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

Denis Pitzalis[1], Patrick Sinclair[2], Christian Lahanier[1], Matthew Addis[3], Richard Lowe[3], Shahbaz Hafeez[3], Paul Lewis[2], Kirk Martinez[2], mc schraefel[2], Ruven Pillay[1], Geneviève Aitken[1], Alistair Russell[2], and Daniel A. Smith[2]

[1] Centre de Recherche et Restauration des Musèes de France,
Palais du Louvre,
Porte de Lions,
14 quai F.Mitterrand,
75001 Paris,
France
{denis.pitzalis,christian.lahanier,ruven.pillay,genevieve.aitken}@culture.fr
[2] Electronics and Computer Science,
University of Southampton,
SO17 1BJ,
United Kingdom
{pass,phl,km,mc,ar5,das05}@ecs.soton.ac.uk
[3] IT Innovation Centre,
Southampton,
SO16 7NP,
United Kingdom
{mja,rl,szh}@it-innovation.soton.ac.uk

Abstract. In this paper we describe the practical implications of semantically exposing a collection management system (EROS [1][2][3][4]) for large cultural heritage multimedia repositories. This is achieved through a Search and Retrieve Web Service (SRW[5]) that maps the EROS metadata schema to the CIDOC Conceptual Reference Model[6], a core ontology for describing the semantics of schema and data structure elements in cultural heritage documentation. The mSpace([18]) framework, an interaction model to help people access and explore information, has been integrated with the SRW to provide powerful navigation facilities on the rich multimedia collections served by EROS. We also present some of the lessons learnt whilst adapting the technologies to the EROS system and offer some insight into the performance issues with a system of this size.

1 Introduction

Semantic web technologies have the potential to greatly benefit the cultural heritage domain. Cultural heritage institutions, such as museums and photographic archives, are rich resources of heterogeneous multimedia content, depicting people, objects, events, places and so on. This material, along with any supporting metadata, 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. Open interfaces to the collections are also rarely provided, leaving search engine access to HTML pages as the only way in.

The use of semantic web technologies could make an impact on several levels. Interoperability between disparate multimedia collections can be improved, allowing users to search for related material across different collections. Richer semantics can greatly improve the information systems used by conservators, curators and historians by enhancing the retrieval and browsing facilities. Relationships between artifacts can be explored in greater detail and become better understood. Using semantic markup for the metadata format also enables richer presentation possibilities. For example, it becomes possible to represent artists as people rather than simply a name so that the richness of the biographical data about that artist can be revealed. Finally, making the data available through semantic web services could provide opportunities for tackling complex research problems in the cultural heritage domain, perhaps by allowing the data to be exploited by agent-based systems.

However, there are still barriers for applying semantic web technologies directly. Many cultural 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 might bring semantics to traditional content management systems are desirable in this context. We have been developing tools that allow us to semantically expose information stored in relational databases. In this paper we describe how we applied these tools to a real world content management system.

2 Background Work

D2R Server [15] provides SPARQL access to relational databases by transforming SPARQL queries into SQL via mappings specified in the D2RQ mapping language. This allows RDF-based applications to access the content of relational databases directly, without having to export to an RDF triple store.

D2RQ [16], the mapping language used in D2R Server, describes the relation between relational database schemas and OWL/RDFS ontologies. An ontology is mapped to a database schema using class maps, which typically map database tables to classes in the ontology, object property bridges, which typically map relations between instances of classes and datatype bridges, which construct RDF literals from database values.

R2O [17] is a declarative XML-based language that describes mappings between ontology elements (classes, attributes and relations) and relational elements (relations and attributes). The R2O approach acknowledges the trouble with poorly structured information in existing databases, and provides powerful mechanisms to select and transform instance data in the mapping specification. R2O has been applied using the ODEMapster processor, which offers two modes of execution: query driven, that is on the fly query translation, and massive upgrade batch process which exports all instances expressed in the ontology.

