The eCHASE System for Cross-border Use of European Multimedia Cultural Heritage Content in Education and Publishing

M. Addis, S. Hafeez, D. Prideaux, R. Lowe,
IT Innovation Centre, Southampton, SO16 7NP, UK
[mja,szh, djp,rl]@it-innovation.soton.ac.uk

P.Lewis, K. Martinez, P. Sinclair
ECS, University of Southampton, SO17 1BJ, UK
[phl, km, pass]@ecs.soton.ac.uk

Abstract

Europe's digital cultural heritage content has tremendous exploitation potential in applications such as Education, Publishing, e-Commerce, Public-Access and Tourism. Value is hugely amplified if the content can be aggregated repurposed and distributed at a European level. The eCHASE project seeks to demonstrate that public-private partnerships between content holders and commercial service providers can create new services and a sustainable business based on access and exploitation of digital cultural heritage content.

This paper describes the eCHASE demonstrator from a technical perspective, briefly detailing the tools and components which make up the system and the use of open standards.

1. Overview

Content holders such as museums and galleries, especially small to medium organisations, often generate digital content through internal activities such as collection management and curation, or art object conservation and restoration. These activities typically generate high quality multimedia digital content (images, video, 3D models, metadata), which have significant exploitation potential outside of the organisation. However, this content is typically not in a form suitable for external access and is often 'locked away' in internal legacy systems, for example in collection management tools. Furthermore, due to the terminologies, data structures, and legacy systems used for content creation and management, significant work is often needed to semantically integrate content from multiple sources in a way that addresses the contextualisation, aggregation and localisation needed for specific end-user applications, e.g. education or publishing.

The eCHASE project (www.echase.org) has developed a software system for semantic integration and access to content from libraries and archives across Europe and is using this system to experiment with business models for exploiting digital cultural heritage content.

The system consists of a centralised portal where authors of content products (e.g. books, DVDs or electronic teaching materials) can search and browse content collections for the media they require. Facilities are provided to collect, annotate and export groups of relevant objects. The media and metadata about the objects selected by the user can then be imported into various authoring packages, e.g. web design or desktop publishing tools, where the high quality, editorial product can be developed. The process for content import, aggregation, search, retrieval, export and use is shown in *Figure 1*.

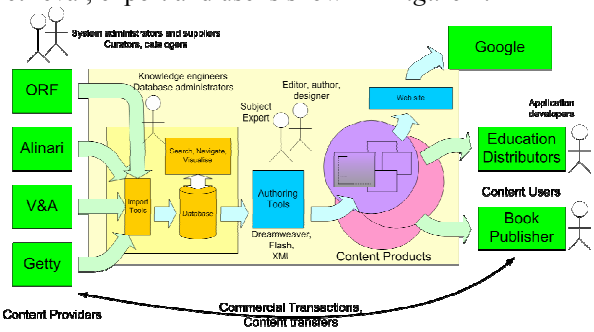

Figure 1 eCHASE workflow

Metadata and media from content-holders is imported into the eCHASE system using a workflow-based approach. After data cleaning and transformation, the metadata is semantically integrated by mapping to a common structure using the CRM Core version of the CIDOC CRM [2]. The media is processed to generate thumbnails and browse resolution images/video as well as to generate various content-based descriptors that can be used for content-based retrieval. The integrated dataset is then exposed through a series of Web

Services including a search and retrieval interface based on the SRW protocol [4].

These services are used by both a web application layer that provides a human usable interface and also by external applications, e.g. eLearning, which remotely access the content in the system. An overview of the system architecture is presented in *Figure 2*

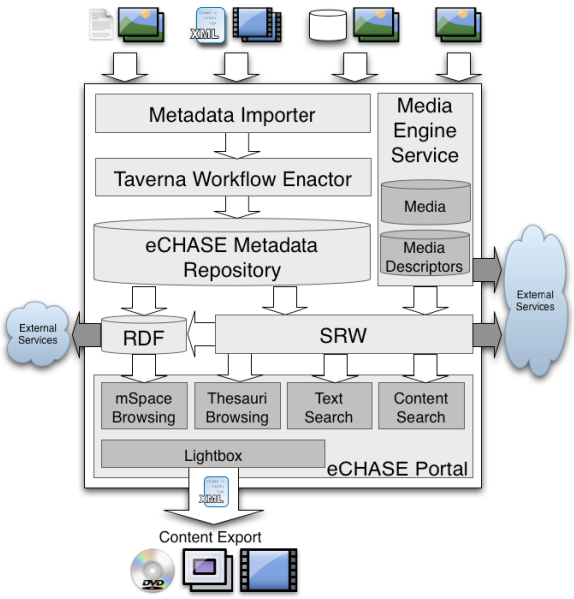

Figure 2 System overview

2. Metadata and media import process

The metadata is imported into a relational database format so that cleaning and transformation techniques can be applied. The cleaning and transformation process is defined and orchestrated using the Taverna Workbench workflow system [1][5].

