Ontology-Driven Interoperability for Cultural Heritage Objects

WORKING NOTES

DELOS – MultiMatch Workshop
Tirrenia, Italy, 15 February 2007

Editors:
Vittore Casarosa ISTI-CNR
Carol Peters ISTI-CNR

DELOS: a Network of Excellence on Digital Libraries
www.delos.info

MultiMatch: Multilingual / Multimedia Access To Cultural Heritage
www.multimatch.eu

Information Society Technologies

Table of Contents

Introduction

This workshop on ontology driven interoperability for cultural heritage digital objects is a joint DELOS – MultiMatch event. Both DELOS and MultiMatch are supported by the unit for Content, Learning and Cultural Heritage (Digicult) under the Sixth Framework Programme. DELOS is a Network of Excellence aimed at coordinating a joint programme of research activities in digital library related areas[1]. MultiMatch, a 30 month specific targeted research project, is developing a multilingual search engine designed specifically for the access, organisation and personalised presentation of cultural heritage (CH) information[2].

The goal of MultiMatch is to build a system that will enable users to explore and interact with online CH content, across media types and language boundaries. The MultiMatch search engine will provide specialised search facilities: it will be updated via regular focused crawling of domain-specific web sites and will automatically classify the information acquired in a semantic-web compliant fashion, based on document content, metadata, context, and the occurrence of relevant CH concepts. Search results will be organised and displayed in an integrated manner, according to the user needs or profile. The system will be designed to support diverse user classes but also to assist CH institutions to disseminate their content widely and raise their visibility. Clearly, being able to handle, interpret, match and/or merge the diverse metadata schemas and knowledge representation formats currently found in the CH field is a crucial issue for MultiMatch.

The original idea for organising this workshop thus resulted from early discussions within MultiMatch and more specifically in the group working on the definition of the most appropriate metadata schema and overall conceptual framework for the diverse types of information that would be handled by the project. The intention is to define the basis for maximum interoperability. The first activity of this group was to study the most relevant standards for each CH sub-domain, together with the most relevant generic schemas (Dublin Core, MPEG-7, MPEG-21) and reference models (FRBR and CIDOC-CRM). The results of this study are publicly available on the MultiMatch website[3]. Following this initial analysis of the state-of-the-art, the MultiMatch metadata group is now in the process of making the final decisions with respect to the metadata schemas, knowledge representation formalisms and/or ontologies, etc, that will best serve the diverse needs of the project. However, it was felt that before making any definite commitment it would be extremely useful to be able to discuss and compare ideas with experts in the field and with others who have already faced or are currently addressing the same issues.

DELOS has long been concerned with questions concerning semantic interoperability and has published a comprehensive report on Semantic Interoperability in Digital Library Systems[4]. The report (written as part of the activities of the cluster on “Knowledge Extraction and Semantic Interoperability”) defines interoperability very broadly as enabling any form of inter-system communication, or the ability of a system to make use of data from a previously unforeseen source. This report was one of the sources of information for MultiMatch. It thus seemed obvious to arrange a meeting between people in both DELOS and MultiMatch interested in questions concerning semantic interoperability for a useful exchange of opinions. However, as we began to organise this meeting we became increasingly aware of the fact that there is currently a lot of interest in this area. In fact, last year, BRICKS and EPOCH organised a workshop on “Semantic Interoperability”, held in connection with the VAST 2006 Conference[5], and the European Commission has just set up working group on Digital Library Interoperability aimed at identifying the major obstacles hindering interoperability.

We thus decided to organise a one-day workshop on this topic and the DELOS Conference provided us with the perfect venue. The purpose of the workshop is to bring together experts in the field (both theoreticians and practitioners) to investigate the current state-of-the-art, identify common problems and issues, and hopefully formulate recommendations aimed at encouraging (i) adoption of standards, (ii) interoperability.

[1] See www.delos.info

[2] See www.multimatch.eu

[3] MultiMatch – D2.1.1: First Analysis of Metadata in the Cultural Heritage Domain

[4] DELOS - D5.3.1: Semantic Interoperability in Digital Library Systems, 29th June 2005.

[5] See http://public-repository.epoch-net.org/workshops/semantic_interop.pdf

Our keynote speakers, Martin Doerr, Maja Žumer and Chrisa Tsinaraki, will present three of the best known existing conceptual frameworks (CIDOC-CRM, FRBR and MPEG-7) and some of the relationships between them. These talks will be followed by a panel discussion, moderated by Stavros Christodoulakis, aimed at investigating how these frameworks can be made interoperable.

After lunch, we will have a series of position statements by a number of projects and institutions working in the CH domain. The EDLProject, TEL, MICHAEL, BRICKS, IMAGINATION, EPOCH plus the Dutch Cultural heritage Institution will briefly present the problems they are currently facing in this area and/or the solutions they are adopting. The last two speakers will give the perspective of the Text Encoding Initiative and present the recently formed EC Interoperability Group. The final session of the workshop will be dedicated to a discussion of the main issues that have emerged during the day. All participants will be expected to contribute and present their opinions.

We should like to end this brief introduction by expressing our gratitude to our programme committee for their assistance in the organisation and the identification of themes for discussion, and to our speakers for agreeing to present their ongoing work. We conclude by thanking all the participants in advance for their valuable contributions to the discussions and wishing everyone an enjoyable workshop.

Carol Peters and Vittore Casarosa

Programme Committee
DELOS**:** Martin Doerr, *FORTH, Greece*; Carlo Meghini, *ISTI-CNR*, Italy; Stavros Christodoulakis, *Technical University of Crete, Greece*.
MultiMatch: Giuseppe Amato, *ISTI-CNR, Italy;* Neil Ireson, *University of Sheffield, UK*; Johan Oomen, *Sound and Vision, The Netherlands*.
External Expert**:** Stefan Gradmann, *University of Hamburg, Germany*.

Organization
DELOS: Vittore Casarosa, *ISTI-CNR, Italy*
MultiMatch: Carol Peters & Alessandro Nardi, *ISTI-CNR, Italy*

The CIDOC CRM and an Integrated Approach to Semantic Interoperability

Martin Doerr

Decades of research have been devoted to the goal of creating systems which integrate information into a global knowledge network, yet we still face problems of cross-repository interoperability, lack of public infrastructure, and a coherent research agenda - both theoretical and practical – to face these challenges. Interest in the Semantic Web has revived the dream, but many are sceptical. The presentation will address semantic problems and requirements to integrate digital information into large scale, meaningful networks of knowledge that support not only access to source documents but also information use and reuse. We present a new approach based on (i) interdisciplinary research of scholarly and scientific discourse, (ii) a generic global ontological model based on relations and co-reference rather than objects, (iii) semi-automatic maintenance of co-reference links, and (iv) public engagement in the creation and development of the network.

We regard Digital Libraries or better *Digital Memories* as : information systems preserving and providing access to source material, scientific and scholarly information, such as libraries of publications, experimental data collections, scholarly and scientific encyclopedic or thematic databases or knowledge bases. There is still a widely held and traditional view of the task of libraries as institutions limited to the collection and preservation of documents and to providing assistance in finding specific items of literature or information. In this view, the library's role is completed when the (one, best) document is handed out: 'All you want is in this document.'

This view has not helped much in raising the level of new functionality that semantic interoperability of resources would permit. There is little or no support for the searches to produce new and informed responses from aggregated sources or to retrieve them by contexts (e.g. "Which excavation drawings show the finding of this object?"). There is little or no support to allow integration of complementary information in multiple sources into new insight (e.g., "What is known about the people who participated in this excavation"). Finally there is typically no support for cross-disciplinary search (e.g. to find relevant related information from the many disciplines that contribute to archaeological knowledge, such as ecology, ethnology, biodiversity, etc.). Central to our approach is a suitable knowledge management. We distinguish:

1. Core ontological relationships for "schema semantics", such as: "part-of","located at","used for", "made from" which are localized atomic relationships, but and rich in potential structural information, relating to content.
2. Categorical data": taxonomies used for reference to and agreement on sets of things, rather than as means of reasoning, such as: "basket ball shoe", "whiskey tumbler", "burmese cat", "terramycine". These terms define and order concepts rather than providing structural information. They aggregate categories as opposed to integrating sources. The leaves of the taxonomic structure would be entries in a thesaurus.
3. Factual background knowledge for reference and agreement as objects of discourse, such as particular persons, places, material and immaterial objects, events, periods, names. These would be elements of the taxonomic classes.

Global core ontologies play a fundamentally different role to that played by specialist terminologies in practical knowledge management. They are small and can be manually created. They support schema mediation, data transformation and migration.

The CIDOC Conceptual Reference Model is presented as an example of such a global model. It is a core ontology and new ISO standard (ISO 21127, accepted Sept. 2006) designed for the semantic integration of information from museums, libraries, and archives. It has been developed by CIDOC, the International Committee for Documentation of the International Council of Museums (ICOM), and an international multidisciplinary team of experts. The CIDOC CRM concentrates on the definition of relationships, rather than terminology, in order to support mediation, transformation and integration between heterogeneous database schemata and metadata structures. It is a product of re-engineering the dominant common meanings from the most characteristic schema elements in use in these institutions. It is not prescriptive, but provides a controlled language to describe common high-level semantics that allow for information integration at the schema level. This integration has been demonstrated in a large range of different domains including cultural heritage, e-science and biodiversity.

The CRM foresees domain-specific extensions, such as the integration of the conceptual model contained in the Functional Requirements for Bibliographic Records (FRBR), developed by the International Federation of Library Associations (IFLA), with the CIDOC CRM.

Whereas the second level, the "categorical" data, have been extensively treated by information science and the Semantic Web likewise, the third level of factual knowledge, which is orders of magnitude larger, is widely overseen as topic of semantic interoperability. There is a growing awareness of the need for information systems which provide reasoning capability, but before any "reasoning" can be done over integrated knowledge resources, the data must be connected in a "global network of knowledge." This requires:

- A sufficiently generic global model as presented above
- Methods of knowledge extraction / data transformation to populate the network.
- Massive, distributed, semiautomatic detection of co-reference relations (data cleaning) across contexts.
- Referential integrity of co-referencing needs to be curated in order to create, maintain and improve the consistency of global networks of knowledge as a continuous process.

Further research is needed on co-referencing to make global information integration a reality. Advocating a global model does not mean using a common schema. The global model needs only to be used only on a virtual level in integrated information management. A common schema is counter productive and hinders evolution. A global core ontology is a question on agreeing on a common understanding of basic concepts. We argue that metadata for digital memories are based on a similar enough discourse for most domains. This is demonstrated in a few examples, extensions to the presented ontology not withstanding.

FRBR: Is this the beginning of a beautiful friendship?

Maja Žumer
University of Ljubljana
Slovenia

1. Background

Functional Requirement for Bibliographic Records (FRBR) is the conceptual model of the bibliographic universe developed by IFLA (International Federation of Library Associations and Institutions). It was approved in 1997 and published in 1998. Interestingly enough, we still mostly refer to is as 'the new library model'. This can be explained by the fact that there are not many real-life implementations of FRBR. The model is often seen as (only) an interesting intellectual exercise, but the cataloguing practice stays the same. Current cataloguing rules are still based on Paris principles (from 1961) and our catalogues are in fact only an electronic replica of a card catalogue.
But it is encouraging to see some implementations of FRBR, for example FictionFinder of OCLC, developments of VTLS, etc. The problem of legacy data, millions of existing bibliographic records, is addressed by the so-called FRBRisation algorithms.
FRBR, which is a general framework, is complemented by two additional segments: FRAR (Functional Requirements for Authority Records) and FRSAR (Functional Requirements for Subject Authority Records). FRAR is ready for final review, while FRSAR is still under development. When finished, they will offer a comprehensive conceptual model, covering all aspects, important for library users.
The IFLA FRBR Review group, established in 2004 is taking a proactive role in development and use of FRBR. Newly formed working groups are focusing on particular problems. Two have to be mentioned in particular: WG on expression entity and WG on aggregates. The first has already provided a pragmatic interpretation of the expression entity and the latter will deal with anthologies, series, augmentations, journals (all composites of individually created dependent/independent works), which have not been treated in detail in the original model. The Review Group has already decided to review the attributes of all entities.

