalogue, they will employ easy-to-use forms, and the SW graph will be changed according to their input.

The third type of user, the catalogue designer/modeler/programmer, will use a programming interface. The problem considered in this paper is: how to design an interface that allows users of this type to adapt the application to their needs, in a way that follows best-practice information engineering principles such as the encapsulation of complexity, the separation of concerns and the non-repetition of declarations. To achieve this we propose a Semantic Component Architecture (SCA), that is, an adaptation of component architecture principles to a SW application context, in which data structure rules and application logic are closely linked. In addition, we propose mechanisms for encapsulating low-level graph structures—which should also help catalogue designers/modelers/programmers follow best-practice principles—and discuss catalogue record template systems.

[1] The situation of historical archives varies greatly from one archive to another. Other recurring difficulties include access restrictions and insufficient funding; the first of these is also a major focus of the project described in this article. See [8] and [2].

To date, two incomplete versions Pescador have been created. Both are currently used for Web sites that offer simple consultation functions for on-line archives (available at [11] and [5]). Our proposals stem from the experience of developing these versions of the application. Though the project IPISAR may yet generate new archival Web sites using the second version, it is clear that to implement all proposed features, a major rewrite is unavoidable. It should be noted that work on the rewrite and on the detailed design and implementation of the SCA has not yet begun. The general nature of the proposals outlined here is a reflection of this.

All versions of Pescador are provided under the terms of the free GNU GPL license.

2 Previous related development experiences

Since before work on Pescador began, it was clear that to integrate data in various formats and adapt to diverse scenarios, the system would have to offer ways of selecting vocabulary, establishing graph constraints[2] and configuring the algorithms used to transform information as it passed between the model and the UI. In retrospect, we can say that we have consistently underestimated the complexity of the mechanisms required for these and related processes. The first version of Pescador (version 0.1) allowed catalogue designers/modelers/programmers to select RDF schemas and set presentation information, which was modeled using an augmented version of the Fresnel Display Vocabulary [4]. In the subsequent version (0.2), we abandoned Fresnel due to its limited scope (it was designed only for specifying how to display SW data) and text manipulation features, and instead created a domain-specific language (DSL) that provided an integrated way of establishing vocabulary, graph constraints, inference rules, path definitions and display specifications. Our DSL also allowed the creation of executable code fragments that could be hooked into various parts of the system. Though never implemented in full, the DSL was, we believe, an important step in the right direction, as it recognized the need to take into account best-practice principles when setting out configuration information [6]. However, our version 0.2 design was flawed precisely because it defined a configuration system—it offered a very limited range of possibilities for creating arbitrary logic and encapsulating complexity, as compared to our current proposal, the SCA, which by its very nature calls for opening as widely as possible the catalogue designer/modeler/programmer's options. The difficulties we faced in organizing and re-using code written in our own DSL highlight the need for mechanisms that do not impose the limits that our DSL did.

[2] By graph constraints, in this context and in the rest of the paper, we mean the specifications of how to structure and link subgraphs that describe elements in the catalogue. This includes, but is not limited to, establishing which properties may be used with resources of a which classes, and the properties' allowed cardinality.

3 SCA

3.1 General conception

The SCA we envision would coordinate pluggable "components" or "bundles"[3] that would wrap interrelated elements of any of the following types: schemas, constraints, inference rules, ontologies, path definitions, executable code, display specifications, Abox configuration information, and links to external data sources. As in other component frameworks, bundles would be able to provide hooks of various kinds and use the hooks exposed by other bundles. These hooks would be the bundle's means of controlling and simplifying access to the elements it wraps. A bundle could also "depend on" other bundles, and would thereby know which external hooks are available to it. The standard definition of a component as "a software artifact consisting of three parts: a service interface, a client interface and an implementation"[13] might, with little or no modification, provide adequate theoretical grounding for an SCA—though we would have to ask how a SW application context would modify our understanding of what constitutes an interface and what constitutes an implementation.

3.2 Justification

An SCA would help solve issues of encapsulation and re-usability. Consider, for example, an archive that includes books, photographic prints, negatives and handwritten manuscripts. Catalogue record formats for each type of object would certainly vary, though they would also have elements in common. For example, all objects would probably be associated with an author and a creation date. But only the negatives, photographic prints and books could be associated in a chain of image reproductions (as in, a negative could be reproduced on a print, and the print could be touched up and cropped, then reproduced in a book). Many objects might have a place of creation, but the precise meaning of this concept would not be the same for a book (place of publication) and for a negative (place the photo was taken). Of course, the SW excels in the alignment of data structures like these. However, setting up a digital archive to integrate records of such objects involves much more than establishing data structures and their intersections; bits of logic have to be defined, and it makes sense to define them alongside related data structure information in an encapsulated way—in other words, to "keep together what goes together".

Thus, in an SCA, the logic needed to transform the SW model of a book's metadata into a standard bibliographic citation might be stored in the same bundle as the vocabulary, graph constraints, and remaining display information for books. Similarly, a component could be created for chains of image reproductions (like the one described above). The component could enclose vocabulary, constraints, path definitions, and code for generating visual representations of

[3] Herein the terms "bundle" and "component" are used as synonyms.

these chains. Other components—such as those for books and photographs—could depend on it, interfacing with the hooks provided and thus employing in a simplified manner the enclosed complexity.

The advantages of component architectures, and of the modularity and reusability they offer, have long been recognized, and countless frameworks and related technologies are used in applications of all shapes and sizes.[4] Component architectures are increasingly viewed as a programming paradigm in their own right.

Our experience shows that a SW-based system that relies, as Pescador 0.1 did, on partially overlapping RDF data sets coupled with fragments of related logic scattered throughout the application, ends up grossly violating best-practice principles in many respects, to the detriment of flexibility and scalability. An SCA would address these problems in a thorough manner, and provide a basis for the re-use of ontologies coupled with application logic.

4 Path Definitions in the Context of an SCA

In both versions of Pescador, the algorithms that extract information from the graph frequently traverse paths from one resource to another. These paths vary greatly in length and complexity. To allow for encapsulation and separation of concerns as proposed, an SCA must include a means of defining path patterns that supports these principles. (We first ran into path-related encapsulation issues in version 0.1 of Pescador, which used SPARQL to extract information from the graph. In that version, we found ourselves writing SPARQL queries that repeated information already set out in several other parts of the application—quite the opposite of a maintainable application structure.) The solution we suggest is to adapt an existing path definition mechanism (of which there are many [14]) to allow bundles to define paths as aggregates of other paths, which may in turn be defined opaquely in other bundles.

Let us illustrate this proposal with reference to the hypothetical archive described in 4 Justification. Consider: one bundle might define paths from any photographic image to its original negative, along a chain of image reproductions; another might define paths from cities to the continents on which they are located. In such a scenario, it should be possible, in a third bundle, to aggregate paths and define a route from any photographic positive to the continent on which the negative that created the photo was snapped. Fig. 1 presents this example.

A preliminary survey of existing path definition mechanisms indicates that the SPARQLeR extension to SPARQL may be the best candidate for adaptation to our SCA and Pescador. The authors of SPARQLeR, working in the life sciences domain, found they needed a means of detecting "semantic associations"—undirected paths "that connect two entities in the knowledge base using named relationships" [10]—and that the associations they had to detect could not be

[4] Examples of frameworks and technologies include OSGi, Enterprise Java Beans, COM and CORBA. See [1] for the history of the concept.

"explicitly defined by the fully specified structure of a graph pattern". In modeling catalogues and Social Science data, we have noticed a similar requirement, which SPARQLeR has the potential to fulfill.

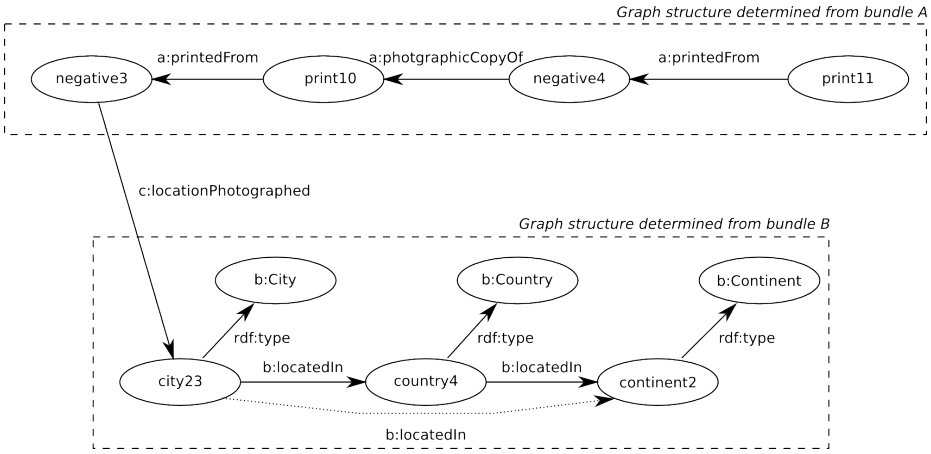


Fig. 1. Encapsulation and Separation of Concerns in Path Definitions

In this example, bundle A defines vocabulary and related logic for chains of image reproductions, bundle B does the same for geographic locations, and bundle C manages photograph catalogue records at a higher level, and depends on bundles A and B. Bundle A might define a path `originalNegative` to go from any photographic print to its original negative, and bundle B might define a path `onContinent` to traverse from any place to the continent on which that place is located. Bundle C could then define an aggregate path that traverses first `originalNegative`, then the property `c:locationPhotographed`, and finally `onContinent`, to go from a photographic print to the continent on which the photo was taken (`print11` to `continent2`). Bundle C would not need to know the inner workings of `originalNegative` and `onContinent`, and in the event that the specifics of the graph structures determined by bundles A or B were to change, the required changes to path definitions would be limited to one bundle.

5 Low-Level Encapsulation: SW Abstraction Layer

This section describes a mechanism called "the SW graph abstraction layer", which seeks to apply, in low-level graph operations, the same principles that underlie the SCA. The abstraction layer will offer simplified access to and correct processing of very basic graph structures that occur repeatedly in a wide variety of models that we have worked with. It will be integrated directly with the Pescador's triple store; graph extraction and modification operations will access it in a transparent manner. The abstraction layer will be logically above a SW-conformant graph. Both the abstraction layer and the SW-conformant graph will

represent exactly the same information; each will simply provide a different view of that information. For interchange with other SW applications, the Pescador will be able to export the SW-conformant view of the contents of its triple store in standard SW serialization formats (like RDF/XML).

Below are some examples of simple graph structures as seen by the abstraction layer and as reflected in the underlying SW specifications-compliant graph. The first two examples are presented in detail and the rest are summarized.

Note that version 0.2 of Pescador already implements many of the features described here. Unlike the SCA, the abstraction layer has demonstrated its usefulness in a functioning implementation, and can be considered a relatively mature proposal. Nonetheless, substantial work remains to be done to develop a theoretical grounding for this solution, analyze its implications in formal terms, and design a new implementation.

Multilingual Values A common idiom in SW graphs is the multilingual value: conceptually a single value with alternate versions for different languages. It is often modeled with a blank node that is an instance of `rdf:Alt`. The problem we have encountered is the recurring need to select an appropriate version of a multilingual value on the basis of information available at runtime. Fig. 2 shows the abstraction layer's view of this idiom, together with the underlying standards-compliant model.

This type of structure is created only when specifically requested for a given property and a given subject resource. The query system takes the abstraction layer view of the model; queries are executed in a given language context, and only one of the "switchable" triples is available during a given operation. This makes for cleaner path definitions, which need only specify the main property to be traversed (in this example, `sys:name`). Without this feature, the sub-pattern for selecting the correct language alternative would have to be repeated in many path definitions.

Ordered Multiple Values In catalogue records, it is often necessary to present several values of the same field in a specific order—for examples, the authors of a publication must be listed in the same order in which they appear on the publication itself. Similarly, if several topics from a thesaurus are associated with a given entity, the most relevant topics are listed first. These fixed orders must be reflected in the model. Since in RDF multiple, identical properties on a single subject are unordered, when a fixed order is needed, multiple values may be grouped using an `rdf:Seq`. Fig. 3 shows how the abstraction layer understands this low-level structure.

As in the case of multilingual values, this structure is only created when it is requested for a given property and a given subject. The query system traverses from the main subject (`photo2`) directly to the objects proper (`topic21`, `topic42` and `topic25`) and extracts them in the correct order. Again, no special sub-pattern need be repeatedly included in the path definitions that traverse this type of structure.

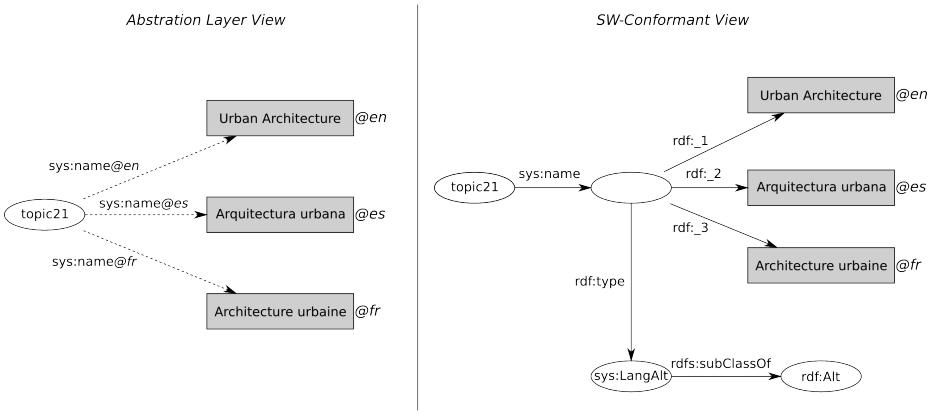


Fig. 2. Multilingual Values
In the abstraction layer view, the arcs shown with a dotted line are seen by extraction operations as "switchable" versions of a single triple. Queries are executed in a given language context, and only one arc may be "switched on" during a given operation.

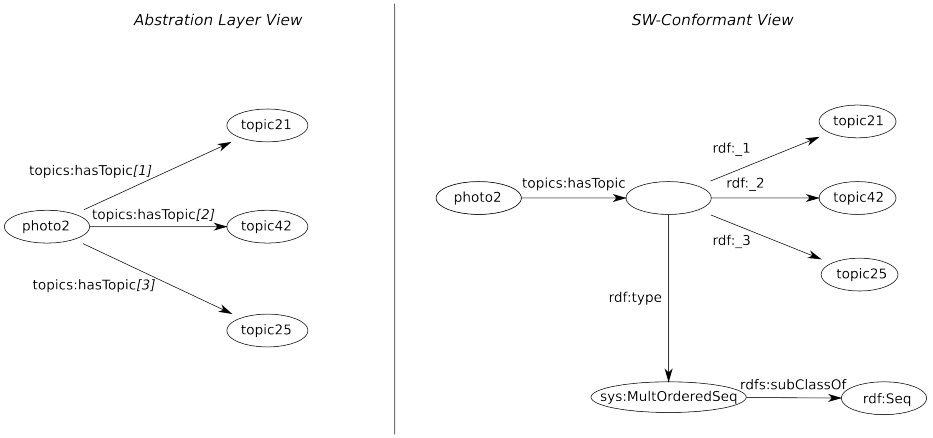


Fig. 3. Ordered Multiple Values
In the abstraction layer view, only the main property appears (`topics:hasTopic`), and queries automatically extract the object resources in the correct order.

Missing Value Explanations It is frequently useful to include in a catalogue an explanation for a value's absence (common reasons are "unknown" or "field does not apply"). The abstraction layer provides a special mechanism for modeling such explanations. When a path query does not produce any results, it may expose a pertinent missing value explanation, if one is expressed in the model.

N-ary Relations As explained in [12], there are many ways of modeling n-ary relationships with the SW. The abstraction layer integrates facilities for working with structures that model these relationships, allowing clean path definitions and encapsulating code that deals with corresponding low-level subgraphs.

Structured Values The RDF specification establishes an idiom called a "structured value", which consists of a blank node that is at once the object of a property and the subject of additional triples used to model the aforementioned property's value. The abstraction layer builds on this concept by offering a mechanism for explicitly identifying nodes that are structured values, thus allowing the implementation of special features for constructs of this type.

6 Display Templates

We now turn our attention another problem related to the catalogue designer/modeler/programmer interface: the definition of algorithms for processing information as it flows between the model and the user interface. This is a vast issue; here, we focus only on systems for record display specification, which we will call "display template systems".

There exist several general systems of this type, and many SW applications use internal template mechanisms. We agree with the definition of the problem given by the authors of Fresnel (an important proposal in this area), who state that "presenting Semantic Web content in a human-readable way consists in addressing two issues: specifying what information contained in an RDF graph should be presented and how this information should be presented." [4] However, this definition is deceptively simple, as both parts of the problem—the selection of information from the model and its transformation into a presentable format—can be quite complex.

Clearly there is a need for templates in SW applications: models often do not contain all the information required to create user-friendly descriptions, and even when they do, it is not always be desirable to show users all available information. The most basic kind of SW template involves a selection and ordering of properties; when the template is "applied" to a resource, label-value pairs are created from the properties' labels (often set using `rdfs:label`) and values for that resource. On this foundation, numerous advanced features may be built, such as:

- Facilities for creating sections, subsections and similar structures within records. This is often required for lengthy description; see, for example, full records in [5] and [11].
- Ways of including additional elements in records, such as images and text that is not part of a label-value pair.
- Facilities for defining short, human-readable labels for resources—normally used for values in label-value pairs, to refer to resources that are the objects of the properties displayed.
- Ways of setting special, context-appropriate property labels. (Consider, for example, a property with the `rdfs:label` "Location Photographed". In a photograph's catalogue record, one might wish to call the property "Location", since in this context, the full label would be needlessly long.)
- Means of embedding the result of one template in the result of another one.
- Means of retrieving information from diverse parts of the model—not just over the direct properties of the resource being described. This may be accomplished using path definitions.
- A hierarchy of templates and inheritance of templates' characteristics over the hierarchy.
- Media-agnostic template definitions, or a separation of record content specifications from media-specific formatting details.
- Facilities for embedding arbitrary logic—in other words, executable code—in templates, in a manner similar to languages for creating dynamic Web pages (JSP, ASP, PHP, RHTML, etc.). This allows templates to run loops, generate text and modify their output on the basis of conditions described in the executable code.
- Programmatic template creation and modification. For example, at runtime, a search component may create temporary templates that display only fields containing hits.
- Vocabulary and conventions for modeling the templates themselves.

Fresnel, Pescador 0.1 and Pescador 0.2 all implement different subsets of these possible features. A challenge for the next version of Pescador is to determine which features are required, and how to integrate them with the SCA, and maintain support for encapsulation and separation of concerns. In additional, we must take into account a lesson learned in work on Pescador 0.2, namely: that the scope of a templating system is wider than record display itself. This is because numerous elements of a user interface must be coordinated with record display. To illustrate this, let us consider a catalogue in which photographs are described with the fields "photographer", "title", "date", "location" and "topics". A user interface that provides access to such a catalogue would refer to these fields in several places, not just when displaying records. For example, a menu might offer the option of listing items ordered by date or title. Another might offer the possibility of grouping items by photographer, location or topic. An advanced search interface could include options for searching only within one or more of these fields. On the screen displaying the catalogue records themselves, diverse functions may be available in fields' context menus. Last but not least, the

interface for adding, deleting and modifying items in the catalogue will mention fields in various ways. In all these parts of the interface, references to fields must be consistent, clear and appropriate for their respective contexts. To achieve this, specialized display specification mechanisms are required, and it makes sense to integrate these mechanisms with the template system.

7 Other Issues

There are important issues facing Pescador development that we have not discussed here. They include:

Model generation, concurrency and validation To create a Pescador Web interface for creating, deleting and modifying catalogue items requires the study of algorithms for changing the model in a multiuser environment. Input validation logic and graph constraint checking must also be considered and integrated with the SCA and other proposals discussed here. Our intention to offer an "undo" mechanism further complicates the matter.

Abox setup and permissions In this paper, we have said relatively little about organizing and configuring the Abox—the part of the model where most assertions about individuals reside—however, there are various unsolved problems in this area. Versions 0.1 and 0.2 of Pescador divided the Abox into "regions", in anticipation of future versions of the system that would allow the combination of multiple repositories in a single model, and would thus require the establishment of distinct read and write permissions for different parts of the model. This remains a part of our proposal for the next version of the system, but many aspects have yet to be determined, and personalization mechanisms for Abox structure and permissions must be integrated with the SCA.

Inference rules In many situations it is desirable to establish custom inference rules for a model. Yet again, the details of how to do this must be established, and whatever solution we choose must integrate with the SCA.

Networking A proposed feature of the system is to allow meta-searches in the repositories of multiple Pescador Web servers linked over a peer-to-peer network. To make this a reality we must study, among other things, model alignment and the sharing of SCA bundles over a network.

Optimization The diverse operations that the system will carry out must be performed quickly, even in large repositories with hundreds of thousands records. This will require a concerted investigation of optimization procedures, especially for operations that extract information from the model.

Patron and cataloguer interface design Patrons and cataloguers will access Pescador via a Web interface. Implementation of the many features planned for these users will require interfaces that link what is possible using a SW backend with users' habits, expectations, needs and imaginable ways of using the system.

Effective transdisciplinary research methodology The IPISAR project faces obstacles posed by disciplinary limits, as the research questions it puts

forth span several fields, including Computer Science, Social Science and Library and Archive Science. To date, the results of our attempt at transdisciplinary research are positive, however a true transdisciplinary integration of activities and theoretical groundings remains elusive.

8 Conclusion

In this paper we have considered some issues encountered in the development of an adaptable SW-based digital library application. Specifically, the issues discussed relate to modularization, display specifications, the structuring of application logic and a programming interface for catalogue designers/modelers/programmers that allows the use of best-practice information engineering principles. We have also outlined possible solutions. Although the implementation and detailed design of these solutions is far from complete, our proposals as they stand establish research directions and provide starting points for further work towards the creation of the application we envisage.

References

1. Atkinson C., Paech B., Reinhold J., Sander T.: Developing and Applying Component-Based Model-Driven Architectures in KobrA. IEEE: Alamitos, California (2001) http://ieeexplore.ieee.org/iel5/7549/20561/00950441.pdf
2. Aguayo, F., Roca, L.: Estudio introductorio. In: Aguayo, F., Roca, L. (eds.): Imágenes e investigación social. Instituto Mora, México (2005) 9-28 http://durito.nongnu.org/docs/Aguayo_Roca_2.html
3. Ambler, S. W.: Object Relational Mapping Strategies http://www.objectmatter.com/vbsf/docs/maptool/ormapping.html
4. Bizer, C., Lee, R., Pietriga, E.: Fresnel Display Vocabulary for RDF: User's Manual. World Wide Web Consortium (2005) http://www.w3.org/2005/04/fresnel-info/manual-20050726/
5. Fototeca Digital: Fotógrafos y Editores Franceses en México. Siglo XIX. Instituto Mora and Instituto de Investigaciones Estéticas, National Autónomous University of Mexico (2007) http://afmt.esteticas.unam.mx
6. Green, A.: Logic and a Little Language for Heritage Resource on the Semantic Web. Poster accompanying a system demonstration, presented at the 4th European Semantic Web Conference (June, 2007) http://durito.nongnu.org/docs/innsbruck2.pdf
7. Green, A. R.: Metadatos transformados: Archivos digitales, la Web Semántica y el nuevo paradigma de la catalogación. In: Amador C., P., Robledano A., J., Ruiz F., R. (eds): Quintas Jornadas: Imagen, Cultura y Tecnología. Universidad Carlos III de Madrid: Madrid (2007) 11-22 http://durito.nongnu.org/docs/metadatos_transformados_green.pdf
8. Green, A. R.: Rescate de la memoria. Ciencia y Desarrollo (Sept. 2006). Consejo Nacional de Ciencia y Tecnología, Mexico
9. Oren, E., Delbru, R. Gerke, S., Haller, A., Decker, S. ActiveRDF: Object-Oriented Semantic Web Programming (2007) http://www.eyaloren.org/pubs/www2007.pdf

10. Kochut, K. and Janik, M., SPARQLeR: Extended Sparql for Semantic Association Discovery (2007) http://www.eswc2007.org/pdf/eswc07-kochut.pdf
11. Marcas de Fuego de la Biblioteca "José María Lafragua" de la BUAP. Autonomous University of Puebla (2006) http://www.marcasdefuego.buap.mx/
12. Noy, N., Rector, A.: Defining n-ary relations on the semantic web. Working Draft for the W3C Semantic Web best practices group (2005)
13. Parrish, A., Dixon, B., Hale, D.: Component-Based Software Engineering: A Broad-Based Model is Needed (1999) http://www.kiv.zcu.cz/publications/
14. RDF Path Languages And Templating. In: ESW Wiki. World Wide Web Consortium http://esw.w3.org/topic/RdfPath

Semantics-Aware Querying of Web-Distributed RDF(S) Repositories

Georgia D. Solomou, Dimitrios A. Koutsomitropoulos, Theodore S. Papatheodorou

High Performance Systems Laboratory, School of Engineering University of Patras, Building B, 26500, Patras-Rio, Greece
{solomou, kotsomit, tsp}@hpclab.ceid.upatras.gr

Abstract. Because of the scattered nature of the Semantic Web, the existence of an integrated framework for storing, querying and managing distributed RDF data is of great importance. Already existing and widespread systems that share similar goals, like Jena and Sesame, do not appear to support distributed storage and retrieval of RDF data for the time being. In this paper we present a mechanism, based on Sesame that facilitates querying of distributed RDF repositories using the SPARQL protocol. This mechanism achieves to successfully retrieve requested RDF data; at the same time it aggregates web-distributed ontological knowledge in an attempt to exploit potential inferences, a fundamental aspect of the Semantic Web. We present its architecture and the method used for the analysis of SPARQL queries, trying to implement retrieval of RDF data in an optimized way.

Keywords: RDF, Semantic Web, Distributed Querying, SPARQL, Optimization, Reasoning

1 Introduction

Resource Description Framework (RDF) [10] and its related technologies appear in the core of the Semantic Web technologies stack. As a means to describe web resources in a consistent manner, RDF can be of great value in representing, unifying and possibly interpreting information hidden in disparate databases, information management systems and portals. This kind of description is for example greatly valuated in the manipulation and interoperability of metadata and information about resources stored and disseminated through digital libraries.

Since RDF descriptions have become commonplace in such scenarios, the need for their management, maintenance and exploitation has risen to spawn the development of integrated software systems known as RDF repositories. Such systems typically allow for the efficient administration of RDF data lifecycle, their preservation and most importantly, their querying in data-intensive contexts. Amongst the most prominent, Sesame [3] and Jena [11] appear dominant in production as well as research-driven applications.

Although the development of RDF repositories comes with the notion of coping with the scattered nature of web descriptions, it is inevitable that the simultaneous existence of multiple such repositories leads to the standard information integration problem; that is, the transparent treatment of distributed RDF data hidden now inside disparate repositories. This condition becomes even worse when reasoning about RDF information is to be considered, which is a rather complex process, since it may involve the combination of triples stored in arbitrary sources that ultimately form the distributed knowledge.

In this paper we therefore suggest a mechanism that achieves distributed querying of RDF(S) repositories, a feature not currently supported by any of the related systems. This method builds around a *mediate* repository that attempts to fetch and store from each remote system only the most relevant triples to each particular query.

In addition and when reasoning is to be taken into account, we show how, through a careful preprocessing of the query we may avoid fetching unnecessary triples that would not contribute to the distributed knowledge whatsoever. In this way we suggest a method towards the optimization of RDF distributed querying when reasoning is involved.

At the same time, when inferences are not required, possibly in need of rapid results to massive querying, this method exhibits desirable scaling behavior: These triples are fetched that are exactly the ones to participate in the query evaluation, a quite time-consuming process if conducted against each remote repository separately.

The rest of the paper is organized as follows: In section 2 we mention some of the most popular query languages for RDF as well as some existing frameworks supporting them. Section 3 deals with the mechanism that was developed for querying distributed RDF repositories, presenting its architecture and its particular techniques used for achieving its main purpose in an optimized way. In section 4 we detail the implementation process by showing some results, whereas in the last section we present the derived conclusions and possible future work

2 Background

The mechanism's development is based on a kind of analysis of the RDF query language SPARQL [14]. SPARQL is a W3C recommendation which has borrowed many elements from previous existing RDF languages. It fulfils all requirements stated in [6] as being necessary in order to query RDF data, like compositionality, schema awareness, optional path expressions and datatyping. The results of SPARQL queries can be result sets or RDF graphs.

Generally speaking, the most prominent query languages are those that were conceived as first generation tryouts of RDF querying, with little or no RDF-specific implementation and use experience to guide design, and based on an ever-changing set of syntactic and semantic specifications [8]. Many of these languages were created having in mind SQL, so they share many features in common. Some languages of this kind are RDQL[12], RQL[9] and of course SeRQL[5], the fundamental query language of the RDF framework Sesame. Because RDQL and RQL are first generation languages, they seem to be less powerful than, for example, their successor

SeRQL. As far as SeRQL and SPARQL are concerned, they appear to share many features in common. However, our preference to use SPARQL as a basic query language for our mechanism came from its substantially improved structure related to older languages of this kind, plus the fact that SPARQL is an official W3C recommendation.

SPARQL is fully supported by both Jena and Sesame. These open source systems offer full reasoning for the semantic extension of RDF, RDF Schema (RDFS) [2], and many different ways for storing RDF data. They evaluate queries against local or remote repositories, but they do this only for a single storage each time. They are not able to query distributed RDF data and so they lack the essential ability to infer new knowledge by combining distributed information.

Sesame forms the base of our mechanism, too. It is a mechanism that allows querying at the RDF Schema level. It makes use of a forward-chaining inference engine to compute and store the closure of its knowledge base [4]. Whenever a transaction adds data to the repository, the inferencing algorithm applies RDF Model Theory entailment rules in an optimized way, making use of the dependencies between them to eliminate most redundant inferencing steps. Furthermore, Sesame's modular architecture allows for an easy combination of different repositories, using a standardized repository access API called SAIL. The modularity of Sesame's architecture and its proven scalability are two of the most important features that encouraged our choice to use this particular system as our mechanism's basis.

Beyond the actual storage system that may be used in the backend, it seems that a number of applications that use multiple RDF sources in practice have been developed, such as the DOPE Project (Drug Ontology Project for Elsevier) [13] which accomplishes integration of information in a customized way. Moreover, Kowari[1] – another storage and retrieval system for RDF – appears to offer a kind of querying capabilities over multiple sources. Its basic drawback, though, is that it doesn't provide data independence, as the target repositories for such a distributed querying process must be explicitly defined, so it requires the users to know in advance where to get certain information.