3 Case Study

During the Sculpteur([7][8]) project we investigated the semantic interoperability of Cultural Heritage digital libraries through the use of a z39.50 Search and Retrieve Web Service (SRW) and mapping legacy metadata schemas to the CIDOC-CRM, a core ontology for describing the semantics of schema and data structure elements in Cultural Heritage documentation.

Researchers at Southampton and at C2RMF have been collaborating on applying the technology initially developed for SCULPTEUR, and now being further developed in the eCHASE([9][10]) project, to the multimedia repository at C2RMF. Furthermore, recent developments on mSpace, an interaction model and software framework to help people access and explore information, has enabled the direct integration of this software to the SRW. This allows the C2RMFs valuable multimedia collection to be explored through the rich mSpace interface.

This work has also allowed us to further explore the issues that emerge when a database to semantic web mapping tools are applied to a large, real world multimedia collection.

3.1 C2RMF

The Centre de Recherche et de Restauration des Musées de France (C2RMF) was created in 1931 with a mission to study, catalogue and help preserve works of art kept within all of the museums of France. The main laboratory is housed within the Louvre Museum in Paris with a smaller laboratory within the grounds of the Palace of Versailles. The laboratory performs not only restoration, but scientific studies, especially chemical and materials analysis, imaging (X-ray, infrared, multispectral etc) and even has a particle accelerator.

The C2RMF has been an active participant in 11 EU projects (within the ESPRIT, IMPACT, RAPHAEL, IST and e-CULTURE frameworks) since 1989. These projects have helped transform the way the C2RMF gathers, handles and uses its digital data while providing real user-input into the semantic web research at Southampton. These projects have involved a wide range of topics, encompassing many aspects of the work carried out in the Centre. Many projects have also tried to build on and extend the work carried out in previous projects, resulting in a coherent and inter-dependent body of work.

The C2RMF was a pioneer in 1990 in high definition digitisation of images for research and conservation and has a huge archive of museum-related scientific photography and documentation. We have developed an open source relational database management system named EROS (European Research Open System) to retrieve images, reports and analysis. The system is now used by several institutions involved in conservation. Translation of the interfaces, terms for indexing and their definitions allows the user to access the content in several languages. The data in the system are composed of a wide range of types:

- metadata related to the works of art, images, reports, analysis, analytical reports, restoration reports, conservation surveys, chemical, structural, isotopic and molecular quantitative and qualitative analytical results and published papers;
- high definition digital images: photographic films taken with different techniques such as infra-red, X-ray and ultra-violet light, detailed cross-sections, electron microscopy views, graphs, spectra, multispectral images, panoramic views, movies, audio and 3D models;
- feature vectors for 2D and 3D image content recognition for automatic classification and image category retrieval.

3.2 EROS - European Research Open System

Currently, over 300000 photographic and radiographic images, 10,000 technical reports, 1000 3D objects, 200000 quantitative analyses related to more than 60000 works of art are accessible online in digital form. This heterogeneous group of data is common in real world applications.

The system is organized in different parts: the storage back-end, the relational database, the image server, the middleware and the web server (Fig. 1).