2. Interoperability within the library community

This seems to be the crucial issue for ‘survival’ of FRBR: unless the problem of legacy data is solved, the model will not be accepted by the library community.
Theoretically, assuming that we have complete and consistent MARC records, it is possible to extract enough information to identify FRBR entities (work, expression, manifestation). Several FRBRisation experiments have confirmed that, but also the fact that the records we normally have are neither complete nor consistent. In addition, quite a lot of important information is recorded as unstructured text, mostly as notes, and not appropriate for computer processing.
Most librarians are also not encouraged by statistics reported by Hickey and O’Neill: a large percentage of works have only one expression and manifestation and only a relatively small number have more than one expression. Superficially, these results bring into question the economy of FRBR. What is usually overlooked is the fact that the latter works are central for the users…
What is the solution? First, librarians have to accept a less-than-perfect result of FRBRisation and allow for subsequent corrections. It might be even worth to try user annotation and social tagging. Second: cataloguing rules according to FRBR have to be developedi. RDA (Resorce Description and Access) seems to be going in that direction and there are reports of new Italian cataloguing rules. Third: we need to develop a data model and interface prototypes. As long as we only have a theoretical discussions there will be no real breakthrough.

3. Interoperability with other communities

In the study there is a claim that all types of materials are covered by the model. Most critics point out that the focus is really on traditional, mostly textual publications. The interoperability with other communities has not really been the focus of discussions.
But there is no doubt that there is potential, particularly with other cultural heritage institutions: museums and archives.

An important development is the work of FRBR/CRM Harmonisation Group, a joint effort of IFLA and ICOM CIDOC (International Council of Museums – International Committee for Documentation). Libraries and museums share users and types of materials, it is therefore important that a common view of cultural heritage information is developed. The goal is to bring together (harmonise) the library model (FRBR) and museum model (CRM: Conceptual Reference Model). While preparing an object-oriented version of FRBR, additional goals are to check FRBR's internal consistency, enable interoperability and integration, to extend the scope of both conceptual models, and open the way to future applications. The first complete draft of the object-oriented FRBR has been published for public comment as "FRBRoo". The harmonised model will be further developed and refined.
Another area of possible (and so far overlooked) cooperation outside the library community is intellectual rights management. The FRBR entities can be linked to intellectual rights and if also take into account the wealth of name authority files the benefit is obvious.

This may be the turning point for FRBR implementation. Several parallel current developments seem to be very favourable to FRBR: development of new cataloguing rules (RDA) and, at the same time, International Meetings of Experts for an International Cataloguing Code (IME-ICC) under the auspices of IFLA, in charge of the definition of new International Cataloguing Principles to replace the "Paris Principles". New FRBRisation tools are being developed and tested and we may expect more and more prototypes of new catalogues.
After a relatively slow start FRBR has recently gained some momentum. To foster further development we have to emphasise the model's biggest potential: access to distributed bibliographic information in union catalogues and portals such as The European Library. For such portals FRBR offers meaningful clustering of search results and navigation. The same approach could then be applied to access to all cultural information.

Ontology-Driven Interoperability for MPEG-7

Chrisa Tsinaraki
TUC/MUSIC, Technical University of Crete Campus, 73100 Kounoupidiana, Crete, Greece
chrisa @ced.tuc.gr

Abstract

Research efforts for interoperability support in the multimedia domain are presented here. Interoperability, both at the syntactic and the semantic level, is necessary in the open multimedia consumption environment formed in the Internet today, so that the multimedia content services offered by different vendors may interoperate. Syntactic interoperability support is achieved in the multimedia domain through the adoption of the dominant MPEG-7 standard for multimedia content and service description. Domain knowledge is then integrated, in the form of domain ontologies, in the MPEG-7 constructs, in order to achieve semantic interoperability. Finally, the utilization of the interoperability support in real world applications is discussed.

1 Introduction

Multimedia content services are becoming increasingly popular in the open multimedia consumption environment formed in the Internet today due to the advanced network infrastructures that allow for the fast and efficient multimedia content delivery and the availability of cheap consumer electronic devices that allow the consumption and management of multimedia content. Interoperability, both at the syntactic and the semantic level, is necessary in this open multimedia consumption environment, so that the multimedia content services offered by different vendors may interoperate.

The syntactic interoperability is usually achieved through the adoption of standards, while the semantic interoperability is achieved through the integration of domain knowledge that is expressed in domain ontologies.

The dominant standard in multimedia content and service description is the MPEG-7 (Chang, Sikora and Puri 2001). The adoption of the MPEG-7 in the multimedia domain guarantees syntactic interoperability and the integration of domain knowledge in the MPEG-7 constructs achieves semantic interoperability.

In this paper the ontology-driven interoperability support for MPEG-7 is discussed. The rest of the paper is structured as follows: Domain knowledge representation using pure MPEG-7 constructs is presented in section 2, OWL ontology driven interoperability for MPEG-7 is discussed in section 3, interoperable multimedia application support is described in section 4 and the paper concludes in section 5.

2 Domain Knowledge Representation using pure MPEG-7 Constructs

We present here how domain knowledge, in the form of domain ontologies, can be expressed using MPEG-7 constructs and can then be integrated in the MPEG-7 semantic descriptions. This is achieved through the methodology described in (Tsinaraki et al 2005).

According to this methodology, the domain ontology classes are represented as abstract semantic entities and the domain ontology individuals are represented as concrete semantic entities. The *AbstractionLevel* element of the *SemanticBaseType* (which represents the semantic entities) specifies if a semantic entity is abstract or concrete. If the *Dimension* attribute of *AbstractionLevel* has the value 0, the semantic entity is concrete and if it has a non-zero value, the semantic entity is abstract.

An abstract semantic entity that represents a domain-specific class is related with each of the semantic entities representing its subclasses through: *(a)* A relationship of type *generalizes*, which has as source the semantic entity that represents the class and as target the semantic entity that represents the subclass; and *(b)* A relationship of type *specializes*, which has as source the semantic entity that represents the subclass and as target the semantic entity that represents the class. In addition, an abstract semantic entity that represents a class is related with the concrete semantic entities representing the class individuals through pairs of *exemplifies*/*exemplifiedBy* relationships.

The properties of the domain-specific classes are represented as *Property* elements (if they are of simple type) or as pairs of *property*/*propertyOf* relationships that associate semantic entities (if they are of complex type).

3 OWL Ontology Driven Interoperability for MPEG-7

It was shown in the previous section that the MPEG-7 allows for the representation of domain ontologies using pure MPEG-7 constructs. The domain ontologies are usually expressed in OWL (McGuinness and F. van Harmelen 2004) syntax, as OWL is the dominant standardization effort in ontology description. It is therefore very important for the multimedia community to have a methodology for the interoperability of OWL with MPEG-7 and for the integration of domain knowledge expressed in OWL within MPEG-7. This way, the Semantic Web tools (such as reasoners) and methodologies may be used with MPEG-7.

The first research effort in this direction was presented in (Hunter 2001), where the DAML+OIL (McGuinness, Fikes, Hendler and Stein. 2002) ontology definition language has been used to partially describe the MPEG-7 MDS and Visual metadata structures. The ontology has been recently translated in OWL. An important shortcoming of this ontology is the limited coverage of the MPEG-7 constructs.

An Upper OWL-DL ontology that fully captures the MPEG-7 MDS and the parts of the MPEG-7 Visual and Audio that are necessary for the complete representation of the MPEG-7 MDS has been presented in (Tsinaraki, Polydoros and Christodoulakis 2007). The ontology was manually developed, according to a methodology that allows the transformation of the XML Schema constructs of MPEG-7 in OWL-DL.

A methodology for the definition of OWL domain ontologies integrated in the MPEG-7 semantic model has been described in (Tsinaraki, Polydoros and Christodoulakis 2007). In this methodology, the domain-specific entities are represented as domain ontology classes. These classes are (direct or indirect) subclasses of the OWL classes that represent the subtypes of *SemanticBaseType* (*EventType*, *ObjectType*, *AgentObjectType*, *SemanticPlaceType*, *SemanticTimeType*, *SemanticStateType* and *ConceptType*) in the OWL Upper ontology defined in (Tsinaraki, Polydoros and Christodoulakis 2007).

Interoperation of the multimedia content descriptions with applications using pure MPEG-7 is achieved through a set of transformation rules (Tsinaraki, Polydoros and Christodoulakis 2007) that allow the transformation of domain ontologies and semantic content descriptions to valid MPEG-7 descriptions. They allow the transformation of domain ontologies defined according to the methodology described in section 2 into abstract MPEG-7 semantic descriptions as well as the transformation of OWL individuals that belong to the domain ontology classes into MPEG-7 semantic descriptions. The produced descriptions are valid MPEG-7 (parts of) documents.

During the metadata transformation from OWL to MPEG-7, the individuals representing MPEG-7 constructs are transformed into XML elements. The object properties are transformed into elements and the datatype properties are transformed into the constructs they represent in the original MPEG-7 schemas (attributes, elements or simple values). In order to produce valid MPEG-7 descriptions, information regarding the MPEG-7 XML element order, the default values and the original MPEG-7 representation of the datatype properties is needed. This information is kept in a transformation rule ontology and is utilized during both ontology and metadata transformations.

The generalization and the automation of the methodology for ontology-driven interoperability for MPEG-7 described in (Tsinaraki, Polydoros and Christodoulakis 2007) has led to the development of a generic methodology and its software implementation that allow the expression of the XML Schema (Fallside 2001) semantics in OWL-DL, as described in (Tsinaraki and Christodoulakis 2007).

4 Application Support

We show in this section how the semantic information integrated with MPEG-7 can be utilized in specific applications.

A challenging issue is the consistent description of multimedia content that depicts cultural heritage artifacts. It can be achieved through the alignment of the MPEG-7 semantics with the semantics of the CIDOC/CRM (ISO/IEC 2004), which is a model that subsumes the semantics of the different cultural heritage standards. The alignment of the CIDOC/CRM with MPEG-7 is under way using both the standards expressed using OWL syntax (Doussias 2007).

Important multimedia content services that are useful in several different domains and form the basis for complex semantic based services are the semantic multimedia content retrieval and filtering. These services rely on the existence of semantic knowledge integrated with the multimedia content descriptions and the capability of the users to express their preferences on the multimedia content semantics. In order to allow the expression of the user preferences regarding multimedia retrieval and filtering on every aspect of the MPEG-7 descriptions, the MP7QL query language and its compatible filtering and search preference model have been developed (Tsinaraki and Christodoulakis 2006).

5 Conclusions – Future Work

We have presented in this paper research efforts in ontology-driven interoperability support for MPEG-7 and we have outlined some important applications that will benefit from the infrastructures developed in this context.

References

Chang S.F., Sikora T. and Puri A. 2001. Overview of the MPEG-7 standard. In *IEEE Transactions on Circuits and Systems for Video Technology* 11:688–695.

Doussias A. 2007. *Alignment of MPEG-7 with the CIDOC/CRM*. Diploma Thesis, TUC/MUSIC, 2007.

Fallside D. 2001. *XML Schema Part 0: Primer*. W3C Recommendation, http://www.w3.org/TR/xmlschema-0/.

Hunter J. 2001. *Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology*. In Proc. of the International Semantic Web Working Symposium (SWWS), July 30 - August 1, 2001.

McGuinness D. L. and van Harmelen F. (eds.) 2004. OWL Web Ontology Language: Overview. W3C Recommendation, http://www.w3.org/TR/owl-features.

ISO/IEC, ISO/DIS 21127 -- Information and documentation -- A reference ontology for the interchange of cultural heritage information, 2004.

Tsinaraki C., Christodoulakis S. 2006. *A User Preference Model and a Query Language that allow Semantic Retrieval and Filtering of Multimedia Content*. In the proceedings of the Semantic Media Adaptation and Personalization Workshop (SMAP 2006), pp. 121-128, December 2006, Athens, Greece.

Tsinaraki C. and Christodoulakis S. 2007. XS2OWL: *A Formal Model and a System for enabling XML Schema Applications to interoperate with OWL-DL Domain Knowledge and Semantic Web Tools*. In Proc. of the DELOS Conference 2007.

Tsinaraki C., Polydoros P. and Christodoulakis S. 2007. *Interoperability support between MPEG-7/21 and OWL in DS-MIRF*. In Transactions on Knowledge and Data Engineering (TKDE), Special Issue on the Semantic Web Era, 2007.

