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".