A series of metadata importer components are used to perform cleanup and integration tasks on the legacy metadata collections so that they can be combined into a unified metadata repository. Metadata processing involves multiple steps, e.g. date homogenisation, mapping to a common gazetteer, identifying missing/incorrect/truncated metadata, processing of thesauri, automated language translation, and cross referencing metadata to the images/videos.

Applying the eCHASE import process to different metadata systems, which have a variety of approaches to structuring information, in order to create a consistent unified structure is a complex task involving format and encoding issues, data cleanup, schema transformations and identity consolidation across different collections.

The use of a workflow system has allowed us to break down the complex problems of metadata conversion and mapping into a series of reusable modular services that can be configured into a customised workflow for transforming each collection.

The Taverna Workbench is a service oriented workflow system that facilitates the composition of distributed services for processing information through a workflow and enables us to integrate local tools as well as a variety of third party web services into the import process, for example automated language translation

The use of a workflow approach encourages flexibility and extensibility, as existing workflows can be modified to cope with new data sources or entirely new workflows can be created and added to a workflow library for future re-use.

The processed datasets are then mapped into a consistent eCHASE metadata structure using the CIDOC Conceptual Reference Model (CIDOC CRM) [2] which is described below in more detail. This is a common metadata schema to cover the different metadata repositories from our partners' collections.

The media, for example collections of images and videos, is loaded into a media engine system which provides media transformation (e.g. thumbnail generation) needed for basic web access, as well as algorithms for content-based retrieval which allow for searches based on colour or texture [9][10]. Currently, only 2D image algorithms have been implemented in the system, however future development will allow the addition of algorithms able to deal with different types of media including 3D objects, audio and video but also application specific algorithms, such as a face recognition system that could attempt to images containing people, e.g. portraits.

The media engine itself 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. It is able to provide access to the media via the web application, or can be configured so that the media is hosted on another web server.

3. CIDOC CRM

The CIDOC CRM is a core ontology for the semantic integration of cultural information, including library, archive and other information. Since 1996 it has been developed and supported by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). More recently, it has been accepted as an ISO standard and is

available as a Final Draft ISO standard (ISO/FDIS 21127). The CIDOC CRM concentrates on the definition of relationships, rather than classes, to capture the underlying semantics of multiple data and metadata structures. This has led to a compact and easy to comprehend model of 80 classes and 130 relationships, comprising the most characteristic concepts required for museum, archive and library documentation. The CIDOC CRM enjoys rapidly increasing uptake by information systems designers all around the world. An ongoing collaboration of the ICOM and IFLA committees has resulted in the harmonization of CIDOC CRM with the FRBR model, a standard for conceptualizing bibliographic information. This process has demonstrated that CIDOC CRM subsumes all of the relevant FRBR concepts. The model is available as an XML DTD, and it has also been formulated as RDFS and OWL ontologies.

The CRM Core is a recent proposal from CIDOC for a highly condensed set of metadata elements, designed for resource discovery and as such is ideal for use in eCHASE. CRM Core captures the basic functions of identification, classification, participation, part decomposition, references and similarity. In other words, it describes the most fundamental relationships that connect things, concepts, people, time and place.

CRM Core is not only a metadata format for resource discovery, but also a simple schema for summarization of historical facts. It allows for exploiting the fact that metadata about the creation, use and discovery constitute historical facts comparable to the information found in documents themselves. An example is shown in Figure 3 which shows how the CRM Core can be used to represent complex relationships, for example the production of a self portrait by Van Gogh and how the physical artefact is represented by a digital image.

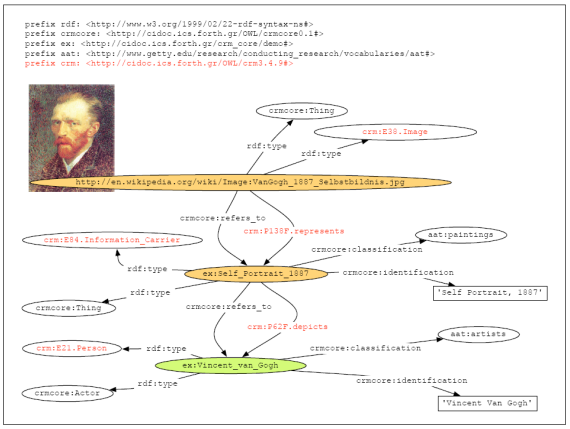

Figure 3 Example of modeling using the CRM Core

The CIDOC CRM defines the underlying semantics of cultural heritage information in terms of a formal ontology, and thus it does not specify any of the terminology appearing typically as data in the respective data structures. CRM Core defines characteristic relationships for the use of controlled terminology by allowing CRM Core records to be classified according to entries in controlled thesauri.

