eCHASE: SUSTAINABLE EXPLOITATION OF ELECTRONIC CULTURAL HERITAGE

P. Sinclair[1], P. Lewis[1], K. Martinez[1], M. Addis[2], D. Prideaux[2], D. Fina[3] and G. Da Bormida[3]

1 Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom, [pass,phl,km]@ecs.soton.ac.uk
2 IT Innovation Centre, Southampton, SO16 7NP, United Kingdom, [mja,djp]@it-innovation.soton.ac.uk
3 Istituto Geografico De Agostini, Corso della Vittoria 91, 28100 Novara, Italy, daniela.fina@deagostini.it, g.dabormida@inwind.it

Keywords: Semantic Web, Cultural Heritage, Content based image retrieval

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 these issues and introduces the eCHASE architecture that is being developed to showcase the business models created for the project.

1 Introduction

European cultural heritage content holders (museums, galleries, audiovisual archives, photo archives, art libraries) are rich in high-quality multimedia digital content. This content is readily exploitable in a range of applications such as Education (e-Learning and distance learning, creation of training and course material), Publishing (books, videos, newspapers, magazines, television for the public market; and advertising, graphic design, web design for the corporate market), e-Commerce (on-site shops at museum or galleries, on-line and third-party B2B and B2C web sites), and Public Access and Tourism (promotional material, web sites, museum and gallery terminals).

Large commercial archives, for example picture libraries, already store, catalogue and make their content accessible for the purpose of commercial sales and have established brands in the market place. Whilst commercial exploitation of their holdings is part of their core business, it is still typical for the majority of customers to be at the national level; international penetration is often limited due to lack of multilingual textual descriptions (annotation) of their holdings.

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, the content is typically not in a form suitable for external access and is often 'locked away' in internal legacy systems. Furthermore, commercial exploitation of the content is typically not regarded as a core part of the organisation's business and hence receives little attention or investment despite the potential revenues that can be generated.

The exploitation potential of digital cultural heritage content can be amplified if it can be aggregated and accessed at a European level. For example, imagine having a unified way of accessing all the digital representations, textual information and audiovisual material relating to the works of art of Rembrandt. Currently, this information is highly distributed across Europe in a range of museums, galleries, art libraries, and audiovisual archives and much of it is not accessible online at all! Alternatively, imagine being able to access images, documentaries and news programmes about a particular event in European history. Again this information is highly distributed across Europe in a range of national audiovisual archives and photo libraries. There is significant value in having content aggregated and presented in a particular context, for example in an editorial framework that allows a user to look at content relevant to a specific period, place or author.

The key to this problem is improved access and exploitation of the content through reuse and repurposing in commercial applications such as e-Learning, publishing, eCommerce and tourism. Such services, business models and markets have the potential to provide the revenue streams necessary to expand and sustain the digital content chain and realise the exploitation potential of Europe's cultural heritage information.

2 Objectives

The following scenario is an exemplar of the driving vision for eCHASE.

"An international multimedia publishing organisation wishes to produce a book, DVD and interactive software module (e-Learning) on renaissance art to be sold in several countries across Europe, including Italy and the Czech Republic.

Material is sourced from cultural heritage and content holding organisations across Europe. Images, virtual models and detailed textual information are retrieved from the National Gallery in London, the Uffizi in Florence and the Louvre in Paris. Relevant books, manuscripts and documents are identified, for example the Bibliothèque nationale de France and the UK National Art Library. Audiovisual archives are searched for relevant news items and documentaries. Links are made to eCommerce sites containing high-quality digital material, for example the Alinari photo library, the Hulton Archive and the Louvre on- line shop.

The results of the search are presented graphically in an integrated and easy to navigate way according to concepts such as artist, place, time, work of art, technique, language, and accessibility of further material including cost and rights. Multimedia information, for example images and textual descriptions are presented as a set of interlinked web pages allowing rapid and transparent navigation between content sourced from different content holders. Cross-language searching and results-presentation provide the multilingual support needed when dealing with content distributed across Europe.