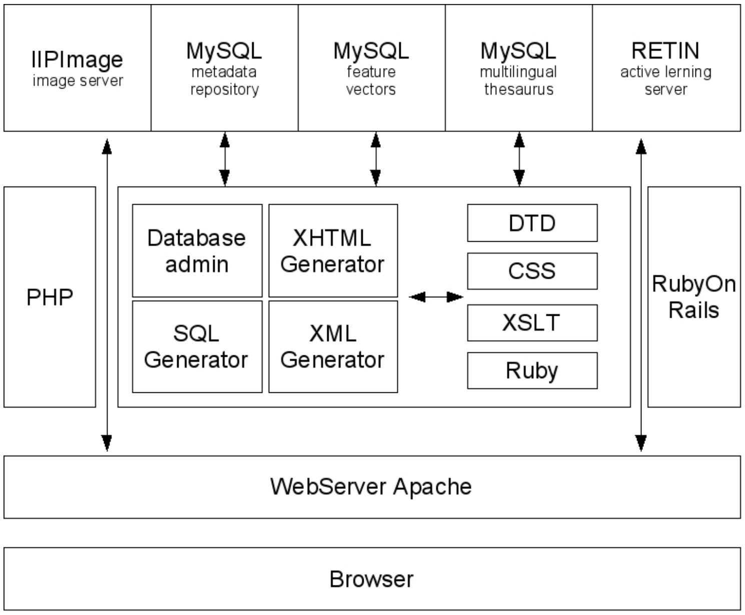


Fig. 1. Structure of the EROS System

The tables in the database are organized as:

- administration, historical and material data related to the objects;
- photographic films and digital images;
- documents such as reports, articles, multimedia etc.;
- quantitative and qualitative analyses of the material composition.

The multimedia content like audio, video, pdf is not stored directly in the database but it is accessible by a link. This could be dangerous in structural changes, but more efficient in term of performance. The images are accessible by an image server developed in order to provide quick access to huge images (IIPImage [11] available as GPLv2 software from http://iipimage.sourceforge.net).

The EROS system is in effect composed of two different databases, the main one contains all the data in the form of alphanumerical codes and links, and the other one is the Thesaurus which contain all the translations in 13 languages (not all the terms are translated, some free text notes are left as is). The links point to external sources like the various images available (for example normal light, UV, IR, X-Ray, raking light, cross sections, scanning electron microscopy images, spectra, graphs, panoramic views, multi-spectral images and 3D models etc.) and documents that are available in numerous formats such as DOC, HTML, PDF etc.

3.3 Semantic Mapping Through SRW

Semantic interoperability of Cultural Heritage digital libraries has been investigated in the SCULPTEUR and eCHASE projects by using a z39.50 search and retrieve web service (SRW) and by mapping legacy metadata schemas to the CIDOC CRM. 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 [14]) expressions. The SRW publishes the mapping information in XML through the SRW explain operation. The SRW is able to dynamically map Common Query Language (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.

The CRM ontology itself is available in RDFS and may be used by client applications to manipulate the mappings and query results expressed in the CRM. In this way, legacy datasets can be mapped and exposed in a semantically interoperable way that allows the data to be searched and retrieved by client applications. The use of CRM mappings to establish common field semantics, the use of SRW as a Web Service based search and retrieval protocol, the use of CQL to provide a simple query language, and the use of XML for syntactic interoperability all combine to hide the user from the complexities and heterogeneity of the multitude of different data structures used by museums and galleries for their metadata.

Our SRW implementation is available as open source in the form of OpenMKS. This provides an SRW implementation that allows relational data to be mapped to an XML representation, including CRM Core. The OpenMKS software is freely available from http://openmks.sourceforge.net and the easy configuriguration system allow the user to simply adapt to his database, whitch was the case for the EROS content and metadata within C2RMF. The OpenMKS SRW allows the mSpace system (see below) to access the relational EROS data without having to duplicate that data in RDF (Fig. 2).