Tsinaraki C., Polydoros P., Kazasis F. and Christodoulakis S. 2005. *Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content*. In Multimedia Tools and Application Journal (MTAP), Special Issue of on Video Segmentation for Semantic Annotation and Transcoding, 26:299-325.

EDLproject: challenges of multilingual access to multilingual European content

Maja Žumer
University of Ljubljana
Slovenia

1. Background

EDLproject is a Targeted Project funded by the European Commission under the eContentplus Programme, within the area of Cultural content and scientific/scholarly content. It is coordinated by the German National Library. EDLproject builds on the existing The European Library, a service funded by CENL, the Conference of European National Librarians, providing unified access to the electronic resources of the main European National Libraries as well as to other library services. The project is also a continuation of the TEL-ME-MOR project, which has supported The European Library with the inclusion in the service of the ten New Member States National Libraries. EDLproject integrates the bibliographic catalogues and digital collections of the National Libraries of Belgium, Greece, Iceland, Ireland, Liechtenstein, Luxembourg, Norway, Spain and Sweden, into The European Library: by the end of 2007 ALL EU countries will be members of the European Library service. EDLproject further enhances access to the European Library portal, by continuing to develop its multi-lingual capacity. EDLproject leverages the influence and resources of CENL as a key player and stakeholder in the content field to work towards a consensual resolution of certain issues raised by the Communication "i2010: Digital Libraries", such as potential availability of digital content from national libraries and the scope for collaboration between The European Library and other content providers funded by eContentplus.

EDLproject has a total budget of € 2.114 million EUR, of which € 1 million contribution by the eContentplus programme. The project started in September 2006 and will last for 18 months.

2. Multilinguality

The research in this area started in the TEL-ME-MOR project. From the end-user viewpoint there are several layers of multilinguality:

- Language of the interface
- Language of bibliographic records/cataloguing language/language of metadata
- Language of the resources

Ideally, each user should be able to:

- Use his/her own language when communicating with the system
- Use his/her own language to formulate a query and, as the result, retrieve all relevant (digital) objects in any language

The first part, the interface in all languages, has been solved. The portal interface has been translated into all partner languages, updates are being translated and appropriate tools are available. Therefore this only remains as an organisational issue.
The second part presents a big challenge. There are (too) many language combinations and an all-to-all mapping of free-text may not be feasible.
Therefore some smaller steps were investigated in TEL-ME-MOR, particularly mapping of subject access tools (subject headings, classification). The final recommendations are incorporated into the EDLproject workplan:

- Recommendations for improving subject access interoperability
  - Test a selection of cross-language approaches to subject data, including MACS, MSAC and CrissCross
  - Investigate the feasibility of loading into the MACS system the Luxembourg subject headings (Laval English / French) and the Spanish subject headings (Spanish – English), and if feasible, load the data
  - Provide an updated overview of recent European projects and initiatives that may have relevance to cross-language access to various types of collections.

        - Test cross-language searching in the European Library using the 70'000 RAMEAU/LCSH links against data from British Library and Bibliothèque nationale de France
        - Test large-scale linking (Sports and Theater – 1000 terms)
            - Load MSAC data in to the European Library portal
            - Incorporate more languages into the EL cross-language interface
        - Use the MACS Link Management system (replicated as required) central source (clearinghouse) of mapping results between subject headings used in European national libraries (classifications and subject headings)
- Testing automated mapping
    - Investigate HILT, TermSciences and WebDewey as tools to speed up link creation
    - [In TELplus, plans to test Stitch]
    - Contact OCLC to investigate their project exploring new web-based services for dynamic mapping between subject authority lists http://www.oclc.org/research/projects/termservices/
    - Contact the MULIR project:
        - "Our current Multi-Lingual Information Retrieval (MULIR) entry vocabulary index maps from English Library of Congress Subject Headings (LCSH) to words and phrases in over 100 languages and vice versa. This prototype was created from over ten million records of the University of California MELVYL online library catalog."
- Recommendations for improving interoperability across authorities
    - Extending the theoretical analysis of D3.5 of TEL-ME-MOR and building on projects such as LEAF, VIAF and ONESAC, design and test the feasibility of a name authority control tool, implemented by (automatically) clustering existing variant headings, thus enabling searching on any of possible names for a person, persona, corporate body or geographic name. Result: A prototype of common (name) authority tool.

In addition to tasks proposed as recommendations of TEL-ME-MOR, the plan in EDLproject is to investigate free-text searching. For that an overview of existing tools would be necessary: language recognition tools, stemmers, automated translation tools etc. At this point we are not aiming at a fully functional cross-language IR for all languages, but at least a prototype of a limited application. Therefore cooperation with any relevant research is needed.

Interoperability in The European Library: 'the devil is in the details'

Sjoerd Siebinga
The European Library Office
the Netherlands

1. Background: What is the European Library

The European Library (http://www.TheEuropeanLibrary.org) is a hybrid portal, which provides a unified access-point of the combined resources of Europe's national Libraries. From its inception, it has aimed to provide a pan-European cross-collection search which would otherwise have been impossible. To achieve this, overlapping techniques/protocols are used to connect from the portal to the different libraries and return results from the users queries to the browser. The main techniques/protocols used are OAI-PMH, SRU, Z39.50 and AJAX (Asynchronous Javascript And XML). Currently, 246 collections are available through the portal from 22 different libraries.

To make these different libraries interoperable the TEL Application Profiles (henceforth TEL-AP) were designed, one for applications and one for collection descriptions. The TEL application profile for applications can be seen as the largest possible set of common metadatafields amongst the libraries, that trigger a function in the portal. Each collection made available via the European Library has it own TEL Application Profile for collection descriptions, to allow the user to select their own subsets. The TEL-AP approach is perfectly suitable for a basic disclosure library resources, but when further integration is desired a more refined method of querying the library sources is required. The European Library office is currently reviewing the option of developing a server-side query-building-engine which composes target-specific queries and clusters the target responses. This new strategy moves away from the current architecture in which the integration of the targets is done in the browser. However, the transition is not as great a departure from the original philosophy as it might appear at first glance. The SRU/Z39.50 gateway, for example, already mitigates much of the library-side diversity.

Even though, we are in the early stages of research, it might still be interesting to go into some of the nitty gritty problems we need to tackle in order to create a more fine-grained level of interoperability between European National Libraries. In the following sections, I will present some of the main focus areas that need to be processed before the query can be send to the target.

2. Interoperability through adaptive query-construction

One of the most time-consuming consequences of this approach is that the full breadth of the detailed differences between the collections need to be charted and added to the collection descriptions. We have chosen to use a collection view rather than a library, because we have found that even collections within a single library can have wide ranging differences. Even if they appear identical in the specs, hidden features such as normalisation rules or different use of the datefield can still give unexpected results. In the following list, a number of these focus areas are listed.

- normalisation rules
- character encoding
  - if unicode-variant, are composed or decomposed diacritics used?
- main language, auxiliary language
  - is the spelling system used consistent and brought up to standards?
- access-protocol
- recordscheme
  - has there been any non-standard use of metadatafields, e.g. '*fourteenth century'* in date-field

- What controlled vocabulary has been used and on which fields
- Which metadatafields can be seen as free text fields
- Are the identifiers used persistent, or in any case suitable for deep-linking.
- etc.

In the query building component as we currently envision it, we aim address the following issues in the near future:

- Target adaptive normalisation
- better handling of composed vs. decomposed unicode glyphs
- better handling of target-specific non-unicode character encoding requirements
- Multi-lingual subject access
- Mapping of various cataloguing schemes to provide unified access for the user on record level
- Multi-lingual search in free text fields (from any to all)

3. Conclusion

Although much can be accomplished via computational interoperability, full scale accurate coverage is an extremely difficult ideal to work towards. Making library catalogues remains a manual exercise with all its inherent inconsistencies. And even when one takes these inconsistencies aside, the sheer number of variables that need to be taken into account increases exponentially with every collection that is added. Therefore, the interoperability within the European Library, especially controlled vocabulary, will always remain a work in progress. That being said, the user benefit is unmistakable. In-depth access to large sets of collections via the European Library portal, will provide the user with new vistas and cross-sections for research to Europe's cultural heritage. Through continued efforts of collaboration from various fields of expertise, it is my opinion that, this ideal can be approached.

Semantic interoperability via ontology mapping

Andrea D'Andrea[1], Franco Niccolucci[2]
EPOCH
[1] Università di Napoli "L'Orientale" – [2] PIN, Prato

1. Introduction.

According to a well-known definition [Staab, Studer 2004], an ontology is "*a formal explicit specification of a shared conceptualization for a domain of interest*" – in other words, it is related to a community of users. As long as such users share the same concept organization, namely the same ontology, there is no issue and interoperability is guaranteed. Problems arise, on the contrary, when different domains and the related communities of interest (and practice) try to co-operate and aim at interoperable archives without any common ontology. In such cases, one might believe that they generate a more general (perhaps interdisciplinary) domain, and a more general community, and that there exists some overarching ontology, of which the original domain ontologies are just specializations. So, for example, if one deals with Renaissance paintings and another one with 19th century stamps, a 2D image ontology could be expected to guarantee interoperability among them, since both domains are subsets of the set of 2D images. It is a tempting approach, particularly for those neither involved with paintings nor with stamps, who can claim to solve the problems of both in this way. In many cases, however, such a generalizing approach leads to build a very complex theoretical structure based on a bare-bone ontology (e.g. Dublin Core), unfortunately of rather little utility for the individual domains. In fact, there are very few (if any) applications where this approach has proved to be successful in practice.
The opposite, bottom-up approach is often based on ad-hoc solutions, strongly relying on peculiar features of the domain, which loose significance when another domain is considered. This approach is sometimes strongly supported by heritage professionals, perhaps scared of loosing the control on their discipline if it is contaminated with concepts coming from elsewhere. The usual argument is that interoperability is unnecessary in cultural applications, and the unavoidable loss of specialization that accompanies it has no gain in improved knowledge. Paradoxically, the ineffectiveness of the top-down approach brings support to this point.
This argument has become weaker with the request for trans-national interoperability and the extensive use for cultural purposes of non-traditional data, as images, 3D models and movies. This has led to the need of incorporating such data into cultural databases, previously generally limited to text and reference to pictures, in general stored in external files or archived as blobs. Nowadays digital cultural objects have a complex nature and the organization and management of digital cultural archives (the so-called digital libraries) must reflect this complexity. A "librarian" approach is unsatisfactory, because it tends to ignore the peculiarity of each domain, as explained above for the top-down approach. In conclusion, digital libraries are no librarian's business at all.

To avoid this dilemma we have adopted a third approach, focusing on the following aspects:

- Determine which is the most effective way of storing the many facets of digital cultural objects; find the best "container" format and give it a sound theoretical basis.
- Establish guidelines for mapping existing data structures (and ontologies) on some established standard.
- Accompany each step with real examples based on extensive datasets, and provide tools for their management.

2. The features of digital cultural objects

Work on this issue is still in progress. At present, a preliminary list of features has been established and for each of them a standard has been chosen. The overall container will probably be MPEG-7 or METS.
Reconciliation of the different ontologies involved has been analyzed in some special, but rather general, cases, for example as far as 3D models are concerned [Niccolucci, D'Andrea 2006], using X3D as standard for the 3D part. Whether the geometry must be considered as a feature of a cultural object, or, vice versa, cultural information is to be considered as a set of attributes of a physical artefact, is in fact irrelevant, as both approaches have been shown to be viable [Niccolucci, D'Andrea 2006, Niccolucci 2007].

2. Mapping

The mapping process has been investigated for the case of archaeological data [D'Andrea et al 2006]. This is only apparently a simplification, both for the complexity of such data and the very large amount of legacy

archives, currently in use for research, management and other purposes. The institutions in charge of maintaining such archives, usually at a national or regional level, are reluctant to convert to a different system for several reasons, including national regulations that have not yet been superseded by some European norm. Mapping to some intermediate international standard appears therefore to be the only possibility to guarantee interoperability and maintain the semantic richness of the archives. Our choice for the common standard has been CIDOC-CRM.
Semantic interoperability may be really useful in the archaeological domain. For example, information concerning prehistoric "cultures" spanning over vast areas is usually spread through the archives of several modern states, and is stored not only in different languages, but also according to different methods reflecting different national regulations. Multilingualism appears also to be a key issue, and work is now addressing multilingual thesauri. As it is well known, problems here come not only from the translation, but also from the diverse history of Europe. For example, the term "Iron Age" has a different time span in European countries. As a consequence, the year 600 AD would belong to Middle Ages in Poland, to Iron Age in Norway, to Early Medieval in UK, and to (very late) Classical period for Romania. It would be Byzantine in Greece, Asia Minor, the Levant, and other parts of the Mediterranean region, but not everywhere (e.g. Spain). In Italy, it would depend from the region. So it is becoming clear that simple concepts as *who*, *when*, *where* are in fact all time and space dependent – and there is no such thing as a universal calendar or gazetteer.

