Semantically Exposing Existing Knowledge Repositories: A Case Study in Cultural Heritage

Denis Pitzalis, Patrick Sinclair, Christian Lahanier, Matthew Addis, Richard Lowe, Shahbaz Hafeez, Paul Lewis, Kirk Martinez, mc schraefel, Ruven Pillay, Geneviève Aitken, Alistair Russell and Daniel A. Smith

***Abstract*— In this paper we describe the practical implications of semantically exposing a cultural heritage multimedia collection system (EROS) through a Search and Retrieve Web Service (SRW).**

***Index Terms*— multimedia system, semantic web, cultural heritage**

I. INTRODUCTION

Semantic web technologies have the potential to greatly benefit the Cultural Heritage (CH) domain. CH institutions, such as museums and photographic archives, are rich resources of heterogeneous multimedia content, depicting people, objects, events, places, etc. This material, along with any supporting metadata, tends to be locked away in internal legacy systems, and open interfaces to the collections are rarely provided.

The use of semantic web technologies could make an impact on several levels. Richer semantics can greatly improve the information systems used by conservators, curators and historians by enhancing the retrieval and browsing facilities. Making the data available through semantic web services could provide opportunities for tackling complex research problems in the CH domain.

However, there are still barriers for applying semantic web technologies directly. Many CH institutions are tied in to their commercial content management systems. There are also high costs in converting and mapping all of their existing material to semantic representations such as RDF. Although some of the technical issues such as triple store scalability are being overcome, many still have doubts about the applicability of semantic web technologies in practice. Alternatives that bring semantics to traditional content management systems are desirable in this context.

II. CASE STUDY

The C2RMF is the Research and Restoration Centre of French Museum located in the Louvre. It's mission is to analyse, restore and document the works of art kept within all museums in France. This work requires the management of huge quantities of different kinds of data. To organise our digital library we developed the EROS system[1][2]. This system consists of a relational multilingual database that allows us to organise different media: at the moment over $250,000$ photographic and radiographic images, $10,000$ technical reports, $1,000$ 3D objects, $200,000$ quantitative chemical and physical analyses related to more than $60,000$ works of art are accessible in digital form. This heterogeneous group of data is common in real world applications.

Semantic interoperability of CH digital libraries has been investigated in the SCULPTEUR[5] and eCHASE[6] projects by using a z39.50 search and retrieve web service (SRW[3]) and by mapping legacy metadata schemas to the CIDOC Conceptual Reference Model (CRM[4]), an ontology for describing the semantics of CH documentation. Additional semantics are attached to the legacy database attributes in order to more fully define their meaning in the context of the CRM framework. The CRM mapped attributes are exposed through the SRW as a flat list that can be queried by using Common Query Language (CQL) expressions. The SRW publishes the mapping information in XML through the SRW explain operation. The SRW is able to dynamically map CQL queries expressed in terms of the CRM mappings to the relevant legacy database fields (in our case using SQL against a relational database) and return the results as XML structured according to the CRM mappings.

Our SRW implementation is available as open source in the form of OpenMKS (http://openmks.sourceforge.net), which provides an SRW implementation that allows relational data to be mapped to an XML representation. It also provides a web-based user interface to the SRW that allows end users to search and browse the content. Through the configuration system we were able to adapt the system to the EROS content and metadata within C2RMF.

mSpace[7] is an interaction model and software framework to help people access and explore information. mSpace helps people build knowledge from exploring relationships in data. mSpace does this by offering several powerful tools for organising an information space to suit a persons interest: slicing, sorting, swapping, information views and multimedia preview cues. When we access a subset of the EROS data set through the mSpace interface each category in the information space is displayed in a separate column, and the selection in each column narrows down the results presented in the next

D. Pitzalis, C. Lahanier, R. Pillay and G. Aitken are with Centre de Recherche et de Restauration des Musèes de France, Palais du Louvre, Paris, France. Email: {name.surname}@culture.fr

P. Sinclair, P. Lewis, K. Martinez, mc schraefel, A. Russell and D.A. Smith are with Electronics and Computer Science, University of Southampton, UK. Email: {pass,phl,km,mc,ar5,das05}@ecs.soton.ac.uk

M. Addis, R. Lowe and S. Hafeez are with IT Innovation Centre, Southampton, UK. Email: {mja,rl,szh}@it-innovation.soton.ac.uk

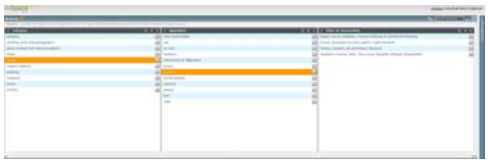

Fig. 1. Subset of the EROS data set displayed through the mSpace interface

column. mSpace has been designed to be independent of the backend database and while the original mSpace server relied on an RDF triplestore, the flexibility of mSpaces data access protocol has been utilised in this project to provide an mSpace to a relational database exposed through the SRW.