Therefore, the problem becomes one of identifying or creating suitable authority files and then referencing their namespaces in the CRM Core records. For example, in cultural heritage several domain vocabularies are widely used, e.g. the Art and Architecture Thesaurus (AAT) and ICONCLASS and the Union List of Artist Names (ULAN) for artist names. Outside of the cultural heritage domain there are a wide variety of vocabularies that can be used, such as IPTC news codes. Location information is handled through a gazetteer, such as the Getty Thesaurus of Geographical Names (TGN).

In this way, all references to people, institutions, places, events and things are performed by referencing authority data. The CRM Core then provides the semantic glue to capture how instances in these authorities relate to each other in a particular context, e.g. the photographing of a historic event or the production of a specific work of art.

4. Search, retrieval and semantic interoperability

Semantic interoperability of Cultural Heritage digital libraries has been investigated in the SCULPTEUR [3][8][10] and eCHASE projects by using a z39.50 search and retrieve web service (SRW) and by mapping legacy metadata schemas to the CIDOC CRM.

This allows additional semantics to be attached to 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 [11]) 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[6]. This provides an SRW implementation that allows relational data to be mapped to an XML representation, including CRM Core.

For efficiency and scalability reasons, especially in handling free text searching, we retain the relational database used in the data import process for the storage and management of the metadata in eCHASE. Having already transformed the legacy data into a consistent, well-structured schema, the task of converting to a semantic web format such as RDF is straight forward, for example the CIDOC CRM structured XML can then be converted to RDF through the use of XSLT.

5. Web interface

The user interface in eCHASE is provided by a server-side web application that builds upon the SRW web service. A screenshot of the user interface is shown in *Figure 4*.

The Explain response from the SRW is used to automatically generate a set of search filters that can be used to search the content exposed by the SRW. The search response XML received from the SRW is transformed using XSLT into a form suitable for use in the end application. This does not have to be for immediate visual display, but can also be used for non-visual processing using whatever data format is required, e.g. as a different XML structure, HTML, JSON, plain text or RDF markup.

At its core, the OpenMKS Web application essentially provides a RESTful interface to the SRW, allowing the different 'views' of the output from the SRW to be used in virtually any application.

The eCHASE web application allows users to search and browse the collections that have been added to the system. Users can 'filter results' by searching by specific dates, places and concepts (e.g. using gazetteers and thesauri). Filters also include the ability to specific content based queries, e.g. based on colour, where the user can supply a query image or ask for images that are similar to one they have already found in the eCHASE system. Content-based search is currently only available for image based searches and not for video, but does already provide a useful way to further refine searches or to find things in the database that are difficult to define based on keywords and textual descriptions alone.

Navigation from the results of a query to related items in the database, e.g. those with the same keywords is supported by automatic generation and embedding of hypertext links that initiate new or refined searches.

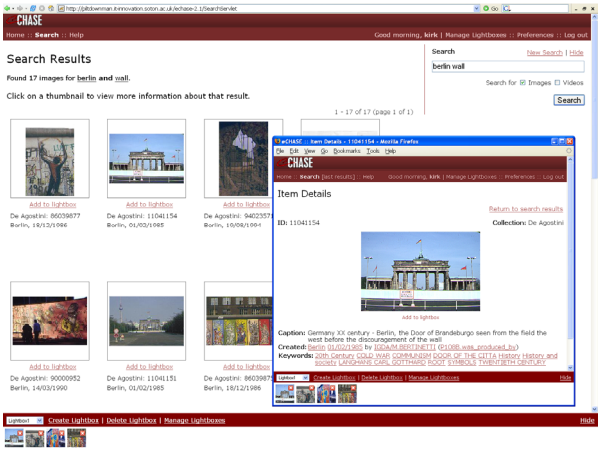

Figure 4 eCHASE web interface

There are facilities which allow users to group and annotate sets of images that are of interest to them via the Light box. Lightboxes can be shared between users with different levels of permissions that range from read only to add/delete/annotate items The ability to share lightboxes including multi-user annotation addresses requirements from both picture researchers, e.g. when soliciting comments on suitability of images from their clients, and from educational usage scenario. For example, a lecturer teaching a particular topic in a classroom assignment, e.g. the fall of the Berlin Wall, could create a lightbox containing all the relevant images and videos they find in eCHASE. They could share this with other teachers/lecturers covering the same topic as well make the lightbox available for student annotation during project work within the classroom.