3. Tools and applications

As yet, two tools have been provided, both still as prototypes.
The first one, AMA, is a help to create the mapping. It accepts as input the description of two ontologies e.g. in RDF (other ways are possible) and provides a graphical interface for establishing the correspondence. The mapping is then saved as a "template", i.e. an XML description. The tool creates also an XSL for the automatic conversion of the data, which need to be XML encoded to be processed in this way. An additional advantage of AMA is the possibility of editing an existing template in order to define a mapping which differs from an existing one only for some details. The AMA tool is being tested on an extensive number of datasets, provided by several national agencies in charge of archaeological data management. The AMA team is also developing a tool to manage poorly structured or unstructured text documents.
For more details on AMA, including the composition of the research team, visit the EPOCH web site www.epoch.eu: AMA is described in the chapter on NEWTONS, accessible via the tag "Research".
The second tool, MAD, is a data management system based on an XML native DBMS [Felicetti 2006]. The system can work on separate collections, stored on different servers, regardless of their structure, which is nonetheless very important to retrieve significant results. MAD can accept the output of the AMA conversion and thus offers a solution for practical cases when data conversion is performed.. Queries in MAD are based on XQUERY. Semantic (i.e. RDF-based) queries are presently being experimented. MAD has been used for a number of archaeological archives, and could be used for any archive where records consist of XML documents.
Both systems are distributed as Open Source and work on a number of platforms (Windows, Linux, Mac OS). They are available for download (or will be in short) from the above mentioned EPOCH web site.

Acknowledgement

The present research has been partially funded through the project EPOCH by the European Commission, under the Community's Sixth Framework Programme (contract no. 507382). However, this paper reflects only the authors' views and the European Community is not liable for any use that may be made of the information contained herein.

References

Staab S., Studer R., 2004. (eds.) *Handbook on Ontologies*. Berlin, Springer, vi.

Niccolucci F., D'Andrea A., 2006. An Ontology for 3D Cultural Objects. In M. Ioannides, D. Arnold, F. Niccolucci, K. Mania (eds.) *Proceedings of VAST 2006*. Aire-La-Ville, Eurographics, 203 – 210.

Niccolucci, F. 2007. Standards, credibility and philology of virtual models in archaeology. In B. Frisher (ed.) *Beyond Illustration*, in press.

[D'Andrea A., Marchese G., Zoppi T., 2006. Ontological Modelling for Archaeological Data. In M. Ioannides, D. Arnold, F. Niccolucci, K. Mania (eds.) *Proceedings of VAST 2006*. Aire-La-Ville, Eurographics, 211 – 218.

Felicetti A., 2006. 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, 124 – 131.

Achieving Interoperability in the MichaelPlus Project

Anna Christaki[1], Vassilis Tzouvaras[1], Antonella Fresa[2], Rosella Caffo[2], Pier Giacomo Sola[3] and Stefanos Kollias[1]

[1]Image, Video and Multimedia Laboratory
National Technical University of Athens, Greece
{achristaki, tzouvaras, stefanos}@image.ntua.gr,

[2]Ministero per i Beni e le Attività Culturali (MiBAC)
Via Michelangelo Caetani 32, I-00186 Roma
rcaffo@beniculturali.it, fresa@promoter.it

[3]Amitié
Centro di Ricerche e Servizi Avanzati per la Formazione
pgsola@amitie.it

Abstract

In this paper we present how interoperability is achieved in the MichaelPlus project. MichaelPlus uses the Michael platform for knowledge organization and presentation of the content provided by the archives participating in the project. The Michael platform supports interoperability in the schema, record and repository levels. The end user has the ability to make cross-lingual queries to all the archives through the controlled vocabularies embedded in the platform. Scalability of the platform is achieved through schema mapping techniques and metadata can be harvested using the OAI-PMH standard. However, since the Michael platform is based on XML, we can not ensure semantic interoperability. The future plan of the MichaelPlus project is to upgrade the platform using Semantic Web technologies. These technologies will provide declarative and procedural semantics of the metadata records aiming at semantic interoperability. We present two methods in which semantic interoperability can be achieved. The first method is through the use of the SKOS vocabularies. The second method is to apply ontology alignment techniques.

1 The MichaelPlus Project

Michael-Multilingual Inventory of Cultural Heritage in Europe is a deployment initiative supported by two projects co-financed by the eTEN programme, namely: Michael (2004-2007) and MichaelPlus (2006-2008). The scope of the initiative is to celebrate the richness, breadth and diversity of the European cultural heritage by promoting it to a world-wide audience through the Internet.

The MichaelPlus consortium brings together public and private bodies: national and regional cultural ministries, state agencies, major cultural institutions and technical partners with specific expertise. The consortium was born in 2004 with partners from three countries (Italy, France and the UK) and extended in 2006 to eleven more countries (Czech Republic, Finland, Germany, Greece, Hungary, Malta, The Netherlands, Poland, Portugal, Spain, Sweden).

The project started in June 2004 with the aim of implementing an innovative multi-lingual open source platform, equipped with a search engine which provides the ability to search, browse and examine multiple national cultural portals from a single point of access. The target audience of MichaelPlus is broad, for example students and researchers are able to discover information about European collections that might previously have been difficult to find. The services will also support cultural tourism, the creative industries and other interests.

The architecture of the service is based on national implementations that are created, managed and maintained by the Michael national partners and on a European portal that is able to harvest data from the national instances. The national instances are already functioning in Italy, France and the UK and the European portal is currently accessible online. Eleven new partner countries are currently implementing their national instances in the scope of the MichaelPlus project.

The Michael software platform consists of two modules that work together to provide data management and publishing services:

- A module targeted to the cataloguers that allows users to create, modify, import and manage records that describe the digital collections. These functions are available using a standard Web browser. Data is stored using an XML database.
- A module that offers the public interface to end-users to search for digital cultural heritage within their Web browsers. The module is based on an XML search. Institutions and countries can customise their own display engine to adapt the interface to meet their particular needs.

The two Michael modules act as data repositories that are consistent with the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH).

The platform is distributed as open source software and is built upon other open-source components, i.e.: Apache Cocoon, eXist, Xdepo, SDX. It is based on the Michael data model, which derives from the work done by MINERVA,

closely related to the RSLP collection description schema and to the Dublin Core Metadata initiative on collection description. It is available under open source licence which enables the platform and data model to be used by other projects that want to create, for example, regional cultural inventories and other added value services.

This paper will examine closely the way the Michael infrastructure is dealing with interoperability issues at different levels of operation.

2 Interoperability in MichaelPlus

The concept of Interoperability can be defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [3]. At the most fundamental level the interoperability concept is simply about making things work together.

Today's vast Internet growth has led to the development of numerous digital libraries and repositories. These repositories however have been built around specific project needs, user communities, subject domains etc. The diversity of these resources has led to interoperability challenges that need to be addressed urgently. The implementation of interoperability can be considered from a methodological point of view in three different levels of operation: the schema level, the record level and the repository level [1]. These three levels of operation and the way they have been implemented in the Michael infrastructure will be analysed in the following sections.

2.1 Schema level Interoperability

A metadata schema establishes and defines data elements and the rules governing the use of data elements to describe a resource. The selection of a metadata schema to be used in digital collection representation should be made on the basis that the digital collection built on that schema will be interoperable with other collections or repositories. To achieve interoperability on the Schema Level actions must be taken during the design of the system's Data Model before any metadata records are created. The focus is therefore on the elements of the Data Model [1].

Derivation is one of the methods used to achieve interoperability at this stage. Derivation involves a new schema creation based on an existing one. In a collection of digital repositories where components vary based on different needs an existing schema can be used as the source from which individual schemas will derive.

The Michael data model is a derivation of the RSLP collection description schema and the Dublin Core Metadata initiative on collection description. The RSLP Collection Description is encoding collection descriptions using the XML encoding of RDF. The RSLP encoding syntax follows the draft recommendations for encoding Dublin Core metadata within RDF particularly in the area of how to encode the *scheme* associated with a particular value. By encoding descriptions in RDF/XML and by making use of Dublin Core properties as far as possible, the RSLP collection description aim to be positioned very closely alongside other emerging descriptive practices on the Web [4]. The use of the RSLP Collection Description for the creation of the Michael metadata schema ensures a high level of interoperability between the Michael digital repository and many existing or future digital collections.

Another approach to schema level interoperability is the creation of Crosswalks. A Crosswalk is "a mapping of elements, semantics, and syntax from one metadata scheme to those of another" [5]. Crosswalks are broadly used as a method of achieving interoperability between different metadata schemas. In the scope of the Michael project crosswalks have been implemented mapping the Michael model to other popular metadata schemas (LOM, DC, ISAD(G) etc.) as well as other metadata schemas used by the cultural institutions participating in the project (TEL, KB etc). These Michael Crosswalks are used for the creation of migration tools for importing existing content into the Michael databases, minimizing the manual input that needs to be carried out by the cultural institutions and building upon existing annotations as much as possible.

2.2 Record Level Interoperability

In the Record Level efforts are focused on integrating metadata records through the mapping of the elements according to their semantic meanings. Common results include converted records and new records resulting from combining values of existing records [2]. Metadata record conversion is essential when a project is dealing with established data repositories. The biggest challenge in this process is to avoid data loss and distortion. Since MichaelPLUS aims to reuse national digital collections, existing records have to be mapped to the MichaelPLUS metadata schema before they can be imported to the MichaelPLUS repository. Crosswalks that have been developed in the scope of the project assist in the development of mapping tools for the conversion of records to the common Michael data format.

The degrees of equivalence however in mapping the metadata fields between records of different digital collections can vary between: one-to-one, one-to-many, many-to-one [6]. There are cases where record fields from target repositories have to be broken down to smaller units or grouped to one bigger unit in order to implement the mappings successfully. The degree of complexity in the record conversion is even higher when there's a need to map field values to values from controlled vocabularies. The Michael platform makes use of such controlled vocabularies to improve consistency in recording each collection or item.

In order to facilitate the conversion process, the mapping of existing national vocabulary lists to the MichaelPlus lists is taking place before the conversion of metadata records to the Michael common format. This procedure is followed to lower the conversion complexity and to assure data integrity and minimal data loss.

2.3 Repository level Interoperability

Efforts at this level focus on mapping value strings associated with particular elements of integrated or harvested records from varying sources (e.g. terms associated with *subject* element). The results enable cross-collection searching. The processes related to ensuring interoperability at the repository level include metadata harvesting, supporting multiple formats, aggregation, cross-walking services, value-based mapping for cross-collection searching and value-based co-occurrence mapping [2].

Repository level interoperability can be achieved through the use of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) framework. The goal of OAI-PMH is to supply and promote an application-independent framework that can be used by a variety of communities engaged in publishing content on the Web [7]. The Michael data repositories are built to be consistent with the OAI-PMH making metadata available in both standard Dublin Core and Michael format. Records from distributed sources can be gathered into a Michael instance and published together. The MichaelPlus European service makes use of these harvesting facilities to bring together the contents developed by the separate national instances.

The Michael publication module also includes a REST-based API for searching and retrieving records using simple HTTP requests and XML responses. These facilities have been put in place to enable other cultural information service providers to incorporate MichaelPlus search services within their websites. Furthermore, the project has used the method of value-based mapping for cross-collection multilingual searches.

A number of controlled vocabularies have been constructed in order to support multilingualism. These vocabulary lists have derived from a number of standards and projects (e.g. UNESCO thesaurus for “subject” list, ISO 3166.1 for “spatial coverage” list, ISO 639-2 and ISO 639-3 for “language” list etc.). By utilising controlled vocabularies for specific fields, mappings between the national Michael repositories are implemented not on a field-by-field basis but by translating these controlled vocabularies to the languages of the partner countries. These mappings are then used to effectively combine digital collections from different national Michael repositories when browsing the central European service.