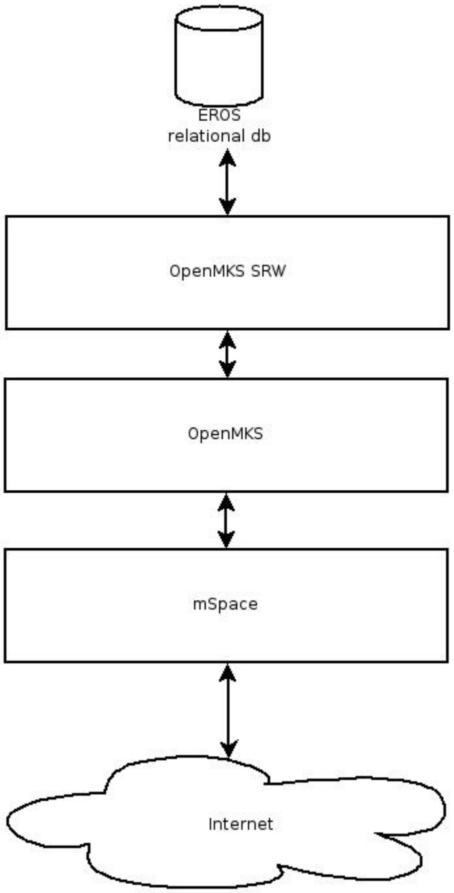
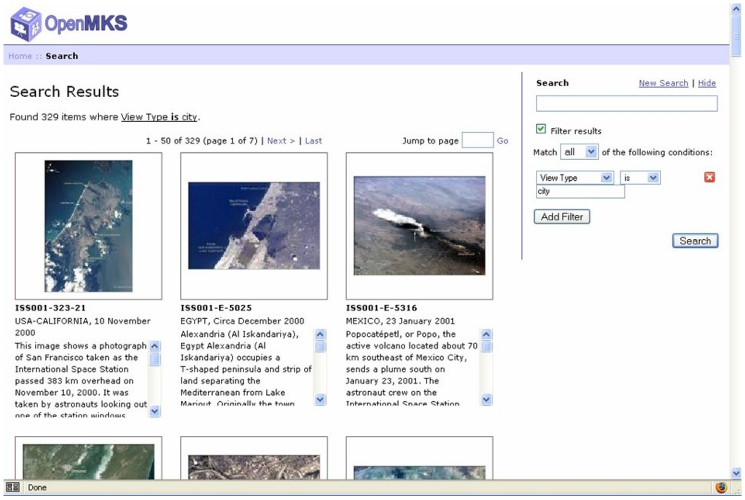


Fig. 2. System Diagram

3.4 OpenMKS Search Interface

OpenMKS also provides a Web-based user interface to the SRW to allow end users to search and browse the content that is made available. A demo application is provided that uses the “Earth From Space” dataset from NASA (Fig. 3):



Search Results

Found 329 items where View Type is city.

1 - 50 of 329 (page 1 of 7) | Next > | Last   Jump to page Go

ISS001-323-21
USA-CALIFORNIA, 10 November 2000
This image shows a photograph of San Francisco taken as the International Space Station passed 383 km overhead on November 10, 2000. It was taken by astronauts looking out one of the station windows

ISS001-E-5025
EGYPT, Circa December 2000
Alexandria (Al Iskandariya), Egypt Alexandria (Al Iskandariya) occupies a T-shaped peninsula and strip of land separating the Mediterranean from Lake Mariout. Originally the town

ISS001-E-5316
MEXICO, 23 January 2001
Popocatépetl, or Popo, the active volcano located about 70 km southeast of Mexico City, sends a plume south on January 23, 2001. The astronaut crew on the International Space Station

☑ Filter results

Match all of the following conditions:

View Type is city

Fig. 3. OpenMKS: default search results view, including thumbnails and metadata for each result and a search interface to refine the search.

The interface builds upon the SRW web service and is highly-configurable. It uses the Explain response from the SRW to automatically generate the filters that can be used to search the content exposed by the SRW. The search response XML that is received from the SRW is transformed using XSLT. This does not have to be intended for immediate visual display, but can also be used for non-visual display using whatever data format is required a different XML structure, HTML, JSON, plain text or RDF, such as the CRM markup. At its core, the OpenMKS Web application provides a RESTful interface to the SRW, allowing the different ’views’ of the output from the SRW to be used in virtually any application, such as the mSpace interface described below.

3.5 mSpace

mSpace 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.

An mSpace presents several associated categories from an information space, and lets users manipulate how many of these categories are presented and how theyre arranged. In this way, people can organize the information to suit their interests, while concurrently having available to them multiple other complementary paths through that information. Users are presented with multimedia examples of the possible selections before they commit to exploring them, these preview cues allow people to select what they like based on their evaluation of the multimedia.