eCHASE includes the ability to launch mSpace as a way to navigate and explore the content in the eCHASE database. mSpace[7] is an interface service that includes an interaction model and software framework to help people access and explore information. mSpace presents several associated

categories from an information space, and then lets users manipulate how many of these categories are presented and how they are 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. An example of the mSpace interface is shown in *Figure 5*.

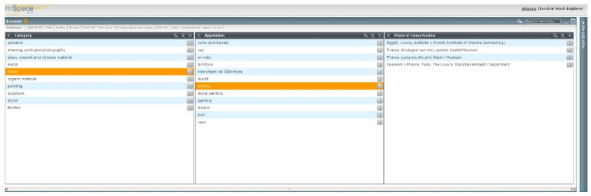

Figure 5 mSpace interface

Using the mSpace explorer, the information domain is explored by selecting elements in a column. The arrangement of columns is called a slice; selecting an element in a column will affect the content of the next column (to the right) in the slice. The first column (left-most) is always populated. Columns can be added, removed and moved around to explore the information in different ways.

In this way, the navigation and exploration modality enabled by mSpace complements the targeted search and retrieval modality of the main eCHASE user interface.

6. Discussion

The main benefits of eCHASE are three fold. Firstly, eCHASE supports an end-to-end process for multimedia content to be taken from multiple sources and semantically integrated and exposed for use in a variety of applications. Secondly, eCHASE combines several different search modalities into one user interface. These include: targeted search and retrieval using authorities and free text searching in the web application; exploration and navigation using mSpace; and content-based retrieval using the media engine. Thirdly, eCHASE provides the ability to select, aggregate, annotate and share groups of media items in one or more lightboxes which facilitates the process of using content in end-user applications, for example teaching or publishing.

Of these, the most technically challenging and fundamental to the system is our guiding principle that content holders shouldn't need to restructure or re-catalogue their contents when they want to integrate or provide access to their collections. Instead, we use semantic mapping techniques to relate the contents of different collections to common standards (CRM Core, DublinCore, AAT, ULAN etc.).

It is this semantic mapping and integration that allows new paths and links to be created through diverse collections of cultural content, which in turn allows users to ask those questions that are hard to answer using conventional systems. The semantic integration also underpins the navigation and exploration of the data in mSpace.

However, our approach is not without tradeoffs. On the one hand, our approach makes it easy for the user to 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. On the other hand, there 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. Furthermore, the 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, for example through mSpace.

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. Therefore, we have found it necessary to implement result set caching at several levels to enhance retrieval time when successive searches take place with a subset of the data.

7. Acknowledgements

The 3.5MEuro eCHASE project is co-funded by the European Commission under the eContent programme (EDC 11262). The eCHASE consortium includes Istituto Geografico De Agostini S.p.A., IT Innovation

and IAM within Electronics and Computer Science at the University of Southampton, Fratelli Alinari, Giunti Interactive Lab S.r.l., Hewlett Packard, Österreichischer Rundfunk, System Simulation Ltd and Getty Images.

3. References

[1] T. Oinn, M. Addis, J. Ferris, D Marvin, M. Senger, G. Greenwood, T. Carver, K. Glover, M. Pocock, A. Wipat, and Li. P. "Taverna: A tool for the composition and enactment of bioinformatics workflows", Bioinformatics Journal, 20(17):3045–3054, 2004.

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

[3] 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.

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

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

[6] OpenMKS http://openmks.sourceforge.net

[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

[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] Goodall, S., Lewis, P. H., Martinez, K., Sinclair, P. A. S., Giorgini, F., Addis, M. J., Boniface, M. J., Lahanier, C. and Stevenson, J. (2004) SCULPTEUR: Multimedia Retrieval for Museums. Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings 3115/2004 pp. 638-646.

[10] Lewis, P. H., Martinez, K., Abas, F. S., Ahmad Fauzi, M. F., Addis, M., Lahanier, C., Stevenson, J., Chan, S. C. Y., Mike J., B. and Paul, G. (2004) An Integrated Content and Metadata based Retrieval System for Art. IEEE Transactions on Image Processing 13(3) pp. 302-313.

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