3 Future Extensions

In the previous sections we presented how we ensure interoperability among the different archives that participate in the MichaelPlus project as well as a way to perform cross-lingual retrieval and presentation. The data model of MichaelPlus has been based in the XML technology. XML is the adopted standard for exchanging metadata on the Web. However the lack of formal semantics in XML creates obstacles in providing semantic interoperability. Ontologies have proved that they can enable semantic interoperability. An ontology is a controlled vocabulary that describes objects, and relations between them, in a formal way and has a grammar for using the vocabulary terms to express something meaningful, within a specified domain of interest. Ontologies in an application organize data used to describe other data, called metadata, in a machine understandable way giving the opportunity to agents to (semi)automatically carry out complex tasks assigned by humans in a meaningful (semantic) way.

The use of formal semantics in metadata representation in MichaelPlus will enable semantic interoperability. In order to provide the extra semantic layer SKOS vocabularies can be utilised [8]. SKOS or Simple Knowledge Organisation System is a family of formal languages designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. SKOS Core (Simple Knowledge Organisation System) Core is a model and an RDF vocabulary for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, 'folksonomies', other types of controlled vocabularies, and also concept schemes embedded in glossaries and terminologies.

The SKOS Core Vocabulary is an application of the Resource Description Framework (RDF). RDF provides a simple data formalism for talking about things, their properties, inter-relationships, and categories (classes). Using RDF allows data to be linked to and/or merged with other RDF data by Semantic Web applications. In practice, this means that data sources can be distributed across the web in a decentralised way, but still be meaningfully composed and integrated by applications, often in novel and unanticipated ways.

SKOS is enabling semantic interoperability through the creation of a harmonising semantic layer for all the existing metadata standards. Another way to enable semantic interoperability is ontology alignment. The rapid growth of the Semantic Web has as a counterpart the development of a large number of ontologies in the area of cultural heritage. These ontologies try to provide knowledge bases, the capability for knowledge sharing and reusability. Reusing heterogeneous and partly overlapping ontologies requires a tremendous and considerable amount of effort [9]. Before we can reuse our source ontologies their meanings have to be fully understood [10]. Afterwards ontologies have to be combined by using integration, merging or application reuse resulting in new more complete, powerful and complex ontologies [11].

Defining the problem of ontology alignment, which is often called ontology matching, we can say that alignment of ontologies is the process of finding and providing semantic mappings among ontologies to overcome semantic heterogeneity and to provide interoperability among ontologies. In order to identify which are the most relevant concepts between ontologies and to provide relevant and acceptable mappings between concepts one has to provide measures that give the similarity between concepts and relations resided within ontologies.

4 Conclusions

In this paper we presented the methods used the Michael platform to achieve interoperability among the archives participating in the project. According to [1], there are three levels of interoperability, 1) the schema, 2) the record, and 3) the repository levels. However semantic interoperability is not achieved in the existing platform. The future plan for the MichaelPlus project is to upgrade the platform using Semantic Web technologies. We have also presented two ways for achieving semantic interoperability.

5 References

[1] L. M. Chan, M. L. Zeng, "Metadata Interoperability and Standardization – A Study of Methodology Part I", *D-Lib Magazine*, 2006.

[2] M. L. Zeng, L. M. Chan, "Metadata Interoperability and Standardization – A Study of Methodology Part II", *D-Lib Magazine*, 2006.

[3] IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries, New York: IEEE, 1990.

[4] A. Powell, M. Heaney, L. Dempsey, "RSLP Collection Description", *D-Lib Magazine*, 2000.

[5] "Understanding metadata", NISO (National Information Standards Organization), 2004.

[6] M. L. Zeng and L. Xiao, "Mapping metadata elements of different format", E-*Libraries*, 2001.

[7] C. Lagoze, "Open Archives Initiative progress and practice", 2002.

[8] "Simple Knowledge Organisation Systems (SKOS)", http*://www.w3.org/2004/02/skos*.

[9] M. Uschold, M. Healy, P. Clark, S. Woods, "Ontology Reuse and Application".

[10] M. Klein, "Combining and Relating Ontologies: An Analysis of Problems and Solutions".

[11] S. Pinto, A. Gomez-Perez, J. Martins, "Some Issues on Ontology Integration", 1999.

[12] J. Euzenat, P. Valtchev, "An Integrative Proximity Measure for Ontology Alignment", 2003.

[13] J. Sowa, "Electronic Communication in the Onto-Std mailing list", 1997.

[14] W-S. Li, C. Clifton, "Semantic Integration in Heterogeneous Databases Using Neural Networks", 1993.

[15] T. Kohonen, "Adaptive, associative, and self-organizing functions in neural computing", 1978.

[16] D. McGuinness, R. Fikes, J. Rice, S. Wilder, "An Environment for Merging and Testing Large Ontologies".

[17] F. Noy, M. Musen, "SMART: Automated Support for Ontology Merging and Alignment".

[18] F. Noy, M. Musen, "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment".

[19] D. Lin, "An Information - Theoretic Definition of Similarity".

[20] P. Resnik, "Using Information Content to Evaluate Semantic Similarity", 1995.

[21] Z. Wu, M. Palmer, "Verb Semantics and Lexical Selection", 1994.

[22] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos , A. Halevy, "Learning to Match Ontologies on the Semantic Web", 2002.

[23] A. Meadche, S. Staab, "Comparing Ontologies – Similarity Measures and a Comparison Study", 2001.

[24] F. Hakimpour, A. Geppert, "Resolving Semantic Heterogeneity in Schema Integration: An Ontology Based Approach".

[25] N. Guarino, "Formal Ontology in Information Systems".

[26] A. Fraquhar, R. Fikes, J. Rice, "The Ontolingua Server: A Tool for Collaborative Ontology Construction", 1997.

[27] P. Weinstein, W. Birmingham, "Comparing Concepts in Differentiated Ontologies".

Semantical interoperability with IMAGINATION content using standardized ontologies

Andreas Walter[1], Gabor Nagypal[2]

[1]IPE, FZI Forschungszentrum Informatik, Haid-und-Neu-Straße 10-14, 76131 Karlsruhe
awalter@fzi.de

[2]disy Informationssysteme GmbH, Erbprinzenstraße 4-12, 76133 Karlsruhe
nagypal@disy.net

Abstract: The IMAGINATION project provides image-based navigation for digital cultural and scientific resources. Users can click on parts of an image to find other, interesting images to a given context. The combined application of object detection and identification in images together with text mining techniques exploiting domain specific ontologies will help generate high-quality semantic metadata. We want to share this metadata with other information systems, e.g. in the domain of cultural heritage. This paper describes the requirements of IMAGINATION that must be fulfilled to reach this goal and analyzes CIDOC-CRM, FRBR and MPEG-7 based on these requirements.

1 The IMAGINATION Project

The main objective of IMAGINATION[1] is to bring digital cultural and scientific resources closer to their users. Our aim is to enable image-based navigation for such resources. Users can receive meaningful contextual information about images and image parts, which makes images easier to understand. Moreover, IMAGINATION allows users to navigate to other relevant images and texts in the knowledge repository just by clicking on interesting image parts.

The major instrument to provide context-sensitive, relevant information is the use of semantic metadata. The highest possible quality level of automatically generated metadata is achieved by iteratively combining the results of three processing steps (Figure 1). Step 1 extracts semantic information from the surrounding text and textual metadata of the image (such as its caption) using text mining techniques. In Step 2, object detection algorithms detect faces and objects. Finally in Step 3, the detected objects are identified by using object identification algorithms. All steps exploit an ontology describing the application domains of IMAGINATION – First World War and contemporary European politics (e.g. Person: *Ferdinand Foch*; Profession: *Marshall*; "participates At" Event: *Sign of armistice at Compiegne*). In addition, each

[1] The IMAGINATION Project – http://www.imagination-project.org

step considers the already available semantic information to improve its results and thus to achieve a synergy effect among the different types of algorithms.

Figure 1 : IMAGINATION technology applied to Wikipedia

2 Our Requirements for semantical interoperability

To enable semantical interoperability with other systems, IMAGINATION needs to exchange technical data, metadata and semantical information.

2.1 Technical information

Technical information needs to be exchanged with other systems to guarantee the correct view of the images. In IMAGINATION, all images are available in digitalized formats (e.g. JPEG) and stored on web-servers. Every image is available in three sizes, a very small one and a medium one for a fast preview of the image and a big size image for commercial usage, e.g. printing. For each image region with interesting content, the coordinates of the corresponding areas are available.

2.2 Textual and semantic metadata

Textual metadata information exchange e.g. text descriptions, the creation date of the image, owner information and usage information, e.g. types of allowed usage. Semantic metadata contain references to instances of the domain ontology of IMAGINATION and information about the relevance of these instances concerning

the whole image or an image region. Especially specifying semantic metadata for image regions is an uncommon requirement that is specific to IMAGINATION.

2.3 Concepts and Instances of the domain specific ontologies

Domain specific ontologies contain concepts, instances and relations. Ontology elements define the possible values for semantic metadata. It is therefore crucial that the ontology contains all information that is needed to (automatically) generate semantic metadata. In our case it is especially includes lots of instances in the WWI and EU politics domains.

3 Using standardised ontologies for semantical interoperability

3.1 CIDOC-CRM

CIDOC-CRM [1] is a conceptual ontology for semantical interoperability. It contains concept and property taxonomies to describe e.g. historical time periods, events, places and persons for the scientific documentation of museum collections. In IMAGINATION, we can use these concepts as a base for our domain specific ontology. Then, we can exchange our concepts based on CIDOC-CRM. CIDOC-CRM does not provide for the semantical interoperability of our technical information because it aims to describe physical objects like paintings or printings in museums and not digitalized images or image regions. Also, we see no possibility to exchange metadata information using CIDOC-CRM.

3.2 FRBR

The scope of FRBR [4] is to exchange data that are recorded in bibliographic records or by national geographic agencies. The integrated entities work, expression, manifestation and items allow the exchange of all required technical information in IMAGINATION except from the definition of image regions. Also, it is possible to exchange metadata. However, the entities are not as powerful as the concepts in CIDOC-CRM. Hence, we would loose a lot of semantical information when exchanging data with FRBR.

3.3 MPEG-7

MPEG-7 [2] provides important functionalities for manipulation and transmission of objects and associated metadata in multimedia content. MPEG-7 allows the definition of all required technical information, including the definition of image regions.

Seungyup et al [5] proposed an extension of the MPEG-7 standard, a description scheme for image content. Using this extension, it is also possible to exchange textual metadata with MPEG-7.

The extraction of semantic descriptions and annotation of the content with the corresponding metadata, though, is out of the scope of the MPEG-7 standard [3]. That means, MPEG-7 does not allow the semantical interoperability based on our own domain specific ontology. This is a missing feature for our requirement to exchange extensive semantical information based on domain specific ontology elements.

4 Conclusions

Our analysis showed that MPEG-7, as an ontology designed for multimedia content, solves two main requirements for the semantical interoperability of IMAGINATION contents. These are the exchange of technical information and metadata.

CIDOC-CRM allows semantical interoperability for ontologies that are aligned with it. Therefore, we aim to use CIDOC-CRM as a basis for our domain specific ontology in IMAGINATION and extend its concepts and properties where needed.

To summarize: the combination of MPEG-7 and its introduced extensions for the description of images for the exchange of technical information with CIDOC-CRM for the exchange of semantical information leads to the highest possible level of semantical interoperability for IMAGINATION content.