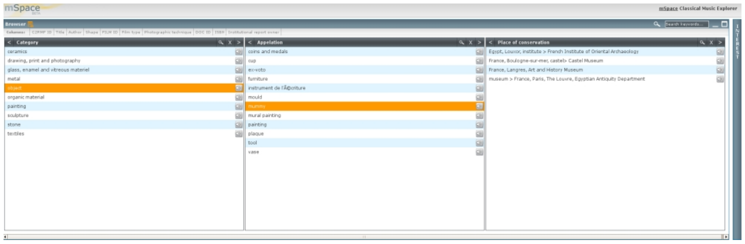
Figure 4 illustrates a subset of the EROS data set presented through the mSpace interface. Each category in the information space is displayed in a separate column, and the selection in each column narrows down the results presented in the next column. Here we can see the different categories of artefacts in the EROS system. Object is selected in the first column and the types of different objects are shown in the appellation column. The user has selected mummy, and the following column lists all institutions where mummys are held.

mSpace has been designed to be indipendent 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.

This has been accomplished by separating the user interface element, i.e. the client, from the component that constructs the queries to the underlying knowledge repository, i.e. the server. The mSpace client communicates with the mSpace server only in terms of the information it requires. Queries might include listing all available categories and listing the items in a given category. It is up to the server to interpret these queries and then construct the appropriate query to the knowledge base. Typically, these are SPARQL queries to an RDF triple store. We have developed an mSpace server that communicates CQL queries to the SRW running on the EROS data.

Components of the OpenMKS web application have been integrated with the mSpace framework. The mSpace client is able to display details about each resource when it is selected. For example, if an artefact is selected users will want to see all of the relevant information relating to the object: an image depicting it, details on where it was made, some descriptive text and so on. However the client does not know which information should be displayed for a particular resource, so it is up to the mSpace server to provide this. In our mSpace SRW server implementation, we configure information detail views for each category in the OpenMKS web application, and we use these to return the detailed information for the selected resource to the mSpace client.

One of the issues that was encountered is that the mSpace framework requests all possible items in a category at once, which it then caches on the client. This allows basic string searching to be performed to allow users find particular items they are interested in. However, the SRW is designed to retrieve small quantities of records at a time, so a paging strategy is required to collect all possible items for each category. This can take some time, especially considering the large amounts of data stored in EROS. To overcome this problem 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.



Category	Appelation	Place of conservation
ceramics	coins and medals	Egypt, Louxor, institute > French Institute of Oriental Archaeology
drawing, print and photography	cup	France, Boulogne-sur-mer, castel> Castel Museum
glass, enamel and vitreous materiel	ex-voto	France, Langres, Art and History Museum
metal	furniture	museum > France, Paris, The Louvre, Egyptian Antiquity Department
object	instrument de l'Ã©criture	
organic material	mould	
painting	mummy	
sculpture	mural painting	
stone	painting	
textiles	plaque	
	tool	
	vase	

Fig. 4. Subset of the EROS data set displayed through the mSpace interface

4 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 Cultural Heritage 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 queries languages such as SPARQL and the ability to use server side reasoning. For example, an SRW user is not able to query for relationships or concepts that are not specified in the mappings, and they cannot use the SRW to perform inferencing over the data. On the other hand, CQL does include support for many constructs that are particularly useful to real users of museum data, for example the ability to specify substring searching in text fields, use of comparators and proximity operators. Furthermore, SRW and CRM do not impose any semantics on data values (they are only concerned with the schema level). Care is needed 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.