Another related attempt is the development of a distributed storage and query infrastructure on top of Sesame, mentioned in [1]. This system extends Sesame in terms of a mediator component that provides centralized access to a collection of local and remote sources, but it doesn't offer inferencing capabilities, a requirement that constitutes the main objective of our mechanism.

3 A Technique for Distributed RDF Querying

Our query mechanism has two main features: the first one is that it queries distributed RDF repositories and retrieves data that may be combined in order to infer new RDF statements. The basic idea is to retrieve utilizable data from remote repositories and not just simple query results. Its second feature is that in order to achieve this data retrieval, our mechanism employs a preprocessing method that analyzes a query into smaller parts and retrieves statements according to this analysis.

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

The preprocessing method exploits the basic structure of RDF data. RDF, in order to describe resources provides a simple tuple model, <*S*,*P*,*O*>. The interpretation of this *statement* is that subject *S* has property *P* with value *O*, where *S* and *P* are resource URIs and *O* is either a URI or a literal value. This simplified approach of knowledge expression, namely *triples*, is handled in an efficient way so as to achieve optimization of query processing of RDF data.

The architecture and the main processing phases of our mechanism are described in the sections that follow. We detail the query analysis method and the data retrieval method, as these phases play an essential role in the optimization of our query processing technique.

3.1 Architecture

The basic architecture of our SPARQL query mechanism is composed of a central storage system where all retrieved RDF triples are gathered and a central processing system which facilitates the communication with remote repositories (Fig. 1).

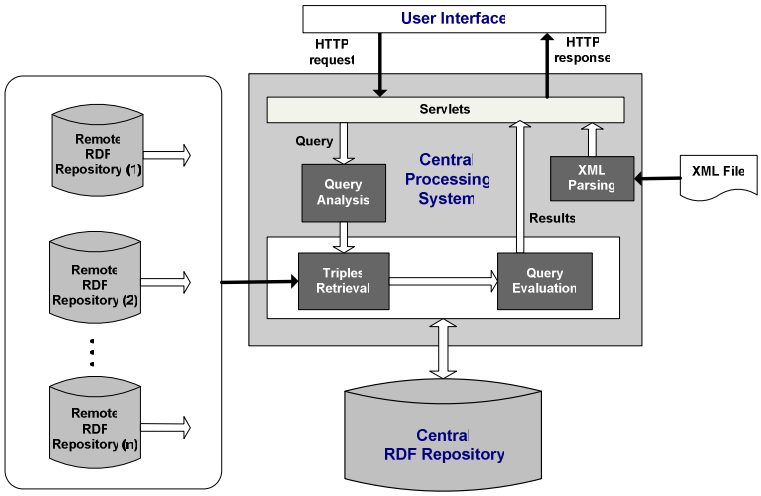


Fig. 1. Architecture of the SPARQL query mechanism for distributed RDF repositories. Central Processing System is divided into smaller, interconnected subsystems, each playing a fundamental role in the query evaluation process.

The processing system has, among others, the essential role of analyzing any query into smaller parts, as well as that of retrieving data from each remote repository, according to each resulting part. As a final step, it has to evaluate initial query against the central, local RDF repository.

The query is given as a string in a text box of a simple user interface and is passed to the central processing system through an HTTP connection. All necessary information, needed to establish connection to the available Sesame servers and RDF

repositories, is kept in a XML file, which is parsed by the corresponding mechanism in the central processing system.

3.2 Processing Phases

The basic idea of our mechanism is not just to fetch query results from each remote repository but to retrieve from them all the triples that take part in the formulation of final results. This kind of approach gives us the opportunity to combine distributed knowledge and obtain new inferred RDF statements.

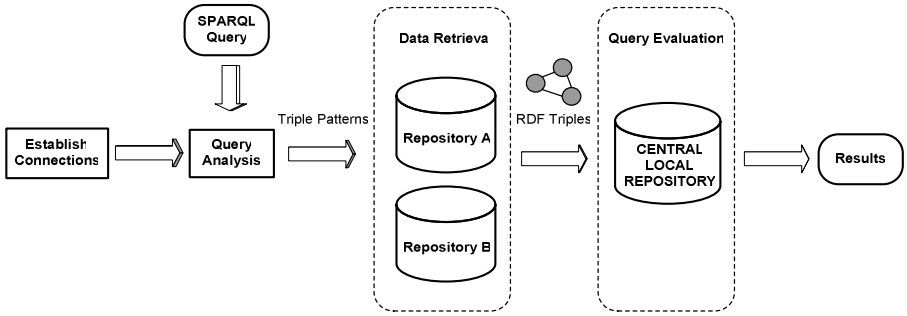


Fig. 2 Processing phases.

The first step is focused on establishing the connection with every available remote RDF repository (see Fig. 2). Afterwards, a query preprocessing is done, according to a method that involves a kind of syntactic analysis of query strings and their breaking into separate triple patterns and is detailed in section 3.3. Following this analysis, from each remote repository we retrieve RDF statements that adhere to these obtained triple patterns. All retrieved statements are placed in a central, local RDF repository, that also supports reasoning about RDFS. The central repository is emptied every time a new query evaluation is taking place. As soon as all RDF triples have been gathered from all available repositories, evaluation of the initial query follows. Evaluation is done against central repository's data, producing final results.

The motivation of our query analysis approach is better explained through the following example, shown in Fig. 3 and Fig. 4.

If we request all super-properties of <*subjectA*>, then, by evaluating such a query against each repository separately and by merging the results, we would get nothing more than single entities answering to the initial query, namely <*objectA1*>, <*objectA2*> and <*objectB4*> which cannot be further utilized (Fig. 3).

If we try, instead, to retrieve from each remote repository, not just the results, but statements that take part in the formulation of the results, the evaluation of the initial query against the central repository leads to the acquisition of new data, namely to new inferred information. Hence, in the following example, it is due to this analysis that we end up to obtain two extra statements, talking about super-properties of other entities. Furthermore, the central repository's capability to reason about RDF data leads to the proper combination of distributed data and to the disclosure of possible

"hidden" knowledge. As shown in Fig. 4, final answers to the example's initial query come from some extra information, apart from explicitly stored objects.

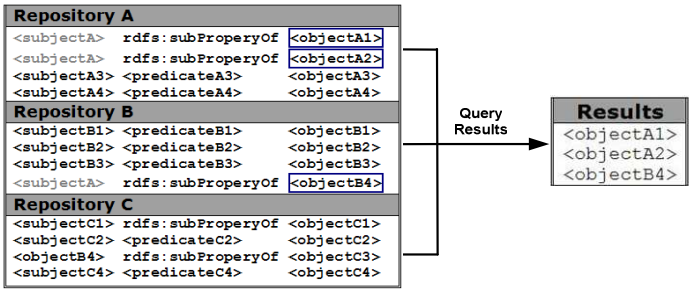


Fig. 3. Example of objects retrieved during a query evaluation against each of the three available distributed RDF repositories.

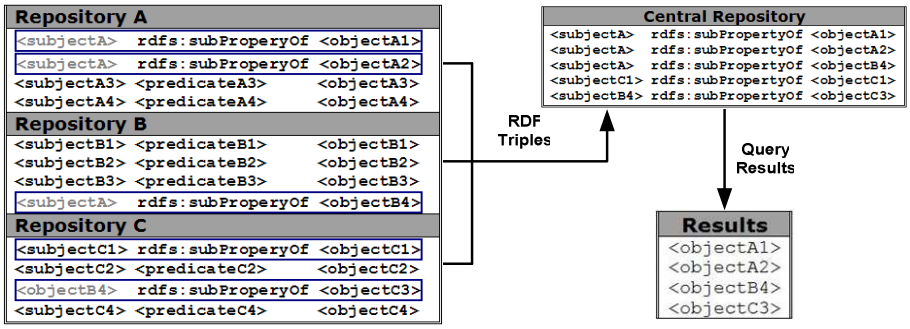


Fig. 4. Example of triples retrieved during a query evaluation against three distributed RDF repositories, using a central, local repository as an immediate store.

3.3 Query Analysis and Optimization

Query analysis is based on the idea that SPARQL queries are composed of one or more triple patterns, which create a graph pattern. A triple pattern is a RDF triple of the type <*S*, *P*, *O*> having the difference that some of its terms may have been replaced by variables. The objective of a query is to obtain results that match each triple pattern stated in its graph pattern. Hence, the main goal of the analysis is to recognize triple patterns found in a query's main body and retrieve RDF statements adhering to these patterns.

As shown in Table 1, where the structure of a SPARQL query is described, triple patterns are located in the fourth field of the query, namely in the WHERE clause. Therefore, analysis is focused in this particular field. It is also necessary to analyze the first field, the prologue, where all used base and prefix URIs are stated.

In the current phase of query analysis, we consider optional matches as obligatory, whereas we ignore any restriction imposed on retrieved data, stated by the keyword

FILTER. The first consideration has to do with our demand to avoid losing any possible utilizable information. As far as filter constraints are concerned, they mainly refer to the matching of data to certain numeric or string values: although these values would have restricted us to the retrieval of fewer RDF triples, the consideration of filter constraints can be rather time consuming, when applied to each repository separately. Finally, we don't take into account possible use of datasets, as this feature of SPARQL can't be utilized yet in a way that could benefit our mechanism.

Table 1. Structure of SPARQL query.

1. Prologue (*optional*)	`BASE <iri>`
	`PREFIX prefix: <iri> (repeatable)`
2. Query Result Forms (*required*)	`SELECT (DISTINCT)` sequence of ?variable
	`SELECT (DISTINCT) *`
	`CONSTRUCT {graph pattern}`
	`ASK`
	`DESCRIBE` sequence of ?variable or `<iri>`
	`DESCRIBE *`
3. **Query Dataset Sources** (*optional*)	Add triples to the background graph (repeatable): `FROM <iri>`
	Add a named graph (repeatable): `FROM NAMED <iri>`
4. Graph pattern (*optional, required for ASK*)	`WHERE {graph pattern [FILTER expression]}`
5. Query results ordering (*optional*)	`ORDER BY …`
6. Query results selection (*optional*)	`LIMIT n, OFFSET m`

By treating each query as a simple string, our method starts its analysis by searching the prologue field in order to find used URI bases and prefixes, which are necessary for later steps. We continue by analyzing the WHERE clause so as to find the stated triple patterns. Triple patterns are written as a whitespace-separated list of subject, predicate and object and the process of breaking them into smaller pieces is relatively complicated and based on the specification of SPARQL. The detailed description of this process is beyond the scope of this paper.

Following some simple rules and restrictions we reach to a point where each triple pattern found in the query's graph pattern has been isolated and seen as an individual triple with a subject, a predicate and an object. The RDF statement retrieval is done according to these obtained triple patterns.

3. 4 Data Retrieval

As mentioned above, a triple pattern is just a RDF triple (<*S*,*P*,*O*>) where some of its terms may have been replaced by variables. Another option for these terms is to be blank nodes or either to have a certain value (e.g. a URI or a literal). The data

retrieval method depends on whether or not inferred statements are requested to be included in the results, as well as on the type (variable, blank node, URI or literal) of the terms found in the triple patterns.

When concrete values (URI or literal) are used for triple terms, and provided that no inferencing is required, then we retrieve RDF statements that match these given values. When, instead, a triple term is represented by a variable or a blank node, we translate it as having to retrieve triples with any value at this particular place.

When inferencing constitutes the main objective of a query, data retrieval is done in a slightly different way. In this case, for each term found in a triple pattern of the query, we initially define its role (subject, predicate or object) and then we seek statements matching its value for this particular role. We also retrieve inferred statements related to this term. That means that we retrieve data in a way that ignores combinations of a certain subject-predicate-object triple. Instead, each term is treated as a separate entity and the retrieval process adheres only to this term's restrictions. Obviously, terms represented by variables or blank nodes are excluded, as they are seen as entities which can take "any value".

Table 2. Triples stored in each repository.

Sesame Server 1
Repository A
ex:Animal rdf:type rdfs:Class ex:Bear rdfs:subClassOf ex:Mammal
Repository B
ex:BrownBear rdfs:subClassOf ex:Bear
Sesame Server 2
Repository C
ex:Mammal rdfs:subClassOf ex:Animal ex:Fish rdfs:subClassOf ex:Animal ex:Bird rdfs:subClassOf ex:Animal
Repository D
ex:PolarBear rdfs:subClassOf ex:Bear
Sesame Server 3
Repository E
ex:Salmon rdfs:subClassOf ex:Fish ex:PolarBear ex:colour "white" ex:BrownBear ex:colour "brown" ex:PolarBear ex:eat ex:Seal ex:BrownBear ex:eat ex:Salmon ex:BrownBear ex:eat ex:Honey ex:PolarBear ex:weight "600"^^xsd:integer ex:BrownBear ex:weight "250"^^xsd:integer

Use Cases and Results

Based on Sesame's available API for storing RDF data and querying local and remote RDF repositories, our SPARQL query mechanism exploits all Sesame features related to reasoning about RDFS data. In this section we show a use case which demonstrates how this mechanism extends RDFS inferencing capabilities in distributed repositories.

We tested our mechanism using five different RDF repositories, located in three Sesame servers. Each repository's content was carefully chosen so that, if seen in combination with the stored triples in the other repositories, new RDF statements could be inferred. The distribution of RDF data in the available repositories is shown in Table 2. In this table, RDF data are represented as triples in their short form, using prefixes defined as follows:

1. prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2. prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3. prefix ex: <http://anemos.sesameweb.org/ex#>
4. prefix xsd: <http://www.w3.org/2001/XMLSchema#>

A more detailed look in the content of available repositories shows that a combination of their data implies a simple hierarchy of *Animal* class, as depicted in Fig. 5.

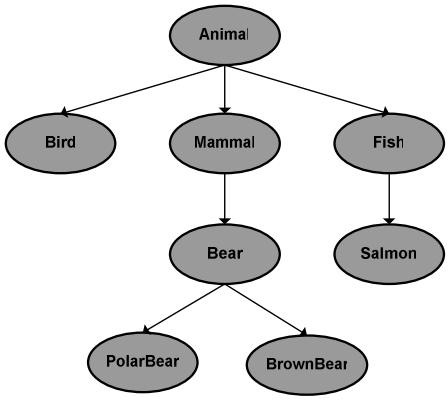


Fig. 5. Hierarchy of the Animal Class

To show our mechanism's special features in combining distributed knowledge, we evaluated a query for the simple case where no inferencing is required. We did the same when inferred statements is asked to be included in the formulation of the final results. In the next paragraphs we comment on the results.

Our first example query requests the projection of all possible superclasses of *PolarBear* class and is structured as follows:

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX ex:<http://example.org/ex#>
SELECT DISTINCT ?class
WHERE {ex:PolarBear rdfs:subClassOf ?class }
```

In the first case, where inferred statements are excluded from the query evaluation, the only answer we take is that *PolarBear* is subclass of *Bear*.Taking a detailed look at repositories' contents in Table 2, we indeed see that the only information about *PolarBear* superclasses is given in Repository D. The latter contains only one statement which corresponds to the necessary answer. No other repository contains explicit information about superclasses of *PolarBear*.

As a second step, we evaluate the previous query but this time we request the participation of inferred statements in the results. The answers we take contain some extra objects due to the effects of the reasoning process. Therefore, beyond the obvious result that *Bear* is superclass of *PolarBear*, we get the extra information that *PolarBear* is also subclass of *Mammal* and *Animal*, as shown in the Table 3. This is because in Repository A there is a triple stating that *Bear* is subclass of *Mammal*, whereas in Repository C we find the information that *Mammal* is subclass of *Animal* class. The combination of these three triples explains the final answers.

The fact that the results also contain the information that *PolarBear* is subclass of *Resource* and of itself comes from RDFS rules, stating that every class is subclass of itself and of *Resource* class.

Table 3. Results of the first query, including inferred statements.

class
http://example.org/ex#Bear
http://example.org/ex#Mammal
http://example.org/ex#Animal
http://example.org/ex#PolarBear
http://www.w3.org/2000/01/rdf-schema#Resource

A more complicated example, asks for the projection of the animal class, its corresponding weight and nominated food class as well as of all possible superclasses of the latter, restricting the results to those animals weighting less than a certain number. This query can be stated as follows:

```
PREFIX ex:<http://anemos.sesameweb.org/ex#>
SELECT DISTINCT ?animal ?weight ?food ?subClass
WHERE {   ?animal ex:eat ?food .
          ?animal ex:weight ?weight .
          OPTIONAL {?food rdfs:subClassOf ?subClass}.
          FILTER (?weight < 500)  }
```

Results, in the first case, when no inferencing is asked, correspond to the repositories explicit information, as shown in Table 4 (triple terms are depicted in their short form).

In the second case where inferred statements are taking part in the query process, results differ, as they contain new knowledge about superclasses of *Salmon* class, coming from the combination of data located in the other repositories (Table 5). Hence, new results inform us that *Salmon* is also subclass of *Animal* class, based on information coming from Repository C: *Fish*, an explicitly stated superclass of *Salmon*, is also subclass of *Animal*.

In this example, it is important to mention that results are indeed correct *and* complete, as optional matches and restrictions on numeric values were taken into account, omitting information about *PolarBear* class that has a weight value that is more than the requested one. These limitation conditions, as described earlier, are omitted during the phase of query analysis and data retrieval, but are always included in the last phase of query evaluation against the central repository. Therefore, our

choice to exclude them from the preprocessing phase in order to avoid extra overhead seems to have no negative effect in the final results.

Table 4. Results of the second query, where inferred statements are excluded from the results.

class	Weight	food	subClass
ex:BrownBear	"250"xsd:integer	ex:Salmon	ex:Fish
ex:BrownBear	"250"xsd:integer	ex:Honey	-no binding-

Table 5. Results of the second query, where inferred statements are included in the results

class	Weight	food	subClass
ex:BrownBear	"250"xsd:integer	ex:Salmon	ex:Fish
ex:BrownBear	"250"xsd:integer	ex:Salmon	rdfs:Resource
ex:BrownBear	"250"xsd:integer	ex:Salmon	ex:Salmon
ex:BrownBear	"250"xsd:integer	ex:Salmon	ex:Animal
ex:BrownBear	"250"xsd:integer	ex:Honey	-no binding-

5 Conclusions and Future Work

RDF distributed querying appears to be of crucial importance, having in mind the scattered nature of web resources descriptions. To this end, we have shown how distributed query answering can be achieved, in multiple RDF repositories. We have done so following a standards based approach, namely adopting the recent W3C standard SPARQL as querying protocol.

In addition, care has been shown on how to treat queries that involve reasoning as well. In this case, we have suggested a concrete preprocessing technique that amounts to a careful lexical analysis of the query and aims ultimately at the optimization of the response by amortizing related overheads, such as fetching and reasoning times.

As a proof of concept, a web-based querying mechanism has been developed, building upon and extending Sesame. Taking advantage of this service, we have presented a series of results that clearly demonstrate the potential of our method. At the same time this mechanism can be seen as enabling for richer interfaces that can be used in querying and harvesting metadata from distributed repositories and digital libraries. In particular, by gathering, combining and querying distributed information in a transparent way, it contributes in communicative interoperability of such digital stores. Furthermore, possible applied inferencing, leads to implied associations and information, thus gaining in the field of semantic interoperability.

Potential improvements to this approach may involve: first, the in-depth exploration of the various overheads occurring during the querying process, in an attempt to devise an upper bound on the optimization that is possible to be achieved; second the implementation and support for more expressive Semantic Web languages, namely OWL, a situation where reasoning-based query analysis will be even more subtle and demanding.

References

1 Adamku, G., Stuckenschmidt, H.,: Implementation and Evaluation of a Distributed RDF Storage and Retrieval System. In: Proceedings of of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (2005)

2 Brickley, D., Guha, R.V., (eds): RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, http://www.w3.org/TR/rdf-schema/

3 Broekstra, J., Kampman, A., Harmelen, van F. : Sesame: A generic architecture for storing and querying rdf and rdf schema. In: The Semantic Web - ISWC 2002, volume 2342 of LNCS, pages 54–68. Springer (2002)

4 Broekstra, J., Kampman, A.: Inferencing and truth maintenance in RDF Schema: exploring a naive practical approach. In: Workshop on Practical and Scalable Semantic Systems (PSSS) 2003, Second International Semantic Web Conference (ISWC), Sanibel Island, Florida, USA (2003)

5 Broekstra, J., Kampman, A.: SeRQL: An RDF Query and Transformation Language. In: the Proceedings of the Third International Semantic Web Conference (2004)

6 Broekstra, J., Kampman, A.: Serql: A second generation RDF query language, Technical report (2003)

7 Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K.: Jena: implementing the semantic web recommendations. In: Proceedings of the 13th international World Wide Web Conference on Alternate Track Papers & Posters (New York, NY, USA, 2004). WWW Alt. '04. ACM Press, New York, NY, pp. 74-83.

8 Haase P., Broekstra, J., Eberhart, A., Volz, R.: A Comparison of RDF Query Languages. In: Proceedings of the Third International Semantic Web Conference (2004)

9 Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., Schol, M.: RQL: A Declarative Query Language for RDF. In: Proceedings of the Eleventh International World Wide Web Conference (WWW'02), Honolulu, Hawaii, USA (2002)

10 Klyne, G., Carroll, J. J., (eds): Resource Description Framework (RDF):Concepts and Abstract Syntax. W3C Recommendation, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

11 McBride, B.: Jena: Implementing the RDF model and syntax specification. In S. Decker et al., editors, Second International Workshop on the Semantic Web, Hong Kong (2001)

12 Seaborne A.: RDQL - a query language for RDF, W3C member submission (2004), http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/

13 Stuckenschmidt, H., Waard, A. de, Bhogal, R., Fluit, C., Kampman, A., Buel, J. van, Mulligen, E. van, Broekstra, J., Crowlesmith, I., Harmelen, F. van, Scerri, T.: Exploring large document repositories with rdf technology - the dope project. IEEE Intelligent Systems (2004)

14 SPARQL query language for RDF. W3C Recommendation, http://www.w3.org/TR/rdf-sparql-query/

Semantic Maps and Digital Islands: Semantic Web technologies for the future of Cultural Heritage Digital Libraries

Achille Felicetti[1], Hubert Mara[1]

[1] PIN, University of Florence, Italy
{ achille.felicetti, hubert.mara}@pin.unifi.it

Abstract. This paper provides an overview regarding the application of the Semantic Web oriented technologies we have developed as part of the EPOCH and AMA projects for Cultural Heritage Digital Libraries. We wanted to enhance interoperability among diverse archives and to make disperse digital information available through the web in a standard format. Our toolset includes an application for mapping existing archive schema to ontology schema (AMA Mapping Tool), a tool to recursively markup unstructured text documents (AMA Text Tool) and a Semantic Web Database able to store, query and return simple and complex semantic information (MAD). We used the CIDOC-CRM core ontology to define the entities we dealt with and to describe concepts and relations among them. The framework has been tested on the Arrigo VII Digital Kiosk, a multimedia application combining 3D models, images, movies, sounds, free texts and HTML pages. FEDORA Digital Object Repository was thus used to create our test digital archive.

Keywords: Semantic Web, Mapping, Ontologies, CIDOC-CRM, Cultural Heritage, Digital Libraries

1 Introduction

Digital Libraries provide access to large amounts of resources in the form of digital objects and a huge variety of tools to search, browse and use the digital content in many ways. Complex and flexible systems of metadata schema have also been developed and used to describe digital objects in collections (i.e. Dublin Core, MODS, METS).

Unfortunately every system and metadata schema has a closed structure and uses different approaches to deal with digital data. Thus every system tends to become a wonderful dream island, rich in wonders to explore and treasures to discover. Most of the time, however, if one does not possess the right map, discovery and exploration are difficult.

The European Commission is funding numerous projects related to Digital Libraries to enhance the availability, usability and the long term preservation of relevant information of any kind. The biggest challenge will be to guarantee all these

features in the future. This will occur, in our opinion, through the implementation of strong semantic layers settled on top of Digital Libraries to unify the description of objects and concepts coming from different digital archives.

Semantic Web tools allow digital library maintainers to link their schema and automatically translate their terms, expanding mutual comprehensibility.

The semantic layers will be the Treasure Maps allowing users to discover Treasure Islands: any "sailor" can easily find his way to the information if the "coordinates system" of his map is designed in a uniform, logic and accurate way and if "X" clearly marks the spot.

This paper tries to outline the importance of using Semantic Web technologies to enhance the usability and interoperability of Digital Libraries. It describes the tools and the technology we developed as results of our research on standards and semantic web technology application in the framework of the EPOCH [1] and AMA [2] projects to implement semantic capabilities on digital data.

2 Semantic integration

Digital Libraries are currently very complex entities. It is often difficult to explain what they actually are, but we could think of them as big indexes designed to serve as guides for retrieving and returning digital information stored on the Web in different formats and archives. The first problem to face is that every guide has its retrieval system and uses a metadata grammar to describe and index data so specifically that it would never work on other systems. None of these metadata systems can analyze all the information on the Web, unless we make them available through a machine understandable format using RDF [3].

The second relevant problem concerns the information itself: the huge variety of formats used to index data is a big obstacle to integration and must be seriously analyzed. Even if we limit our efforts solely to Cultural Heritage archives (i.e. databases of museums and collections, archaeological excavation records, reports and other unstructured data) we have to admit that information is as dispersed as the material culture it refers to.

To create a uniform conceptual layer, semantic information should be extracted from databases, HTML pages, descriptive texts, metadata tags and put into a standard format in order to capture the conceptual meanings and to create correspondences at a higher level (conceptual mapping). These operations are facilitated today by the constant activity of the W3 Consortium in defining new standards for web information encoding, such as RDF and SPARQL, by the use of ontologies such CIDOC-CRM [4], explicitly created for Cultural Heritage and by the power of the available semantic tools, able to physically extract semantic information in a semi-automatic way [5].

Once the conceptual layer for both data and metadata is ready, the semantic information will be stored in a container based on RDF and ontologies. This will be the real integration place, where unified conceptual information from different digital archives will be browsed and searched as a unique Digital Library. The RDF language is robust and flexible enough to guarantee interoperability and to provide a common

denominator not only among Digital Libraries, but also with other systems and services.

3 Creating the Semantic Layer

3.1 Mapping

For every adventure, for every exploration, or any kind of quest for treasures, what is really needed is a map showing us how to reach the island and where to find the buried gold. Similarly the creation of logical maps is one of the most important activities towards data integration, even if the mapping process requires uncommon skills and knowledge among those who are supposed do the job.

The mapping process consists of defining the elements of a given data schema (starting schema) using the entities provided by another data schema (target schema) in order to establish a direct correspondence between the two different element sets. Usually the mapping occurs between a closed or personal data schema and a more general and widely used standard.

The standards presented in the past were not widely accepted by the community. Culture professionals and heritage policy makers found it difficult to map the existing data structures to them, which effectively impeded the preservation of existing archives. In fact, it is not the absence of a facilitating tool, but the existence of practices, legal obligations and the lack of a clear motivation that has delayed or reduced the creation of such mappings to a handful cases.

Already many mapping templates exist among common ontologies and metadata standards used for Digital Libraries (i.e. Dublin Core and FRBR metadata schema are already mapped to CIDOC-CRM ontology) and many others will be interconnected to create graphs of concepts that can leverage the semantic use of digital content [6]. In addition, the namespace mechanism provided by the RDF syntax can be used to extend existing ontologies by incorporating concepts coming from other schema to describe, for instance, 3D objects or geographic entities [7]. In the following RDF example, a GML polygon object has been used to extend the CIDOC-CRM information object describing a place:

```
<crm:E53.Place rdf:about="US1020">
 <crm:P67B.is_referred_to_by>
  <crm:E73.Information_Object
rdf:about="gmlModel_US1020"> ƒ
   <gml:Polygon srsName="osgb:BNG">
    <gml:outerBoundaryIs>
     <gml:LinearRing>
      <gml:coordinates>
          278534.100,187424.700
          278529.250,187430.900
          278528.700,187431.650
```