Appropriate material is then selected and repurposed for development of the book and DVD. For example, the e-Learning module developer creates a virtual exhibition of renaissance art, which itself contains links back to the original sources of information as well as a structured series of learning tasks to accompany the book and DVD.

The revenue streams from the commercial exploitation of content held by the cultural heritage organisations is used to fund further digital content creation and access activities, which improves availability for the community as a whole."

This scenario demonstrates the value of being able to access, aggregate and repurpose digital cultural heritage content at a European level. The objectives of this project are to progress the realisation of this vision by solving many of the business and technological barriers involved. Furthermore, we will demonstrate the business viability of our approach by establishing a sustainable business model for commercial sale of European digital cultural heritage content.

The specific objectives are:

- **Develop sustainable business models for exploitation of digital cultural heritage content at a European level.** There is no 'one size fits all' solution in the cultural heritage sector. The diversity of content holding organisations and the varying degree to which they consider commercial exploitation to be part of their core business means that different business models, distribution chains and revenue streams need to be applied. For example, a small museum or gallery might want to outsource the whole infrastructure and services for commercial sales of their digital collection to a third-party and would receive royalties on any sale. On the other hand, commercial picture and audiovisual libraries would want to build upon their already established brands and are interested in widening their distribution channels to achieve increased volumes of sale. Therefore, a range of business models needs to be developed, proven and integrated.

- **Remove the technological barriers associated with cross-border exploitation of digital content.** From a technical perspective, digital cultural heritage content is highly fragmented, distributed across Europe, and is stored and managed using bespoke systems. A lack of common standards, protocols and semantics for accessibility make content hard to search and navigate for those not already intimately familiar with its organisation and annotation. Furthermore, textual descriptions (annotation) are typically only in one language, which is a barrier to international exploitation. Finally, content will be filtered for different usage contexts and rights carefully managed.

- **Promote the value and benefits of public-private partnerships** in the cultural heritage sector for increasing accessibility and generating revenue streams from European digital cultural heritage content.

3 eCHASE Framework

As part of the project we are developing a demonstrator system where we will be able to experiment with the eCHASE business models for exploiting digital cultural heritage content. This system consists of 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.

An overview of the system is shown in Figure 1. Cultural heritage content providers deliver media collections along with the accompanying metadata. The metadata is imported into a relational database format so that cleaning and transformation techniques can be applied using a workflow enactor system. The metadata is converted into a consistent eCHASE metadata structure. The media, such as sets of images and videos, is loaded into our media engine system that provides basic web access to the media as well as mechanisms for content-based retrieval. The eCHASE portal allows users to search and browse the collections that have been added to the system. Users are able to group and annotate sets of images that they are interested in through the Lightbox. Finally, the system exports images and associated metadata so it can be used in content authoring packages.

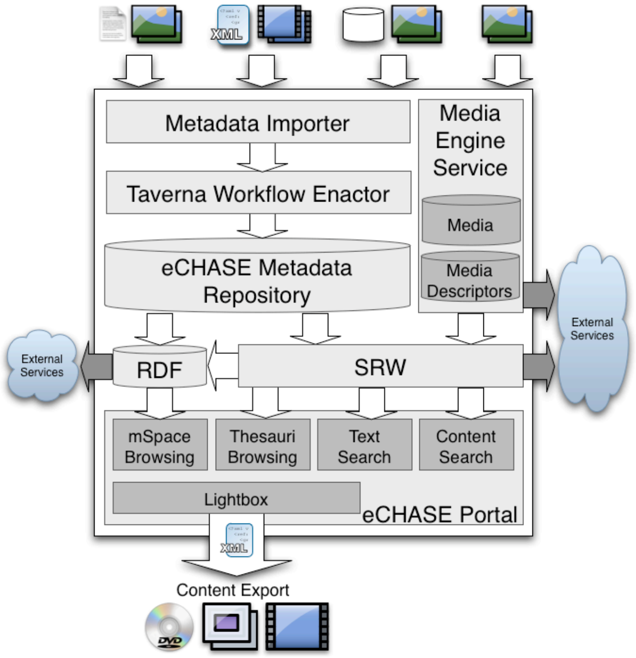


Figure 1: System Overview.