Whilst the use of SRW on top of relational legacy data sources is scalable to the large datasets often held by cultural 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. For example, careful design of the user interface to ensure that frequent requests are not made to the SRW for large volumes of data, or extraction of subsets of data from the SRW that can then be held in a more appropriate and local form for subsequent investigation (e.g. using optimised in-memory datastructures), or optimisation of the underlying legacy data source for high performance querying (e.g. denormalisation of the relational tables). 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, e.g. using an RDBMS. The use of multi-tier architectures and inefficient data exchange structures (e.g. XML) does not help either. This is a common problem in many domains, and sees a variety of solutions, e.g. the use of multidimensional datacubes as part of On Line Analytical Processing (OLAP [12]), or the use of in-memory data stores and hardware based graphic processing in interactive visualisation ([13]).

5 Conclusion and Future Work

We have described how we have semantically exposed a cultural heritage 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 cultural heritage 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. As such we will be investigating optimisations of the SRW, as well as how the underlying database schema could be optimised and improved without causing a huge impact.

We are also interested in performing a quantitative evaluation of the SRW running against the EROS data set, and perhaps comparing these results to other systems such as D2R Server. Further work might also consider the performance of the SRW for satisfying the low response times required to support highly interactive interfaces such as mSpace.

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. At C2RMF we are considering using the mSpace interface to support a more intuitive, way to test, validate, debug and check consistency of the metadata maintained in EROS.

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 cultural heritage 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 cultural heritage institutions, including picture libraries, television archives, publishers and we hope to attract museums and galleries over the coming months. This requires dealing 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.

6 Acknowledgements

eCHASE is a two year European Commision funded eContent project which started in January 2005, within the "European digital content for the global networks" programme. The eCHASE consortium includes Istituto Geografico De Agostini S.p.A., IT Innovation and the School of Electronics and Computer Science at the University of Southampton, Fratelli Alinari, Giunti Interactive Lab S.r.l., Hewlett Packard, sterreichischer Rundfunk, System Simulation Ltd and Getty Images.

References

[1] Pillay, R.,Pitzalis, D., Lahanier, C., Aitken, G.: "EROS 2006: a fine vintage" Electronics and Visual Arts 2006 Florence, Pitagora, April 2006, Florence, Italy

[2] 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

[3] 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"

[4] Pillay, R.,Pitzalis, D., Lahanier, C., Aitken, G.: "EROS : New development in Digital Archiving for Research in Conservation" Electronics and Visual Arts 2005 Florence, Pitagora, April 2005, Florence, Italy

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

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

[7] Addis, M., Boniface, M., Goodall, S., Grimwood, P., Kim, S., Lewis, P., Martinez, K. and Stevenson, A.: "SCULPTEUR: Towards a New Paradigm for Multimedia Museum Information Handling" In Proceedings of Semantic Web ISWC 2870, pages 582–596

[8] 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)

[9] Sinclair, P., Lewis, P., Martinez, K., Addis, M., Pillinger, A. and Prideaux, D.: "eCHASE: Exploiting Cultural Heritage using the Semantic Web" In Proceedings of 4th International Semantic Web Conference, ISWC 2005, Galway.

[10] "eCHASE project": 2004-2006 eContent no. 11262. www.echase.org.

[11] Pitzalis, D., Pillay, R., Lahanier, C.: "A New Concept in High Resolution Internet Image Browsing" ELPUB2006. Digital Spectrum: Integrating Technology and Culture - Proceedings of the 10th International Conference on Electronic Publishing held in Bansko, Bulgaria 14-16 June 2006 / Edited by: Bob Martens, Milena Dobreva. ISBN 978-954-16-0040-5, 2006, pp. 291-298

[12] OLAP: http://www.olapreport.com/fasmi.htm

[13] Fekete, J., Plaisant, C.: "Interactive Information Visualization of a Million Items" INFOVIS 2002. IEEE Symposium on Information Visualization, 2002, Page(s): 117 -124, Boston, October 2002

[14] CQL: http://www.loc.gov/standards/sru/cql/

[15] D2R server: http://www.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/

[16] D2RQ: http://sourceforge.net/projects/d2rq-map/

[17] R2O: http://www2006.org/programme/item.php?id=p160

[18] 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