```
          278527.250,187433.600
        </gml:coordinates>
       </gml:LinearRing>
      </gml:outerBoundaryIs>
     </gml:Polygon>
    </crm:E73.Information_Object>
   </crm:E53.Place>
</rdf:RDF>
```

3.2 The AMA Mapping Tool

To make the mapping process easy and accessible, we created the AMA Mapping Tool, a flexible tool developed as part of the AMA Project to facilitate the mapping of different archaeological and museum collection data models (with various structured, as well as non-structured data, i.e. text description) to a common standard based on CIDOC-CRM ontology.

The tool is based on the concept of "template", this being the instantiation of the abstract mapping between the source data structure and the standard one. Templates capture the semantic structure, as well as the intellectual content, of the sources and their transformations.

The AMA Mapping Tool comes with a web application developed according to the open source principle (see Fig. 1). We chose to implement the tool in PHP5 using the DOM extension for parsing XML. A common request made by our EPOCH partners was to develop in the widely used PHP4, but for future sustainability we decided against it, as its end of life was announced on July, 13th 2007. Even though the capabilities of XML and XSD already allow a vast number of possibilities for database schema, we focused on adding support to RDF-Schema (RDFS). Therefore we choose to use the RAP - RDF API for PHP – of the Freie Universität Berlin, Germany [8], which is available under the GNU Lesser General Public License (GPL). Another reason to choose RAP, besides GPL and functionality, is the long-term existence and regular update policy of its authors. As PHP and RAP are used on (web-)server-side you need only a HTML-browser to access it – no Java(script) or any other plug-in, Flash or ActiveX is required.

To use the AMA tool you must upload a source and a target schema in any XML format, while RDFS is recommended, at least for the target schema (e.g. CIDOC-CRM), due to its semantic nature and its easier mapping capabilities. The next step is mapping pairs of classes (1:1). If it is necessary to map classes on 1:n or n:m ratio, this can be achieved by giving the same name to sets of pairs. Furthermore you can specify if a mapping is an alias (A) or an inheritance (I). The third and final step of the mapping is the definition of the relations between classes in the target schema using the mapped class' properties. This can be achieved either by choosing the property/range and then the object/class or vice versa: choose the object/class and then the property/range connecting them. The application also provides the possibility to add new classes, in case the existing ones are not sufficient to define a relation.

You can download your mapping schema as XML containing the mapping and the relations of classes. In addition a graphical representation of mapping and relations can be generated and directly browsed on-line.

The final result of the AMA mapping process is the creation of a mapping file to be used as a template for the data conversion process, directly in the original databases or along with conversion tools like D2R. Semantic data extracted from archives will be ready for integration with other CIDOC-CRM compliant archives and for use in Semantic Web contexts.

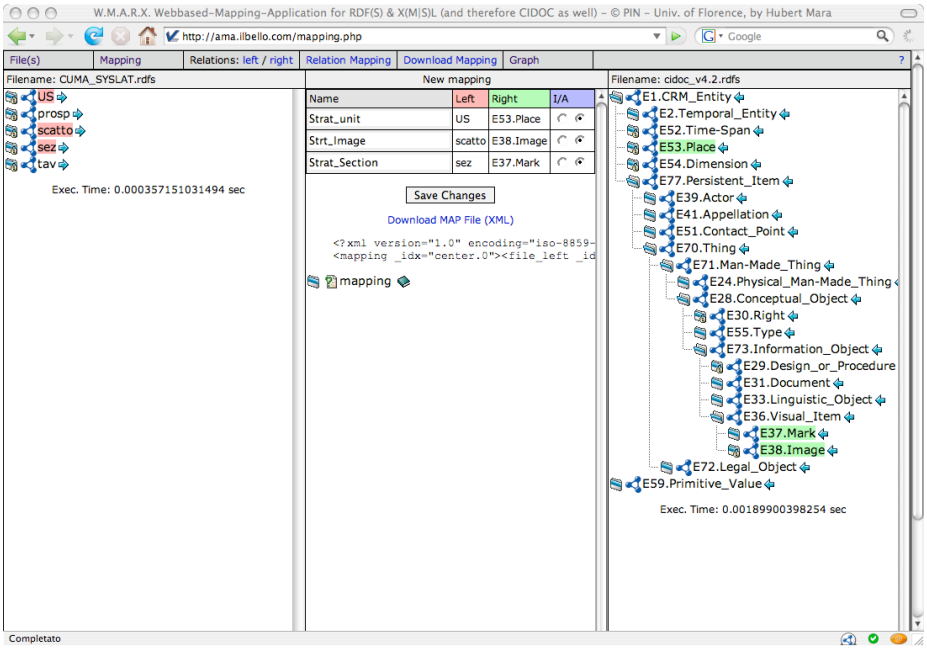


Fig. 1. The AMA Mapping Tool web interface in action to create mapping templates between a user ontology (left frame) and the CIDOC-CRM ontology (on the right frame). Mapped elements are shown in the central frame together with the mapping file generated on the fly.

3.3 Getting semantic information from databases

D2R is a tool providing a powerful set of features for real time conversion and query without modify legacy RDBMS.

The D2R tool can be used for publishing our relational databases on the Semantic Web, enabling RDF and HTML browsers to navigate the content of the database and allowing applications to query the database. SPARQL is the query language used. Then a customizable mapping framework called D2RQ is used to map database content into a set of virtual RDF datasets that can be browsed and searched by semantic web applications. Requests from the Web are rewritten into SQL queries via

the mapping. This on-the-fly translation allows to publish RDF from large live databases and eliminates the need for replicating the data into a dedicated RDF triple store.

D2R Server's Linked Data interface makes RDF descriptions of individual resources available over the HTTP protocol. An RDF description can be retrieved simply by accessing the resource's URI (or URL) over the Web. Using a Semantic Web browser like Tabulator or BrownSauce you can follow links from one resource to the next, surfing the Web of Data. D2R Server uses the D2RQ Mapping Language to map the content of a relational database to RDF. A D2RQ mapping specifies how resources are identified and which properties are used to describe the resources. The main advantage in using D2RQ is the possibility to customize the mapping by using elements from widely accepted ontologies. The mapping file can be edited with any text editor or automatically generated by modifying the AMA Tool middleware mapping files [9].

3.4 Unstructured documents: AMA TextTool

Another important step towards the process of data integration in Cultural Heritage concerns the encoding of free texts made available for processing by semantic engines. Dealing with this kind of document means dealing with unstructured information that would be impossible to extract and put in a semantic format using fully automatic procedures. Most of the data extraction or encoding is usually carried out manually by reading and keying or marking up text information using XML tags. However semi-automatic tools can assist the users during the encoding process and simplify their work by providing control mechanisms able to validate the markup and manage co-references in the text.

The Unit for Digital Documentation (EDD) of the University of Oslo, also involved in the AMA Project, has developed the AMA TextTool, a semi-automatic tool aimed at speeding up and improving the encoding process of archaeological texts in CIDOC-CRM. The tool is written in Java to be cross platform and is based on concepts and techniques from computational linguistics. The AMA TextTool implements a KWIC (Key Word In Context) concordance tool directly connected to the electronic text(s) used to find a word, a phrase or a pattern of words and possibly XML-markup already in the text. Users can then analyze the text and mark which occurrences in the KWIC concordance they want to tag. The system then inserts the mark up in the text file(s). This semi-automatic “search and replace” feature makes it possible for the user to create and include new algorithms, both for text retrieval and for text mark up, and new features of electronic text processing.

The AMA TextTool includes functions to enrich texts with a TEI-header, providing bibliographical information about the original text and the current electronic one, and structural markups to identify, for instance, chapters and paragraphs. The structural information can then be extended by adding the semantic layer with the identification and markup of the conceptual elements present in the texts (actors, objects, places, events). The latter operation can be accomplished using the set of tags and elements provided by ontologies like CIDOC-CRM.

After the tagging process is complete, the documents are ready to use in different Digital Libraries, both for their structural and semantic encoding. It is vital in this case to preserve the textual nature of this kind of document, avoiding the need to extract relevant pieces of information to be stored in relational databases. This operation would cause the destruction of the original document flow. HTML and TEI are suitable formats to make them accessible and easy to share over the Web. What can be extracted instead is the conceptual information to be used for the description of the documents' semantic maps and indexes. Conceptual descriptions can then be written using the RDF syntax and shared with other semantic information coming from databases and linked to the original documents stored in Digital Libraries using the URI mechanism [10].

4 MAD: a Semantic Web Database

4.1 Overview

A powerful container is needed to manage the complexity arising from the data encoded using ontologies and the RDF syntax. It must deal with semantic data, just as relational databases deal with data stored in tables and records and with their relationships. A Semantic Database is needed for this, a big container where semantic information is stored and maintained to provide users with the necessary tools for their semantic search. For this purpose we created MAD (Managing Archaeological Data), a framework originally designed to manage structured and unstructured archaeological excavation datasets encoded using XML syntax, including free text documents marked up in XML.

The latest release of MAD comes with a multipurpose semantic engine able to store and manage ontology encoded information, i.e. data structured in CIDOC-CRM compliant format, a semantic query set of interfaces based on SPARQL and RQL query languages and a Firefox plug-in implementing a semantic browser for RDF graphs.

MAD can be used to store, browse and query semantic data in many powerful ways, but also to transform and supply semantic information on demand. The whole framework has been developed as part of an EPOCH activity for creation of information management systems for the Semantic Web and is entirely based on Open Source software, XML standards and W3C technology [11].

4.2 The MAD Semantic Web Database

MAD is built around a powerful Java native XML Database providing technology to store and index XML documents in a file-system-like structure of folders and sub-

folders (collections). This container can be used to store information coming from mapped and digital metadata and content, along with annotations and tag sets created by users, mapping templates, schema, ontologies and everything can be expressed using the RDF language. Users can browse and query this integrated archive to get semantic information on Cultural Heritage objects described herein, or references to remote digital objects stored elsewhere (i.e. URIs and URLs linking specific resources). In this sense MAD acts as a Semantic Digital Library.

4.3 Semantic Queries in MAD

In order to query RDF data we are using the two most important languages available at present: SPARQL [12] and RQL [13], two semantic query languages designed for retrieving information from RDF graphs. These languages provide a clear, easy query syntax and the ability to obtain information from encoded documents without knowing its explicit syntactical structure (i.e. elements and properties names).

RQL is used due to its ability to combine schema and data querying using advanced pattern-matching facilities. SPARQL and RQL combined have been used for the creation of a group of semantic query interfaces able to parse the RDF documents in different ways. The ability of RQL to query ontology structures make the retrieval of classes, subclasses and properties from the models very simple allowing the building of structure-based semantic queries.

Classes and properties can be clicked to visualize subclasses and sub-properties and to define the elements that will participate to the query definition. While clicking on the different elements, an RQL query is constructed and then submitted, to be evaluated by the Semantic engine (Fig. 2). A human readable version of the query is also shown to make it understandable. An RQL query serialized on the CIDOC-CRM structure, for instance, may appear like this:

```
  select $class0
  from  {instance : $class0} @p {value : $class1}
  where $class0 in subClassOf( kyme:E28.Conceptual_Object
)
  and @p = kyme:P70B.is_documented_in
  and $class1 in subClassOf( kyme:E31.Document )
  and value like "*1022*"
```

It is the corresponding machine readable version of the clearer spoken language request: “I am looking for a *E28.Conceptual_Object* which *P70B.is_documented_in* the *E31.Document* containing '1022' “.

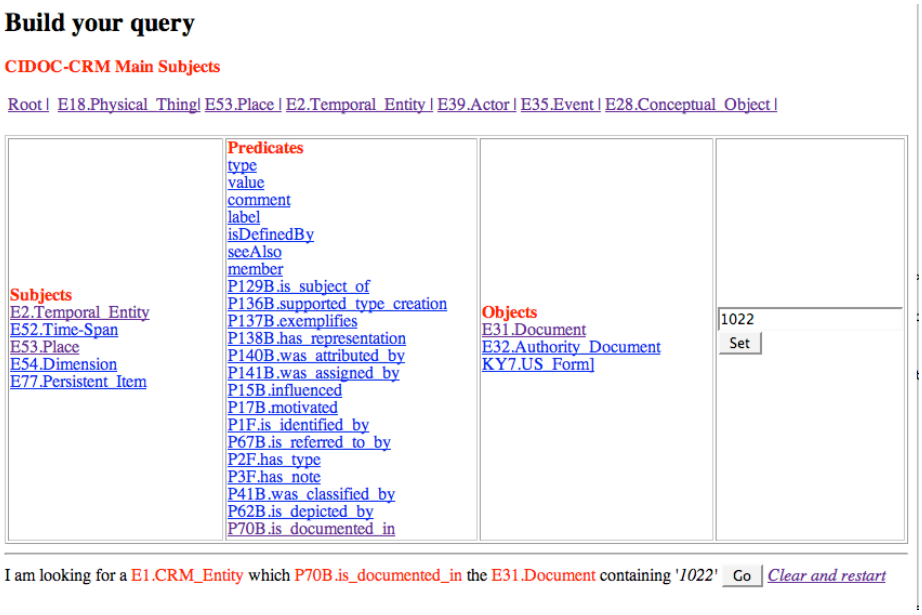


Fig. 2. The Semantic Query Interface created to allow users to build queries based on the combination of CIDOC-CRM classes and properties with free text.

4.4 MAD: The Semantic Web Browser plug-in

The richness of the RDF graph model in which data are distributed, makes it often difficult for users to get effective and meaningful data retrieval when only a simple or complex query interface is used, particularly when the graph grows in dimensions and complexity.

Sometimes it would be simpler and faster to browse the multidimensional graph structure allowing users to chose a starting point and move along different paths through the graph to reach the desired data. To allow this kind of data navigation we have developed a Mozilla Firefox plug-in based on SIMILE technology [14]. The plug-in turns the Firefox browser into a semantic browser able to open, search and save RDF datasets. Browsing is based on the faceted browsing UI paradigm.

A facet in this view is a particular metadata element considered important for the dataset browsed. Users can select a starting point which they consider relevant for their search. The browser extracts a list of facets, their values, and the number of times each facet value occurs in the dataset. Then it's possible to add or remove restrictions in order to focus on more specific or more general slices of the model. A "free text" restriction can be also added to reduce the browsed dataset to all items that contain the required string (Fig. 3). The interface was also configured to interact with the MAD container: all the semantic information stored therein can be browsed in order to retrieve relevant semantic objects and references to external resources. RDF

resources on the Web can also be visualized by providing their URLs and saving them in the MAD Semantic Database.

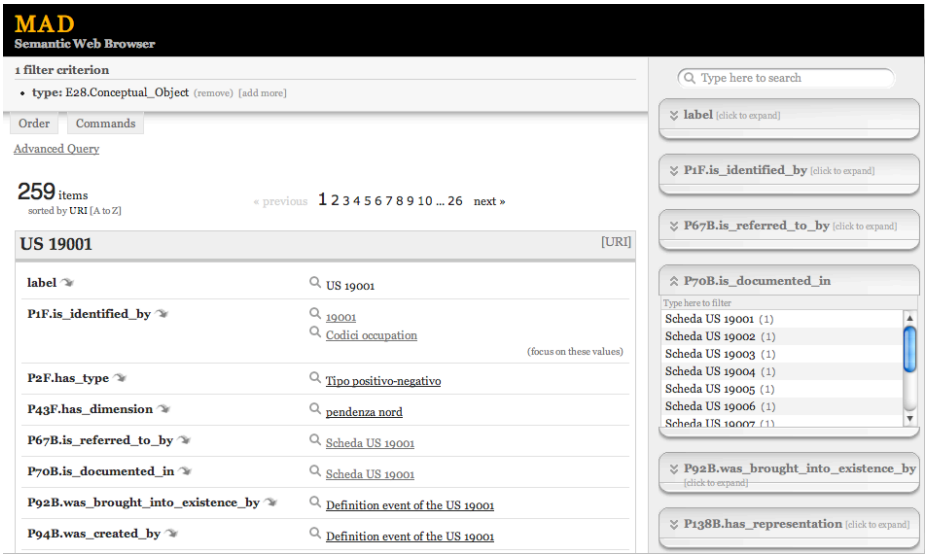


Fig. 3. MAD Semantic Browser Interface with faceted browsing capabilities showing entities and properties of the CIDOC-CRM.

We have tested the MAD framework to build an on-line version of the archaeological dataset recorded during the excavation of the ancient city of Cuma, containing information on stratigraphical units and other related resources. We are also using MAD with the AMA toolset to create an on-line application for the complete semantic management of coin collections for the COINS Project [15]. All this information is stored in RDF format and is ready to be queried, integrated and shared in the Semantic Web framework [16].

5 Case Study: the Arrigo VII funerary monument application

The *Arrigo VII Mausoleum* is a multimedia application created by the CNR of Pisa for the Museum of the Cathedral of Pisa. It reconstructs the history and the form of the funerary complex of the Holy Roman Emperor Henry (Arrigo in Italian) VII of Luxembourg, who died near Siena in 1313 and was buried in Pisa [17]. The Arrigo application has been included in the Interactive Salon, a touring exhibition created for the EPOCH Project and used for the creation of a demo shown during the EPOCH final event in Rome [18].

The application combines 3D reconstructions of the monuments, 3D models of the

statues still present in the Cathedral of Pisa, multimedia content including sounds and movies and descriptions, either in free text or hypertext formats. This case study was created to demonstrate the possibility of creating connections among digital objects, free texts and semantic information and was built in collaboration with the University of Leuven and the University of Graz.

To build a test Digital Library, 3D objects were extracted from the application, digitally annotated and stored in the Fedora container of the University of Leuven. In the same container we placed texts containing the descriptions of the various elements, marked up using the AMA TextTool to generate a set of TEI encoded HTML documents and the RDF CIDOC-CRM encoded information describing the content. All the RDF from digital annotations of 3D objects and from tagged texts were stored in the MAD Semantic Database to build a semantic index (semantic map) of our digital archive. Semantic information and digital objects were identified using the URI mechanism and linked via specific URLs.

Once the Digital Library and the Semantic Index are created, it is possible to retrieve digital objects by simply using the semantic interfaces provided by MAD to query the RDF information. It is possible, for instance, to find an object (i.e. a statue) made of marble and composed of different parts (hands, shoulders, head and so on), identified by an appellation and positioned in a specific place (i.e. a corner of the Cathedral of Pisa). The result of this query would be the RDF description of this entity (the statue), its URI and the URL pointer to the digital object stored in our Digital Library.

6 Conclusion and future work

Right from the start, the Semantic Web encountered a lot of resistance from developers and users, mainly for the lack of knowledge in this field and for the absence of applications that put this vision into practice. During the last years Semantic Web technologies have become mature enough to be used in a fruitful way and we are now able to take advantage of the power provided by them. The Arrigo case study demonstrated how simple the integration can be. The tools we developed proved to be powerful and flexible enough to be used for enhancing interoperability in different context, either in pure Semantic Web implementations or in Digital Libraries scenarios. Thanks to the CIDOC-CRM encoding of Arrigo's semantic information, the integration with information coming from other archives, already stored in MAD, was natural and immediate, as it is very simple to integrate elements when they are represented at a conceptual level (physical objects, places, actors and so on) [19]. The door towards Digital Libraries integration can now be fully opened.

Future work will focus on enhancing interoperability among our tools to guarantee a more fully integrated and usable system. Currently, for instance, the mapping files created by AMA needs further processing operations to be used for real data conversion. For these reasons we are going to provide MAD with a mapping framework able to connect the MAD engine directly to legacy databases and to extract information automatically, according to the high level mapping files created using the AMA Mapping Tool. Other important improvements will concern the

interfaces allowing users to perform more intuitive queries on semantic data, thus reducing the gap among query languages and natural languages.

References

1. EPOCH, European Network of Excellence in Open Cultural Heritage, http://www.epoch.eu
2. AMA, Archive Mapper for Archaeology, http://www.epoch.eu/AMA/
3. RDF, Resource Description Framework, http://www.w3.org/RDF/
4. Croft, S. N., Doerr, M., Gill, T., Stead S., (eds.): Definition of the CIDOC Conceptual Reference Model, http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.doc
5. Niccolucci, F., Felicetti, A.: Digital Libraries and Virtual Museums. In: Cappellini, V., Hemsley, J. (eds.) Electronic Imaging & the Visual Arts. EVA 2007 Florence (2007)
6. Doerr, M., Le Boeuf, P.: Modelling Intellectual Processes: the FRBR – CRM Harmonization. First DELOS Conference on Digital Libraries, February 2007 Tirrenia, Pisa, Italy (2007)
7. Felicetti, A., Lorenzini, M.: Open Source and Open Standards for using integrated geographic data on the web. In: Arnold, D., Chalmers, A., Niccolucci, F. (eds.) Future Technologies to Empower Heritage Professionals. The 8th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2007). Brighton 26-30 November (2007)
8. RAP - RDF API for PHP, http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/
9. D2R Server. Publishing relational databases on the semantic, http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
10. Eide, Ø., Felicetti, A., Ore, C.E., D'Andrea, A., Holmen, J.: Encoding Cultural Heritage Information for the Semantic Web. Procedures for Data Integration through CIDOC-CRM Mapping, in press
11. Felicetti, A.: MAD: Managing Archaeological Data. In: Ioannides, M., Arnold, D., Niccolucci, F., Mania, K. (eds) The e-volution of Information Communication Technology in Cultural Heritage. The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2006), Nicosia, Cyprus, 30 October – 4 November (2006)
12. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
13. Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., Scholl, M.: RQL: A Declarative Query Language for RDF. The Eleventh International World Wide Web Conference (WWW'02), Honolulu, Hawaii, USA, May 7-11 (2002)
14. The SIMILE Project, http://simile.mit.edu/
15. COINS: Combat On-line Illegal Numismatic Sales, http://www.coins-project.eu
16. MAD, Managing Archaeological Data, http://www.epoch.eu/MAD/
17. Baracchini, C., Brogi, A., Callieri, M., Capitani, L., Cignoni, P., Fasano, A., Montani, C., Nenci, C., Novello, R. P., Pingi, P., Ponchio, F., Scopigno, R.: Digital reconstruction of the Arrigo VII funerary complex. VAST 2004, Bruxelles 6-10 Dec., pp. 145-154 (2004)
18. Arrigo in Fedora: EPOCH project common infrastructure tool chain demo, http://partners.epoch--net.org/common_infrastructure/wiki/index.php/Arrigo_in_Fedora
19. Sugimoto, G., Felicetti, A., Perlingieri, C., Hermon, S.: CIDOC-CRM Spider. Stonehenge as a case study of semantic data integration. In: Arnold, D., Chalmers, A., Niccolucci, F. (eds.) Future Technologies to Empower Heritage Professionals. The 8th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2007). Brighton 26-30 November (2007)

Mapping, Embedding and Extending: Pathways to Semantic Interoperability The Case of Numismatic Collections

Andrea D'Andrea[1] and Franco Niccolucci[2]

[1] CISA, Università di Napoli „L'Orientale" ,
Napoli, Italy
dandrea@unior.it
[1] STARC, The Cyprus Institute ,
Nicosia, Cyprus
f.niccolucci@cyi.ac.cy

Abstract. This paper illustrates current mappings of Cultural Heritage data structures to CIDOC-CRM. After discussing the features of such mappings and evidencing the problems, it illustrates an example concerning coins collections.

Keywords: Ontologies, mapping, semantic interoperability

1 Introduction

Using ontologies to describe the "fundamentals" of knowledge is paramount to semantic interoperability. It does not guarantee it, though, because different sources of information, even if pertaining to the same domain, may obviously have different reference ontologies. To establish a correspondence among such diverse sets of basic concepts is therefore indispensable. This kinds of exercis is known under different names, such as alignment, harmonization, extension, mapping, and so on [1], which usually refer to different ways of pursuing interoperability. In some cases the result is indeed powerful and enriches the knowledge that can be extracted from the joint result. In others, the interoperability obtained in this way is illusory, because only a poor common organization can be superimposed on the joint information.

This may be the case when digital repositories pre-exist, and there is an effort to join them in a digital library. This operation requires a common basis and should result in preserving – as far as possible – the richness of parent repositories also in the sibling. Sometimes this is achieved; sometimes it is impossible; sometimes this issue is overlooked and the quality of the result may be measured only in terms of "digital objects" made jointly available, regardless of the actual utility of having them in the same repository. Often in the latter case Dublin Core is taken as a panacea, a minimal level of interoperabilty for any data collection and a catch-all alibi against objections of limited usability and usefulness.

More effective harmonization efforts in the Cultural Heritage domain have dealt with disparate archives, pushing the search for commonalities as far as possible. It is

indeed plausible that information on similar topics and within the same discipline is organized in a very similar way, despite of apparently different organization and denominations deriving from historic, cultural and administrative reasons. CIDOC-CRM, the well-known ISO standard for Cultural Heritage data [2], may be effectively used as common glue for preserving the richness of the original sources.

In this paper we will survey and summarize such harmonization efforts, trying to classify them, showing problems and, hopefully, how to manage them. Work recently undertaken within a EU project on numismatics will be summarily reported as an example of mapping exercises.

2 Mappings to CIDOC-CRM

The potential of CIDOC-CRM in serving as underlying common reasoning system to Cultural Heritage knowledge has been shown in more than one example.

Many mappings to CIDOC-CRM are already available for data encoded according to Dublin Core [3], EAD [4], AMICO [5], TEI [6] and FRBR [7]. Besides, CIDOC-CRM is aligned to DOLCE, a foundational top-level ontology.

Examples of extensions to other domain/task ontologies have also been developed, for example the one for MPEG7 [8], and more recently one for X3D [9].

National standards, for instance the English monument inventory system (MIDAS [10]), and English [11] or Italian [12] archaeological forms, have already been already mapped onto CIDOC-CRM. An on-going investigation aims at integrating the digital library of Perseus Project with Arachne, the central database for archaeological objects of the German Archaeological Institute (DAI) and the University of Cologne (FA): as both archives are encoded according to different information systems, the mapping on CIDOC-CRM is expected to facilitate mediation and integration of the resources [13].

The above quoted examples show different categories of “mappings”, using this term in a rather generic way. Some are qualified as harmonization, i.e. aim at reconciling different approaches. Others in fact embed within CIDOC-CRM alien concepts such as three-dimensionality or multimediality, showing how well they may fit in. Others, finally, transfer pre-existing data structures to the new standard and offer a way to overcome limitations, allowing for example the dialogue between museum information – when organized according to CIDOC-CRM – and on-field archaeological investigations, so far documented in a plethora of different ways that CIDOC-CRM might unify.

The process is neither straightforward nor painless, as the examples cited below show. Nevertheless, it is indispensable to proceed in this direction even at the price of slowing down the achievement of ambitious targets – the tenth of millions of digital objects stated in EU policies as a term of reference for digitization initiatives. It is a price, however, that perhaps it is not necessary to pay, thanks to new tools facilitating these mappings (as the AMA tool developed within the EPOCH project [14]) and to the enlarged commitment of the scientific community to develop the mapping first, and then proceed to merge digital repositories.

3 Ontologies and mappings

Different "mapping" exercises may be required to provide interoperability. Therefore, it may be useful to recall the most important concepts about ontologies and the related mapping types.

An *ontology* is a logical theory accounting for the *intended meaning* of a formal vocabulary, i.e. its *ontological commitment* to a particular *conceptualization* of the world.

According to the level of generality, there are different types of ontologies [15]:

- *Top-level ontologies*, describing very general concepts like space, time, matter, object, event, action, etc., which are independent of a particular problem or domain.
- *Domain ontologies* and *task ontologies*, describing, respectively, the vocabulary related to a generic domain or a generic task or activity, by specializing the terms introduced in the top-level ontology.
- *Application ontologies*, describing concepts depending both on a particular domain and task, which are often specializations of *both* the related ontologies. These concepts often correspond to *roles* played by domain entities while performing a certain activity.

Each domain uses a "local" ontology, producing data according to its own conceptualization. This process produces heterogeneous sources. In order to merge them, it is necessary to do more than a simple labelling mechanism identifying corresponding objects, classes or meanings. In other words, one cannot expect that it is sufficient to recognize that the object *A* in the ontology *X* is the same as the object *B* in the ontology *Y*: were this the case for all concepts, *X* and *Y* would be just the *same* ontology, with objects/classes disguised under different names. Actually it is often the case that concepts do not exactly overlap; that there are multiple correspondences and properties do not correspond; that some concepts of either are more specialized or differentiated than those in the other one. So the process of reconciling inhomogeneous sources specializes in different cases:

- *Extension*: it implies a specialization of the domain ontology, linking some concepts between the two original ontologies. The two conceptual models are in fact complementary, one of them detailing concepts from the other and covering elements or attributes ignored by the second one. This is the case, for instance, of the embedding of the X3D ontology into CIDOC-CRM developed in [9]: in this case it was shown how X3D may be used to extend the concept represented by the CIDOC-CRM element *E36.Visual_item* to detail information about 3D visualization, the scope of X3D but beyond the goal of CIDOC-CRM.
- *Harmonisation*: it implies a semantic equivalence between domain and application ontologies relating to the same ontological commitment. In this case a domain may be regarded as a specialization of the other, which represents a more general or abstract formal level.
- *Alignment*: it implies a generalization of the domain ontology through general axioms and concepts. The two models have (many/a few) general concepts in common.

According to their level of commitment there may be different level of integration (Fig. 1).

- *Extension*: through one or two similar concepts it is possible to set the equivalence between one or two classes in order to specialize the domain ontology with a complementary task-ontology.
- *Harmonization*: two domain-ontologies are harmonised when for a specific commitment it is possible to overlap many concepts or properties mutually.
- *Alignment*: a top-ontology alignments or incorporates a domain ontology through some equivalent and general classes corresponding to the ontological primitiveness, like time, space, and so on.

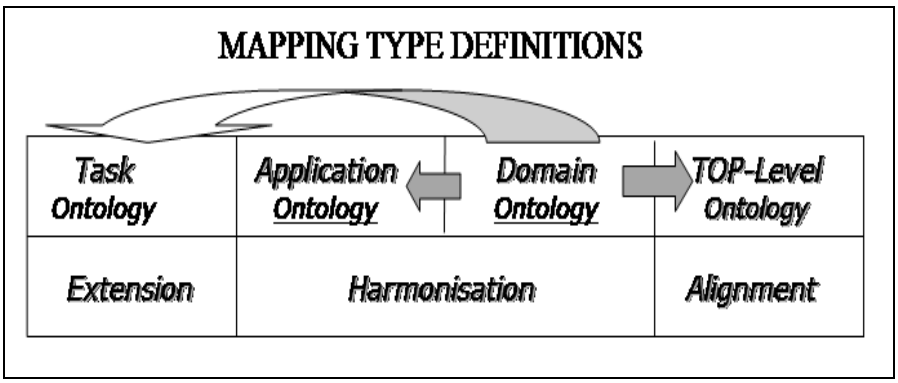


Fig. 1. Mapping types of domain ontology with other types of ontologies.

4 CIDOC-CRM as the Semantic Glue for Cultural Heritage

As far as archaeological documentation is concerned, CIDOC-CRM may indeed be considered as the universal glue to merge heterogeneous sources.

It does not define any of the terms appearing typically as data in the diverse data structures, rather it foresees the characteristic relationships for their use. It does not aim at proposing what cultural institutions should document, rather it explains the logic of what they actually currently document, and thereby enables semantic interoperability.

So far there are several attempts to conform and interpret digital content on the conceptual schema of CIDOC-CRM. However, at a more detailed exam of the different solutions, both from a technical and a theoretical point of view, they appear to adopt substantially different mapping mechanisms. Some solutions, based on an extension of the model, come alongside with integrations relying upon harmonization; in some cases the original model has been mapped on CIDOC-CRM classes; in others the model has been made richer creating new sub-classes and sub-properties. In the former case, it is possible to guarantee a higher compatibility with CIDOC-CRM; in the latter, based on specialization, it is possible to enrich the semantic content at the price of a lower compatibility. In any case, it is a standard that should be adopted since the very beginning of the repository design. But, as the AMA project [14] has

shown, excellent results can be obtained also from legacy archives. With this tools, not only different archaeological forms were made compatible, but also such archives were merged with unstructured text descriptions offering a unique global documentation system.

5 Mapping to CIDOC-CRM

In the introduction to CIDOC-CRM handbook its authors point out that "since the intended scope of the CRM is a subset of the "real" world and is therefore potentially infinite, the model has been designed to be extensible through the linkage of compatible external type hierarchies."

To explain correctly the method to be followed for such an extension, they precise that "a sufficient condition for the compatibility of an extension with the CRM is that CRM classes subsume all classes of the extension, and all properties of the extension are either subsumed by CRM properties, or are part of a path for which a CRM property is a shortcut."

In this sense "compatibility of extensions with the CRM means that data structured according to an extension must also remain valid as a CRM instance."

The CIDOC-CRM documentation gives a number of procedures that can be followed to allow that coverage of the intended scope is complete:

- Existing high-level classes can be extended, either structurally as subclasses or dynamically using the type hierarchy.
- Existing high-level properties can be extended, either structurally as sub-properties, or in some cases, dynamically, using properties of properties that allow sub-typing.
- Additional information that falls outside the semantics formally defined by the CRM can be recorded as unstructured data using *E1.CRM_Entity. P3.has_note: E62.String*.

Using mechanisms 1 or 2, the CRM concepts subsume and thereby cover the extensions. On the contrary, in mechanism 3, the information is accessible at the appropriate point in the respective knowledge base. CRM authors highlight that "this approach is preferable when detailed, targeted queries are not expected; in general, only those concepts used for formal querying need to be explicitly modelled: in some ontologies entities can have the same name, but different meaning or, vice versa, have different labels but refer to the same concept; in some cases an entity can be complementary to another or partially combine".

Starting and target structures may not correspond exactly: in this case a new (source) model fitting with the CIDOC-CRM hierarchy will be created, by making explicit some concepts that are implicit in the source, in order to parallel the axioms/paths characterizing CIDOC-CRM structure and hierarchy. This is a sort of anisomorphic process modifying the apparent structure of the starting source.

A typical case is the *event*, which is central to the structure of CIDOC-CRM but often is ignored in other models. However, also in these models there is some event determining a change in the state of the object being documented, be it the creation,

destruction or any other state change. To map correctly this model to the CRM, such events may be stated explicitly.

There may be many other cases of anisomorphism, for example when a unit (grams, centimetres, etc.) must be specified or a disambiguation is needed. We will show in section 7 other examples of this situation.

This shows that mapping is neither an easy matter nor a linear process and requires disciplinary competence as well as deep understanding of the implicit assumptions of the source model.

In conclusion, the process of creating the mapping among heterogeneous sources and different schemas may change in case of multiple conditions [KDP06]. Some basic rules have been identified:

- Introducing an intermediate node, in order to precise the whole path for the mapping of the source path. The most common such addition is an event.
- Contracting several classes, e.g. in address (names, coordinates), in one entity.
- Making explicit the causal connection between some related classes in order to allow the interoperability between information from other sources.
- As regards the previous rule, two relations in the source schema has to be mapped with the same intermediate node.
- Some elements may appear in multiple mappings.

The next paragraphs describe how mapping to CIDOC-CRM has been implemented as a data transfer mechanism. Many of the examples come from work made some years ago and rely upon an old CIDOC-CRM version, so they would need updating. Nevertheless they are an attempt to point out the overlapping between various domain-ontologies having in common CIDOC-CRM as an inter-lingua.

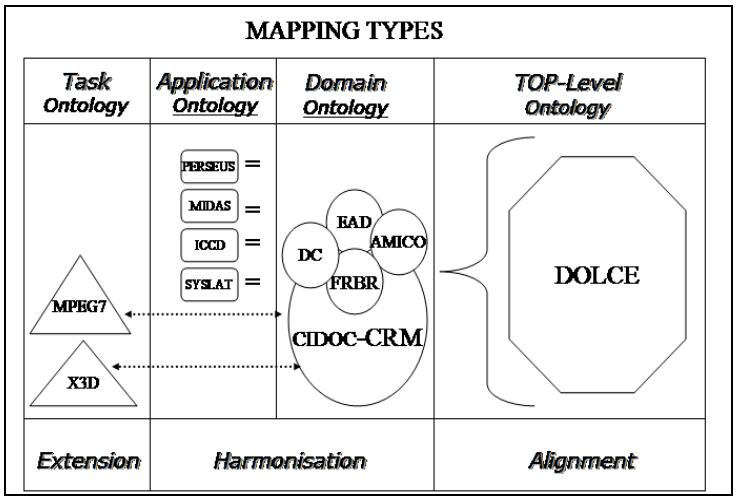


Fig. 2. Existing mappings to CIDOC-CRM.

Unfortunately, such experiments have not received so far sufficient attention by practitioners and have had no enough impact on current practice. They can however be regarded as experimental evidence of the fitness of CIDOC-ROM to its intended goal. Fig. 2 represents the linkages between CIDOC-CRM on one side and other ontologies on the other.

5.1 Harmonising *domain ontologies*: AMICO, DC, EAD, FRBR, TEI and mappings to CIDOC-CRM

AMICO, DC, EAD, FRBR and TEI deal with various domain ontologies reflecting different commitments not always completely covering the Cultural Heritage domain. In same cases the coverage is only partial, while in other cases the model can integrate and extend CIDOC-CRM original scope.

These models aim at:

- Managing standard for museum multimedia (AMICO – the standard is no more supported since 2005);
- Structuring metadata for cross-domain information resource such as video, sound, image, text, and composite media like web pages (DC);
- Designing intellectual or physical parts of archival finding aids (EAD) such as inventories, registers, indexes, and other documents created by archives, libraries, museums, and manuscript repositories (EAD);
- Capturing and representing the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information (FRBR);
- Implementing an interdisciplinary standard in order to enable libraries, museums, publishers, and individual scholars to represent a variety of literary and linguistic texts for online research, teaching, and preservation (TEI).

All these standards describe a structured record using elements, but any kind of relation or property is expressed between attributes or records; moreover discrete entities are not explicitly identified. In order to create a semantic equivalence the existence of common classes with the same meaning has been recognized and highlighted. Besides, according to CIDOC-CRM paths, one or more intermediate nodes need to be added to make the source structure compatible with the target one.

In these cases, beyond the specification of equivalent entities, a structure has been introduced in the mapping process; only adding these paths it has been possible to extract the implicit knowledge embedded in the sources.

5.2 Extending/Specializing CIDOC-CRM with complementary *application ontologies* as X3D and MPEG7

X3D and MPEG7 are both task ontologies adopted in order within specific domains: 3D content description the former, and latter multimedia the latter.

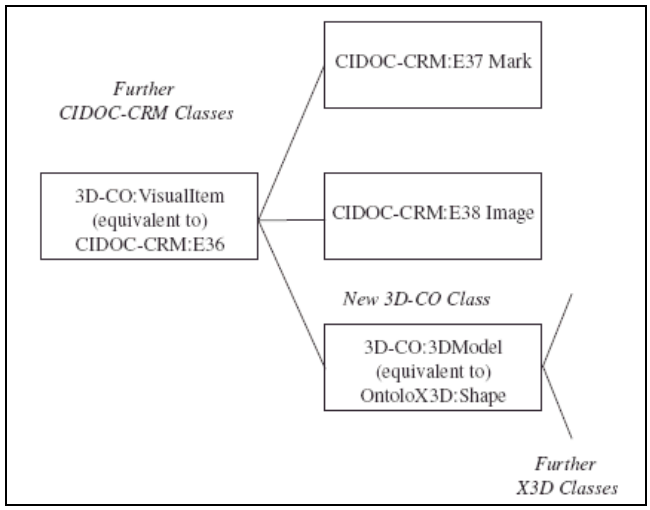


Fig. 3. Diagram of the 3D model extension (from [9] fig.3).

As such specialized aspects are out the scope of CIDOC-CRM, the extension has been realized linking two corresponding classes as shown in figure 3 above. In those cases models remain separated and are joined to each other only through a node, whose task is to specialize a general concept of CIDOC-CRM; in this way it isn't necessary to transform the structure of the source adjusting it to the CIDOC-CRM model. No intermediate node has therefore been added to the original ontologies.

5.2 Mapping *task ontologies* to CIDOC-CRM: MIDAS, English Heritage, ICCD, PERSEUS

Some task ontologies cover fully the CIDOC-CRM scope. They deal with local national standard implemented for managing specific tasks refer to cataloguing, inventorying and documenting archaeological data.

They arose above all from various attempts made at normalizing the production of archaeological documents, mainly coming out from excavations; in order to reduce the subjectivity in the recording process the normalization have lead to the elaboration of a large number of forms. These recommendations or best practices have been translated in tables and relational databases. Several information systems, implemented for the managing of the archaeological forms, are already mapped on CIDOC-CRM.

In such mappings, the original sources, containing data from an autonomous domain or sub-domain within Cultural Heritage, are read and re-interpreted in the light of CIDOC-CRM philosophy, adding only the nodes that are necessary to link entities and properties to events.

Considering some of these mapping procedures, it may be noticed that there are alternative ways of representing the same conceptual archiving practice. While the English Heritage mapping chose to base on the creation of new sub-classes of "IsA"

type specializing the original CIDOC-CRM and making it richer, the Italian ICCD mapping preferred to maintain a full compatibility with CIDOC-CRM, fitting the starting source with the destination ontology only through the semantic equivalence between corresponding classes.

Actually, from a theoretical and methodological point of view, both these mechanisms are formally correct, although they may be mutually integrated only at a general level. Perhaps, it would be appropriate to reach a consensus on the preferred way of performing this mapping exercise.

6 Tools for mapping CIDOC-CRM

As already mentioned, the EPOCH project has produced a mapping tool, which is being maintained and improved even after the conclusion of the project. It is denominated AMA (Archaeological Mapping Tool) [14], an acronym that hints to the reconciliation of different sources and, hopefully, to mutual understanding and "love" (*amor*, in Latin). The tool may in fact be used for any kind of data structures, not only for archaeological ones. AMA does not substitute human intelligence in the identification of corresponding concepts and relations. It simplifies the registration of such identification and facilitates the task of transcribing the correspondences. A proposed standardized way of registering the mapping is under way following the indications appeared in [16].

AMA is accessible on-line from the EPOCH web site and is available under a Creative Commons license.

7 The COINS mapping: work in progress and lessons learnt

COINS, another EU-funded project [17], is facing the task of reconciling different data organizations in the same domain, in this case numismatics. It was expected that this goal could be obtained with no great difficulty as numismatists have already achieved a high degree of homogeneity and standardization in their descriptions. The case studies are the archives of the numismatic collection at the Cambridge Fitzwilliam Museum, the coins collection of MNIR, the Romanian National Museum in Bucharest, and the huge numismatic collection of the Museum of Palazzo Massimo in Rome of the Soprintendenza Archeologica of Rome (SAR). These digital archives have been chosen as representatives of three different traditions, also geographically, which have nevertheless a long practice of scientific collaboration. Previous work on a similar subject was performed in [18], with a different approach.

Actually the data organization was similar, but differences were obviously present in the three digital repositories.

The coin mapping is a good example for concepts enounced in section 3. They show the need of expanding the data structure to take into account implicit concepts, which are necessary to a correct mapping.

In this case the two relevant events are the "coinage" and the "cataloguing". All the concepts may be related to one of these two events, implementing the event-based

approach of CIDOC-CRM. The insertion of the event "coinage" supersedes the slight differences between data structures, which become fully compatible with each other. For example:

Fitzwiliam	CIDOC-CRM	SAR
Maker_mint Maker_place Maker_State Maker Maker_role Maker_qualifier	Coinage	Zecca Zecca-Luogo Zecchiere_nome Zecchiere_luogo Stato Autorità Tipo_autorità

A more complex example refers to the authority minting the coin, usually managed as

AUTHORITY = *value*

for example

AUTHORITY = ATHENAI

meaning that the coin was minted by the town-state of Athens.

To map these concepts on CIDOC-CRM it is necessary to expand this relationship as follows:

P11F.has_participant	E40.Legal_Body (Authority)	P131F.is_identified_by	E82.Actor_Appellation (Athenai)

Even more complex is the apparently simple case of the weight, which is usually recorded as a figure without the indication of the unit, implicit in the source model and possibly, but not regularly, referred in some documentation or a standard one – unlikely to be for coins, which are never weighted in Kg.

P43F.has_dimension	E54.Dimension (Weight)	P90F.has_value	E60.Number	P91F.has_unit	E58.Measurement_Unit (gr)

In this case the field AUTHORITY is the instance of Legal_Body, while the value is the instance of the class E82. So the mapping has to include sometimes not only the value of a specific field but also the same fielding order to create a correct and understandable path.

It would probably be too long to enter here into other examples of further details. Mappings are available as project deliverables on the project web site.

The idea is to abstract from individual mappings – having, as already mentioned, slight differences from each other – to arrive at the construction of a general

numismatic reference model, fully compatible with CIDOC-CRM. The latter was chosen as universal glue for the systems, each of which is mapped to it. Such mappings allow semantic interoperability preserving all the richness of original sources. The project has also developed a collection management system enabling semantic searches, aptly named MAD (Management of Archaeological Data).

From the mappings the project team is progressively reconstructing the underlying ontology. Comparisons with other digital repositories and discussing intermediate results with the users' community are planned to extend the validity of this inductive approach.

8 Conclusions

Undertaking a standardization process involving archaeologists and archaeological data may perhaps be considered as a symptom of naivety. Few scientific communities are more individualistic than this, the result being an extreme fragmentation of systems and data models. It is always the case that a minute detail, ignored by the majority, is vital to some individual researcher, two such details never being the same for different investigators. In the end, all these small differences might lead to a paradoxical situation of many different data models that differ very little from each other when considered in couples, but have a rather limited intersection when taken all together.

Mapping to CIDOC-CRM is probably the only way to reconcile such disparate approaches, providing a very large common conceptual area but safeguarding the interests of individual researchers to personalize what they consider relevant information.

The above examples show the complexity and difficulty of the task of reconciling different sources of information and data models. In general, just pairing concepts does not suffice, but further source interpretation, disambiguation and resolution of co-references is required. On this regard, harmonization and alignment result simpler, because their scope is high-level reconciling. Being based on general and abstract concepts, it is usually easier to find correspondences and commonalities. Difficulties come when mapping is undertaken, as it has to cope with reality and practice. Here differences and lack of homogeneity are commonplace, and reconciling different archives may require much more effort and ingenuity.

At least, it is hoped that the above examples show that apparently easy shortcuts (AKA crosswalks) usually are trivialization of mappings, and often, rather than saving time, lead to an undesired destination. Only concrete practice of real archive mapping will eventually show the superiority of CIDOC-CRM as a global schema for the organization of Cultural Heritage knowledge and information.

References

1 Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. *The Knowledge Engineering Review* 18(1), 1-31.

2 Crofts N., Doerr M., Gill T., Stead S., Stiff M.: *CIDOC CRM Special Interest Group*, Version 4.2, June 2005. http://cidoc.ics.forth.gr/docs/cidoc_crm_version _4.2. pdf.
3 Kakali C., Lourdi I., Stasinopoulou T., Bountouri L., Papatheodorou Ch., Doerr M., Gergatsoulis M. In Integrating Dublin Core metadata for cultural heritage collections using ontologies, 2007 Proc. Int'l Conf. on Dublin Core and Metadata Applications.
4 Theodoridou M., Doerr M.: Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM. *Technical Report FORTH-ICS/TR-289*, June 2001.
http://www.ics.forth.gr/proj/isst/Publications/paperlink/mappingamicotocrm.pdf.
5 Doerr M.: Mapping of the AMICO data dictionary to the CIDOC CRM. *Technical Report FORTH-ICS/TR-288*, June 2000.
http://www.ics.forth.gr/proj/isst/Publications/paperlink/mappingamicotocrm.pdf.
6 Eide O., Ore C.E.: TEI, CIDOC-CRM and a Possible Interface between the Two.
http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/ore.pdf.
7 Doerr M., LeBoeuf P., Modelling Intellectual Processes: The FRBR - CRM Harmonization. http://cidoc.ics.forth.gr/docs/doer_le_boeuf.pdf.
8 Hunter J.: Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums, In *Museums on the Web 2002*, Boston, April 2002.
http://www.metadata.net/harmony/MW2002_paper.pdf.
9 Niccolucci, F., D'Andrea, A.: An Ontology for 3D Cultural Objects. in Ioannides, M., Arnold, D., Niccolucci, F., Mania F. (Eds), The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST2006, 203-210.
10 Crofts N.: MDA Spectrum CIDOC CRM mapping.
http://cidoc.ics.forth.gr/docs/MDA%20Spectrum_CIDOC_CRM_mapping.pdf.
11 Cripps P., Greenhalgh A., Fellows D., May K., Robinson D.: Ontological Modelling of the work of the Centre for Archaeology, September 2004.
http://cidoc.ics.forth.gr/docs/Ontological_Modelling_Project_Report_%20Sep2004 .pdf.
12 D'Andrea, A., Marchese G., Zoppi T.: Ontological Modelling for Archaeological Data. in Ioannides, M., Arnold, D., Niccolucci, F., Mania F. (Eds), The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, VAST2006, 211-218.
13 Kummer R.: Integrating data from The Perseus Project and Arachne using the CIDOC CRM.
http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/kummer.pdf.
14 Felicetti A.: MAD – Management of Archaeological Data. In M. Ioannides, D. Arnold, F. Niccolucci, K. Mania (eds.) *The e-volution of Information Communication Technology in Cultural Heritage – Project papers*. Budapest, Archaeolingua, 2006, 124 – 131.
15 Guarino N.: Formal Ontology in Information Systems. In *Proceedings of FOIS'98,* Trento, Italy, 6-8 June 1998, Amsterdam, IOS Press, 3-15.
16 Kondylakis H., Doerr M., Plexousakis D.: Mapping Language for Information Integration. *Technical Report 385, ICS-FORTH*, December 2006.
http://cidoc.ics.forth.gr/docs/Mapping_TR385_December06.pdf.
17 http://www.coins-project.eu/.

18 Nussbaumer P. Haslhofer B.: CIDOC-CRM in Action – Experiences and Challenges. ECDL 2007: 532-533.

A Methodological Framework for Thesaurus Semantic Interoperability*

E. Francesconi, S. Faro, E. Marinai, and G. Peruginelli

Institute of Legal Information Theory and Techniques
Italian National Research Council (ITTIG-CNR)
{francesconi,faro,marinai,peruginelli}@ittig.cnr.it
http://www.ittig.cnr.it

Abstract Thesaurus interoperability is an important property which garantees quality in indexing and retrieval of heterogeneous data sources in a distributed environment. This paper presents a methodological framework for semantic mapping between thesauri as well as a specific approach within such framework on a case study aimed at mapping five thesauri of interest for European Union institutions having only schema information available.

1 Introduction

In the last few years accessing heterogeneous data sources in a distributed environment has become a problem of increasing interest. In this scenario the availability of thesauri or ontologies able to provide a controlled source of terms or concepts is an essential pre-condition to guarantee quality in document indexing and retrieval. Currently a further necessity is growing, related to the availability of services able to guarantee cross-collections retrieval which means providing a query from a single interface and retrieving pertinent documents from different collections. Since the quality of the retrieval in single collections is often linked to the availability of specific thesauri, the quality of cross-collections retrieval is linked to the interoperability among thesauri of different collections.

