Semantic Web Integration of Cultural Heritage Sources

P. Sinclair, P. Lewis, K. Martinez
Electronics and Computer Science,
University of Southampton,
SO17 1BJ,
United Kingdom
+44 (0) 23 8059 3255
{pass,phl,km}@ecs.soton.ac.uk

M. Addis, D. Prideaux
IT Innovation Centre,
Southampton,
SO16 7NP,
United Kingdom
+44 (0) 23 8076 0834
{mja,djp}@it-innovation.soton.ac.uk

ABSTRACT

In this paper, we describe research into the use of ontologies to integrate access to cultural heritage and photographic archives. The use of the CIDOC CRM and CRM Core ontologies are described together with the metadata mapping methodology. A system integrating data from four content providers will be demonstrated.

Categories and Subject Descriptors

H.3.3: Information Search and Retrieval.

General Terms

Design, Experimentation, Standardization.

Keywords

Multimedia, ontologies, semantic web, interoperability

1. INTRODUCTION

Cultural heritage institutions and photographic libraries are rich content resources, depicting people, objects, events, places and monuments. Making this material accessible requires rich metadata structures, able to capture the diversity of the media, the subject matter and the historical context around each information asset. This information tends to be 'locked away' in internal legacy systems, each with its own metadata format that has been designed to deal with a specific collection or set of objects.

In the eCHASE project [1] the CIDOC Conceptual Reference Model (CRM) [2], in particular the recent CRM Core proposal is being used as the common model for different multimedia collections. By mapping the metadata which exists in each collection to a common ontology interoperability has been achieved across diverse collections. This allows not only the unified access sought by users but also introduces new capabilities due to the preservation of the rich interrelationships between information. This paper concentrates on the issues of mapping metadata to the ontology and the poster/demo shows the running system.

2. SEMANTIC INTEROPERABILITY

2.1 Ontologies

The CIDOC CRM is an extensive ontology for the semantic integration of cultural information, including library, archive and other information. It has been in development by members of the International Committee of Documentation (CIDOC) of the International Council if Museums (ICOM) since 1996 and has been accepted as an ISO standard (ISO/FDIS 21127). The CIDOC CRM defines 80 classes and 130 relationships that comprise the most characteristic concepts required for museum, archive and library documentation.

CRM Core is a recent proposal from CIDOC for a highly condensed set of metadata elements that capture the most fundamental relationships connecting things, concepts, people, time and place. CRM Core can be expressed in a Dublin Core compatible format that, unlike Dublin Core, is able to precisely model the complex, event-based cultural heritage information.

2.2 Accessibility

One of the key issues being addressed in the eCHASE project is that cultural heritage media is not accessible in a unified way. Information is distributed across different museums, galleries, art libraries and audiovisual archives, typically 'locked away' in internal legacy systems. When available, online access to the material is generally restricted to web sites, each with its own user interface and exposing different levels of searching and browsing functionality. Being able to obtain metadata in a machine-readable format is uncommon and often metadata formats are used which hide the rich semantics of the information.

In the eCHASE project, we are using the Z39.50 Search and Retrieve Web service (SRW) [3] extended in the Sculpteur project [4]. The SRW provides a web service interface to the information via the CRM ontology. If it is installed in a content-provider's site it can be used to harvest their information. In our demonstrator we also use it as the main interface into a central system which has harvested data or has regular uploads from the providers.

Although the CRM was expressive enough to capture the semantics of the data, the museum partners had great difficulty understanding how their metadata should be linked to the full CRM, as this results in a highly interconnected network of concepts and properties. Mapping to the CRM-code is simpler because there are fewer concepts to understand so that is our current approach. In the cultural heritage domain, there is often ambiguity on how the metadata fields for each museum should be mapped to the CRM. If different museums and galleries make different interpretations of the CRM even if their data is semantically equivalent, then there is the potential for interoperability to be lost when the data is exposed through the SRW using different mappings. In this case, a burden is placed on the client application to reason over the CRM mappings used by different museums to determine if they are semantically equivalent. The SRW does not support the ability to do this reasoning, so it is up to the client application to implement this functionality. In the same way that difficulties can arise with achieving consistent data value semantics, this is not a inherent problem with the SRW and CRM per se, rather it is an issue of ensuring that everyone uses an agreed mapping and transformation process. Again, with a semantic web approach there are potentially more options for dealing with this problem if it does occur through post processing and correction of the RDF and identification of equivalent concepts and assertion as such. Semantic web query languages, such as RDQL and SPARQL, are more expressive than CQL (used in the SRW) and this would also overcome some of the problems by providing more support for client applications. Ontological information can also be used such as concept inheritance, where a concept subclass in a mapping chain can be regarded as equivalent to the concept in another chain.