III. DISCUSSION

The user can explore the CRM ontology and then use the SRW/CQL to retrieve corresponding instances. In this way we leverage Semantic Web techniques to describe the complex space of CH information, whilst using XML and Web Service standards to provide an easy to use search and retrieval service to access this information. This is a trade-off between the complexity of queries that can be formulated and the need for a simple query language that makes it easy for third-parties to develop their own client applications. Whilst the SRW/CRM solution is relatively easy for both content-providers and end-user application developers to understand and use, this is at the expense of the expressivity of semantic query languages and the ability to use server side reasoning.

Whilst the use of SRW on top of relational legacy data sources is scalable to the large datasets often held by CH institutions, it does not necessarily provide the performance needed for highly interactive user querying of this data. In other words, our use of the SRW and CRM is geared towards semantic interoperability of multiple heterogeneous datasets, not high performance retrieval needed for interactive data exploration of these datasets. If a high degree of user interactivity is required for large datasets, for example by using mSpaces to explore the EROS database, then specific additional optimisations are typically necessary. The need for, and the choice of, a suitable performance optimisation strategy is not a result of our decision to use SRW, CRM mapping or CQL per se, but is more a reflection on the way that the underlying legacy data is structured, stored and searched.

IV. CONCLUSION AND FUTURE WORK

We have described how we have semantically exposed a CH multimedia repository, EROS, through the SRW and how we integrated the mSpace interaction framework. There are still barriers to the practical use of semantic web technologies in the CH domain, and this approach enables some of the benefits to be explored whilst still supporting the existing infrastructure.

Many of the issues we have encountered are due to the scale of real world collections, such as the EROS system. To overcome part of these problems we decided to implement a simple caching mechanism on the mSpace SRW server, which improved overall performance once a query had been made. Unfortunately, due to the vast size of the EROS data set, some of the queries take a long time to complete by the SRW so further optimisation will be investigated in the future. As such we will be investigating optimisations of the SRW, and study how the underlying database schema could be optimised and improved without causing a huge impact.

We believe that the integration of semantically-based interaction paradigms, such as the mSpace framework, with legacy data management systems is extremely valuable. Not only does this provide rich browsing and navigation functionality that tends to be overlooked in many traditional systems, it show cases the benefits of semantically marked up information in a tangible way. This allows users to serendipitously discover artefacts and media that they would never have found through a traditional search box. It is also a great way of illustrating many of the data quality issues present in many metadata systems, as errors and inconsistencies are highlighted when the data is presented in an interface such as mSpace.

As part of our future work, we are investigating the integration of the EROS system with the bibliographic records in the C2RMF library. This will draw on the work by the CIDOC CRM working group on the alignment of the UNIMARC standard to the CIDOC CRM. In the context of our longer term goals, that is providing cross-collection searching and browsing of disparate multimedia sources in the CH domain, we are working on the harmonization of the data from different collections. In the eCHASE project, we are integrating the collections of several large CH institutions, including picture libraries, television archives, publishers and we hope to attract museums and galleries over the coming months. This requires aligning the different data representations, ranging from time and date, places, identifying the people across collections and categorization schemes such as controlled lists and thesauri.

ACKNOWLEDGEMENT

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. We would also like to acknoweldege the EPOCH network of excellence (IST-2002-507382).

REFERENCES

[1] Aitken, G., Lahanier, C., Pillay, R.,Pitzalis, D.: “Database Management and Innovative Applications for Imaging within Museum Laboratories” 7th European Commission Conference ”SAUVEUR”, June 2006, Prague, Czech Republic

[2] Aitken, G., Lahanier, C., Pillay, R.,Pitzalis, D.: “EROS : An Open Source Database For Museum Conservation Restoration Preprints for the 14Th Triennial Meeting ICOM-CC, J&J London, 2005, The Hague, Netherlands”

[3] z39.50 SRW: http://www.loc.gov/z3950/agency/zing/srw/ (2005)

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

[5] Addis, M. J., Martinez, K., Lewis, P., Stevenson, J. and Giorgini, F.: “New Ways to Search, Navigate and Use Multimedia Museum Collections over the Web” In Proceedings of Museums and the Web 2005, Vancouver, Canada. Trant, J. and Bearman, D., Eds. z39.50 SRW: http://www.loc.gov/z3950/agency/zing/srw/ (2005)

[6] “eCHASE project”: 2004-2006 eContent no. 11262. www.echase.org.

[7] m. c. schraefel, D. A. Smith, A. Owens, A. Russell, C. Harris and M. Wilson: “The evolving mSpace platform: leveraging the semantic web on the trail of the memex” Proceedings of the sixteenth ACM conference on Hypertext and Hypermedia, ACM Press, Salzburg, Austria, 2005