In this context interoperability means using a particular thesaurus for users' query and mapping it to thesauri in other languages, to more specialized vocabularies, or to different versions of the thesaurus [1], in order to obtain a retrieval from different digital collections which is coherent to the original query.

This work proposes a methodological framework for semantic mapping between thesauri as well as a specific approach within such framework on a case study aimed at mapping five thesauri (EUROVOC, ECLAS, GEMET, UNESCO Thesaurus and ETT) of interest for European Union institutions having only schema information available.

This paper is organized as follows: in Section 2 the schema-based thesaurus mapping methodologies are summarized; in Section 3 a formal characterization

* This work has been developed within the tender n. 10118 "EUROVOC Studies" of the Office for Official Publications of the European Communities (OPOCE).

of the schema-based thesaurus mapping problem is proposed; in Section 4 the standards to be used for the proposed framework are introduced; in Section 5 the case study implementation of the proposed methodological framework is presented; finally in Sections 6 and 7 the interoperability assessment criteria and some experimental results for the case study are shown.

2 Overview of thesaurus mapping methodologies

Thesaurus mapping can be seen as the process of identifying terms, concepts and hierarchical relationships that are approximately equivalent [1]. The problem therefore is the definition of "equivalence" between concepts.

In literature "concept equivalence" is defined in terms of set theory: according to this vision two concepts are deemed to be equivalent if they are associated with, or classify the same set of objects [2] (*Instance-based mapping* [3]). This approach is characterized by the availability of data instances giving important insight into the content and the meaning of schema elements.

On the other hand concepts may also be associated with semantic features and mappings can be based on equivalences between feature sets (*Schema-based mapping* [3]). This approach is the only possible when only schema information is available, and it represents the case of interest for our study (mapping of five thesauri for which only schema information is available).

The most complete classification of state-of-the-art schema-based matching approaches can be found in [4], where schema-based methods are organized in two taxonomies with respect to the: *Granularity/Input interpretation* (classification based on the granularity of match (element or structure level)), *Kind of Input* (classification based on the kind of input (terminological, structural, semantic)).

Elementary matchers are distinguished by the *Granularity/Input interpretation* layer in terms of: a) element-level vs. structure-level (entities are analyzed in isolation or together in a structure) b) Syntactic vs. external vs. semantic (input is interpreted in function of its sole structure, exploiting auxiliary (external) resources, using some formal semantics to interpret the input and justify their results.

According to the *Kind of Input*, elementary matchers can be distinguished in terms of the kind of data the algorithms work on: a) strings (terminological or linguistic methods); b) structure (entities internal structure as well as relations with other entities); c) models (semantic interpretation of the knowledge structure, using also compliant reasoner to deduce the correspondences).

Taking into account this classification and elementary techniques described in literature, a framework for schema-based thesaurus mapping is proposed along with a methodology to implement schema-based thesaurus mapping for the case study.

3 A formal characterization of the schema-based thesaurus mapping problem

As introduced in Section 1 thesaurus mapping for the case-study is a problem of term alignments, where only schema information is available (*Schema-based mapping*). It can be considered a problem where to measure the conceptual/semantic similarity between a term (simple or complex)[1] in the source thesaurus and candidate terms in a target thesaurus, in case ranked according to the similarity degree.

These arguments allow us to propose a characterization of the schema-based Thesaurus Mapping ($\mathcal{TM}$) problem as a problem of Information Retrieval ($\mathcal{IR}$). As in $\mathcal{IR}$ the aim is to find the documents, in a document collection, better matching the semantics of a query, similarly in $\mathcal{TM}$ the aim is to find the terms, in a term collection (target thesaurus), better matching the semantics of a term in a source thesaurus.

The $\mathcal{IR}$ problem is usually formalized in literature as a 4-upla $\mathcal{IR} = [D, Q, F, R(q_i, d_j)]$ [5], where:

1. D is a set of the possible representations (*logical views*) of a document in a collection;
2. Q is a set of the possible representations (*logical views*) of a document with respect to the user information needs. Such representations are called *queries*;
3. F is a framework for modeling document representations, queries, and their relationships;
4. $R(q_i, d_j)$ is a ranking function, which associates a real number with (q_i, d_j) where $q_i \in Q$, $d_j \in D$. Such ranking defines an ordering among documents with respect to the query q_i.

Similarly the $\mathcal{TM}$ problem can be viewed and formalized as a problem of $\mathcal{IR}$, therefore $\mathcal{TM} = [D, Q, F, R(q_i, d_j)]$ where:

1. D is a the set of the possible representations (*logical views*) of a term in a target thesaurus (in $\mathcal{IR}$ it corresponds to the representation of documents to be retrieved in a collection);
2. Q is the set of the possible representations (*logical views*) of a term in a source thesaurus (in $\mathcal{IR}$ it corresponds to the representation of queries to be matched with documents of the collections);
3. F is the framework of term representations in source and target thesauri;
4. $R(q_i, d_j)$ is a ranking function, which associates a real number with (q_i, d_j) where $q_i \in Q$, $d_j \in D$, giving an order of relevance to the terms in a target thesaurus with respect to a term of the source thesaurus.

[1] hereinafter when the expression "term" is used for addressing the elements of a thesaurus, we refer to a "simple or complex term": for example *Parliament* is a simple term, *President of the Republic* is a complex term.

Having identified an isomorphism between $\mathcal{IR}$ and $\mathcal{TM}$, the implementation of a specific methodology for $\mathcal{TM}$ is connected with the instantiation of $\mathcal{TM}$ using different approaches for:

- the identification of suitable frameworks F for $\mathcal{TM}$, namely the representations of terms in source and target thesauri (Q and D respectively)
  - for better representing the semantics of thesaurus terms;
  - in a way amenable for computation;
- the identification of the ranking function $R(q_i, d_j)$ between source and target terms able to provide a similarity measure between the "semantics" of such terms.

4 Standards for implementing mapping methodologies

The framework discussed in Section 3 can be implemented following recent directives for using open-standards coming from the W3C community and for the need to propose semantic web oriented solutions.

In particular RDF/OWL standards are available to represent ontology concepts and relationships. For representing mapping relationships among thesauri the standards SKOS (Simple Knowledge Organisation System) can be used [6]. It provides a standard way to represent knowledge organisation systems using RDF to describe concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies and other types of controlled vocabulary, thus guaranteeing interoperability among applications. SKOS is a modular and extensible family of languages organized in three components: SKOS Core, SKOS Mapping and SKOS Extensions.

5 A Thesaurus Mapping Case Study

The methodology framework previously described has been implemented in a case study of interest for OPOCE[2]; such case study is aimed at testing the interoperability among five thesauri: EUROVOC, ECLAS, GEMET, UNESCO Thesaurus, ETT.

EUROVOC is the main EU thesaurus containing a hierarchical structure with inter-lingual relations. It helps a coherent and effective managing, indexing and searching information of EU documentary collections, covering 21 fields. ECLAS is the European Commission Central Libraries thesaurus, covering 19 domains. GEMET, the GEneral Multilingual Environmental Thesaurus is utilised by the European Environment Agency (EEA). UNESCO Thesaurus is a controlled vocabulary developed by the United Nations Educational, Scientific and Cultural Organisation which includes subject terms for several areas of knowledge. ETT is the European Training Thesaurus providing support to indexing and retrieval vocational education and training documentation in the European Union.

[2] Office for Official Publications of the European Communities.

From the point of view of their structure, all these thesauri are based on the same standards ISO-5964 and ISO-2788. From the point of view of the domain, EUROVOC, ECLAS and UNESCO Thesaurus can be identified as multidisciplinary thesauri, while GEMET and ETT can be considered as specialized or domain thesauri.

According to the project specifications, a mapping between EUROVOC and the other thesauri of interest are expected. Other different mappings in fact might not be meaningful since some of them pertain to different domains. Therefore in the proposed mapping strategy, EUROVOC acts as a reference, and the mapping strategies are tested to and from EUROVOC terms. This technique reduces the computational complexity of the problem of multi-thesaurus interoperability (N-to-N mapping) from a factor N^2 to a factor $2N$.

The basic mapping methodologies are applied to *descriptors* within corresponding microthesauri in their *English version* as a pivot language.

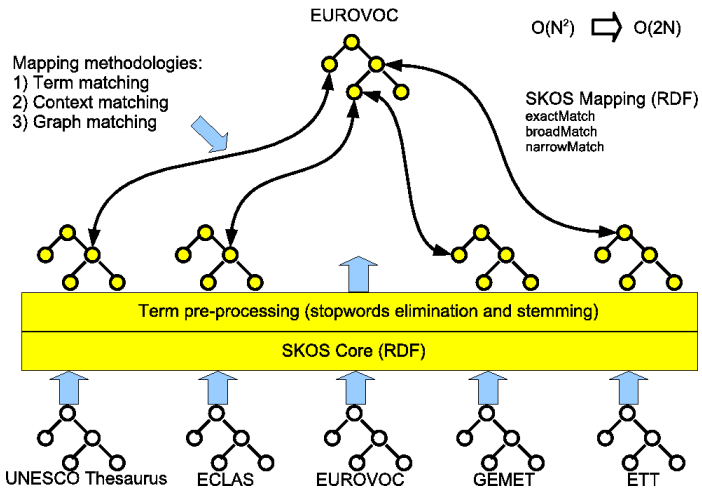


Figure 1. Thesaurus mapping workflow

The steps of the system workflow is here below described (Fig. 1 can be considered as reference).

5.1 SKOS Core transformation and terms pre-processing

The first step consists in transforming the thesauri of interest from their XML proprietary format into an RDF SKOS Core representation using XSLT techniques. Moreover, to reduce the computational complexity of the problem, thesaurus terms are normalized so that digit characters and non-alphabetic characters (if any) are represented by a special character; then other operations as *stemming* or the use of word stoplists (*stopwords elimination*) are performed.

Once basic terms have been identified, a vocabulary of terms for each thesaurus is created, containing all the terms which occur at least once in the set.

The following steps consist in the instantiation of the components of our $\mathcal{TM}$ formal characterization.

5.2 Logical views of terms in source (Q) and target (D) thesauri

Term mapping between thesauri, rather than a process which finds formal (lexical) equivalences between terms, is mainly a process which aims at matching their meanings, namely the *semantics* of the terms from source to target thesauri. In traditional thesauri, in fact, *descriptors* and related *non-descriptors* are terms expressing the same meaning (entry). More precisely, each meaning is expressed by one or more terms (linguistic expressions by single or multi words) in the same language (for instance 'pollution', 'contamination', 'discharge of pollutants'), as well as in different languages (for instance, the EN term 'water' and the IT term 'acqua', etc.). Moreover each word can have more than one sense, i.e. it can express more than one concept. In this view, in order to effectively solve, automatically or semi-automatically, the $\mathcal{TM}$ problem, term (simple or complex) semantics has to be captured.

According to the isomorphism introduced in Section 3 such a problem can be approached in a similar fashion as in the $\mathcal{IR}$ problem.

In $\mathcal{IR}$ a mapping between queries and documents is expected. The more a query is semantically characterized, the more the system will be able to match query semantics to document semantics. A query is usually constructed as a context (set of keywords) able to provide a more precise semantics to the query terms, or using metadata, if any. Similarly, for the $\mathcal{TM}$ problem the aim is to map a term of a source thesaurus (our "query") to a term in the target thesaurus (our "document"). The more terms in source and target thesauri are semantically characterized, the more the system will be able to match them according to their meanings.

The semantics of a term is conveyed not only by its morphological characteristics, but also by the context in which the term is used as well as by the relations with other terms. In the $\mathcal{TM}$ problem F is exactly aimed at identifying the framework for term representations able to better capture the semantics of terms in source and target thesauri.

We propose to represent the semantics of a term in a thesaurus, according to an ascending degree of expressiveness, by: its *Lexical Manifestation*, its *Lexical Context*, its *Lexical Network*.

Lexical Manifestation of a thesaurus term is its expression as a string of characters, normalized according to the pre-processing steps discussed in Section 5.1.

Lexical Context of a thesaurus term is represented by a vector d of term binary entries (statistics on terms to obtain weighted entries are not possible since document collections are not available) composed by the term itself, relevant terms in its definition and linked terms. Firstly a vocabulary of normalized

terms is constructed from a target thesaurus. The dimension T of such vocabulary is the dimension of the vector representing a term. For each term in fact a T-dimensional vector will be constructed, whose entries represent information on the corresponding vocabulary terms characterizing the current term lexical context (such vector can be viewed as a document of the $\mathcal{IR}$ problem). Similarly a term in a source thesaurus is represented by a T-dimensional vector whose entries represent information on the corresponding vocabulary terms characterizing the current term lexical context (such vector can be viewed as a query of the $\mathcal{IR}$ problem).

Lexical Network of a thesaurus term is a *direct graph* where nodes are terms along with related ones, and the labeled edges are semantically characterized relations between terms.

For *Lexical Contexts* and *Lexical Networks* the terms connection degree is based on a strict adjacency relations with each descriptor used for mapping implementation.

5.3 The proposed Framework (F)

Having identified thesaurus terms logical views, the frameworks in which the $\mathcal{TM}$ problem can be modeled are also identified.

For term *Lexical Manifestations*, the framework is composed of strings representing terms and the standard operations on strings. For term *Lexical Contexts*, the framework is composed of T-dimensional vectorial space and linear algebra operations on vectors. For term *Lexical Networks*, the framework is composed by graphs (described by nodes, edges and related labels) and the algebra operations on graphs.

The frameworks identified can also provide the intuition for constructing a ranking function R, which will be linked to the chosen representation of the space elements (terms).

5.4 The proposed Ranking Functions (R)

The ranking function R will be able to provide a similarity measure between a term in a source thesaurus and an associated one in a target thesaurus; when extended to a set of target terms such a function may provide a matching order among such terms. With respect to the three logical views on terms identified in Section 5.2, here below possible ranking functions to measure the degree of mapping between thesaurus terms are proposed.

Ranking function for Lexical Manifestations. In this case *String-based techniques* are used, in particular the *Edit distance/similarity* (or Levenshtein distance/similarity) applied on pre-processed strings through *language-based techniques* (as Tokenization, Lemmatization (*Stemming*) and Elimination (*Stopword elimination*)) normalized with respect to the longest string (therefore this measure varies in the interval [0,1]).

Ranking function for Lexical Contexts. Having represented the semantics of a thesaurus term as a *Lexical Context*, namely a vector representing

the term to be mapped, its "definition", if any, and related terms, a similarity between contexts can be measured as the correlation between such vectors, quantified, for instance, as the cosine of the angle between these two vectors

$$sim_v = sim(\boldsymbol{d_j}, q) = \frac{\boldsymbol{d_j} \times \boldsymbol{q}}{|\boldsymbol{d_j}| \cdot |\boldsymbol{q}|}$$

where $|\boldsymbol{d_j}|$ and $|\boldsymbol{q}|$ are the norms of the vectors representing terms in target and source thesauri, respectively.

Ranking function for Lexical Networks. Having represented the semantics of a thesaurus term as a *Lexical Network*, basically a direct graph, a vast literature exists in graph theory [7] and methods to measure the distance between two graphs [8] [9]. In literature the more frequently addressed graph similarity measure is the *Graph Edit Distance*, namely the minimum number of nodes and edges deletions/insertions/substitutions to transform a graph g_1 into a graph g_2 [10]. Because of computational complexity we have considered three variants of the Graph Edit Distance: *Conceptual similarity*, *Relational similarity* and *Graph similarity*.

The *Conceptual similarity* sim_c expresses how many concepts two graphs g_1 and g_2 have in common by an expression analogous to the Dice coefficient [11]:

$$sim_c = \frac{2n(g_c)}{n(g_1) + n(g_2)}$$

where g_c is the *maximum common subgraph* of g_1 and g_2 (it denes the parts of both graphs which are identical to one another, and, intuitively, it describes the the intersection between the two graphs) and $n(g)$ is the number of concept nodes of a graph g. This expression varies from 0 when the two graphs have no concepts in common to 1 when the two graphs consist of the same set of concepts.

The *Relational similarity* sim_r indicates how similar the relations between the same concepts in both graphs are, that is, how similar the information communicated in both texts about these concepts is. In a way, it shows how similar the contexts of the common concepts in both graphs are. The Relational similarity sim_r is aiming to measure the proportion between the degree of connection of the concept nodes in g_c, on the one hand, and the degree of connection of the same concept nodes in the original graphs g_1 and g_2, on the other hand. With this idea, a relation between two concept nodes conveys less information about the context of these concepts if they are highly connected in the original graphs, and conveys more information when they are weakly connected in the original graphs. Using a modified formula for the Dice coefficient, sim_r can be obtained as:

$$sim_r = \frac{2m(g_c)}{m_{g_c}(g_1) + m_{g_c}(g_2)}$$

where $m(g_c)$ is the number of the arcs (the relation nodes in the case of conceptual graphs) in the graph g_c, and $m_{g_c}(g_i)$ is the number of the arcs in

the immediate neighborhood of the graph g_c in the graph g_i. The immediate neighborhood of $g_c \subseteq g_i$ in g_i consists of the arcs of g_i with at least one end belonging to g_c.

Considering a graph g to be matched with a graph g_T as reference, a possible similarity measure able to sum-up the previous two is the *Graph similarity* [12]:

$$sim_g = \frac{N_c(g, g_T) + E_c(g, g_T)}{N(g_T) + E(g_T)}$$

where $N_c(g, g_T)$ is the number of nodes shared by graph g and g_T; $E_c(g, g_T)$ is the number of edges common to g and g_T; $N(g_T)$ is the number of nodes in graph g_T; $E(g_T)$ the number is of edges in g_T.

5.5 Ranking among candidate terms and mapping implementation

Terms of the target thesaurus, represented according to one of the discussed models, are matched with the chosen term in a source thesaurus, represented with the same model, using a proper similarity measure. Candidate terms of the target thesaurus are ranked according to the similarity measure values $sim \in [0, 1]$ and a semantics to the mapping relation is assigned using proper heuristic threshold values (T_1, $T_2 \in [0, 1]$)

if $sim < T_1$ $\Rightarrow$ exactMatch
if $T_1 < sim < T_2$ $\Rightarrow$ partial match (broadMatch or narrowMatch)
if $T_2 < sim$ $\Rightarrow$ No Match

Then the representation of the established relations between thesaurus terms is expressed using RDF SKOS Mapping standard.

6 Interoperability assessment through a "gold standard"

Interoperability between thesauri is specifically assessed on a data sample selected from the thesauri of interest. In order to evaluate the performance of automatic mapping algorithms an intellectual activity is needed to create a "gold standard", namely a groundtruth file of thesauri term mapping examples (one for each couple of thesauri having EUROVOC as pivot) which represents the ideal set of expected correct mappings. It is aimed at 1) tuning system heuristics (similarity measure threshold values are tuned to obtain the best results with respect to the gold standard (performance convergence)), 2) evaluating the performances of automatic mapping algorithms, comparing the ideal set of mapping examples with system predictions.

For the purpose of comparison between the "gold standard" and the algorithms prediction, mapping relations of the "gold standard" is described using SKOS Mapping, limited to the *exactMatch*, *broadMatch* and *narrowMatch* relations. In particular when a concept in EUROVOC corresponds exactly to one or more concepts in a target thesaurus according to the expert judgment, the relation is an exact match. A broad equivalence is one where the source term is more specific in some sense than the target term. Similarly a narrow equivalence

is one where the source term is more general in some sense than the target term or expression [13]. Following [1], when it is possible, mappings that are at least complete, and ideally optimal, have to be established.

A *complete mapping* [13] is one where a source term, having no exact equivalence in the target, is matched to *at least* one target term that is semantically broader and *at least* one target term that is semantically narrower.

An *optimal mapping* [13] is one where the aforementioned broader target term is the nearest broader term to the source term, and where the aforementioned narrower target term is the nearest narrower term to the source term.

In our case study the "gold standard" construction intellectual activity has been carried out by experts of the specific chosen domains, using tools to make their work easier. Some solutions have been evaluated to implement the "gold standard", in particular Protégé PROMPT merging tool[3], Vocabulary Integration Environment (VINE)[4] and AIDA[5]. All these software have been tested. For different reasons they have been considered too problematic and not enough user-friendly. Therefore we have developed a specific application (THesauri ALigning

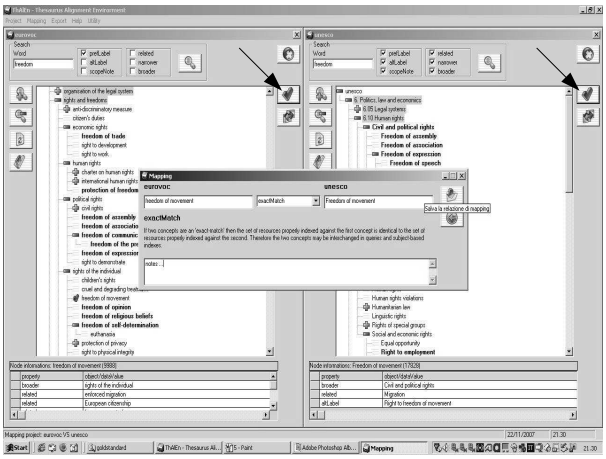


Figure 2. A screenshot of THALEN showing a parallel view of two thesauri and a form where to establish mapping relations.

ENvironment (THALEN)) able to provide a user-friendly access to thesauri and simple functionalities to establish term relations for mapping.

THALEN has been developed on an MS-Access relational database and provides simple functionalities of user authentication, thesaurus loading, parallel view of two thesauri, search modalities (term browsing and searching), manual mapping implementation, summary of the established term mapping as well as

[3] http://protege.stanford.edu/plugins/prompt/prompt.html
[4] http://marinemetadata.org/examples/mmihostedwork/ontologieswork/vine/index.html
[5] http://www.vl-e.nl

exportation of mapping relations in RDF SKOS. Moreover this application can be used for human validation of the automatic mapping. In Fig. 2 a screenshot of THALEN as developed for the project is shown.

7 Preliminary test of the proposed $\mathcal{TM}$ approaches

The "gold standard" produced by experts has been used to assess the proposed methodologies for automatic thesaurus mapping: it includes 624 relations, of which 346 are exactMatch relations.

The system mapping performances are assessed with respect to the "gold standard" for each single mapping relation type, using the typical *Precision* and *Recall* measures. In particular for the project case study the system *Recall* has been assessed since the automatic mapping is addressed to identify matching concepts within the system predictions, to be validated by humans. In our case study an important measure is represented by *Recall* with respect to a less specified relation (untypedMatch) able to express a generic association between concepts.

Preliminary experiments have shown satisfactory performances as regards the identification of untypedMatch relations between terms, and, as a consequence, the identification of noMatch relations; on the contrary, good performances have been obtained as regards the selection of term exactMatch relations, while the distinction between narrowMatch and broadMatch relations revealed a high degree of uncertainty. Global performances on specific types of relations are highly affected by this uncertainty, therefore meaningful results can be given with respect to untypedMatch relations as well as the system ability to identify exactMatch relations with respect to the "gold standards".

The proposed logical views for thesaurus terms and the related ranking functions outperformed a simple string matching among thesaurus terms. In particular for the following couples of thesauri (EUROVOC vs. {ETT, ECLAS, GEMET}) the Lexical Manifestation logical view and the Levenshtein Similarity ranking function has given the best results (untypedMatch Recall = 66.2%, exactMatch Recall = 82.3%), while for the couple EUROVOC vs. UNESCO Thesaurus the best results have been obtained using the Lexical Network logical view and the Conceptual Similarity ranking function (untypedMatch Recall = 73.7%, exactMatch Recall = 80.8%).

8 Conclusions

This paper presents a methodological framework and a specific implementation of schema-based thesaurus mapping. The approach has been assessed on a case study focused on five thesauri of interest for the EU institutions.

Different thesaurus terms logical views and related ranking functions to establish terms conceptual similarity have been proposed and tested. The preliminary results revealed that a simple Lexical Manifestation logical view and

Levenshtein Similarity ranking function produced the best results on most of the matches between the thesauri couples of interest.

More complex descriptions of thesaurus concepts (Lexical Contexts, Lexical Networks) suffer from problems of computational tractability (in particular as regards Lexical Networks), moreover the use of higher number of features for concepts description provides a higher variability of the similarity measures, affecting the algorithms performances.

To improve the performances of the system using Lexical Contexts and Lexical Networks, which usually provide a higher degree of expressiveness in concept descriptions, different criteria of features selection can be tested, aiming at reducing the computational complexity, as well as the variability of the similarity measures.

References

1. M. Doerr, "Semantic problems of thesaurus mapping," *Journal of Digital Information*, vol. 1, no. 8, 2001.
2. A. Miles and B. Matthews, "Deliverable 8.4: Inter-thesaurus mapping," 2004. http://www.w3c.rl.ac.uk/SWAD/deliverables/8.4.html.
3. E. Rahm and P. Bernstein, "A survey of approaches to automatic schema matching," *The International Journal on Very Large Data Bases*, vol. 10, no. 4, pp. 334–350, 2001.
4. J. Euzenat and P. Shvaiko, *Ontology Matching*. Springer, 2007.
5. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.
6. A. Miles and D. Brickley, "Skos simple knowledge organization system reference," 2008. http://www.w3.org/TR/skos-reference.
7. J. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, 1984.
8. J. Zhong, H. Zhu, J. Li, and Y. Yu, "Conceptual graph matching for semantic search," in *Proceedings of* 10^{th} *International Conference on Conceptual Structures (ICCS)*, Lecture Notes in Artificial Intelligence 2393, Springer-Verlag, 2002.
9. M. Montes-y-Gómez, A. Gelhukh, A. Lopez-Lopez, and R. Baeza-Yates, *Flexible Comparison of Conceptual Graphs*. Lecture Notes in Computer Science 2113, Springer-Verlag, 2001.
10. C.-A. M. Irniger, *Graph Matching – Filtering Databases of Graphs Using Machine Learning Techniques*. PhD thesis, Institut für Informatik und angewandte Mathematik, Universität Bern, 2005.
11. E. Rasmussen, "Clustering algorithms," in *Information Retrieval: Data Structures & Algorithms* (W. B. Frakes and R. Baeza-Yates, eds.), 1992.
12. W.-T. Cai, S.-R. Wang, and Q.-S. Jiang, "Address extraction: a graph matching and ontology-based approach to conceptual information retrieval," in *Proceedings of the Third Intemational Conference on Machine Learning and Cybernetics*, pp. 1571–1576, 2004.
13. A. C. Liang and M. Sini, "Mapping agrovoc and the chinese agricultural thesaurus: Definitions, tools, procedures," *New Review of Hypermedia and Multimedia*, vol. 12, no. 1, pp. 51–62, 2006.

Semantic interoperability issues from a case study in archaeology

D. Tudhope, C. Binding, University of Glamorgan [1]

K. May, English Heritage

Abstract. This paper addresses issues arising from the first steps in mapping different (a) datasets and (b) vocabularies to the CIDOC CRM, within an RDF implementation. We first discuss practical implementation issues for mapping datasets to the CRM-EH and then discuss practical issues converting domain thesauri to the SKOS Core standard representation. We finally discuss, at a more theoretical level, issues concerning the mapping of domain thesauri to upper (core) ontologies.

1 Introduction

The general aim of our research is to investigate the potential of semantic terminology tools for improving access to digital archaeology resources, including disparate data sets and associated grey literature. The immediate goal discussed here concerns describing and accessing cultural objects using the CIDOC CRM core ontology [3, 7], as an overarching common schema. Different datasets must be mapped to the CIDOC CRM, where the datasets are indexed by domain thesauri and other vocabularies. Thus semantic interoperability is central.

This paper addresses issues arising from the first steps in mapping different (a) datasets and (b) vocabularies to the CIDOC CRM, within an RDF implementation. The work, in collaboration with English Heritage (EH)[7], formed part of the JPA activities of the DELOS FP6 Network of Excellence, Cluster on Knowledge Extraction and Semantic Interoperability [6] and the AHRC funded project on Semantic Technologies for Archaeological Resources (STAR) [17].

Some previous work in the European DL context (BRICKS) has reported difficulties when mapping different cultural heritage datasets to the CIDOC CRM due to the abstractness of the concepts resulting in consistency problems for the mapping work and also a need for additional technical specifications for CRM implementations [11, 12]. The CRM is a high level conceptual framework, which is intended to be specialised when warranted for particular purposes. We also found a need to provide additional implementation constructs and these are outlined below. For mapping to datasets at a detailed level, we worked with an extension of the CIDOC CRM (the CRM-EH) developed by our collaborators (May) in English Heritage [5, 10]. The CRM-EH models the archaeological excavation and analysis workflow. Working with

[1] contact address dstudhope@glam.ac.uk

May, an implementation of the CRM-EH has been produced as a modular RDF extension referencing the published (v4.2) RDFS implementation of the CRM. Additional extensions to the CIDOC CRM, necessary to our implementation, are also available as separate RDF files.

We go on to first discuss practical implementation issues for mapping datasets to the CRM-EH and then discuss practical issues converting domain thesauri to the SKOS Core standard representation. We finally discuss, at a more theoretical level, issues concerning the mapping of domain thesauri to upper (core) ontologies.

2 Data extraction and mapping process and conversion to RDF

Initial mappings were made from the CRM-EH to three different database formats, where the data has been extracted to RDF and the mapping expressed as an RDF relationship. The data extraction process involved selected data from the following archaeological datasets:

- Raunds Roman Analytical Database (RRAD)
- Raunds Prehistoric Database (RPRE)
- York Archaeological Trust (YAT) Integrated Archaeological Database (IADB)

The approach taken for the exercise was to extract modular parts of the larger data model from the RRAD, RPRE and IADB databases via SQL queries, and store the data retrieved in a series of RDF files. This allowed data instances to be later selectively combined as required, and avoided the data extraction process from becoming unnecessarily complex and unwieldy.

The intellectual mapping requires some expert knowledge of the data and the CRM-EH. Initial mappings were performed by May and communicated via spreadsheets. Some subsequent mappings were performed by the project team using the initial mappings as a guide, with validation by May. This process is time consuming and a data mapping and extraction utility was developed to assist the process. The utility consists of a form allowing the user to build up a SQL query incorporating selectable consistent URIs representing specific RDF entity and property types (including CRM, CRM-EH, SKOS, Dublin Core and others). The output is an RDF format file, with query parameters saved in XML format for subsequent reuse. Details will be available shortly on the STAR project website.

2.1 ID format adopted

RDF entities require unique identifiers. Some of the data being extracted was an amalgamation of records from separate tables – e.g. *EHE0009.ContextFind* actually contained records from RRAD.Object & RRAD.Ceramics tables. It was therefore necessary to devise a unique ID for all RDF entities beyond just using the record ID

from an individual table.. The format adopted to deal with all these issues was a simple dot delimited notation as follows:

[URI prefix]entity.database.table.column.ID
e.g. "EHE0008.rrad.context.contextno.100999"

This format (although verbose) allowed the use of existing DB record ID values without introducing ambiguities. In RRAD database, Ceramics and Objects were both instances of *EHE0009.ContextFind*. This therefore involved the combination of data from two tables:

• *EHE0009.rrad.object.objectno.105432 [an EHE0009.ContextFind record from the RRAD object table]*
• *EHE0009.rrad.ceramics.ceramicsno.105432 [an EHE0009.ContextFind record from the RRAD Ceramics table, with a coincidental ID value]*

The format also allowed the same base record ID to be used for both *EHE0009.ContextFind* and *EHE1004.ContextFindDepositionEvent* (these records actually originated from the same table and had a 1:1 relationship), using a different entity prefix to disambiguate the records:

• *EHE0009.rrad.object.objectno.105432 [The ContextFind record ID]*
• *EHE1004.rrad.object.objectno.105432 [The ContextFindDepositionEvent record ID]*

Finally an arbitrary URI prefix (http://tempuri/) was added to all ID values. According to need, this can be replaced with a more persistent prefix.

2.2 Date/Time format adopted

There is nothing dictated in CRM or CRM-EH about date/time representation formats, however we clearly needed to maintain a consistent format throughout the data. For the purposes of the data extraction to keep all data consistent we used a "big endian" (i.e. from most to least significant) format compatible with both W3C standards and ISO8601 ("Data elements and interchange formats – Information interchange – Representation of dates and times"). The format is as follows:

CCYY-MM-DDThh:mm:ss e.g. "2007-05-03T16:19:23"

This format does not introduce any restrictions on how dates & times are eventually displayed or used within applications; it merely provides a common string representation mechanism for interoperability of data.

2.3 Co-ordinate format adopted

Spatial co-ordinates appeared in various formats within the datasets. RRAD co-ordinates were 6 digit numeric values in separate "Easting" and "Northing" columns. RPRE coordinates were slash separated string values, sometimes with an extra 4 digit value appended (i.e. either *nnnnnn/nnnnnn/nnnn* or *nnnnnn/nnnnnn*). IADB co-ordinates were numeric values in separate "Easting" and "Northing" columns (and appeared to be relative to a site local reference datum). CRM/CRM-EH requires a single string to represent a spatial co-ordinate value. The consistent format chosen for output was 6 digit space delimited Easting and Northing values, with an optional Height value (Above Ordnance Datum). These values were all assumed to be in metres:

nnnnnnE nnnnnnN [nn.nnnAOD] e.g. "105858E 237435N 125.282AOD"

2.4 Modelling notes/annotations

The CRM has a modelling construct in the form of "properties of properties. For example, property *P3.has_note* has a further property *P3.1.has_type* – intended to model the distinction between different types of note. However, this construct does not translate well to RDF. As evidence of this, property *P3.1.has_type* is not actually part of the current RDFS encoding of CRM on the CIDOC website (in the comment header there is a suggestion to create specific sub properties of *P3.has_note* instead). The more recent OWL encoding of CRM also avoids including the construct.
The EH recording manuals and the current datasets contain several kinds of note fields. Through discussion with EH it is possible to distil these down to a common core set of note types, such as
• Comments
• Method of excavation
• Interpretation
• Siting description
• Site treatment
While it might potentially be restrictive to model notes as strings (notes have other implicit attributes such as language, author/source etc.), this is the current position within the CRM (*E1.CRM Entity _ P3.has_note _ E62.String)*. However, taking the RDFS encoding of CIDOC CRM recommendation, we intend to create sub properties of *P3.has_note* e.g. *EHPxx1.has_interpretation*, as part of future work.

2.5 Modelling of Events

The CRM-EH and CRM are event based models. Events defined in the models and used to interconnect objects and places etc. were often only implicit within the original relational database structures and in the mappings created. In the translation from relational data structures to an RDF graph structure it was necessary to create this event information by the formation of intermediate 'virtual' entities.

2.6 Modelling of Data Instance Values

Being a higher level conceptual model the CRM has little intrinsic provision for the representation of actual data instance values. The approach adopted for the STAR data extraction process was to create rdf:value relationships as an additional property to model instance data for entities wherever appropriate.

2.7 Initial mapping of data fields to extended CRM

The extracted data represented a subset of the full English Heritage extended CRM (CRM-EH) model. For the initial phase we limited the scope of the data extraction work to data concerning contexts and their associated finds. The relationships between entities extracted and modelled in RDF are shown in Figure 1.

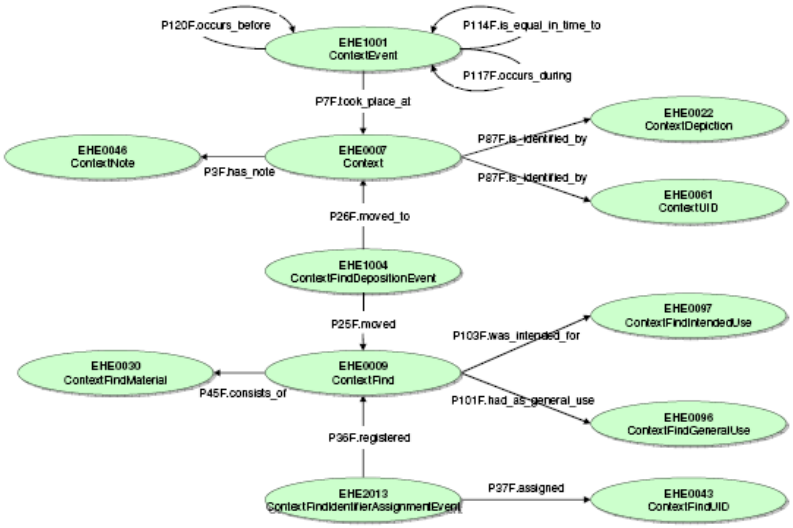


Figure 1: CRM-EH entities initially modelled

The number of statements (triples) contained in the resultant RDF files is **1,080,913.** Some triples (e.g. *rdf:type* statements) were duplicated due to entities occurring within multiple files, but any duplication was removed during the aggregation process.

A number of separate RDF files were combined in the aggregation process including the CRM itself, the CRM-EH extension, alternative language labels for the CRM, and various EH domain thesauri.

2.8 Validation of extracted data

The data files produced were each validated against the W3C RDF validation service. Whilst this did not prove the validity of the data relationships or even conformance to CRM-EH, it did at least give confidence in the validity of the basic RDF syntax.

2.9 Aggregation of extracted data

The SemWeb library [14] was employed to aggregate the extracted data files into a single SQLITE database. The extended CRM ontology plus the English Heritage SKOS thesauri were also imported. The resultant database of aggregated data was 193MB overall and consisted of 268,947 RDF entities, 168,886 RDF literals and 796,227 RDF statements (triples). The SemWeb library supports SPARQL querying against the database, but the SQLITE database itself also supports direct SQL queries.

2.10 Use of aggregated data

This simple, initial example illustrates a SPARQL search via the CRM model relationships for a *Dish* made of *Pewter*. The search is case sensitive and returned 5 records within 1 second. It is possible to deduce the origin of the result records due to the ID convention adopted for the data export process. All are *EHE0009.ContextFind* objects - 3 originated from Raunds Roman (RRAD) *object* table, 1 from the Raunds Prehistoric (RPRE) *flint* table and 1 from the RPRE *objects* table. Merging the exported datasets into the RDF data store facilitates cross searching and location of records from multiple databases.

