eCHASE: Exploiting Cultural Heritage using the Semantic Web

P. Sinclair, P. Lewis and **K. Martinez**
Electronics and Computer Science,
University of Southampton,
SO17 1BJ,
United Kingdom
pass,phl,km@ecs.soton.ac.uk

M. Addis, A. Pillinger and **D. Prideaux**
IT Innovation Centre,
Southampton,
SO16 7NP,
United Kingdom
mja,agp,djp@it-innovation.soton.ac.uk

Abstract

The eCHASE project is using semantic web technologies to demonstrate sustainable business models based on access and exploitation of digital cultural heritage content at a European level. In this paper we describe the eCHASE project and outline the system architecture.

1 Introduction

The European Commission supported eCHASE (electronic Cultural Heritage made Accessible for Sustainable Exploitation) project is developing sustainable models for accessing and using public sector cultural heritage content. We use Semantic Web technology to add value through aggregation and contextualisation of cultural heritage content from multiple sources. Aggregation in eCHASE means creating one or more narrative threads that link multiple items of content from multiple sources together into an overall context. For example, this might be a richly connected set of images, video and text that covers the life-story of a particular artist including the works of art they created, where they worked, who they worked with, and the influence of the society in which they lived. This richly connected multimedia collection then forms the basis for adding further value through the creation of appealing editorial content products in education and publishing.

Currently, our content providers include two photo libraries (Fratelli Alinari and Getty Images), a publisher (De Agostini) and a television broadcaster (ORF). We are also engaging with other cultural heritage institutions including museums and libraries to involve them in the project. All these institutions are providing content according to various interpretations of a theme entitled 'wandering borders in Eastern Europe'. This provides an interesting and challenging set of multimedia and multilingual content with which we are exploring how semantic web and knowledge technologies can provide new ways for subject experts and creative professionals to explore, navigate, link and annotate the content into editorial products.

2 Demonstrator

We are developing a centralised portal where editorial product authors can search and browse our content partners collections for media they require to produce a content product. By providing facilities to collect and annotate groups of relevant objects, media and metadata about these objects can then be exported into various content authoring packages where the high quality, editorial product can be developed.

From our experiences in the Sculpteur project [Sinclair *et al.*, 2005], the ability to explore and navigate relationships is an important feature of the semantic web for the cultural heritage domain. Collections from different institutions often overlap, with media relating to the same people, places, themes, periods and events. Due to the heterogeneous nature of different collections and metadata systems, exploiting this overlap raises serious technical issues: metadata schemas must be mapped and legacy data must be cleaned and transformed. Moreover, not only are advanced visualisation techniques let down by badly structured metadata, they often highlight and reinforce the problems.

The Sculpteur architecture included a Search and Retrieval Web Service (SRW) [z39.50 SRW, 2005] that exposed museum metadata schemas through the CIDOC Conceptual Reference Model (CRM) [Doerr, 2003], a reference model for the interchange of information in the cultural heritage domain, by dynamically applying mappings to the legacy data. In eCHASE, we are providing a framework for cleaning and transforming the different legacy metadata systems into a well structured, unified knowledge base. Processing and indexing the legacy metadata into a consistent format will improve the effectiveness of innovative visualisation techniques accessing the repository through the SRW.

Various sources of authority data, such as gazetteer and domain thesauri, are used to support the indexing and mapping processes. These involve semantic web technologies, including SKOS [SKOS, 2005] for structuring and serving thesauri information, and we have converted gazetteer information into CRM-modelled RDF. We are also considering existing automatic and semi-automatic thesauri and classification mapping and matching approaches for consolidating the different classifications used by our content partners.

Facilities for collecting and annotating objects and groups of objects are key to the eCHASE architecture. We are extending the Sculpteur light box component so that users can add their own descriptions and content, and manage groups of objects. We are investigating strategies for semantically integrating user created annotations back into the metadata

repository.

2.1 Semantic Harmonisation

The initial work on eCHASE has focused on maximising the quality of aggregation of media and metadata content from our partners collections. Our content providers currently deliver media and metadata electronically by uploading (e.g. FTP) or mailing a CD or DVD; we are also considering harvesting techniques, such as OAI. The metadata is provided in various formats, ranging from database dumps and XML to Microsoft Excel spreadsheets and CSV files.