References

1. *Crofts et al*, Definition of the CIDOC Conceptual Reference Model Version 4.2; CIDOC CRM Special Interest Group; June 2005
2. Martinez, J. M., MPEG-7 Overview Version 10; http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm; October 2005; last visited January 2007
3. *Petridis et al*, Combined Domain Specific and Multimedia Ontologies for Image Understanding, Workshop on Mixed-reality as a Challenge to Image Understanding and Artificial Intelligence at the 28th German Conference on Artificial Intelligence, KI 2005, Koblenz, Germany, September 2005
4. *Saur, K.G.*: Functional Requirements for Bibliographic Records, Final Report , UBCIM Publications – New Series Vol. 19; 1998
5. Sengyup et al, Proposal for MPEG-7 Image Description Schema, ISO/IEC JTC1/SC29/WG11; February 1999

SKOS: a model for metadata representation and interoperability – Dutch Cultural Heritage Institution thesaurus conversion use case

Véronique Malaisé & the CHOICE team

Semantic interoperability between descriptions of digital documents from different collections can be achieved by different methods, including *a posteriori* converting or mapping between the vocabularies and models used for indexing these different collections, *a priori* complying to the same model or *a posteriori* converting or mapping the different data to a common generic (and preferably standard) model. Our research project, CHOICE[1], is focusing on applications and issues related to cataloging practices, in collaboration with the Netherlands Institute for Sound and Vision[2], in which indexing and retrieval is done based on a thesaurus called GTAA (a Dutch acronym for Common Thesaurus for Audiovisual Archives).

For interoperability purpose within other projects of the CATCH program[3] and for using this thesaurus in Semantic Web applications, we chose the third approach and converted it to the SKOS model. This paper describes the SKOS model in comparison with the ISO-standard way of representing thesaurus data, based on the Website of the W3C Semantic Web Deployment Working Group[4] and [1], and the advantages and drawbacks to comply with this model. We then detail the thesaurus that we have converted, the GTAA, describing its standard and specific features, and the conversion problems that we faced. A presentation of the conversion method that we followed and more details about this experiment can be found in [2]. We conclude on the advantages of such a conversion in terms of language integration possibility and of software application, mentioning the example of the SKOS Web Browser developed in our project.[5]

SKOS: the model, advantages and drawbacks

SKOS, Simple Knowledge Organisation System "*[…] provides a standard way to represent knowledge organisation systems using the Resource Description Framework (RDF). Encoding this information in RDF allows it to be passed between computer applications in an interoperable way*[6].". The SKOS Specifications are currently published as W3C Working Drafts, which means they are work in progress but on the way to become a W3C recommendation. In this respect, SKOS is interesting as a model for interoperability on the Web. The main advantage is that it proposes an RDF definition of a thesaurus's main construct, making them machine readable and usable in Semantic Web applications like vocabulary integration [3] or thesaurus browsers [4].

The main difference between a standard thesaurus, as described in the ISO norms, and the SKOS model is that the first is term-centered, whereas the latter is concept oriented, following RDF and ontologies' usual modeling features (Concepts or Classes and Properties or Relationships). A thesaurus distinguishes between preferred terms, meant to be used when indexing documents from a collection, and non preferred terms, which are considered as synonyms of the previous ones, but should not be used when indexing. These two entities become strings attached to a concept in a SKOS representation: a PrefLabel and AltLabel, for preferred label and alternative label. From the five core relationships in thesauri, namely broader term, narrower term (both building the thesaurus' hierarchical structure), related term (called associative relationship), use and use for (sometimes referred to as linguistic relationships), only

[1] http://www.nwo.nl/CATCH/CHOICE, CHOICE is one of the 10 projects of the CATCH program (see http://www.nwo.nl/CATCH), focusing on accessing, describing and integrating resources from Dutch Cultural Heritage Institutions.
[2] http://www.beeldengeluid.nl
[3] See footnote 1.
[4] http://www.w3.org/2004/02/skos/
[5] http://ems01.mpi.nl/CHOICE/, see the Demonstration page and [4].
[6] http://www.w3.org/2004/02/skos/

the first three apply to SKOS representation. In this case, they stand between two concepts, which have each preferred and non preferred labels attached to the conceptual root. This root is referred to by a unique identifier.

The picture below, extract from the UKAT thesaurus[7] and borrowed from the SKOS Specifications, features a thesaurus' classic representation and the corresponding SKOS graph:

Term: Economic cooperation

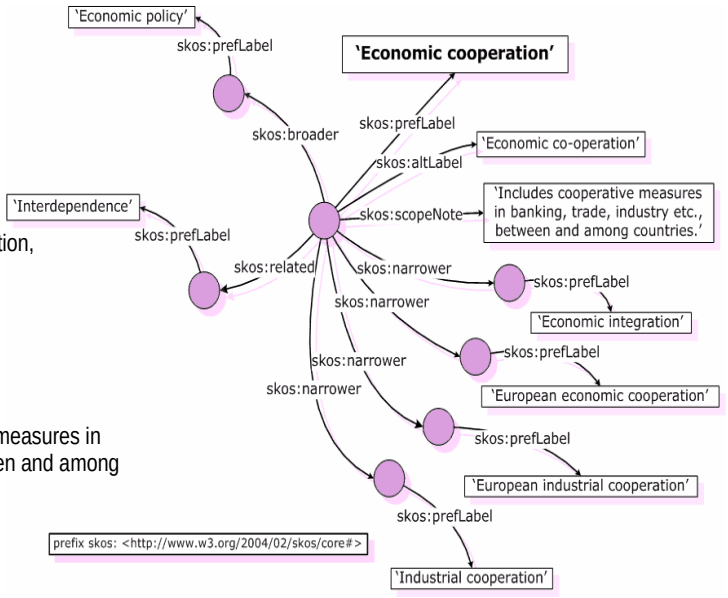
Used For: Economic co-operation

Broader terms: Economic policy

Narrower terms: Economic integration, European economic cooperation, European industrial cooperation, Industrial cooperation

Related terms: Interdependence

Scope Note: Includes cooperative measures in banking, trade, industry etc., between and among countries.



Grouping all the related information under one encapsulating concept makes the thesaurus updates easier: the indexing is done with the concept identifier, and the labels attached to it can easily be changed over the time, may it be to solve polysemy problems or because of terminological evolution. This concept-oriented modeling also helps getting the different versions of the annotations "backwards compatible". But the choice to restrict one thesaurus term to one string (or concept label) makes it impossible to attach additional information to this label, either preferred or not preferred. One such kind of information possibly present in the thesaurus is history notes, which can only be linked to the concept as a whole in SKOS, and not to the specific expressions as in the original thesaurus.

SKOS has been created to answer basic or more sophisticated thesaurus and controlled vocabularies modeling needs, but in some practical cases, it is not sufficient. For example, no standard mechanism is yet defined to express compound concepts or qualifiers. For interoperability purposes, it is suggested to extend the model for specific needs, instead of building local competing models. We will present the thesaurus used at Sound and Vision, the modeling issues that we faced when trying to convert it to SKOS, and in which respect we had to extend it.

The GTAA thesaurus

The GTAA is a faceted thesaurus: its terms are divided into 6 non overlapping groups. These groups are:

- Subject: terms from this facet are used to describe the main topic the TV program is about, or which is mentioned in the program;
- Person: to describe the main people the program is about, or people appearing on the screen;
- Location: to describe the main location the program is about, or the place(s) where it was shot;
- Name: to describe the name of companies, groups, bands etc that the program is about or who appear on screen;
- Genre: to qualify the genre of the program;

[7] http://www.ukat.org.uk/

- Maker: terms indicating the Maker(s) ' name(s).

Subjects and Genres are organised in broader term/narrower term hierarchies, Subjects, Genres and People have a use/use for relationship, Subjects have related terms (associative relationships) and the 6 facets can have scope notes. Besides these standard features, that have straightforward counterparts in SKOS, GTAA has also a number of more specific features and a set of *ad hoc* ones. We list the features of these two categories in the next section, with their SKOS counterpart.

SKOS conversion

The GTAA contains two standard features for which there is no conversion proposal in SKOS Core: facets and qualifiers. The facets can be described at the level of the Metamodel of the thesaurus, according to SKOS[8], but there is no specific construct or property that enables to link a particular instance of a thesaurus concept to one specific facet. Therefore we chose to extend the model and created 6 concepts as sub-classes of the generic skos:concept. Any instance is an instance of a sub-concept of skos:concept, being compatible with the model and keeping the semantics of the original thesaurus. The qualifier's problem is more complex, different modeling possibilities are still under discussion, but they mostly imply the fact that the qualifiers themselves (additional information attached to a term to disambiguate between different possibilities, like in the case of Amsterdam-Netherlands and Amsterdam-US) are entities of the thesaurus. This is not the case for the GTAA, and some qualifiers, like the role of a person in a TV program, are even added at indexing time. We did not choose a definitive modeling solution yet, and are waiting for concrete applications of the vocabulary (in semi-automatic indexing for example) to select the most appropriate one.

The GTAA also contains specific features not described in the ISO norms: "Categories" and a relationship between terms called "linked term". Our first concern was how to interpret the Category relationship: either it is meant to disambiguate different aspects of a term (as a qualifier would do, for example in "Church-institution" *vs* "Church-building"), or it is a way of grouping terms sharing a specific aspect (as with "Milk by animal" and "Cow-milk", "Buffalo-milk", etc.). In the second case, "Milk by animal" is called a node label: it is a way of grouping terms, but the concept itself should not be used for indexing. These node labels are usually part of the term hierarchy. The experts indicated that this option was the intended usage of Categories: to provide a grouping of terms under a label that is not used in the indexing process. Nevertheless, they are meant to provide an alternative grouping of the GTAA terms, and thus are not part of the broader term/narrower term hierarchy. Although we mapped the Categories to the existing SKOS construct for these node labels, namely the skos:Collection, this modeling remains a non standard feature that cannot be processed by SKOS generic softwares. The Categories have explicit identifiers, from which we could infer their hierarchy (01 stands for Philosophy, and 01.01 is one of its subdivisions, for instance).
The linked term relationship connects related terms from different facets, like the name of a Queen with the Subject *Queens* and the country that she rules in the Location facet. These relationships were not instantiated in the original thesaurus, we added the links automatically using Natural Language Processing, and modeled them as sub-properties of the generic skos:related. Thus, we keep the compatibility with SKOS.

Summary: list of the GTAA features with their SKOS counterparts

GTAA unique items:

- Categories in the Subject facet: correspond to node labels and are modeled as skos:Collections, the terms belonging to them being skos:members;
- Linked term relationship: a sub-property of skos:related.

Standard thesaurus features present in GTAA but non addressed in SKOS-Core

- Facets: 6 sub-concepts of skos:concept;

[8] http://www.w3.org/TR/2005/WD-swbp-thesaurus-pubguide-20050517/#secExpressingMetadata

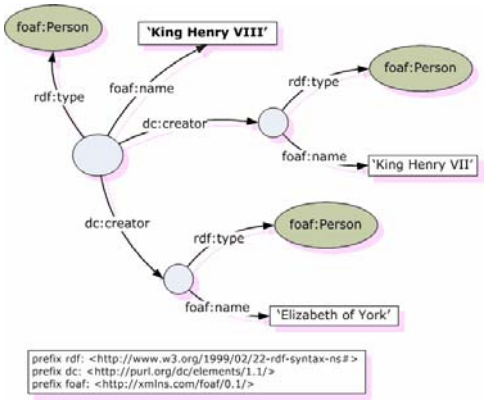
- Qualifiers: we will choose the modeling option that will suit the best our needs in terms of thesaurus usage, still work in progress.

GTAA standard items addressed in SKOS:

- Broader Term, Narrower Term, Related Term: skos:boader, skos:narrower, skos:related;
- Use, Use for: skos:prefLabel, skos:altLabel

Benefit of the conversion

The SKOS model being defined in RDF, turning a thesaurus into a SKOS compliant representation enables one to integrate RDF vocabularies (or other vocabularies also defined in RDF) in the document's description. For example, in the graph below, the thesaurus concept is associated with elements of information defined in Dublin Core[9] and in FOAF[10]:



Complying to SKOS also enables to process the data with generic tools like WebBrowsers (see the project's Demonstration section, for an example of such a Browser displaying the GTAA thesaurus). The page http://esw.w3.org/topic/SkosDev/ToolShed references tools based on or related to SKOS, as a wiki page where authors can freely add their work.

Acknowledgement

This work was partly supported by NWO's CHOICE projects. The authors wish to thank Mark van Assem for his support in the conversion of GTAA and the sharing of his knowledge on SKOS, and also our colleagues at the Netherlands Institute of Sound and Vision for their support and for providing a kind a stimulating working environment.