```
SELECT * WHERE
{
     ?x crm:P103F.was_intended_for "Dish".
     ?x crm:P45F.consists_of "Pewter" .
}
<result>
     <binding name="x">
     <uri>http://tempuri/EHE0009.rrad.object.objectno.12687</uri>
     </binding>
</result>
<result>
     <binding name="x">
     <uri>http://tempuri/EHE0009.rrad.object.objectno.12969</uri>
     </binding>
</result>
<result>
     <binding name="x">
     <uri>http://tempuri/EHE0009.rrad.object.objectno.55006</uri>
     </binding>
</result>
<result>
     <binding name="x">
     <uri>http://tempuri/EHE0009.rpre.flint.recordnumber.55006</uri>
     </binding>
</result>
```

```
<result>
     <binding name="x">
     <uri>http://tempuri/EHE0009.rpre.objects.recordnumber.55006</uri>
     </binding>
</result>
```

3 Conversion of KOS to SKOS/RDF representations

The project has adopted SKOS Core [15] as the representation format for domain thesauri and related Knowledge Organization Systems (KOS). In general, thesauri conforming to the BSI/NISO/ISO standards should map in a fairly straight forward manner to SKOS. However, there may need to be judgments on how to deal with non-standard features. Additionally, the case study illustrates potential problems associated with the use of Guide Terms or facet indicators in some thesauri. Other issues surfaced by the exercise concern the need to create URIs for concept identifiers as part of the conversion and the potential for validation.

3.1 Conversion process

Thesaurus data was received from English Heritage National Monuments Record Centre, in CSV format files [9]. The approach initially adopted was to convert the received files to XML, and an XSL transformation was written to export the data to SKOS RDF format. Although this strategy was successful for the smaller thesauri, XSL transformation of the raw data files proved to be a lengthy and resource intensive operation for the larger thesauri, resulting in the PC running out of memory on some occasions. Therefore the CSV files were subsequently imported into a Microsoft Access database and a small custom C# application was written to export the data from this database into SKOS RDF format.

The major difficulty with the resultant SKOS representations is that we did not model "non-indexing" concepts (guide terms or facet indicators) as *Collections*, the intended equivalent in the SKOS model. Guide terms in SKOS do not form part of the main hierarchical structure, but are groupings of sibling concepts for purposes of clarity in display. It would have entailed changing the existing hierarchical structure of the English Heritage thesauri, in order to utilise the SKOS 'Collections' element. This was not an appropriate decision for the STAR project to take (relevant EH contacts have been informed) and was not a critical issue for the project's research aims. Thus for STAR purposes the distinction between indexing concepts and guide terms is not made, and the (poly) hierarchical relationships in the SKOS files represent those present in the source data.

3.2 Validation process

As a result of running the conversion application, separate RDF files were produced for each thesaurus. The newly created files were first validated using W3C RDF validation service. This is a basic RDF syntax validation test, and all files passed this

initial run with no errors or warnings. The files were then checked using the W3C SKOS validation service [15]. This consists of a series of SKOS compatibility and thesaurus integrity tests, and the output was a set of validation reports. A few minor anomalies arose from these tests, including legacy features such as orphan concepts.

The conversion is efficient and reliable so any updates to thesaurus data at source can be quickly reprocessed. The resultant SKOS files are intended as data inputs to the STAR project and will be used for query expansion and domain navigation tools. It is notable that the validation made possible by the SKOS conversion proved useful to the thesaurus developer for maintenance purposes.

3.3 SKOS based Terminology Services

An initial set of semantic web services have been developed, based upon the SKOS thesaurus representations. These were integrated with the DelosDLMS prototype next-generation Digital Library management system [1]. The services provide term look up, browsing and semantic concept expansion [2]. A pilot SKOS service should shortly be available on a restricted basis from the Glamorgan website. Details of the API and a pilot demonstrator can be found off the STAR website under Semantic Terminology Services [17].

The service is written in C#, running on Microsoft .NET framework and is based on a subset of the SWAD Europe SKOS API, with extensions for concept expansion. The services currently provide term look up across the thesauri held in the system, along with browsing and semantic concept expansion within a chosen thesaurus. This allows search to be augmented by SKOS-based vocabulary and semantic resources (assuming the services are used in conjunction with a search system). Queries may be expanded by synonyms or by semantically related concepts. For example, a query is often expressed at a different level of generalisation from document content or metadata, or a query may employ semantically related concepts. Semantic expansion of concepts for purposes of query expansion yields a ranked list of semantically close concepts [19].

4 Mapping between SKOS and other representations

The next phase of the STAR project involves connecting the thesauri expressed in SKOS to documents or data base items and to an upper ontology, the CIDOC CRM. Figure 2 shows the current model for integrating the thesauri with the CRM. This illustrates two issues concerning the exploitation of SKOS RDF data: (a) the connection between a SKOS concept and the data item it represents and (b) the connection between the CRM and SKOS.

(a) Connecting SKOS concepts and data

The connection between a SKOS concept and an information item is here modeled by a project specific *is represented by* relationship (Figure 2). This is chosen as being the most flexible possibility, which can, if needed, be modified to take account of any standards developments in this area. Another possibility might be the standard *DC: Subject of* if that were appropriate. However, in STAR the application to data items is arguably not quite the same relationship. Another issue is whether, and to what extent, this *concept-referent* relationship should be modeled in SKOS, as opposed to some other indexing or vocabulary use standard. In addition to distinguishing between indexing and classification use cases, there are various other novel DL use cases where KOS are applied to non-traditional data sets for non-traditional purposes. It is important to note the difference between Library Science KOS (intended for information retrieval purposes) and many AI ontology applications, which aim to model a mini-world, where the connection is commonly taken to be a form of *Instance* relationship [18].

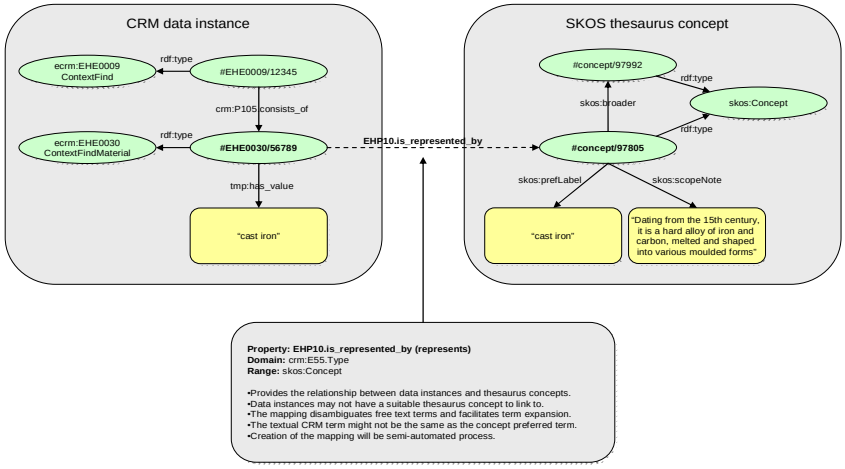


Figure 2: Model for combining SKOS and CIDOC CRM

(b) Connecting SKOS concepts and an upper ontology

The appropriate connection between an upper ontology and domain thesauri or other information retrieval KOS depends upon the intended purpose. It also depends on the alignment of the ontology and domain KOS, the number of different KOS intended to be modeled and the use cases to be supported. Cost benefit issues are highly relevant. This is similar to the considerations and likely success factors for mapping between thesauri or KOS generally (for more details, see the discussion in [13, Section 6.2.1].

In some situations, where the aim is to support automatic inferencing, it may be appropriate to formalize the domain KOS and completely integrate them into a formal ontology, expressing the KOS in OWL, for example. This would allow any benefits of inferencing to be applied to the more specific concepts of the domain KOS. This, however, is likely to be a resource intensive exercise. Since information retrieval KOS and AI ontologies tend to be designed for different purposes, this conversion may change the underlying structure and the rationale should be considered carefully. The conversion may involve facet analysis to distinguish orthogonal facets in the domain KOS, which should be separated to form distinct hierarchical facets. It may involve modeling to much more specific granularity of concepts if the upper ontology is intended to encompass many distinct domain KOS; for example, the need for disambiguation may well not be present in the KOS considered separately but is required when many are integrated together.

Such highly specific modeling should be considered in terms of costs and benefits. It is important to consider the use cases driving full formalisation, since information retrieval KOS, by design, tend to express a level of generality appropriate for search and indexing purposes and driving down to greater specificity may yield little cost benefit for retrieval or annotation use cases. It can be argued that SKOS representation offers a cost effective approach for many annotation, search and browsing oriented applications that don't require first order logic. The SWDWG is currently discussing the recommended best practice for combining SKOS and OWL, following the principle of allowing as many different application perspectives and use cases, as is consistent with the respective underlying principles.

A variant of the above approach, which allows the easier option of SKOS representation, is to consider the domain KOS as leaf nodes of an upper ontology, expressing this, with some form of *subclass* or *type* relationship, depending on the degree of confidence in the mapping. This corresponds to *Leaf Node Linking* in Zeng & Chan's review of mapping [20]. In the CIDOC CRM, for example, one recommended approach is to assert an Instance relationship between a Type property of a CRM class and the top of a thesaurus hierarchy (or the top concept of an entire KOS).

In some cases, including (initial analysis of) the EH case study described above, the domain thesauri may not fit neatly under the upper ontology, the thesauri being designed separately for different purposes. In the STAR project, from the initial discussions with EH collaborators with a subset of the thesauri, the appropriate connection may be a looser SKOS *mapping (broader)* relationship between groups of concepts rather than complete hierarchies. Yet another possibility can be found in Figure 2, which shows a data instance mapped to a CRM entity and where the data items are also indexed with thesaurus concepts. In this case, there is a mapping between data and the integrating upper ontology and another mapping between database fields and the domain thesaurus.

The appropriate mapping between domain thesaurus and the upper ontology ultimately rests upon the use cases to be supported by any explicit connection. In

general, these would tend to be use cases based upon either interactive browsing or automatic expansion (reasoning) of the unified concept space.

Conclusions

The modular approach (coupled with the uniform ID format used) facilitated extraction and storage of relational data into separate RDF files based on CRM-EH structure, and allowed the subsequent merging of selected parts of the data structure originating from multiple data sets. Further combining this data with the CRM-EH ontology (itself a modular unit extending the existing CIDOC CRM) opens up the possibility of automated traversal across known relationships. While more work needs to be done investigating scalability and performance issues, this illustrates potential as a foundation data structure for a rich application.

A CRM based web service has been implemented over the extracted data and model, which offers search capability with subsequent browsing over CRM-EH relationships.

Acknowledgements

The STAR project is funded by the UK Arts and Humanities Research Council (AHRC). Thanks are due to Andrew Houghton (OCLC Research) for helpful input to various parts of the report and to Phil Carlisle & Keith May (English Heritage).

References

1. Binding C., Brettlecker G., Catarci T., Christodoulakis S., Crecelius T., Gioldasis N., Jetter H-C., Kacimi M., Milano D., Ranaldi P., Reiterer H., Santucci G., Schek H-G., Schuldt H., Tudhope D., Weikum G.: DelosDLMS: Infrastructure and Services for Future Digital Library Systems, 2nd DELOS Conference, Pisa (2007) http://www.delos.info/index.php?option=com_content&task=view&id=602&Itemid=334
2. Binding C., Tudhope D.: KOS at your Service: Programmatic Access to Knowledge Organisation Systems. Journal of Digital Information, 4(4), (2004) http://journals.tdl.org/jodi/article/view/jodi-124/109
3. CIDOC Conceptual Reference Model (CRM), http://cidoc.ics.forth.gr
4. CRM-EH Extension to CRM http://hypermedia.research.glam.ac.uk/kos/CRM/
5. Cripps P., Greenhalgh A., Fellows D., May K., Robinson D.: Ontological Modelling of the work of the Centre for Archaeology, CIDOC CRM Technical Paper (2004) http://cidoc.ics.forth.gr/technical_papers.html
6. DELOS Cluster on Knowledge Extraction and Semantic Interoperability, http://delos-wp5.ukoln.ac.uk/
7. Doerr, M.: The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 2493, 75--92 (2003)
8. English Heritage http://www.english-heritage.org.uk/

9. English Heritage Thesauri http://thesaurus.english-heritage.org.uk/
10. May, K.: Integrating Cultural and Scientific Heritage: Archaeological Ontological Modelling for the Field and the Lab. CIDOC CRM Sig Workshop, Heraklion (2006) http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/may.pdf
11. Nußbaumer, P., Haslhofer, B.: CIDOC CRM in Action – Experiences and Challenges. Poster at 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL07), Budapest (2007) http://www.cs.univie.ac.at/upload//550/papers/cidoc_crm_poster_ecdl2007.pdf
12. Nußbaumer, P., Haslhofer, B.: Putting the CIDOC CRM into Practice – Experiences and Challenges. Technical Report, University of Vienna (2007) http://www.cs.univie.ac.at/publication.php?pid=2965
13. Patel M., Koch T., Doerr M., Tsinaraki C.: Report on Semantic Interoperability in Digital Library Systems. DELOS Network of Excellence, WP5 Deliverable D5.3.1. (2005)
14. SEMWEB RDF Library for .NET, http://razor.occams.info/code/semweb
15. SKOS: Simple Knowledge Organization Systems, http://www.w3.org/2004/02/skos
16. SKOS API. SWAD_EUROPE Thesaurus Project Output (2004) http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html
17. STAR Project: Semantic Technologies for Archaeological Resources, http://hypermedia.research.glam.ac.uk/kos/star
18. Tudhope D., Koch T., Heery R.: Terminology Services and Technology: JISC State of the art review (2006) http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf
19. Tudhope D., Binding C., Blocks D., Cunliffe D. Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), 509–533. Emerald (2006)
20. Zeng M, Chan L. Trends and issues in establishing interoperability among knowledge organization systems. Journal of American Society for Information Science and Technology, 55(5): 377 – 395. (2004)

The working group for the Italian Culture Portal: a potential Italian contribution to EDL

Irene Buonazia[1], Maria Emilia Masci[1], Davide Merlitti[1],
Karim Ben Hamida[2], Sara Di Giorgio[2]

[1] Scuola Normale Superiore di Pisa, Piazza dei Cavalieri 7, 56124 Pisa, Italy
{i.buonazia, e.masci, d.merlitti}@sns.it
[2] Ministero per i Beni e le Attività Culturali – Biblioteca di Storia Moderna e Contemporanea, Via Caetani 32, 00186, Roma, Italy
{karim.benhamida, sara.digiorgio}@beniculturali.it

Abstract. The Italian Culture Portal is a project promoted by Italian Ministry of Cultural Heritage and Activities, aiming at making available to the user resources - at item level – pertaining to Italian cultural domain. Metadata of resources coming from many different data providers are harvested through OAI-PMH, according to a DC application profile conceived for the project. The working group, composed by researchers of SNS and personnel of MiBAC, is deepening skills on faceted browsing, multilingual thesauri expressed in SKOS, and ontologies.

Keywords: DC Application Profile; OAI-PMH harvesting; Thesauri; SKOS; faceted browsing.

1 Introduction

The scientific and technical project for the Italian Culture Portal was promoted by the Italian Ministry of Cultural Heritage and Activities (MiBAC) and delivered by Scuola Normale Superiore di Pisa (SNS) during 2005. At the moment SNS works as consultant for MiBAC, flanking the company which is carrying out the Portal, which will be named "CulturaItalia".

1.1 MiBAC working group

The working group of Ministero per i Beni e le Attività Culturali is composed by:

Karim Ben Hamida, Mass Communication Sociologist, presently employed in the MiBAC with the role of executive project manager for the Italian Culture Portal. Focused on information architecture, thesauri and taxonomies, facet browsing

methodologies and semantic interoperabilty.

Sara Di Giorgio, Art Historian, presently employed in the MiBAC with the role of executive project manager for the Italian Culture Portal. Partecipated for the EU Project MINERVA (Ministerial Network for Valorising Activities in digitisation) and OTEBAC (Osservatorio tecnologico per i Beni e le Attività Culturali). Focused on ontology mapping, facet browsing methodology and semantic interoperability.

1.2 SNS working group

The working group of Scuola Normale Superiore di Pisa is composed by:

Maria Emilia Masci, Archaeologist – PHD degree, presently employed in the SNS with a research grant on metadata standards for Cultural Heritage.

Irene Buonazia, Art historian, presently employed in the SNS with a research grant on metadata standards for Cultural Heritage.

Both partecipated in EU projects (DELOS Network of Excellence on Digital Libraries; BRICKS - Building Resources for Integrated Cultural Knowledge Services; CALLAS - Conveying Affectiveness in Leading-Edge Living Adaptive Systems) and in several national research-projects about cultural heritage and archaeology, involving information and multimedia technologies. They partecipated in the technical-scientific Project for the Italian Portal of Culture, and cooperated with MiBAC for the EU Projects MINERVA (Ministerial Network for Valorising Activities in digitisation) and MICHAEL (Multilingual Inventory of Cultural Heritage in Europe).

Davide Merlitti, IT senior analyst. Partecipated in several national projects for digital libraries of archival, librarian, cultural heritage resources. Partecipated in the technical-scientific Project for the Italian Portal of Culture, and cooperated with MiBAC. Presently mainly focussing on ontologies, RDF and faceted browsing.

2 CulturaItalia project: work done

The main mission of the Italian Culture Portal is to communicate to different kinds of users the whole ensemble of Italian culture, as a media conceived for the diffusion of knowledge, promotion and enhancement of cultural heritage. The target of the Portal will be Italian and foreign users, such as tourists and people interested in, and passionate of, culture; business users (publishers, merchandising, etc.); young people, from primary to high school; culture professionals such as scholars, museums curators, researchers, etc. The Domain of “Italian Culture” is a wide concept,

conceived in different ways. MiBAC is responsible for preservation, management, research and exploitation of the Italian cultural patrimony, both tangible and un-tangible. Tangible heritage is composed by architectural and environmental objects; artworks and collections; manuscripts, edited books as well as the current literature; archaeological and demo-ethno-anthropological objects; contemporary art and architecture. Un-tangible heritage deals with events, music, dance, theatre, circuses and performances; cinema, humanities and scientific culture. Thus, CulturaItalia will offer access to the existing resources on cultural contents and will give more exposure to the vast amount of websites pertaining to museums, libraries, archives, universities and other research institutions: users will access resources stored in various repositories browsing by subjects, places, people and time. It will be possible to visualise information from the resources and to further deepen the knowledge directly reaching the websites of each institution.

The Portal harvests metadata from different repositories and can export metadata to other national and international portals. It will also provide contents created and managed by an editorial office, to offer updated news on the main cultural events and to provide thematic itineraries for a guided navigation through the harvested contents.

Resources originating from various data-sources will remain under the control of organizations responsible for their creation, approval, management and maintenance: data will not be duplicated into the Portal's repository but will be retrievable through a unified and interoperable system.

In order to guarantee the interoperability of various kinds of cultural resources and to allow retrieval and indexing functions on their contents, a specific Dublin Core Application Profile (named PICO AP, from the project acronym), has been designed on the basis of the complex domain of "Italian Culture".

The PICO Application Profile has been designed by SNS working group on metadata, on the basis of recommendations, documents and samples published by DCMI. The PICO AP joins in one metadata schema all DC Elements, Element Refinements and Encoding Schemes from the Qualified DC and other refinements and encoding schemes specifically conceived for the project. Therefore, it includes namespaces: 'dc:', 'dcterms:', 'pico:'. While now is more usual to define APs including new elements locally defined, the decision to avoid the introduction of new elements, adding solely new element refinements and encoding schemes, was due to the priority of assuring total interoperability.

The PICO AP is published at the PURL http://purl.org/pico/picoap1.0.xml, according to the DC Application Profile Guidelines, issued as the CEN Workshop Agreement CWA 14855, thus declaring definitions and constraints (Obligation, Condition, Datatype and Occurrence). One of the most relevant extensions of PICO AP is the introduction of three more types, instead of adding new classes and entities (as, for instance, DC Collection Description does), because in this project it was not possible to declare from the beginning all the possible involved classes of agents. For this reason the PICO AP extended the DCMI Type Vocabulary, introducing the PICO Type Vocabulary, a controlled vocabulary which adds the types 'Corporate Body', 'Physical Person' and 'Project'.

Moreover, a PICO Thesaurus has been designed as a specific encoding scheme in order to specify the values of the term "dc:subject", according to which each metadata record is assigned to a specific level of the index tree used for the browsing.

It will be possible to browse the catalogue through the Main Menu or the Theme Menu. According to the 4 High Level Elements of DC Culture, defined by Aquarel project and approved by MINERVA project, the Main Menu of the catalogue is structured in four main facets: Who (people, institutions, etc.); What (art objects, monuments, documents, books, photos, movies, records, theatre and music productions, etc.); When (temporal periods); Where (browsing by region, province, town, through a controlled list or directly through a GIS). User will browse the catalogue using a 'faceted' system: he/she can start the query from one of the four elements and further refine the results range.

Moreover the interface will allow data retrieval on those contents through: a) free search, allowing the user to compose one or more words, using boolean syntax; b) advanced search, to refine queries in the catalogue, selecting if the item to be retrieved is "place", "person", "event", or "object"; c) geographic search, selecting a place on a list or on a map related to a GIS system; d) Themes menu, which groups resources according to the following arguments: Archaeology, Architecture, Visual Arts, Environment and Landscape, Cinema and Media, Music, Entertainment, Traditions, Humanities, Scientific Culture, Education and Research, Libraries, Literature, Archives, Museums, Exhibitions.

Presently about 20 different providers (catalogues by national and regional directorates for Cultural Heritage, museums, collections of digital objects produced by libraries and archives, data bases of public or private bodies) have been involved in the project; mappings have been designed to generate, from different database structure, DC metadata extended according to the PICO AP; metadata have been harvested. Harvested resources presently cover a range from art movable objects (within museums or in the environment), institutions, such as museums or libraries, monuments; books, photographs, archival documents in digital format; library records; research projects, collections (identified within the project MICHAEL).

3 CulturaItalia project: future steps, toward EDL

The work that the group is presently developing, together with carrying on the activity of involving new providers and harvesting more resources, mainly focuses on the release of a new version of the PICO Thesaurus. During the activity of metadata mapping, the first version sometimes shown limitations in the description of resources, therefore many more descriptors have been introduced; moreover, a more consistent hierarchy of broader and narrower terms. With the support of an Italian expert on thesauri design and management, the new version has been designed, taking into consideration both Italian description of cultural heritage expressed in the national law, and international standard thesaural structure (Art and Architecture Thesaurus by Getty Institute, UNESCO Thesaurus, etc.), in order to improve from the beginning the possibility of integration with other international repository. Now the group is working on the development of a SKOS versions, including Italian and English descriptors, scope notes and relations amongst terms.

The second main activity is the adoption of a hierarchical faceted search system, able to cross the different facets (Who, What, When, Where) and, in further steps,

narrower levels of each facet. Some already existing tools and software are under evaluation, together with some projects dealing with such applications (e.g. http://flamenco.berkeley.edu/index.html).

Finally, there is a special task to integrate services Web 2.0-like, such as social tagging and annotation.

4 Interest in joining the SIEDL workshop

As shown above, the project CulturaItalia has many common aspects with Europeana Digital Library: the providing of resources at item level, the adoption of OAI-PMH protocol and of Dublin Core metadata standard (with further extensions) to assure in the same time "syntactic" interoperability amongst different data sources and to maintain the control and management of data by each provider. Therefore, CulturaItalia, to be soon published, could be directly integrated into Europeana, and work as a national metadata aggregator to be harvested by the European repository.

Moreover, the Italian project is now dealing with issues pertaining to semantic interoperability and multilingualism (ontologies, SKOS, RDF).

Therefore the participation of the Italian working group at the SIEDL workshop could be very useful both for the deepening of knowledge of state-of the-art semantic interoperability issues, standards, and tools, and for the potential future integration of the Italian cultural portal into Europeana, specially covering the domain of tangible and un-tangible cultural heritage sector.

Enabling Audiovisual Metadata Interoperability with the European Digital Library

Werner Bailer, Michael Hausenblas and Werner Haas

JOANNEUM RESEARCH Forschungsgesellschaft mbH,
Institute of Information Systems & Information Management,
Steyrergasse 17, 8010 Graz, Austria
firstName.lastName@joanneum.at

Abstract. *Europeana*, the European digital library, museum and archive, will become a reference point for accessing various kinds of cultural contents, including audiovisual contents, in the near future. Thus, establishing interoperability between audiovisual collections and their specific metadata models and the European Digital Library (EDL) is an important issue. We propose a flexible approach for mapping between metadata models and a lightweight method for deploying metadata associated with audiovisual content. Further, we discuss its application for enabling interoperability between audiovisual archives and the EDL.

1 Introduction

There exist a number of portals for accessing cultural digital contents. They target different user groups and many collections are only accessible to professionals. Several initiatives in the domain of audiovisual archives (e.g. INA's Archives pour tous, BBC Open Archive, Beeld en Geluid online public collections[1]) as well as in the domain of libraries are heading in this direction. *Europeana*[2], the European digital library, museum and archive, developed by the EDLnet project, will become in the future a reference point for accessing various kinds of cultural contents, including audiovisual contents. An important issue is to obtain a high interoperability between audiovisual collections and *Europeana*, in order to enable a wide user community unified access to various types of cultural heritage content.

This paper is organised as follows: Section 2 discusses the state of the art of interoperability of audiovisual metadata with the European Digital Library and the Semantic Web. Based on this analysis we propose in Section 3 a flexible approach to mapping audiovisual metadata to EDL compatible target formats and a lightweight deployment mechanism and discuss its application to interfacing with the EDL. Section 4 concludes the discussion.

[1] a comprehensive list can be found at http://del.icio.us/VideoActive
[2] http://www.europeana.eu

2 State of the Art

Interoperable metadata is a key issue for accessing audiovisual content collections. Metadata exchange is hindered by the diversity of metadata formats and standards that exist to cover the specific requirements in certain steps of the audiovisual media production process and in different communities [2]. An overview of existing standards and formats for audiovisual metadata can be found in [11]. There exists also large number of different metadata models and formats for describing the various types of cultural heritage assets (cf. [13]).

Due to this diversity, mapping between different metadata standards is of critical importance. Unfortunately, the "Rosetta Stone" of metadata standards does not exist, i.e. there is no single metadata standard or format that can represent all the different types of metadata, the different structures and granularities of various types of media descriptions [2, 13]. Previous work has shown that this is not even possible in the audiovisual archive domain [3], i.e. there is no single standard or format that satisfactorily covers all aspects of audiovisual content descriptions. A metadata mapping approach using one generic intermediate format is thus not feasible. Due to the number of different formats and standards involved, defining mappings between each pair of formats is also not a feasible approach, as the number of required mappings grows exponentially.

2.1 Audiovisual Content and the Digital Library Community

The digital library community has done significant work in order to establish common metadata models and interoperable metadata representations. However, the EDL and the audiovisual archive community have still not achieved interoperability and the efforts for establishing protocols and formats for interchange are in a very early stage. According to [5] there is ongoing work between the EDL project and the DISMARC[3] and VideoActive[4] projects. DISMARC intends to develop an application profile for audio objects (using metadata terms from the Dublin Core Metadata Initiative plus the Dublin Core Libraries Application Profile [6]). VideoActive uses qualified Dublin Core and MPEG-7. Both projects plan to provide an OAI-PMH[5] interface to their systems.

One important aspect that distinguishes audiovisual content from other media types is the temporal dimension of media items. This is one of the main issues that need to be addressed in order to establish interoperability with other cultural heritage collections. The current EDL metadata model lacks support for representing temporal segments of content and annotating them with specific metadata (support for intra-object descriptions is only one of the long-term

[3] `http://www.dismarc.org`

[4] `http://videoactive.wordpress.com`

[5] Open Archives Initiative, Protocol for Metadata Harvesting, `http://www.openarchives.org`. OAI-PMH is a low-barrier mechanism for repository interoperability. It is a set of six verbs or services that are invoked within HTTP that can be used to exchange documents according to any XML format as long as it is defined by XML schema.

goals [9]), which is commonly done in audiovisual archives, at least for a subset of the collection. For example, major broadcast archives document often more than 50% of their collection analytically [7], i.e. with intra-object metadata. Further aspects which are specific to audiovisual content are among others the number of different objects related to the production of the content (e.g. scripts, story boards), the number of versions that may exist (e.g. rushes, several edited version for different markets and distribution channels), the fact that the semantics of audiovisual content severely depends on the context in which it is used and the much more complex rights situation.

2.2 Interoperability with the Semantic Web

Interoperability of access portals for cultural digital content with the Semantic Web is of growing importance. The EC working group on digital library interoperability [9] defines Semantic Web interoperability with the outside world as one of its goals. In the MultiMatch project[6], OWL is used as a representation of the internal metadata model, which can also serve as a gateway to the Semantic Web. The representation of multimedia metadata in formats that are interoperable with the Semantic Web is still an active research issue. If multimedia objects are described beyond simple cataloging, diverse and partly complex elements need to be represented. A number of multimedia ontologies have been proposed, partly defining new metadata schemes, partly representing existing ones (e.g. MPEG-7). A good overview on the work on multimedia ontologies can be found in [8]. COMM[7] is a new recent proposal for a multimedia ontology. Recently interoperability issues regarding multimedia ontologies have been discussed [15].

One issue that has to be considered is the amount of metadata resulting from the fine-grained description of multimedia content. In a scenario, where just the visual modality of a video is described by low-level descriptors of key frames, one million triples are required to represent only a single hour of video[8]. Given the amount of multimedia data to be accessed in a realistic scenario, this suggests that Semantic Web technologies cannot be easily applied to all metadata in a repository but one has to consider very carefully which part of metadata is suitable for being represented in a Semantic Web compatible format.

3 Flexible Approaches to Mapping and Deployment

In [4] we have recently investigated issues with real-world multimedia assets regarding the Semantic Web. Based on this analysis and the lessons learned from diverse projects in the cultural heritage domain we propose a novel approach to enable interoperability between audiovisual metadata and the European Digital Library. The approach contains two main elements:

[6] http://www.multimatch.org
[7] http://comm.semanticweb.org
[8] http://lists.w3.org/Archives/Public/public-xg-mmsem/2007Jan/0001.html

– A flexible approach to *mapping* that avoids the scalability problem of the currently common mapping approach, i.e. defining hand-crafted mappings between distinct pairs of standards.
– A lightweight approach for the *deployment* of metadata along with the audiovisual content. The term "deployment" is to be understood very broadly as a publication to any target system, e.g. the EDL, some search portal, the Semantic Web.

3.1 Mapping based on Formal Semantics of Metadata Standards