Mapping legacy data schema to a common model such as the CRM is not enough to achieve full semantic interoperability. The data values used in different museum and gallery legacy systems also need to be rationalised and harmonized. This is partly a data cleaning issue to do with misspellings, syntactic differences and poorly structured data. However, part of the problem is also the need for a consensus of agreement on common semantics in the cultural heritage domain for people, places, events and so on. Neither the SRW or the CRM impose any requirements on the semantics of data values since they are only concerned with the schema level. Therefore, care needs to be taken to deal with this issue either at data import time through a data cleaning and value mapping process, or when consuming data from the SRW in a client application. This problem is common across both relational databases and semantic web stores. Arguably, there are more opportunities to tackle the problem when the data is transformed into RDF, both during the export process and using post-export techniques such as co-reference resolution to consolidate the information in different collections. Indeed, this could provide one way to address the problem by building an RDF store of data that is sourced from multiple SRW servers.

2.3 Processing

The reality of mapping the original metadata to terms in an ontology is complicated by the fact that processing is needed to make the incoming text uniform enough. For example dates typically need reformatting but we have also found place names embedded in longer text, which need to be identified. Some archives also maintain a list of keywords which may need contextualising using their own thesaurus or controlled terms list.

Data was delivered as a set of images with corresponding metadata. The metadata was in a variety of formats: XML, Excel Spreadsheets, EXIF metadata embedded in the images and SQL Server dumps. We had to first import all of these formats into a MySQL database so that it could have a consistent mechanism to access the metadata. This required developing importers for each of the different formats.

By relating words in the original metadata to related thesauri it is possible to make a stronger relationship or further links to information elsewhere. For example an artist name can be referred to its entry in the United List of Artiste Names (ULAN) which brings with it relationships to places and dates. Geographic thesauri also allow places to be related properly to their larger region or country for example.

Taverna [5] was used to create a workflow for each data source with the appropriate stages of formatting, cleaning and mapping. This makes importing a new batch of material much easier and also modularises the process.

3. DEMONSTRATOR

In the eCHASE project, we are working with several major content holders, including Fratelli Alinari and Istituto Geografico De Agostini (Italy), the Houlton Archive from Getty Images (UK) and Österreichischer Rundfunk (Austria). We have used the methodology described above to integrate data from each partner, process the metadata and storing it in a common repository. This is accessible via the SRW which presents the information in terms of the CRM ontology.

A web client demonstrator has been designed to allow detailed searching as well as browsing and content-based retrieval of the images. Features to allow user's own sub-collections to be stored and exported while maintaining the semantic links have been implemented.

Currently this search engine is only available to the consortium partners due to issues with the copyrighted content, although we are working on the release of an open access version of the system. The search engine will be demonstrated at the conference.

4. ACKNOWLEDGEMENTS

This research has been supported by the eCHASE project which is co-funded by the European Commission, DG Information Society, under the contract EDC 11262, within the "European digital content for the global networks" programme.

5. REFERENCES

[1] eCHASE; electronic Cultural Heritage made Accessible for Sustainable Exploitation, http://www.echase.org

[2] Doerr, M.: The CIDOC Conceptual Reference Model: An ontological approach to semantic interoperability of metadata. AI Magazine, 24 (2003) 75–92.

[3] Z39.50 SRW: http://www.loc.gov/z3950/agency/zing/srw/ (2005)

[4] Sinclair, P., Goodall, S., Lewis, P., Martinez, K. and Addis, M. (2005) Concept browsing for multimedia retrieval in the SCULPTEUR project. In Proceedings of The 2nd Annual European Semantic Web Conference, Heraklion, Crete.

[5] Taverna: http://taverna.sourceforge.net/

WWW 2006, May 22–26, 2006, Edinburgh, UK.