Bibliography

[1] *SKOS Core Guide*, 2nd W3C Public Working Draft 2 November 2005. Alistair Miles and Dan Brickley eds.
[2] Mark van Assem, Veronique Malaisé, Alistair Miles and Guus Schreiber.(2006). *A method to convert thesauri to SKOS*. In Proc. Third European Semantic Web Conference (ESWC'06), Budvar, Montenegro, June 2006.
[3] 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), Julio Gonzalo, Constantino Thanos, M. Felisa Verdejo and Rafael C. Carrasco (eds.), Springer Verlag, LNCS vol. 4172, pp. 426-437, Alicante, Spain, September 17-22, 2006.
[4] Hennie Brugman, Véronique Malaisé and Luit Gazendam. *A Web Based General Thesaurus Browser to Support Indexing of Television and Radio Programs*. In: Proceedings of the 5th international conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, May 24-26, 2006.

[9] http://dublincore.org/
[10] http://www.foaf-project.org/

TEI

Øyvind Eide, Unit for Digital Documentation, University of Oslo. Convener of the TEI Ontologies SIG together with Christian-Emil Ore.

What is TEI

TEI is a consortium of institutions and individuals from all over the world. The TEI is also a set of guidelines for the encoding of textual material, and it is a set of computer readable files. The guidelines and the computer readable files specify a set of rules documents have to adhere to in order to be accepted as TEI documents.

One of the main goals of TEI is to capture a wide range of intellectual work. It is used extensively in areas such as edition philology, but it is quite possible to create a detailed encoding scheme for e.g. archaeological grey documents and include it as a local extension of TEI, creating documents very different from any other TEI document (Eide 2006).

According to the TEI guidelines (TEI P5 section 1.2.2), there are three primary functions of the guidelines:

- guidance for individual or local practice in text creation and data capture;
- support of data interchange;
- support of application-independent local processing.

TEI is also an important discussion forum, or a set of fora, both through meetings and on-line media, as well as through articles and books. Few encoding problems will be faced that has not been discussed somewhere in the guidelines or in the literature surrounding TEI.

TEI encoding: From text to world

TEI is created for the encoding of texts. This means that the object of a TEI encoding is a text. To be able to find out what to encode and how to do it, the encoder have to read and understand the text. A very simple example is the word "Bergen" in a German text. Based on the interpretation of the context, the word may mean mountains or the place named Bergen.

Traditionally, the object of a TEI encoding has been the text as such. This means that the place name Bergen is encoded as a place name, e.g. <name type="place">Bergen</name>. But there have not been a strong tradition for building up models within TEI for the place to which Bergen is used as a name. From version 3 of the guidelines published in the late 1990's, I quote from the introduction to the chapter on names and dates:

"It should be noted however that no provision is made by the present tag set for the representation of the abstract structures, or “virtual objects” to which names or dates may be said to refer. In simple terms, where the core tag set allows one to represent a *name*, this additional tag set allows one to represent a *personal name*, but neither provides for the direct representation of a *person*. Appropriate mechanisms for the encoding of such interpretative gestures may be found in chapters 15 (‘Simple Analytic Mechanisms’) on p. 378 and 16 (‘Feature Structures’) on p. 394." (TEI P3, p. 483)

But in the almost ten years gone since the final version of P3 was published, the tools suggested for encoding persons and places as opposed to their names was clearly not sufficient for the users of TEI. In the current version of the TEI, P5 version 0.5, a similar paragraph is included at the same place in the guidelines, but with quite different wording:

"Finally, when this module is included, elements are provided to represent the abstract structures, or virtual objects to which names or dates may be said to refer. In simple terms, where the core module allows one to

represent a *name*, this module allows one further to represent a *personal name*. It also allows one to represent the *person* being named, and thus to encode biographical or other personal data for a wide range of applications, quite distinct from the names associated with such data." (TEI P5, chapter 20)

The line that is drawn between name and person in the citation from P3 above does not represent an absolute opposition. As an examples of this, present already in P3, consider the element *hand*. One may argue that the *hand* element is used in the description of the process of writing the document, and is thus external to the text in a way not commonly encoded in TEI apart from the specific interpretation chapters (15 and 16 in P3). Seen this way, it may be as close to a person as to a person's name.

A much clearer need for encoding information about persons was demonstrated through work on methods for the encoding of prosopographical data in TEI. Several element was included in P5 to encode information about persons, among them sex, faith and occupation. Elements for encoding events in a persons life is also included, such as birth and death. Work is currently being done to build up an an encoding system for places within TEI along the same lines, and a general event element is also being discussed.

TEI Ontologies SIG

In May and June 2004, there was a discussion on the TEI mailing list about prosopographical tags. This lead to a suggestion that detailed information about persons (physical and legal), dates, events, places, objects etc. and their interpretation could be marked up outside the text, and that this could be connected to on-going ontology work being done e.g. in the Museum community, such as the Conceptual Reference Model (CIDOC-CRM). The result of this was the establishment of a Ontologies SIG at the 4th annual members meeting of TEI in October 2004 (TEI Ontologies SIG).

During the meetings of the Documentation Standards Group of The International Committee for Documentation of the International Council of Museums (ICOM-CIDOC) in Gothenburg in 2006, liaison with other parties was discussed and the work in the TEI Ontologies SIG came up. The group expressed support for the work done by the Unit for Digital Documentation at the University of Oslo on the relationship between CIDOC-CRM and TEI, and hoped the work would continue. This support includes, of course, the work being done by other parties on the same topic.

During the two and a half years since the establishment of the SIG, work has mainly been done in relation to the CIDOC-CRM ontology. Work was reported and discussed at the Digital Humanities conference in Paris in July 2006 (Ore 2006). Based on the discussions in Paris as well as presentations and discussions at a meeting in Heraklion in Greece in October 2006 and at the SIG meeting in Victoria in October 2006, a draft mapping of a set of TEI elements to CIDOC-CRM was published in January this year (Eide 2007). Relevant work in this area has also been reported by Conal Tuohy (2005).

Gregory Crane of the Perseus Project said at the SIG meeting in October 2006 that he would make sure work will be done on the integration between FRBR and TEI, mainly the TEI header. We hope to receive reports on this work on the mailing list of the SIG later this year.

The way forward

The work of the SIG will continue, as well as connected work in other TEI bodies. I will take part in a TEI work meeting next week to discuss a TEI encoding scheme for places.

In our work on a mapping system from TEI to CIDOC-CRM, we have identified some problems reducing the potential benefit from such work. The discussion below is based on a conference poster to be presented at the Digital Humanities conference in June this year.

If a mapping from a TEI document into a model conforming with CRM is to be created, it will naturally be

based on a general mapping of TEI elements to CRM we are currently developing. But in TEI, many elements are defined quite loose, and depending on the way they are used, they may be modelled differently in CRM. According to the TEI guidelines, tag usage may be described in the TEI header. Such descriptions should then help us in deciding which type of modelling is the most appropriate.

Ideally, such a specific mapping should be created based on an automatic reading of the TEI header. But an element description in a *tagUsage* element in the TEI header is in prose and will generally not be stringent enough to be understood by an automatic reading (TEI P5, sec. 5.3.4). Human interaction will be needed. It may be the case that use of the *equiv* element will make automatic creation of mappings possible, as a reference to a certain CRM class may be included as an external link (ibid, sec. 6.3.4).

A CRM conforming model based on the TEI document and populated with all instances of mapped elements should then be created. This model may be used as a query or a data mining system where the user looks for interesting structures in the CRM conforming model alone, as well as in combination with textual information collected from the TEI source document. But this model may also be used in connection with other CRM conforming models, such as museum databases. The connections will be based on regional or global object identification, such as authority lists of names and classification schema. The resulting "super model" may then be used as a data mining tool based on semantic integration between heterogeneous resources.

We are currently developing the building blocks for a system based on this method. We believe this will improve the usability of TEI documents as information sources as well as simplifying the process of adding more information, such as event elements, into such documents. A general observation from our work is that the more relevant information types is marked up in an TEI document, the easier it is to use automatic methods to generate CRM conforming models. But even a limited tagging with only names and dates marked up do help in such an automatic model generation.

Bibliography

CIDOC (2003). *Definition of the CIDOC Conceptual Reference Model* / Produced by the ICOM/CIDOC Documentation Standards Group, continued by the CIDOC CRM Special Interest Group. ISO/DIS 21127. URL: http://cidoc.ics.forth.gr/de nition_cidoc.html (as of 2006-11-13)

Eide, Øyvind and Jon Holmen (2006). "Reading Gray Literature as Texts. Semantic Mark-up of Museum Acquisition Catalogues". CIDOC 2006, Gothenburg. URL: http://www.edd.uio.no/artiklar/teknikk_informatikk/CIDOC2006/EIDE_HOLMEN_Reading_Gray_Literature.pdf (cheked 2007-02-02)

Eide, Øyvind and Christian-Emil Ore (2007). *Mapping of TEI to CIDOC-CRM. Version 0.1 2007-01-02*. URL: http://www.edd.uio.no/artiklar/tekstkoding/tei_crm_mapping.html (checked 2007-02-02)

Ore, Christian-Emil and Øyvind Eide (2006). "TEI, CIDOC-CRM and a Possible Interface between the Two." P. 62-65 in *Digital Humanities 2006. Conference Abstracts.* Paris, 2006.

TEI P3 (1999). *Guidelines for Electronic Text Encoding and Interchange* / Edited by C. M. Sperberg-McQueen and Lou Burnard. Chicago, Oxford, 1994. Revised Reprint, Oxford, May 1999.

TEI P5 (2006). *Guidelines for Electronic Text Encoding and Interchange*. Ver. 0.5. URL: http://www.tei-c.org/release/doc/tei-p5-doc/html/ (checked 2006-11-13)

TEI Ontologies SIG. *SIG:Ontologies Wiki.* URL: http://www.tei-c.org.uk/wiki/index.php/SIG:Ontologies (checked 2007-02-02)

Tuohy, Conal (2005). "Topic Maps @ NZETC." URL: www.nzetc.org/downloads/TM@NZETC.ppt (checked 2006-11-13)

Interoperability in the context of i2010 Digital Libraries initiative

Stefan Gradmann, Ariane Labat

Our presentation at the DELOS-MultiMATCH event will briefly report on the perspective of the working group on Digital Library Interoperability which is part of the process of setting up the European Digital Library, a common access point that will make millions of objects from all kind of cultural institutions and of all types of format easily accessible for all European citizens by 2008 and beyond.

Indeed, there is no "spontaneous" convergence in the exposition of European cultural content from museums, libraries and archives (MLA) to the European citizens, but rather a fragmented access to European cultural content offered to the final users.
Actually, upstream to the users, the strategic orientations and the implementation solutions chosen by MLA institutions to expose their European cultural content to the European citizens are not spontaneously "converging" and especially not always interoperable.

As improving interoperability is one of the key actions to minimise fragmentation of access to European digital cultural heritage, the Commission has brought together and is chairing a group on "Interoperability and Multilingualism". This group, facilitated by Dr. Stefan Gradman, has the mission of producing recommendations for short term action as well as for a longer term strategy regarding Digital Library Interoperability and Multilingualism in Europe.
The group has the task to identify elements for a short term action plan regarding interoperability and multilingualism to achieve the 2008 objective of making 2 million objects amongst Europe's digital cultural heritage accessible online through a common multilingual access point. Besides, the group has to give recommendations for a longer term strategy, because such a longer term strategy is needed to give some stable perspectives to decision makers and developers in the cultural institutions and in the service providers, for them to anticipate and undertake investments or new developments.

In concrete terms, this group is gathered:
- to identify more precisely the determining factors of the "interoperability problems" between MLA legacy systems when aiming at creating a common access to European digital cultural heritage
- to propose a list of prioritised feasible options to lower the barriers towards more interoperability when creating a common access point to Europe's digital cultural heritage, to be released in 2008 and expanded in 2010 and beyond.

The group is therefore intended to be an environment that seed starts more intensive technical discussions within ongoing or future eContentPlus and IST projects:
>> The eContentPlus programme is there to support experimenting the use of existing semantic tools in MLA current environments.
>> The IST programme is there to support experimenting longer term approaches to solve interoperability issues in MLA future complex environments.

Capturing e-culture: Metadata in MultiMatch

Neil Ireson - University of Sheffield - n.ireson@dcs.shef.ac.uk
Johan Oomen - Nederlands Instituut voor Beeld en Geluid – joomen@beeldengeluid.nl

This position paper briefly introduces MultiMatch project and the current state of research regarding handing of heterogeneous metadata and approach towards interoperability. Several standards (Dublin Core, FRBR, MPEG-7, CIDOC) have been studied but none of these could meet the requirements defined by the project.