Instead of defining mappings for each pair, we propose to formalise the semantics of the standards involved (e.g. Dublin Core, EBU P_Meta, MPEG-7, etc.). The formal description relates common concepts of content descriptions to their respective manifestations in the different standards. The formalisations can then be used to derive mappings for certain pairs of standards [14]. This is a more efficient and generic approach to the problem avoiding the need for specific mappings for each pair of standards.

As explained in more detail in [14, 12], the semantics of a multimedia standard or of a profile are described using an ontology and rules. For example, we have formalised parts of the semantic constraints of the MPEG-7 Detailed Audiovisual Profile (DAVP) [1] and of the MPEG-7 format used to represent master shot boundary reference data of the TREC Video Retrieval Evaluation[9]. By relating the concepts in the ontologies of each of the standards to common concepts found in audiovisual content descriptions, mapping can be established.

Using this approach a generic service can be implemented, which provides mappings between different standards (or profiles thereof) based on the formalisations that have been defined.

3.2 Lightweight Deployment of Multimedia Metadata

Although there exists an array of multimedia metadata formats (e.g. MPEG-7) that can be used to describe what a multimedia asset is about [11], interoperability issues regarding the actual consumption arise. To enable true interoperability, we advocate the use of the RDF data model[10] for deploying *existing multimedia metadata formats*. More specifically, we propose a solution that allows hooking existing multimedia metadata formats into the Semantic Web: *RDFa-deployed Multimedia Metadata (ramm.x)*. With ramm.x, media assets published on the Web link existing descriptions represented in a multimedia metadata format to a formal representation (ontology), see also [11, Sec. 4]). We propose to use RDFa[11] to deliver the metadata along with the content being served. RDFa is a serialisation syntax for the RDF data model intended to be used in (X)HTML environments, defining how an RDF graph is embedded in an (X)HTML page using a set of defined attributes such as `@about`, `@href`, `@rel`, etc.

[9] `http://www-nlpir.nist.gov/projects/trecvid`
[10] `http://www.w3.org/TR/rdf-concepts/`
[11] `http://www.w3.org/TR/rdfa-syntax/`

In [10] we have recently shown that ramm.x can serve as an excellent device for dealing with lightweight multimedia metadata deployment in the cultural heritage domain. The ramm.x use cases[12] indicate the potential applicability of the approach; further research is under way.

3.3 Application in the Context of EDL

There are two aspects that need to be considered when applying the proposed approach in the context of EDL: the metadata representation itself and the container format. As RDF is becoming more common in the DL community, it can serve as a suitable exchange format data model. For example, both in the EDL project [5] as well as in the Bricks project[13] the use of RDF/OWL is proposed as a way of mapping between metadata schemes without defining specific converters or a "super-scheme". The most common format is of course the Dublin Core Library Application Profile (DC-Lib)[14]. ramm.x can be used to reference services being capable of producing DC-Lib descriptions.

In terms of deployment OAI-PMH has become the standard protocol for harvesting information from different collections in the digital library domain. As OAI-PMH can incorporate different XML based representations, also RDF can be embedded. Moreover, defining RDFa embedding in OAI-PMH could be considered.

4 Conclusion

We strongly believe that it is possible to overcome the limitations found nowadays in digital libraries and archives by exploiting Semantic Web technologies. Related activities in standardisation bodies, such as the W3C Video on the Web workshop[15], indicate the importance of the issues discussed herein. In this paper we have proposed a flexible approach for mapping between metadata models. Further we have rendered a lightweight method for deploying metadata associated with audiovisual content and discussed its application for enabling interoperability between audiovisual archives and the EDL.

References

1. Werner Bailer and Peter Schallauer. The detailed audiovisual profile: Enabling interoperability between MPEG-7 based systems. In Huamin Feng, Shiqiang Yang, and Yueting Zhuang, editors, *Proceedings of 12th International Multi-Media Modeling Conference*, pages 217–224, Beijing, CN, Jan. 2006.
2. Werner Bailer and Peter Schallauer. Metadata in the audiovisual media production process. In Michael Granitzer, Mathias Lux, and Marc Spaniol, editors, *Multimedia Semantics—The Role of Metadata*, volume 101 of *Studies in Computational Intelligence*. Springer, 2008.

[12] `http://sw.joanneum.at/rammx/usecases/`
[13] `http://www.brickscommunity.org`
[14] `http://dublincore.org/documents/library-application-profile/`
[15] `http://www.w3.org/2007/08/video/`

3. Werner Bailer, Peter Schallauer, Alberto Messina, Laurent Boch, Roberto Basili, Marco Cammisa, and Borislav Popov. Integrating audiovisual and semantic metadata for applications in broadcast archiving. In *Workshop Multimedia Semantics - The Role of Metadata (Datenbanksysteme in Business, Technologie und Web, Workshop Proceedings)*, pages 81–100, Aachen, DE, Mar. 2007.
4. Tobias Bürger and Michael Hausenblas. Why Real-World Multimedia Assets Fail to Enter the Semantic Web. In *International Workshop on Semantic Authoring, Annotation and Knowledge Markup (SAAKM07)*, Whistler, Canada, 2007.
5. Sally Chambers. Towards Metadata Interoperability between Archives, Audio-Visual Archives, Museums and Libraries: What can we learn from The European Library metadata interoperability model? D1.1 ECP-2005-CULT-38074-EDL, EDL project, Aug. 2007.
6. Robina Clayphan and Rebecca Guenther. Library application profile. DCMI working draft. DCMI-Libraries Working Group. `http://dublincore.org/documents/library-application-profile`, Sep. 2004.
7. Beth Delaney and Brigit Hoomans. Preservation and Digitisation Plans: Overview and Analysis, PrestoSpace Deliverable 2.1 User Requirements Final Report. `http://www.prestospace.org/project/deliverables/D2-1_User_Requirements_Final_Report.pdf`, 2004.
8. Hariklia Eleftherohorinou, Vasiliki Zervaki, Anastasios Gounaris, Vasileios Papastathis, Yiannis Kompatsiaris, and Paola Hobson. Towards a Common Multimedia Ontology Framework (Analysis of the Contributions to Call for a Common multimedia Ontology Framework Requirements). Technical report, AceMedia, Apr. 2006.
9. Stefan Gradmann. Interoperability of Digital Libraries - Report on the EC working group on DL interoperability. `http://bnd.bn.pt/seminario-conhecer-preservar/doc/Stefan%20Gradmann.pdf`, Sep. 2007.
10. Michael Hausenblas, Werner Bailer, and Harald Mayer. Deploying Multimedia Metadata in Cultural Heritage on the Semantic Web. In *First International Workshop on Cultural Heritage on the Semantic Web, collocated with the 6^{th} International Semantic Web Conference (ISWC07)*, Busan, South Korea, 2007.
11. Michael Hausenblas, Susanne Boll, Tobias Bürger, Oscar Celma, Christian Halaschek-Wiener, Erik Mannens, and Raphaël Troncy. Multimedia Vocabularies on the Semantic Web. W3C Incubator Group Report, W3C Multimedia Semantics Incubator Group, 2007.
12. Martin Höffernig, Michael Hausenblas, and Werner Bailer. Semantics of temporal media content descriptions. In *Proceedings of Multimedia Metadata Applications Workshop at I-MEDIA'07*, pages 155–162. Journal of Universal Computer Science (J.UCS), Sept. 2007.
13. Johan Oomen and Hanneke Smulders. First Analysis of Metadata in the Cultural Heritage Domain. D2.1, MultiMatch project, Oct. 2006.
14. Raphaël Troncy, Werner Bailer, Michael Hausenblas, Philip Hofmair, and Rudolf Schlatte. Enabling multimedia metadata interoperability by defining formal semantics of MPEG-7 profiles. In *Proceedings of 1st International Conference on Semantic and Digital Media Technologies*, Athens, GR, Dec. 2006.
15. Raphaël Troncy, Oscar Celma, Suzanne Little, Roberto Garcia, and Chrisa Tsinaraki. MPEG-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In *1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies Workshop (MAReSO07)*, Genova, Italy, Dec. 2007.

Promoting Government Controlled Vocabularies for the Semantic Web: the EUROVOC Thesaurus and the CPV Product Classification System

Luis Polo Paredes, Jose María Álvarez Rodríguez, and Emilio Rubiera Azcona

Fundación CTIC, Gijón, Asturias, Spain,
{luis.polo,josem.alvarez,emilio.rubiera}@fundacionctic.org,
WWW home page: http://www.fundacionctic.org

Abstract. The aim of the paper is to promote government controlled vocabularies for the Semantic Web in the context of the European Union. We will propose SKOS as the RDF/OWL common data model and an enclosed conversion method for the knowledge organization systems available in RAMON, the Eurostat Metadata Server. The study cases of this paper will be the Eurovoc Thesaurus and the Common Procurement Vocabulary, a product classification system.

1 Introduction

Knowledge Organization Systems (*KOS*), such as thesauri, taxonomies or classification systems, are developed by specific communities and institutions in order to organize huge collections of information objects: documents, texts, webpages, and multimedia resources as well. These vocabularies allow users to annotate the objects and easily retrieve them, promoting lightweight reasoning in the Semantic Web. Topic or subject indexing is an easy way to introduce machine-readable metadata for a resource's content description.

In the european eGovernment context, there are several conceptual/terminological maps of particular domains available in RAMON[1], the Eurostat's metadata server: in the Health field, the *European Schedule of Occupational Diseases* or the *International Classification of Diseases*; in the Education field, thesauri as *European Education Thesaurus* or the *European Glossary on Education*; in the Employment field, the *International Standard Classification of Occupations* among others. The structure and features of these systems are very heterogeneous, although some common aspects can be found in all of them: 1. Hierarchical relationships between terms or concepts. 2. Multilingual character of the information.

In this paper, we propose a common data model for RDF/OWL encodings of governmental controlled vocabularies and an enclosed generic method to allow straight-forward conversions. The SKOS vocabulary has been selected as

[1] http://ec.europa.eu/eurostat/ramon

the target common data model, thus we avoid to develop an ontology from scratch in order to metamodel thesauri and classification systems. Adopting a common "semantic" format will facilitate semantic interoperability and common understanding for information interchange between European digital libraries, governmental agencies and private third-parties. We will analyze and convert to SKOS two existing EU knowledge organization systems.

1. **The Eurovoc thesaurus**[2] is a multilingual, polythematic thesaurus focusing on the law and legislation of the European Union (EU). It is available in 21 official languages of the EU. Within the EU, the Eurovoc thesaurus is used in the Library of the European Parliament, the Publication Office as well as other information institutions of the EU. Moreover, the Eurovoc thesaurus is used in the libraries and documentation centers of national parliaments (e.g. Spanish Senate) as well as other governmental and private organizations of member (and non-member) countries of the EU.
2. **The Common Procurement Vocabulary**[3] (CPV) is a single product classification for describing the subject matter of public contracts, allowing companies to easily find public procurement notices[4] and increasing competitivity and business opportunities within European market. Its main goal is to standardize the codes used by contracting authorities. The use of the CPV is mandatory in the European Union from February 1, 2006 and it is regulated by Commission Regulation adopted on November 28, 2007 amending Regulation (EC) Nº 2195/2002 of the European Parliament.

This paper is structured as follows: in Section 2, we check different approaches for thesauri and product classification schemes conversions to RDF/OWL format. In Section 3, we propose the minimun set of common requirements for *KOS* sytems conversion, and we select a generic conversion method. In Section 4, we apply the method to the EUROVOC thesaurus and the CPV. Finally, we evaluate the results of our conversions and we present some conclusions.

2 Existing approaches for converting controlled vocabularies to RDF/OWL

This section discusses existing methods to convert *KOS* systems. We distinguish between RDF/OWL conversions methods for thesauri and product classification systems.

2.1 Thesauri Conversion Methods

A thesaurus is a controlled vocabulary, with equivalent terms explicitly identified and with ambiguous words or phrases (e.g. homographs) made unique. This set

[2] http://europa.eu/eurovoc/
[3] http://europa.eu/scadplus/leg/en/lvb/l22008.htm
[4] Published in *Tender Electronical Daily*.

of terms also may include broader-narrower or other relationships. Usually they are considered to be the most complex of controlled vocabularies.

Thesauri as a *KOS* system can be converted to RDF/OWL by means of different procedures. On one hand, there are methods, as the Soergel et al. one in [14] or Van Assem et al. in [15], that propose specific techniques for thesauri conversions into an ontology. However their method does not target a specific output format and it considers the hierarchical structure of thesauri as logical *is-a* relationships. On the other hand, there are some generic methods for thesauri conversions, as the step-wise method defined by Miles et al. in [10]. This method selects a common output data model, the SKOS vocabulary, and is comprised by the following steps: a) generation of the RDF encoding, b) error checking and validation and c) publishing the RDF triples on the Web. In addition, this method has been refined in [16], adding three new substeps for the generation of RDF encoding: 1. analyzing the vocabulary, 2. mapping the vocabulay to SKOS properties and classes and 3. building a conversion program .

2.2 Product Classification Systems Conversion Methods

Product Classification Systems (also known as PCSs) have been developed to organize the marketplace in several vertical sectors that reflect the activity (or some activities) of economy and commerce. They have been built to solve specific problems of interoperability and communication in e-commerce [9] providing a structural organization of different kind of products tied together by some economical criteria. The aim of a PCS is to be used as a *de facto* standard by different agents for information interchange in marketplaces [13,3].

Many approaches for product classification systems adaptation to the Semantic Web, like [2,7,8], present methods with the goal to convert them to domain-ontologies. The tree-based structure between product terms is interpreted then as a logical *is-a* hierarchy. From our point of view and following the discussion about [5,6], hierarchical links between the elements of each economic sector do not have the semantics of subsumption relationships. The next example taken directly from CPV ("term" and its code) shows how the relationship between the element "Parts and accessories for bicycles" (34442000-7) and its direct antecessor,"Bicycles" (34440000-3), does not seem as an *is-a* relation. In this case, an ontological property for object composition like *hasPart* would be much better. Moreover, there are further remarks against the idea of using the PCSs as domain-ontologies. It is difficult to assert that the CPV element, "Tin bars, rods, profiles and wire" (27623100), represents any specific product. Rather it should be regarded as a collection of products. To convert correctly this element into a domain ontology, it should be considered as equivalent to the union of several concepts (e.g. $TinBar \sqcup TinRod \sqcup TinProfiles \sqcup TinWire$).

Our approach instead do not consider PCSs as domain ontologies, but as a specific kind of *KOS* systems. Any PCS, as well as other classification systems (i.e. product classification systems, economic activities classification systems, occupation classification systems, etc.), are interpreted as a conceptual

scheme comprised of conceptual resources. From this point of view, hierarchical relationships are not considered to be any more logical *is-a* relations, but broader/narrower ones.

3 "Greatest common divisor" of controlled vocabularies

As we have just introduced in the previous section, Knowledge Organization Systems are used for organizing large collections of information objects and efficient retrievery. Existing controlled vocabularies are currently available in several formats: XML files, spreadsheets or text. However promoting them to the the Semantic Web is not a mere process of RDF/OWL conversions of data. Conversions need to fullfil some requirements. Firstly, a common RDF/OWL representation is needed to ensure a) semantic compatibility between different vocabularies, b) processing vocabularies in a standard way and c) sharing vocabularies for third-parties adoption. SKOS, presented in the W3C SKOS Reference Working Draft [11], has been selected for these purposes. Secondly, although controlled vocabularies do not share some features, in practice a distinction between them is very hard to draw. We have identified a minimun set of common features for them. Therefore the data model should be expressive enough to preserve as much as possible the original semantics of primary sources for these common features. Thirdly, a generic method is needed to ensure the quality of data conversions to correct SKOS instances.

We have carried out a refinement of the methods [10,16] for thesauri conversions, by extending it to the PCSs field and taking into account their special features commented in section 2.2. These are the common features of *KOS* systems that have to be covered by the conversion method:

URI generation. Controlled structured vocabularies and conceptual resources are interpreted in SKOS as RDF resources: in particular, instances of `skos:ConceptScheme` and `skos:Concept`. Thus they are referenced by means of Uniform Resource Identifiers (URIs). Although namespaces are out of the scope of our analysis, one of the substeps of the method is the generation of the `rdf:IDs` of `skos:Concept` and `skos:ConceptScheme` from the original datasource. Controlled vocabularies usually provide unique identifiers for their terms or concepts. The options are the following:

1. Generating new identifiers for the elements of the vocabulary. This option introduces additional management. A mapping between elements of the original source and identifiers should be maintained for updating purposes.
2. Using the string of the preferred term. We would like to highlight here that multilingual sources introduce a factor of complexity that it is not present in monoligual systems. In european multilingual sources, this solution implies selecting a preferred term in a given natural language, thus promoting one language over the others with a possible non-desired political impact. In addition, a control procedure has to be established to ensure URI updating if the source term changes.

3. Using the identifier code of an element, if any. This solution avoids the problem of selecting one preferred language to encode the concept URIs. Moreover, codes are usually strings composed by a serial number (legal URI characters) and it preserves the original semantics of a multilingual vocabulary, where these codes identify unique terms or concepts and establish mappings between different languagues. This last option has been chosen for our method.

Hierarchy formalization. From our point of view, one of the common aspects shared by *KOS* is a hierarchy-based structure, at least by thesauri, taxonomies and by most of classification schemes [1]. Hierarchical relations establish links between conceptual resources, showing that the semantics of a resource is in some way more general ("broader") than other ("narrower"). In SKOS, the properties `skos:broader` and `skos:narrower` are only used to assert hierarchical statements between two conceptual resources. By the way, these properties are not currently defined ([11]) as transitive properties (as they were in [12]). Nevertheless, third-parties, if they consider valuable, can use an OWL reasoner to infer the transitive closure of the hierarchy by means of the transitive superproperties of `skos:broader` and `skos:narrower`: `skos:broaderTransitive` and `skos:narrowerTransitive` properties.

Multilingual and lexical features. Regarding European controlled vocabularies, multilinguism is a critical issue. Both CPV Vocabulary and Eurovoc Thesaurus are available in 21 official languages of the European Union (Bulgarian, Spanish, Czech, Danish, German, Estonian, Greek, English, French, Italian, Latvian, Lithuanian, Hungarian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene, Finnish and Swedish) and one more (Croatian). In SKOS conceptual resources are labelled with any number of lexical strings, in any given natural language identified by the `xml:lang` attribute, following normative RDF/XML syntax. One of these labels is selected as the `skos:prefLabel` for any given language, and the others as values of `skos:altLabel`.

In thesauri like Eurovoc, the `USE` and `UF` relations between descriptors and non-descriptors are language specific. Both, `USE` and `UF` relations, express term-equivalence relationships (synonymy, antonymy, holonymy, etc.). The W3C Semantic Web Deployment Working Group[5] is currently working on an extension to SKOS, SKOS-XL[6], for describing in detail these term relationships and for modeling lexical entities. A special class of lexical resources, called `xl:Label`, is defined. The use of this class would be much more accurate, as Eurovoc is a term-based thesaurus. However, there are mainly two problems 1) this information is not explicitely represented in the original datasources (thus the type of equivalence relation has to be detected by a human expert) and 2) the approach of the SWD Working Group is still in a preliminar stage. Moreover, notice that there is no additional benefits for the treatment of multilingual features: the language identification is still encoded using the `xml:lang` attribute.

[5] http://www.w3.org/2006/07/SWD/
[6] http://www.w3.org/2006/07/SWD/SKOS/xl/20080414

4 Case Studies

In this section, we apply our method to the *Eurovoc Theusarus* and the *CPV vocabulary*. Firstly, we generate the RDF/OWL encoding:

4.1 Adaptation of the EUROVOC Thesaurus to SKOS

Step 1: analyze controlled vocabulary. We used the XML version in our analysis. Eurovoc/XML structure is specified using a DTD associated with language specific files, providing a multilingual scheme.

The thematic scope of Eurovoc scope is defined by 21 thematic fields (identified by two-digit numbers and titles in words): e.g. *10 EUROPEAN COMMUNITIES*, divided into 127 microtheasuri (identified by four-digit numbers): *1011 COMMUNITY LAW*. The latest version contains 6,645 descriptors, 6,669 hierarchical and 3.636 associative relationships.

Step 2: map data vocabulary to SKOS. (See Table 1) Eurovoc has been developed using the standards ISO 2788 (for monolingual thesauri) and ISO 5964 (for multilingual theasuri). SKOS is compatible with the ISO 2788. In Eurovoc, descriptors are equivalent across the diversity of the 21 natural languages in which the thesaurus is encoded. The "descripteur_id" of descriptors is used to generate the `rdf:ID` of the instances of `skos:Concept`.

Hierarchical relationships between descriptors are expressed in Eurovoc using `BT` and `NT` relations. They are mapped to their equivalent elements in the SKOS data model: `skos:broader` and `skos:narrower` respectively. They are not defined as transitive properties, so there is no need of a transitive closure of the hierarchical relations in our conversion. In ISO 2788, polyhierarchies are not allowed, however certain descriptors in fields *72 GEOGRAPHY* and *76 INTERNATIONAL ORGANIZATIONS* have more than one broader term at the next higher level. SKOS data model allows this structure thus the correspondence will still be complete producing a correct SKOS. The `skos:related` property is used to map associative links between descriptors of hierarchical-trees (`RT` relationships). Both relationships (`skos:related` and `RT`) are symmetrical, not transitive and incompatible with the hierarchical relationship: if two conceptual resources are linked by a hierarchical relationship then there cannot be an associative relationship between them.

As we described above, Eurovoc descriptors are organized in two hierarchical levels: thematic fields and microthesauri. There is no direct translation into SKOS of these non-standard features of the thesaurus. Microthesauri and fields are also semantically related: each descriptor is assigned to a single microthesaurus, which belongs to a single thematic field. However, SKOS does not provide any property to create semantic links between concept schemes. On the one hand, we want to express that the descriptor *3062*-"executive body" is a top term of the microthesaurus *0436*-"executive power and public service" and the microthesaurus belongs to the thematic field *04*-"POLITICS". On the other hand, we want to represent complex internal structures of concept schemes. There are mainly two options here:

1. Interpreting microthesauri as `skos:ConceptScheme` instances. While fields are considered `skos:ConceptScheme` subclasses. Each Eurovoc microthesaurus is an instance of a single field class. Basically, there are two problems: i) it is not possible to link microthesauri and fields to the Eurovoc thesaurus RDF resource in a standard way and in addition ii) there is also no convincing way to relate top-hierarchy terms and fields. Triples of `skos:hasTopConcept` asserting direct links using is not DL compatible (fields are logical classes and descriptors instances) and formulas of the style $\texttt{Field} \equiv \forall \texttt{skos:hasTopConcept}.\ \{\texttt{Descriptor}_1, .., \texttt{Descriptor}_n\}$ introduce reasoning with nominals without evident benefits. This option has been discarded.
2. Interpreting microthesauri and fields as `skos:ConceptScheme` instances. This approach captures more accurately the semantics of microthesauri and fields resources. The `skos:hasTopConcept` and `skos:inScheme` properties can also be correctly used to assert local associations of descriptors to both microthesauri and fields without contradicting SKOS data model. However a new OWL object-property, `hasScheme`, has to be added (extending SKOS) to be able to assert hierarchical links between instances of `skos:ConceptScheme`. This property allows us to express *KOS* systems internal composition. The property has been defined as follows: its domain and range are `skos:ConceptScheme` and it is transitive from the whole to its parts: the Eurovoc "hasScheme" some fields, and every field "hasScheme" some microthesauri. We have chosen this option.

Step 3: convert the data. We have chosen XSL technology to convert the Eurovoc XML source into RDF/SKOS. In the first one, we built the basic skeleton of the document including all of microthesauri (concept schemes) and terms (concepts). Secondly, we decorated the definitions of concepts adding iteratively.

4.2 Adaptation of the CPV Vocabulary to SKOS

The main vocabulary of the CPV contains around $8,200$ numerical codes, each one describing a single product term.

Step 1: analyze controlled vocabulary. The CPV consists of a main vocabulary, and a supplementary vocabulary that can be used for adding further qualitative information to the description of the subject of a contract. Only the main vocabulary will be considered in this analysis. This main vocabulary is composed of product terms identified by an alphanumeric code (an 8 digit code plus a check digit), see Table 2. The alphanumeric codes are shared between different versions of the CPV in each country, thus defining linguistic equivalence across languages of the European Union. The description "lemons" in the English version and the description "limones" in the Spanish one are both considered to be equivalent because both terms are identified with the same code: 01131210-9.

Step 2: map data vocabulary to SKOS. (See Table 3) Product terms have been considered `skos:Concept` and the code has been used to generate their `rdf:ID`. Specific-language literal description of product terms are mapped

Data Item	Feature	SKOS Element
< $Descripteur_id$=X	Concept	skos:Concept with `rdf:ID`=X
< *Descriptor* >=Y in language=L	Preferred Term	`skos:prefLabel`= Y@xml:lang='L'
< $thesaurus_id$ >=X	Concept Scheme	`skos:ConceptScheme` with `rdf:ID`=X
Microthesaurus=Y in language=L	Concept Scheme Label	`skos:prefLabel`= Y@xml:lang='L'
< $domain_id$ >=X	Concept Scheme	`skos:ConceptScheme` with `rdf:ID`=X
Field=Y in language=L	Concept Scheme label	`skos:prefLabel`= Y@xml:lang='L'
Non-descriptor=Y in language=L, `UF` and `USE`	Equivalence term relation	`skos:altLabel`= Y@xml:lang='L'
CPV Non-descriptor=Y in language=L, `PERM`	Equivalence term relation	`skos:hiddenLabel`= Y@xml:lang='L'
`BT` Term	Broader Term	`skos:broader`
`NT` Term	Narower Term	`skos:narrower`
`RT` Term	Related Term	`skos:related`
`SN` Note in language=L	Scope Note	`skos:scopeNote`= Y@xml:lang='L'

Table 1. Mapping of Eurovoc Data Items to concept-based controlled vocabularies features and SKOS/OWL Classes and Properties.

to the SKOS property `skos:prefLabel` and the `xml:lang` attribute is used to identify the language.

Every product term is assigned to exactly one category. As we have considered product terms as `skos:Concept` instances, a decission had to be made about CPV categories formalization. They can not be straightforward mapped to any SKOS feature, thus we have introduced four new classes and we have declared them as `skos:Concept` subclasses using the RDF Schema property, `rdfs:subClassOf`. Product terms are also declared instances of their corresponding product category level. These statements can be realized parsing patterns in the product terms codes, see Table 2. There are 61 members of the top-level category, each one have been considered the top-term of a single vertical product sector comprised in a tree-structure.

The conversion of each subtree (product sector) has been made using a bottom-top transformation and the SKOS semantic property `skos:broader`. Parsing the alphanumeric code of each product term is sufficient to generate its direct parent in the hierarchy. E.g. Code "01112000-5" of the term "potatoes and dried vegetables", its broader term code is the string "01110000", which identifies the element "cereals and other crops" (the control digit can be easily generated then). However, if we try to generate the tree in the other direction (top-bottom), the transformation is computationally more complex. Given the previous code

Product Category	Identifier Code	Example
Division	XX000000-y	01000000-7
Group	XXX00000-y	01100000-8
Class	XXXX0000-y	01110000-1
Cat(L0)	XXXXX000-y	01112000-5
Cat(L1)	XXXXXX00-y	01112200-7
Cat(L2)	XXXXXXX0-y	01112210-0
Cat(L3)	XXXXXXXX-y	01112211-7

Table 2. Common Procurement Vocabulary Structure.

Data Item	Feature	SKOS Elements
Code=X	Concept	skos:Concept with `rdf:ID`=X
Product Term=Y in language=L	Preferred Term	`skos:prefLabel`= Y@xml:lang='L'
Product Category=C	Product Category	`rdf:ID`=C (subclass of `skos:Concept`)
Code=X for categorization	Product Category	X instance of Class C
Code=X for tree structure	Broader Term	`skos:broader`

Table 3. Mapping of CPV Data Items to concept-based controlled vocabularies features and SKOS/OWL Classes and Properties.

"01112000-5", 10 new alphanumeric codes (01112[0-9]00) will be generated as its possible descendants (`skos:narrover` in this case). Moreover another operation is necessary to check the existence of every generated code as the code generation process does not guarantee that the new codes are present in the CPV Vocabulary (e.g. the code "01112400-5" will be generated from "01112000-5", but it does not identify any product term in the CPV Vocabulary).

Therefore, a bottom-top algorithm has been choosen. The `skos:narrower` relations can be inferred using an OWL reasoner as `skos:narrower` is the inverse property of `skos:broader`.

Step 3: convert the data. In this case, we have used the RDF123 [4] tool and XSL to build the CPV in SKOS. Firstly, we created a new spreadsheet with the original MSExcel version of the CPV and all values for the mappings. Secondly, we loaded the new spreadsheet into RDF123 and created the mappings between cols and RDF nodes. Finally, as for Eurovoc, we applied several identity transformations with XSL to include all languages labels in the generated document.

After the conversions of Eurovoc thesaurus and CPV vocabulary, the execution of the complete method finished with the validation of both transformations with the *W3C SKOS Validator*[7]. Finally all the triples have been stored using the RDF repository, Sesame. The SKOS versions of the controlled vocabularies

[7] http://esw.w3.org/txamplesopic/SkosValidator

are publicly available on: Eurovoc (550, 707 triples)- http://idi.fundacionctic-.org/sparql_client/eurovoc and CPV (191, 472 triples)-http://idi.fundacionctic-.org/sparql_client/cpv. They can be queried using SPARQL.

5 Conclusions

We have defined the minimal set of features that a conversion method to RDF/OWL must cover to be applied for *KOS* systems in the EU context. Also, we have selected and tested SKOS as the common data model for the representation of these controlled vocabularies and we have carried out an existing method successfully.

The evaluation of the conversions must be checked from two different levels:

Correctness of SKOS conversions. Our approach has demonstrated to produce correct SKOS for the "greatest common divisor" of controlled vocabularies conversions: URIs generation, hierarchical relations and multilingual features.

Completeness of SKOS conversions. Our approach has demonstrated that just using SKOS does not produce complete conversions of the selected primary sources. Some specific properties and classes have to be added to preserve completely the original semantics. In the case of the Eurovoc, fields and microthesauri organize the set of descriptors. In this case, a new property, `hasScheme`, has to be added to be able to express the internal structure of the thesaurus. On the other hand, in the case of the CPV Vocabulary, every product term is associated with a product category. The output data model was extended with four new subclasses of `skos:Concept` to preserve the product categories hierarchy in the RDF/OWL conversion. Every product term has been interpreted as an instance of `skos:Concept` and as instance of its correspondent product category class.

Finally, we plan to apply the conversion method to other EU controlled vocabularies available in RAMON, the EUROSTAT metadata server. Our aim is to develop a common EU framework of *KOS* systems based on the SKOS vocabulary. A framework based on the same data model will promote semantic interoperability between European digital libraries, governmental agencies and other third-parties. The case studies of this paper, the Eurovoc Thesaurus and the CPV Vocabulary, show how the method can be put in practice.

In addition further research will allow us to convert to RDF/OWL the mappings between different classification systems. This is the case, for example, of product terms from the CPV Vocabularies and other PCSs. There exist some tables showing the correspondence between the CPV and the *Statistical Classification of Products by Activity* (CPA), the *General Industrial Classification of Economic Activities* within the European Communities (NACE Rev. 1) and the *Combined Nomenclature* (CN). The SKOS language provides a set of specific properties to represent semantic links between concepts belonging to different concept schemes. We will explore how this set of SKOS properties can indicate

that a product term from the CPV is sufficiently similar with another product term from the NACE Classification to use them interchangeably in an information retrieval application.

We would like to report the results to the *Office for Official Publications of the European Communities* and the Linking Open Data[8] initiative, specially to the Riese Project[9] (*RDFizing and Interlinking the EuroStat Data Set Effort*).

Acknowledgements The work could not have been developed without the collaboration of the *Office for Official Publications of the European Communities* that granted us the license of the Eurovoc source, ref. 2008-COP-002, for academic purposes.