We have developed a metadata importer that performs cleanup and integration tasks on the legacy metadata collections so that it can be collected in a unified metadata repository. Performing the mapping from different metadata systems, with a variety of approaches to structuring information, to a consistent unified structured is a complex task involving format and encoding issues, data cleanup, schema transformations and identity consolidation across different collections. We are employing workflow enactor system to break down these problems into a series of reusable modular services that can be configured into a workflow for transforming each collection.

In our experiences with the Sculpteur project, much of the rich information in the cultural heritage metadata systems is handled as unstructured textual information, such as free text description fields. We are considering the use of knowledge mining and extraction tools for extracting this information, but for the first prototype we are only providing basic textual search facilities. For efficiency and scalability reasons, especially in handling free text searching, we are using a relational database to manage the the unified knowledge base. We also consider that the bulk of metadata cleaning and transformation processes are well suited to relational database systems. The Sculpteur SRW can dynamically map records to CIDOC CRM structured XML, that can be converted to RDF through the use of XSLT.

The unified metadata repository consists of three areas: legacy, indexes and mapped data. Legacy data is stored in its original structure, which is useful for providing searching and display facilities. We are using several indexing strategies for improving queries on free text description fields; the indexes are stored in the repository to improve the efficiency of searches. A subset of each collection's metadata is mapped into a highly structured unified database schema, the design of which has been strongly influenced by the CIDOC CRM. The type of information mapped involves information on people, places, dates and categorisation information such as domain thesauri and controlled lists. This information is essential to support innovative browsing facilities, and can also be used to improve search results.

2.2 Media Engine

The eCHASE architecture includes a media engine for serving media and providing content-based querying facilities using algorithms from Sculpteur, including searches based on colour or texture. The media engine is self contained, and provides tools and a user interface to support import and maintenance of the media collections, for example the generation of media descriptors for the content-based algorithms. We are also investigating the integration of ongoing work at Southampton on classification and automatic semantic annotation of media.

2.3 eCHASE Portal

The eCHASE portal provides searching and browsing of content and a facility to collect and annotate groups of objects that users are interested in. The purpose of the web application search engine is to assist authors and experts to develop, manage, visualise, navigate, search and exploit valuable digital resources in the eCHASE repository. The system also provides search and retrieval of large multimedia collections by remote third-party applications.

The portal supports several different methods of searching: text and content based queries and a browsing interface. Textual queries can be run on the data in the unified metadata repository, and the portal exposes the content-based searching facilities provided by the media engine system. Browsing is provided by an mSpace interface [m.c. schraefel *et al.*, 2003], an interaction model designed to allow the navigation of multi-dimensional spaces.

The portal supports a search and retrieval protocol based on the SRW specification developed by the z39.50 community, providing a search operation to handle common query language (CQL) queries and an explain operation to tell external systems what schema are supported. The SRW supports queries based on each collection's legacy metadata schema, the unified database schema and is also able to dynamically map from the unified database schema into a CRM-based structure.

3 Conclusion

In this paper we have introduced eCHASE and given an overview of the software framework being developed for the project.

References

[Doerr, 2003] Martin Doerr. The CIDOC Conceptual Reference Model: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3):75–92, September 2003.

[m.c. schraefel *et al.*, 2003] m.c. schraefel, M. Karam, and S. Zhao. mSpace: Interaction design for user-determined, adaptable domain exploration in hypermedia. In P. De Bra, editor, *AH 2003: Workshop on Adaptive Hypermedia and Adaptive Web Based Systems*, pages 217–235, 2003.

[Sinclair *et al.*, 2005] P. A. S. Sinclair, S. Goodall, P. H. Lewis, K. Martinez, and M. J. Addis. Concept browsing for multimedia retrieval in the SCULPTEUR project. In *Proceedings of the Multimedia and the Semantic Web Workshop, European Semantic Web Conference*, 2005.

[SKOS, 2005] SKOS. Simple knowledge organisation system (SKOS) http://www.w3.org/2004/02/skos/, 2005.

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