1. The MultiMatch Project

Our shared cultural heritage (CH) is an essential part of our European identity, transcending cultural and language barriers. The aim of the MultiMatch project is to enable users to explore and interact with online internet-accessible CH content, across media types and language boundaries, in ways that do justice to the multitude of existing perspectives. This will be achieved through the development of a search engine targeted for the access, organisation and personalized presentation of cultural heritage information. The development of the MultiMatch search engine can be divided into four areas:

Data Collection
- crawl the Internet to identify websites with CH information, locating relevant texts, images, audio and videos
- likewise identify relevant material via an in-depth crawling of selected CH institutions, accepting and processing any semantic web encoding of the information retrieved

Data Analysis
- automatically classify the results, in a semantic-web compliant fashion, based on document content, metadata, context, and on the occurrence of relevant CH concepts
- automatically extract relevant information which will then be used to create cross-links between related material, such as biographies, exhibitions of work, critical analyses, etc.

Indexing
- organise and further analyse the material crawled to serve focused queries generated from user-formulated information needs

Search and Retrieval
- interact with the user to obtain a more specific definition of information requirements
- organise and display search results in an integrated, user-friendly manner, allowing users to access and exploit the information retrieved regardless of language barriers

Within the scope of the project 4 languages (Dutch, English, Italian and Spanish) and 4 media types (text, images, audio and video) are considered.

2. Choosing the appropriate metadata representation

The project started to explore how to capture the dimensions of the data by providing an overview of current practice regarding knowledge representation in the cultural heritage domain. As metadata standards enable interoperability between systems and organisations that information can be exchanged and shared, the overview provided the basis for the approach towards interoperability that will be adopted within the MultiMatch project.

This work is documented in the deliverable D2.1: First Analysis of Metadata in the Cultural Heritage Domain. In order to systematically study current practices, the sub-domain definition advocated by the DEN (Digital Heritage Netherlands) and ePSINet (the European Public Sector Information Network[1]) was used. This study included a descriptive overview of the metadata schemas and semantic resources (i.e. thesauri, controlled vocabularies) widely used within the organizations belonging to the specific sub-domains.

A scheme or vocabulary is included only if the following criteria are met:

- it is constructed and maintained by a renowned institute in one of the sub-domains *and*,
- available in electronic form *and*,
- publicly available; in other words, there may be financial but no copyright hindrances to apply them in MultiMatch *and*,
- it is proven an international standard *or* a local standard, in use nationwide.

Forty metadata schemas and semantic resources have been identified and analyzed in a structured fashion. It became clear that the uptake of international established controlled vocabularies is quite limited. Local and nationally established/managed vocabularies are therefore predominant. Part of the reason for this is that the available international controlled vocabularies are still not available in every European language.

The table below lists the most relevant metadata Schema and Controlled vocabularies currently in use in the European cultural heritage sector.

	Schema	**Controlled vocabularies**
Archives	EAD and ISAD(G)	IPTC thesaurus, ISAAR (CPF), Thésaurus architecture et patrimoine, UK Archival Thesaurus
Libraries	FRBR, MARC, MODS and METS	DDC, UDC, LCSH and RAMEAU
Museums	CDWA, Object ID, VRA	AAT, ULAN, TGN
Educational sector	IEEE LOM	ERIC thesaurus
Audiovisual sector	P_META and SMEF-DM	-
Geospatial sector	CSDGM and ISO 19115:2003	-
Generic	CIDOC, DOI, DCMI, MPEG-7/21	URI, RFC1766, ISO3166, …

The methodology from De Sutter (et. al.) in their paper "Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries"[2] was used to select which standard could be used within the MultiMatch project. The following standards were selected for further analysis:

- Dublin Core: because it is in use through the whole of the cultural heritage domain.
- MPEG-7: because it can handle multimedia in a way appropriate for MultiMatch.
- FRBR: because it provides a data model with relationships and a hierarchy that are probably useful for MultiMatch. (Annex 3 includes the graphical representation of the FRBR entity-relationship model).

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

2 Sutter, R de. [et. al.] Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries. Published in: Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, Rafael C. Carrasco (Eds.): Research and Advanced Technology for Digital Libraries, 10th European Conference, ECDL 2006, Alicante, Spain, September 17-22, 2006, Proceedings. Lecture Notes in Computer Science 4172 Springer 2006.

- CIDOC CRM: because it provides a reference model for the cultural heritage domain. (Annex 4 includes the graphical representation CIDOC class hierarchy).

MultiMatch will handle various types of metadata:

1. Descriptive metadata – both metadata that formally describe the object (for example title, creator, creation date) as well as some semantic elements (for example subject keywords, geographic places);
2. Technical metadata – probably mainly concerning the surrogate or the image of the cultural heritage object, and less concerning the physical cultural heritage object itself;
3. Administrative metadata – some metadata to administer the objects concerned (e.g. content provider name, location information, language of record, record number), possibly also some metadata for the rights management. For instance, the extent to which metadata on copyrights are needed within the central metadata schema is at this point not clear.
4. Examples of typical metadata that will probably not be the focus of the MultiMatch metadata schema include administrative and technical data on museum objects that are needed for the internal management of the museum collection (typically gallery and museum information systems)

The next step was to choose/built upon these standards and design the appropriate metadata representation approach for the project. The main factors that influenced this work can be divided into three areas;

- Meeting the specification of the user requirements,
- Representing the concepts which are present in the data and
- Interoperability: mapping from content provider legacy metadata and to current "standard" metadata schemas.

2.1 Meeting the specification of the user requirements

The MultiMatch project is following a user-centred design strategy and so a good deal of the initial effort has been directed at interviewing the end-users of the system to determine their requirements. These user requirements are translated into a functional specification which is used to guide the development of MultiMatch so that the final system will meet the needs (and expectations) of the end-users. Obviously there is not a singular concept of an end-user, and different users having different levels of expertise and information search needs will present different requirements which should be considered in the system development.

In terms of the influence on the metadata, the user requirements give and indication as to the concepts which are required by the user and the relative importance of those concepts.

Many of the findings of the user requirements are, as one would expect, intuitive, such as the fact text and images are the primarily important media types, and that creations and their works are the main concepts in search, with typical queries involving: proper names, places, titles, general subjects. There was also less definite requirements, such as clearer semantics in faceted browsing and clustering so that the categorisation (clustering) of the search results "make sense". This strongly points to the need to use standard (and well defined) vocabularies and subject taxonomies.

Given that in MultiMatch all search facilities will be **translingual**, i.e. the user will formulate queries in a given language and retrieve results in all languages covered by the prototype. Users

expressed a strong desire to have control over the process, i.e. selection of the appropriate translation in the case of a word with multiple meanings and the ability of not translating certain expressions.

The notion of a controllable and transparent search process was seen as a general requirement on the system, a key reason for this is the ability to determine the “authoritativeness” of any information which results from a user search.

From the expert users survey we can conclude that, on average, experts tend to classify searches for information about creators (authors) and creations (works of art and masterpieces) as their most common search tasks. Therefore, in MultiMatch we have initially decided to focus two types of specialized searches on creators and creations, although specialized searches focused on other relevant categories will also be considered.

2.2 Representing the concepts which are present in the data

Whilst it is crucial to provide a representation to fulfil the user requirements, these will not fully specify the requirements of all users; as only a selection are interviewed, needs are dynamically changing and users concepts of what they want are sculpted by their experience of what is available. In addition, by examining the data it is possible to determine the issues which arise when representing the concepts relating to cultural heritage objects.

Therefore the metadata should consider the ability to adequately represent the concepts which are extractable from the data itself, so that concepts (including possibly unforeseen concepts) can be suitably represented. In the MultiMatch project this data includes direct representations of Cultural Heritage objects (images, texts, etc.) and data describing those objects, which is presented in a human rather than machine readable format.

Within MultiMatch there is an obviously need to represent the creator(s) of an art work; this can include information such as: names, birth/death place and time, colleagues, etc. However when examining the data it is seen that values can; be ambiguous (an individual can be known by several names), have varying degrees of precision (i.e. country, province, town) or may be contested (such as whether the artist actually created the work). Also associations can have varying degrees (artists may collaborate over many years or merely on a single art work).

The ambiguity, imprecision and uncertainty of data is accentuated by the multilingual nature of the data and fundamentally by the use of automatic techniques to extract information form the data. Where a concept is seen to have conflicting values this could be due to a genuine difference of opinion in the CH domain, an error in the data or an error in the extraction process.

One of the key features which was highlighted by the user requirement analysis is the need to represent the authoritativeness of the information presented to the user. It is therefore important for the MultiMatch representation to be able to express the ambiguous, imprecise and uncertain nature of the information to the user. Also it is potentially useful/necessary to provide an “audit-trail” to the source(s) and process(es) which have been used to acquire the information.

In addition to the MultiMatch project extracting metadata from textual data, it also extracts metadata from other media typesaudio, still images and video. Although to an extent audio (transcripts) can be seen as noisy text and video is seen as (keyframe) images and audio (text).

The metadata which is extracted from images represents the underlying, low-level features of an image however what the user requires is the linking of this to a meaningful (semantic) representation of the image. Thus the metadata format must cope with low-level and high-level concepts and the links between these representations.

2.3 Metadata Interoperability

Interoperability is concerned with the capability of different information systems to communicate. This communication may take various forms such as the transfer, exchange, transformation, mediation, migration or integration of information.

Cultural heritage organisations employ professional cataloguers to annotate the objects in their collections. A metadata model and cataloguing rules formalises the way objects are described in the catalogue. Thesauri are often the basis to assign keywords to objects. The situation gets more complex if the goal is to create connections between objects and between collections, even more so if this process needs to be automated to the largest extent possible. It involves matters of syntactic (i.e. schema-level) and semantic interoperability; where both are essential.

Semantic interoperability is characterised by the capability of different information systems to communicate information consistent with the intended meaning of the encoded information (as intended by the creators or maintainers of the information system). It involves processing of the shared information so that it is consistent with the intended meaning and encoding of queries and presentation of information so that it conforms to the intended meaning regardless of the source of information.

So, the basic interoperability questions MultiMatch has to deal with are related to:

- Automatic extraction of metadata field. Will the MultiMatch technology, which we are going to build, be able to generate the metadata fields that are in the MultiMatch metadata schema?
- Mapping of original metadata into the MultiMatch schema. The project will have metadata already available in the content providers databases to be integrated with the MultiMatch schema. Is this mapping feasible?
- Support semantic interoperability. As MultiMatch covers a wide domain, this is a big challenge. How will this be reached? How are multiple languages supported?
- Role of ontologies. Generic metadata schema, like Dublin Core, reveal the existence of overarching concepts. This might be the basis of creating an ontology to create interoperability between individual schema.

In the research conducted by MultiMatch, it became clear that next to the Dublin Core element set, also the entity-relationship model FRBR and CIDOC conceptual reference model can prove valuable solutions to provide the level of interoperability required.

3. Conclusion

The factors influencing the design of the MultiMatch metadata schema come (“top-down”) from requirements of the user, (“bottom-up”) from the data and information extraction process and from the need to be interoperable within the cultural heritage domain. There is no single “standard schema” which meets all these requirements. In terms of interoperability Dublin Core

(DC) is the most widely used standard, and whatever metadata schema is implemented the relevant parts must be mapped on the DC. A number of standard refinements of DC are also available (DCMI, VRA, etc.) and providing the ability to map to these increases the information which can be transferred. The area of metadata for low-level multimedia features is an evolving one although the MPEG-7/21 descriptors seem to offer a powerful (and increasingly standard) way of representing and communicating such information. Finally there are reference models such as CIDOC CRM and FRBR which offer an over-arching representation of the CH domain.

In terms of semantic interoperability the desire is to provide a means which allows for flexibility and coverage (providing information to the widest possible user base) and expressiveness (is as informative as possible). Currently mapping to and from the Dublin Core provides the widest coverage however much of the richness of information which is within the MultiMatch metadata will be lost, thus it is necessary to provide the metadata in more expressive representations (MPEG-7, VRA, CIDOC) allowing the receiver to utilise such information, obviously ensuring that the semantics behind the metadata schema are clearly stated and adhered to during the process of populating the metadata.