References

1. Richard Benjamins, Dieter Fensel, and Asunción Gómez Perez. Knowledge Management through Ontologies, 1998.
2. Oscar Corcho and Asunción Gómez Pérez. Integrating E-Commerce Standards and Initiatives in a Multi-Layered Ontology.
3. Dieter Fensel, Ying Ding, and Borys Omelayenko. Product Data Integration in B2B E-commerce. *IEEE-Intelligent E-Business*, pages 54–59, 2001.
4. Lushan Han, Tim Finin, Cynthia Parr, Joel Sachs, and Anupam Joshi. RDF123: A Mechanism to Transform Spreadsheets to RDF. Technical report, University of Maryland, Baltimore County, 1 2007.
5. Martin Hepp. A Methodology for Deriving OWL Ontologies from Products and Services Categorization Standards. (1):72–99, 2006.
6. Martin Hepp. The True Complexity of Product Representacion in the Semantic Web. In *Proceedings of the 14th European Conference on Information System (ECIS 006)*, Gothenburg, Sweden, 2006.
7. Martin Hepp. Possible Ontologies. *IEEE-Internet Computing*, (1):90–96, 2007.
8. Jörg Leukel, Volker Schmitz, and Frank-Dieter Dorloff. A Modeling Approach for Product Classification Systems. In *DEXA '02: Proceedings of the 13th International Workshop on Database and Expert Systems Applications*, pages 868–874, Washington, DC, USA, 2002. IEEE Computer Society.
9. Jörg Leukel, Volker Schmitz, and Frank-Dieter Dorloff. Exchange of Catalog Data in B2B Relationships - Analysis and Improvement. In *ICWI*, pages 403–410, 2002.
10. Brian Matthews, Alistair Miles, and Michael Wilson. Modelling Thesauri for the Semantic Web. In *Paper submission in to 3rd Workshop on Semantic Web and databases*, Humboldt-Universität Berlin, Germany, 2003.
11. Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organisation Systems. W3C working draft, W3C, 2008. http://www.w3.org/TR/skos-primer/.
12. Alistair Miles and Dan Brickley. SKOS Simple Knowledge Organisation Systems. W3C working draft, W3C, 2005. http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/.
13. Borys Omelayenko and Dieter Fensel. An Analysis of B2B Catalogue Integration Problems. In *Proceedings of the International Conference on Enterprise Information Systems (ICEIS-2001)*, Setúbal,Portugal, 2001.

[8] http://linkeddata.org/
[9] http://riese.joanneum.at/

14. D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, and S. Katz. Reengineering Thesauri for New Applications: the AGROVOC example. *Journal of Digital Information*, 2004.
15. M. van Assem, M. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga. A Method for Converting Thesauri to RDF/OWL, 2004.
16. Mark van Assem, Véronique Malaisé, Alistair Miles, and Guus Schreiber. A Method to Convert Thesauri to SKOS. pages 95–109. 2006.

Video Active
Television Heritage in the European Digital Library: A Semantic Interoperability Approach

Johan Oomen (Netherlands Institute for Sound and Vision)
joomen@beeldengeluid.nl, Vassilis Tzouvaras (National Technical University of Athens) tzouvaras@image.ntua.gr

Abstract: In this paper is provided an insight into the background and development of the Video Active Portal which offers access to television heritage material from 14 archives across Europe. The Video Active project has used all the latest advances of the Semantic Web technologies in order to provide expressive representation of the metadata, mapping heterogeneous metadata schema in the common Video Active schema, and sophisticated query services. Using these technologies, Video Active is fully compliant with the EDL interoperability specifications.

Introduction

The greatest promise of the internet as a public knowledge repository is to create seamless access for anyone, anywhere, to all knowledge and cultural products ever produced by mankind. Mainly due to increased bandwidth availability, web sites offering online video material have managed to mature and in a short period have become extremely popular. Web sites like YouTube [2], MySpace [3], Revver [4] and many others show how the idea of making and manipulating images (once mostly the preserve of professionals) has been embraced as a way of broadcasting who we are to anyone prepared to watch. The most popular site to date, YouTube, was launched in early 2005 and serves over 100 million videos daily [5]. The number of user generated video uploads per day is expected to go from 500,000 in 2007 to 4,800,000 in 2011 [6]. Recent studies indicate that the number of U.S. internet users who have visited a video-sharing site increased by 45% in 2007, and the daily traffic to such sites has doubled [7].

Looking at these numbers, it's evident that the potential for releasing material from audiovisual archives online is enormous. To date, however, from the many millions of hours in these archives [8] online a few percent can be found online. Many of the existing online services are based on user generated content. And if professional content is offered (i.e. Joost [9], Miro [10], Blinkx [11]) the focus is rather on recent material.

Audiovisual archives need to overcome several obstacles before they can set up meaningful online services. These include: managing intellectual property rights, technological issues concerning digitisation and metadata standardisation and issues related to the way the sources are presented to users. The latter is even more challenging if the aim is to present material from several countries in a structured way, in fact the starting point of the Video Active project.

The main challenge of Video Active is to remove the main barriers listed above in order to create multilingual access to Europe's television heritage. Video Active achieves this by selecting a balanced collection of television archive content, which reflects the cultural and historical similarities and differences of television from across the European Union, and by complementing this archive content with well-defined contextual metadata. Video Active is invited member of EDLnet, the network was initiated in 2006 to built consensus to create the European Digital Library [12]. Video active will be made available though the Europeana.eu portal.

This article firstly introduces the main challenges and in the second part will provide some details on the technical infrastructure that was created in the first and second year of this three-year project. The focus here will be on semantic interoperability.

The project

Video Active is funded within the eContentplus programme of the European Commission (Content Enrichment Action) and started in September 2006 for a

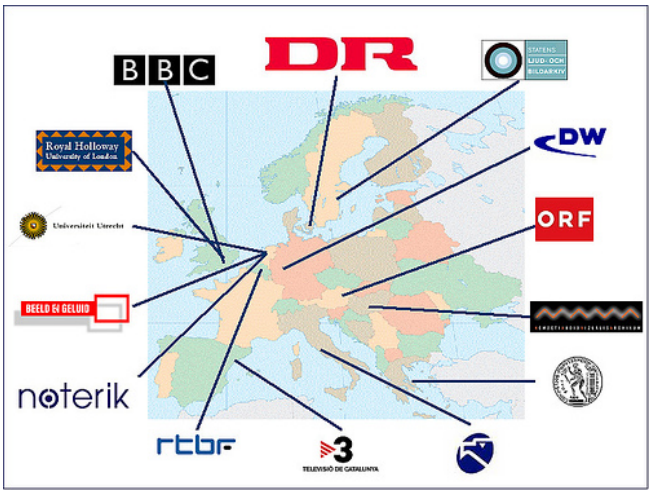
duration of 36 months. The first beta version of the portal was launched in February 2008.

The consortium

14 Major European audiovisual archives, academic partners and ICT developers form the consortium. The archives will supply their digital content; the universities are the link to end-users and play an important role in developing a strategy for selecting the content and in delivering the necessary context information. The ICT developers will be responsible to supply the technology needed. The collections Added up, their collections comprise of five hours of audio and video material from 1890 up to today.

1. British Broadcasting Corporation, UK
2. Danish Broadcasting Corporation, DK
3. Deutsche Welle, DE
4. Hungarian National Audiovisual Archive, HU
5. Institut National de l'Audiovisuel, FR *
6. Istituto Luce, IT
7. Hellenic Audiovisual Archive, GR[1]
8. Moving Image Communications Ltd, UK *
9. Netherlands Institute for Sound and Vision, NL
10. Österreichischer Rundfunk, AT
11. Radio-Télévision Belge de la Communauté Française, BE
12. Swedish Institute for Sound and Image, SE
13. Televisio de Catalunya, ES
14. Vlaamse Radio- en Televisieomroep, BE *

[1] * *indicates that these archives have joined after the start of the project as associate partners*



Amsterdam based Noterik Multimedia is specialised in online video solutions and responsible for the development of the Video Active portal application. The second technical partner is the National Technical University of Athens, expert in knowledge representation and responsible for the metadata management. The media studies faculties of Utrecht University and Royal Holloway, University of London complete the consortium.

Users and their demands

The demand for access to television archive content online has been growing, and this demand has been driven from a number of distinct sectors: education the general public and the heritage sector.

Digitisation of archive content transforms cultural heritage into flexible 'learning objects' that can easily be integrated into today's teaching and learning strategies. For the academic community the rich holdings of television archives are valuable teaching and research resources. Up to now access has been limited with much of the archive content stored on legacy formats and minimum description. Although this is changing with many of the preservation and digitization projects underway in large audiovisual

archives across Europe, the comparative dimension of European television content is not explored yet.

As noted in the introduction, the public demand for archive content has risen with the growth and affordability of the Internet and media publishing tools. Cultural heritage is of interest to everyone, not just specialists and students. The 19th century saw a huge development in museums, libraries, galleries and related heritage institutions, all with public access. Many such institutions have very low charges (or are free) in order to make access truly public and democratic. Audiovisual collections are much less accessible and democratic. Broadcast archives are closed to the public, most 'public' film and video institutions charge by the hour for personal access, and many such institutions are not actually public. Instead, they require proof of research status before allowing access to the general collections.

The digital age also has its impact on the work of professionals in the heritage domain, such as museum curators, organisers of exhibitions, journalists, documentalists, etc. They can conduct their activities and render services faster, better, more efficiently and sometimes at a lower cost. In short, a so-called e-culture is emerging. Additionally, in the digital age, the value of heritage institutions lies increasingly in their role as mediators between networks that produce culture and impart meaning. More and more, they will find themselves contributing their knowledge and content within a cultural arena where a host of highly diverse players are in action, including non-cultural sector institutions, as well as the audience or users. This means that the added value of heritage organisations is increasingly dependent on the extent to which they are able to make knowledge sharing, crossovers, and structural cooperation part of their 'core business'

These user groups have specific expectations and profiles, and the Video Active project has to understand and encompass these to ensure user satisfaction and revisits. Surveys, face to face interviews and desk research have been conducted in the initial stages of the project. The resulting insight in user requirements became fundamental

to define the technical specifications and hence the technical architecture. Further requirements testing will take place on the first release of the portal; comprehensive evaluation with key users will provide the project with input at it develops the second release, planned for the second year of the project.

Content and intellectual property rights

By definition, the selected content on the Video Active portal is heterogeneous in many ways, language being one. A multilingual thesaurus allows multilingual access to the holdings. In the first release of the Video Active portal, ten languages will be supported.

Other challenges regarding the content selection strategy are related to the variety of archive holdings amongst the content providers for the project across both historical periods and genres. Also, the availability of supplementary content (images, television guides etc.) and metadata by the content providers is not spread equally amongst the partners.

In order to tackle these issues, Video Active has developed a content selection strategy that followed a comparative perspective; seeking to explore and show the cultural and historical similarities and differences of television in Europe through various themes [13]. The thematic approach allows for the development of a rich resource that explores the history of Europe using television archive content from across a number of genres and periods. So far 40 different themes have been selected and together with the historical coverage, a matrix for content selection has been created. This comparative approach is also reflected in the data-management and information architecture of the portal. The existing metadata in the archive need not only needed to be syntactically aligned, but also semantically enriched in order to enable the understanding and analysis of the material selected. Several Video Active specific fields were added to the Dublin Core element set [14], including Television Genre, European dimension and National relevance.

A final major factor influencing the content selection are the intellectual property rights (IPR) related to the programmes. In almost all cases, individual rights owners need to be contacted before material can be published online and agreements need to be set up. Material can not be made available until agreements have been set with all relevant parties involved. The project does not have the financial means to finance rights clearances, so needless to say, not all content that was selected in the first instance will find its way to the portal. Every country has different IPR regulations. In some cases for example, it's not allowed to store the video files on a server abroad. The Video Active infrastructure therefore needed to facilitate a distributed solution for content storage; where the central portal links to dispersed servers.

Video Active Architecture

The Video Active system comprises of various modules, all using web technologies. The whole workflow from annotating, uploading material, transcoding material, keyframe extraction, metadata storage and searching is managed by these components. Figure 1 shows the architecture behind the portal.

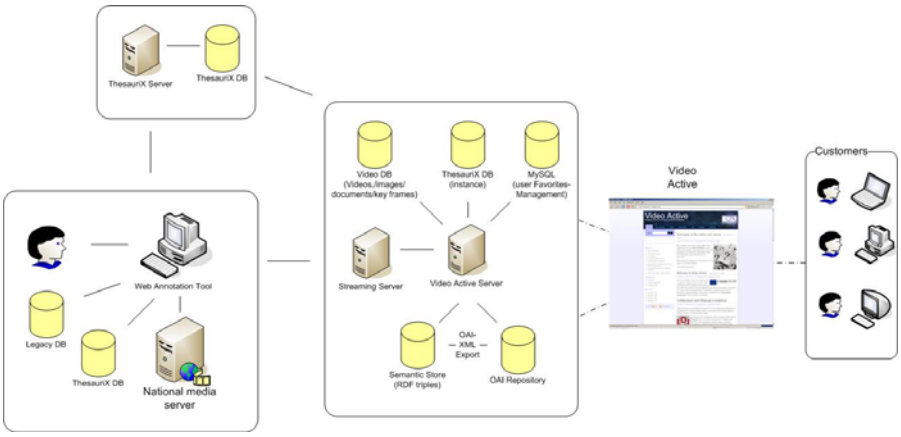


Figure 1: The Video Active Architecture

Video Active provides multilingual annotation, search and retrieval of the digital assets using the ThesauriX technology [15]. ThesauriX is a web-based multilingual thesauri tool based on the IPTC standard [16]. The system also exploits Semantic Web technologies enabling automation, intelligent query services (i.e. sophisticated query) and semantic interoperability with other heterogeneous digital archives. In particular, a semantic layer has been added through the representation of its metadata in Resource Description Framework (RDF) [17]. The expressive power of RDF enables light reasoning services (use of implicit knowledge through subsumption and equivalence relations), merging/aligning metadata from heterogeneous sources and sophisticated query facility based on SPARQL RDF query language [18]. Additionally, XML and Relational database technologies have been used to speed up some process where semantic information is not required. Finally, the Video Active metadata are public and ready to be harvested using the OAI-MPH technology [19].

In the Video Active system each archive has the ability to either insert the metadata manually using the web annotation tool or semi-automatically using a uniform (common for all the archives) XML schema. The Video Active metadata schema has been based on the Dublin Core [20] metadata schema with additional elements essential in capturing the cultural heritage aspect of the resources. The video metadata are produced automatically and are represented in a schema that is based in MPEG-7 [21]. In order to enable semantic services, the metadata are transformed in RDF triples and stored in a semantic metadata repository.

The asset management workflow

The annotation process is either manually or semi-automatically. In the manual process, the archives are using the Web Annotation Tool to insert the metadata. In the semi-automatic process, the archives export their metadata (the ones that have mappings to the Dublin Core elements) using a common XML schema. The elements that cannot be mapped to the Video Active schema (or are missing from the legacy databases, e.g. thesauri terms) are inserted manually (see Figure 2).

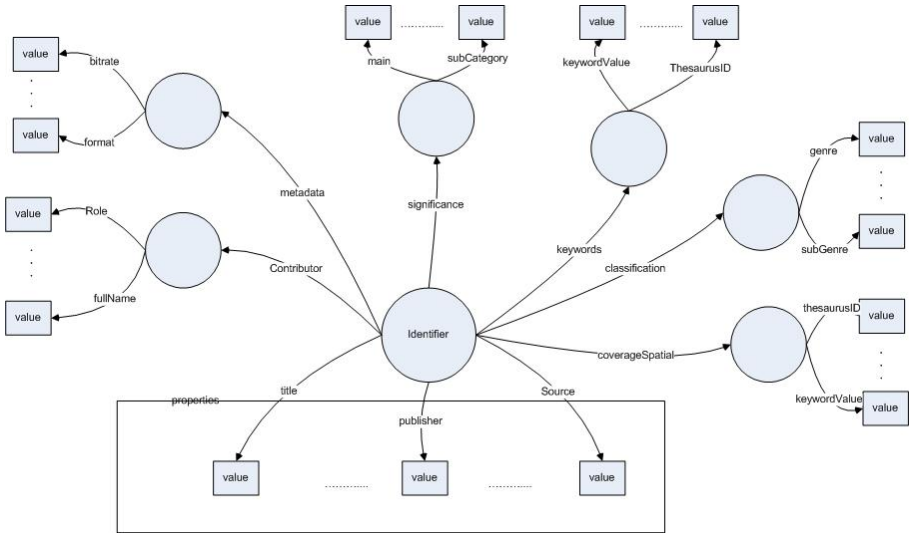


Figure 2: RDF representation of the final Video Active schema

The Web Annotation Tool allows also entering and managing the metadata associated with the media and also handles the preparation of the actual content, i.e. format conversion (low/medium bit rate for streaming service, etc.).

It produces an XML file that contains metadata, based on Dublin Core, as well as content encoding and key frame extraction information. The XML is then transformed in RDF triples (Figure 3) and stored in the semantic repository. The use of an ontology language, such as RDF that has formal semantics enables rich representation and reasoning services that facilitates sophisticated query, automation of processes and semantic interoperability. Semantic interoperability enables common automatic interpretation of the meaning of the exchanged information, i.e. the ability to automatically process the information in a machine-understandable manner. The first step of achieving a certain level of common understanding is a representation language that exchanges the formal semantics of the information. Then, systems that understand these semantics (reasoning tools, ontology querying engines etc) can process the information and provide web services like searching, retrieval etc.

Semantic Web technologies provide the user with a formal framework for the representation and processing of different levels of semantics.

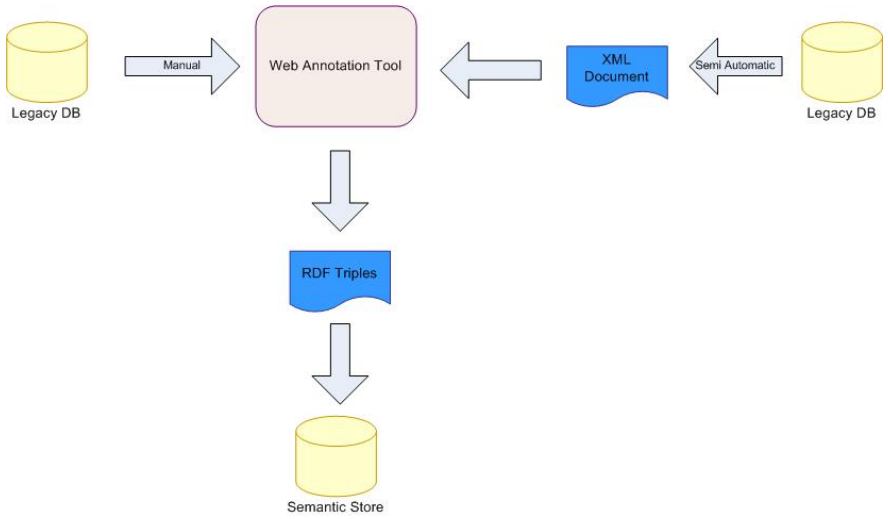


Figure 3: Architecture of Sesame data insertion system

Encoding of the material is done by the archives. Ingest format (notably MPEG 1-2) are transcoded to Flash and Windows Media streaming formats by the so-called Transcoding Factory. The Transcoding Factory is integral part of the so-called Contribution Application; the heart of the asset management of Video Active.

Storing and Querying

The semantic metadata store that is used in Video Active is Sesame [22]. Sesame is an open source Java framework for storing, querying and reasoning with RDF. It can be used as a database for RDF triples, or as a Java library for applications that need to work with RDF internally. It allows storing RDF triples in several storage systems (e.g. Sesame local repository, MySQL database). The procedure for the insertion of the assets into the RDF Store (Sesame) is depicted in Figure 3.

In order to transform the XML documents into RDF triples, Video Actives uses the Jena Semantic Web Framework [23]. Jena is a JAVA API for building semantic web applications. It provides a programmatic environment for RDF, RDFS [10] and OWL [25], and. In this application, Jena mainly for generating the RDF documents from the XML data representation.

The query service of Video Active system has been based on the SPARQL RDF query technology. SPARQL is a W3C Candidate Recommendation towards a standard query language for the Semantic Web. Its focus is on querying RDF triples and has been successfully used to query the Video Active metadata.

The end user has the ability to perform simple Google type searches but also allows browsing through the metadata using predefined filters, a way best compared with the Apple iTunes interface.

Metadata OAI Repository

All the metadata stored in Sesame, with the help of an OAI compliant repository are exposed to external systems/archives. The OAI-Protocol for Metadata Harvesting (OAI-PMH) [19] defines a mechanism for harvesting records containing metadata from repositories. The OAI-PMH gives a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). The metadata that is harvested may be in any format that is agreed by a community (or by any discrete set of data and service providers), although unqualified Dublin Core is specified to provide a basic level of interoperability.

Conclusion: towards a European Digital Library

The European Commission's i2010 Digital Libraries initiative advocates the need for integrated access to the digital items and collections held in Europe's cultural

heritage institutions via a single online access point; The European Digital Library (EDL).

Practical steps towards this goal are currently undertaken in many big and small projects. The EC recently launched a coordinative action to align these efforts, EDLnet. [28] Video Active is an invited member of the 'European Digital Library Thematic Partner Network' within EDLnet. This network aims to bring on board key cultural heritage stakeholders from European countries to prepare the ground for the development of an operational service for the European Digital Library, to be operational in 2008. Through Video Active's participation in EDLnet, semantic interoperability is facilitated between the two systems. Video Active has used all the late advances on Semantic Web technologies in order to exploit the full potential in representing, reasoning, querying and studding the Video Active metadata and content.

As this article has indicated, simply digitising and uploading archive content doesn't release the full potential of audiovisual content. The added value of archives lies in their ability to place material in a context meaningful to different user groups and by enriching the metadata to allow interactive exploration. For a pan-European service, the infrastructure should meet very specific requirements, dealing with semantic and multilingual interoperability, handing of intellectual property rights and so on. As more archives join Video Active, a vital part of our heritage will become available online for everybody to study and enjoy.

References

[1]. Video Active http://www.videoactive.eu

[2]. YouTube http://www.youtube.com

[3]. MySpace http://www.myspace.com

[4]. Revver http://one.revver.com/revver

[5]. YouTube serves up 100 million videos a day online http://www.usatoday.com/tech/news/2006-07-16-youtube-views_x.htm?

[6]. Transcoding Internet and Mobile Video: Solutions for the Long Tail, IDC, September 2007

[7].http://www.pewinternet.org/PPF/r/232/report_display.aspJoost http://www.joost.com/

[8]. Annual Report on Preservation Issues for European Audiovisual Collections (2007)http://www.prestospace.org/project/deliverables/D22-8.pdf

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

[10]. Miro http://www.getmiro.com/

[11]. Blinkx http://www.blinkx.com/

[12]. http://www.europeandigitallibrary.eu/edlnet/

[13]. Dublin Core http://dublincore.org/

[14]. Content selection strategy report http://videoactive.files.wordpress.com/2007/10/23_content_selection_strategy_report.pdf

[15]. Multilingual Thesaurus, http://www.birth-of-tv.org/birth/thesaurix.do?term=4620

[16]. International Press Telecommunications Council, http://www.iptc.org/pages/index.php

[17]. Resource Description Framework (RDF) http://www.w3.org/RDF/.

[18]. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query

[19]. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), http:// www.openarchives.org/OAI/2.0/openarchivesprotocol.htm#Introduction.

[20]. Dublin Core, Dublin Core Metadata Initiative, http://dublincore.org

[21]. MPEG-7 Standard, ISO/IEC 15938-1, Multimedia Content Description Interface, http://mpeg.telecomitalialab.com

[22]. Broekstra, J., Kampman, A., Harmelen, F. (2002). "Sesame: A Generis Architecture for Storing and Querying RDF and RDF Schema". 1st Internacional Semantic Web Conference (ISWC2002).

[23]. Jena – A Semantic Web Framework for Java, http://jena.sourceforge.net/

[24]. Dan Brickley, RDF Vocabulary Description Language 1.0: RDF Schema, http://www.w3.org/TR/rdf-schema/

[25]. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & eds., L. A. S. (2004). OWL web ontology language reference.

[26]. Extensible Markup Language (XML), http://www.w3.org/XML/

[27]. Andy Seaborne, HP Labs Bristol, RDQL - A Query Language for RDF, http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/:

[28]. EDLnet http://digitallibrary.eu/edlnet/

Author Details
Johan Oomen
Policy advisor
Netherlands Institure for Sound and Vision

Email: joomen@beeldengeluid.nl
Web site: http://www.beeldengeluid.nl

Vassilis Tzouvaras
Senior Researcher
National Technical University of Athens

Email: tzouvaras@image.ntua.gr
Web site: http://www.image.ntua.gr/~tzouvaras

Keywords: please list in order of importance the keywords associated with the content of your article (usually up to 8). These will not be made visible in the text of your article but may be added to the embedded metadata:

1. Semantic Web
2. European Digital Library
3. Metadata harvesting
4. Interoperability
5. Audiovisual Archives
6. Streaming media
7. Asset Management
8. Open Archives Initiative

Israel Interest in Semantic Interoperability in the European Digital Library

Dov Winer[1,]
[1] Israel National Library, E. Safra Campus, Givat Ram, The Hebrew University of Jerusalem, Israel, 91904
MAKASH, Applying CMC in Education, Culture and Science, POB 151 Zur Hadasah. Doar Haela , Israel, 99875

dovw@savion.huji.ac.il

Abstract. A brief description of the present interests of the author in Semantic Interoperability is provided with reference to the present MOSAICA, ATHENA and EDLnet projects. Further background information is provided including his role in disseminating the Semantic Web vision in Israel and past projects.

Keywords: Israel Jewish semantic web interoperability digital library digitisation cultural heritage online education metadata

1 Present projects

1.1 MOSAICA

MOSAICA http://www.mosaica-project.eu is developing a toolbox of technologies for intelligent presentation, knowledge-based discovery, and interactive and creative educational experience covering a broad variety of diversified cultural heritages requirements. Jewish heritage have been selected as an environment to test and demonstrate the effectiveness of the approach.
The project draws on two cutting-edge technologies: (1) the Semantic Web together with ontology engineering will be used to integrate semantically-based cultural objects of varying complexity, (2) while distributed content management will facilitate seamless aggregation of highly distributed, diversified content - in order to design and develop a technologically highly advanced solution:
It includes a Web-based portal[1] featuring multifaceted interfaces for knowledge-based exploration and discovery, and a Conceptualisation platform[2] consisting of online

[1] Definition of the MOSAICA Web-based portal http://www.mosaica-project.eu/index.php?id=29

utilities empowering users to collaboratively author and manage cultural resources in the globally distributed environment.
In technology terms, MOSAICA is expected to develop innovation in 4 areas: (1) Semantic annotation (2) Ontology alignment (3) Distributed content management (4) Geographic Information System / Temporal Information System
The author is the coordinator of WP1 - Content selection, aggregation and access – which include tasks like Resources identification and selection; access to MOSAICA resources and related IPR issues; Semantic annotation of selected resources; and definition of interfaces to multiple distributed resources.

1.2 ATHENA

ATHENA is a new project in the framework of eContentPlus presently being negotiated with the Commission. It will tackle content provision to the European Digital Library (EUROPEANA). Among its main goals:
(1) Support for the participation of museums and other institutions from sectors of cultural heritage still not fully involved in Europeana; (2) Coordinate standards and activities of museums across Europe; (3) Identify digital content present in European museums; (4) Develop a set of interoperability tools to be integrated within Europeana to facilitate access to digital contents belonging to European museums

Israel will participate in ATHENA through the MAKASH Institute. The content partners from Israel include the Israel State Archive; the Israel National Library and the Israel Museum in Jerusalem.

1.3 EDLnet

The Israel National Library – http://jnul.huji.ac.il is completing procedures to become an associate partner to EDLnet thematic network and take part of its activities - - http://www.europeandigitallibrary.eu/edlnet/ . This thematic network will build consensus to create the European Digital Library. It will find solutions to the interoperability of the cultural content held by European museums, archives, audio-visual archives and libraries in the context of The European Digital Library. It plans to be active in the work groupss of Work Package 2 which deals with standards, language and technical interoperability.

[2] Definition of the conceptualisation platform: http://www.mosaica-project.eu/index.php?id=30

2 Past activities

2.1 Promoting the Semantic Web Vision

Following the First International Semantic Web Conference in 2002 (http://annotation.semanticweb.org/iswc/documents.html) he introduced and sought to expand interest in the Semantic Web vision in Israel. He organized in 2002 the First Semantic Web Seminar in Israel with the participation of Roberto Cencioni. (see: http://www.jafi.org.il/press/2002/nov/nov3.htm) .
Additional dissemination activities included the seminar on Jewish Networking towards the Semantic Web Age at the 2002 convention o f the Israel Association for Information Technology http://www.jafi.org.il/ph/sweb.htm . As founder of the Israel Internet Society he promoted the establishment of the W3C Office in Israel. It organized an event introducing the Semantic Web (see: http://www.w3c.org.il/events/semWebEvent/semWebEvent.html). In 2004, in collaboration with the CRI Institute for Interdisciplinary Applications of Computer Science he organized a Semantic Web Workshop in Jerusalem with the participation of Jim Hendler (see:
http://www.cri.haifa.ac.il/events/2004/semantic_web2004/semantic2004.htm

2.2 Short Bio

Dov Winer is a psychologist (Hebrew University of Jerusalem) specialized in Online Education and Training (University of London). He is founder of the Israel Internet Association (http://www.isoc.org.il). Dov Winer coordinated the dissemination of Internet use in Israel industry for the Ministry of Trade and Industry (1994/6).
He defined the program for the National Teacher's Colleges Network and trained its development team for Virtual Learning Environments (1996/8)
He established in 1989 the MAKASH Institute for ICT Applications in Education, Culture and Science. MAKASH is the NCP in Israel for the EUN – European Schoolnet, the consortium of the Europeans Ministry of Education for ICT, and participates in its Steering Committee. MAKASH was part of several EC programs: ETB – European Treasury Browser establishment of a comprehensive digital library of educational resources in Europe; VALNET (Schools of the Future); CELEBRATE for collaborative and interoperable Learning Objects in integrated learning environments. It is now collaborating in the integration of Israel educational resources repositories in the Learning Resources Exchange.
He proposed the establishment of the Global Jewish Information Network in 1988 and later planned it (1992) following support from the Parliament and the Ministry of Communication. From 2001 to 2006 he managed the Initiative for Developing Jewish Networking Infrastructures. This later project established a comprehensive database of metadata for Jewish resources in the Web (expressed in RDF) and developed the Thesaurus for its annotation – see: http://www.ejewish.info.

He coordinates the MINERVA network in Israel for digitisation of cultural heritage - http://www.minervaisrael.org.il, part of the broader European cluster of projects http://www.minervaeurope.org for coordination of digitisation in Europe. In this framework he has been active through its expert's working groups and in promoting initiatives like the establishment of the MICHAEL Israel national node to be affiliated to the MICHAEL network - http://www.michael-culture.org - one of components of the envisaged European Digital Library.
He is co-chair of the annual event EVA/MINERVA Jerusalem International Conference on Digitisation of Cultural Heritage - see: http://www.minervaisrael.org.il/program07.html.

SIEDL 2008 – Full Paper

Regine Stein, Carlos Saro, Axel Vitzthum, Monika Hagedorn-Saupe

Title: „museumdat and museumvok – Semantic Interoperability in the German Museum Community"

Keywords: Metadata Standards, CIDOC-CRM application, SKOS application
---

Initiatives on European, national and regional level to make cultural heritage information easily accessible through common portals and repositories have encouraged within the German Museum Association the development of two instruments which make life easier in very practical terms for museums as data providers, for service providers, and for software suppliers: "museumdat" as CIDOC-CRM compliant format for the publication of museum object core data allows for integrating highly expressive metadata of the most varying provenances. "museumvok" as SKOS-based format for the description of controlled vocabularies is used for publishing freely accessible vocabularies on www.museumsvokabular.de but most notably for the delivery of results by the SOAP Web service museumvok-ws.

museumdat: CIDOC-CRM compliant format for the publication of museum object core data

The format "museumdat" describes an XML schema which allows for integrating museum object data of the most varying provenances. Its point of departure was the metadata standard "CDWA Lite" – developed by the J. Paul Getty Trust - , but which is mainly focusing on data of the fine arts and architecture. To be able to properly process also data for other object classes such as cultural and natural history, history of technology – both for retrieval and presentation purposes - CDWA Lite was analyzed using CIDOC CRM, reconfigured towards an event-oriented structure and generalized. The resulting harvesting format "museumdat" now applies for all kinds of object classes and is compatible with CIDOC-CRM. Since its publication in October 2007 it has been widely accepted as metadata standard for the delivery of object data and is properly being implemented in several portals, e.g. the BAM-Portal for Libraries, Archives and Museums or the Museum Network DigiCult Schleswig-Holstein, by most software suppliers, and by some projects on international level. In particular, a working group established by the CDWA Lite Advisory Committee is currently working on a single, CIDOC-CRM compliant schema that harmonizes CDWA Lite and museumdat. For format specification and XML Schema definition see www.museumdat.org.

museumvok: Format for the description of controlled vocabularies

Research provides much better results when controlled vocabularies which are used in the museum object documentation are integrated. Vocabularies whose use is not restricted through copyright fees etc. are at present in German museums often applied and extended in a non-coordinate manner. Since 2006, a number of the more widely known of these have begun to be available on the online platform www.museumsvokabular.de under a Creative Commons Licence, and collaborative management tools that are based on Open Source Software are used for further development of the vocabularies. "museumvok" as the description format for these vocabularies is mainly based on SKOS and uses the SKOS Core and the SKOS Mapping vocabulary. In coordination with museum software suppliers in Germany a SOAP Web service "museumvok-ws" for controlled vocabulary is being built by which the latter can be used for many different applications. The consuming part with 3 main functions – searchConceptsByID, searchConceptsByTerm and fetchHierarchy – is already operational. Format description, interface definition and WSDL are available at www.museumsvokabular.de.

With these instruments, "museumdat" and "museumvok", based on international standards, developed by the Documentation Committee of the German Museum Association (DMB) together with representatives of museums, networks and software suppliers, museums have at hand 2 important keys which ease the integration of object data into portals and which extend the search opportunities towards semantic web applications – a contribution for better interlinking of perhaps not only German museums?

Author Index