From our experiences in the Sculpteur project [4], the ability to explore and navigate relationships is essential when dealing with 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 CIDOC Conceptual Reference Model (CIDOC CRM) [1] is a reference model for the interchange of information in the cultural heritage domain. It has been in development over the last ten years by the museum documentation standards group CIDOC and is in the process of ISO standardisation. The CIDOC CRM is becoming increasingly used in the cultural heritage domain. It is capable of modelling the complex objects and relations within its scope, and can be extended to cover many specialisations. The eCHASE project is using the CIDOC CRM as the common metadata schema to cover the different metadata repositories from our partners' collections.

The Sculpteur architecture included a Search and Retrieval Web Service (SRW) [8] that exposed museum metadata schemas through the CIDOC CRM 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 based on the CIDOC CRM. 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 gazetteers and domain thesauri, are used to support the indexing and mapping processes. These involve semantic web technologies, including the SKOS ontology [6] for structuring and serving thesauri information, and we have converted gazetteer information into CIDOC CRM-modelled RDF. Gazetteer data, including latitude and longitude for a large proportion of place names, has been obtained from the World Gazetteer web site [9]. This data has been converted into the unified knowledge base structure and used in the data mapping stage as well as in the web interface. 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.

3.1 Metadata

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. Metadata has also been provided embedded in

We have developed a series of metadata importer components that perform 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 structure is a complex task involving format and encoding issues, data cleanup, schema transformations and identity consolidation across different collections.

To overcome these issues, we are employing a workflow enactor system. This 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 workflow for transforming each collection. The workflow system we are using is the open source Taverna Workbench [2],[7], which is a service oriented workflow system. Taverna facilitates the composition of distributed services for processing information through a workflow, and we aim to integrate a variety of third party services into the eCHASE framework in this way. The use of a workflow enactor system encourages flexibility and extensibility, as existing workflows can be modified to cope with new data sources or entirely new workflows can be created.

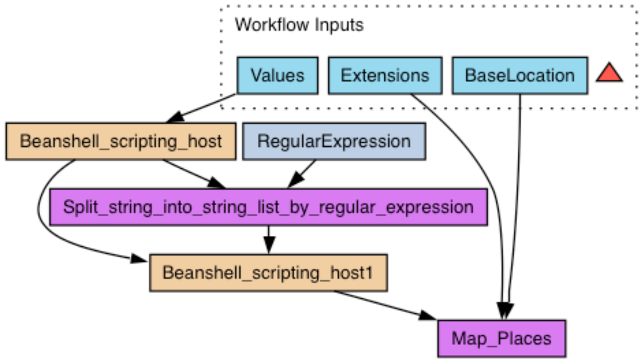


Figure 2: Taverna Workflow for cleaning place data.

Figure 2 shows an example Taverna workflow that performs some data cleaning on the place entries from a legacy system before passing them to a place mapping processor developed for eCHASE. This processor will take in a series of place names for each item and look up the respective place entry in a gazetteer.

As all transforms for each item of metadata are expressed as steps in a workflow, we are able to examine the intermediate status of the data at every step in the process. Besides obvious use in debugging, this allows users to investigate in detail how metadata records presented in the eCHASE portal were transformed from the original legacy information. Another advantage is that by performing all modifications and transforms to the metadata through a workflow, we can reprocess the data faithfully at any time. Often such metadata transforms are performed by a combination of purpose built tools or scripts and manual modifications to the legacy data, which can be hard to reapply in exactly the same way.

The unified metadata repository consists of three areas: legacy data, 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 based on 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.

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 unified knowledge base. We also consider that the bulk of metadata cleaning and transformation processes are well suited to relational database systems. Having transformed the legacy data into a consistent, well-structured schema, the task of converting to a semantic web format such as RDF is straightforward through. The Sculpteur SRW can dynamically map records to CIDOC CRM structured XML that can be converted to RDF through the use of XSLT. For convenience, we are also able to use existing RDF mapping tools that connect directly to the eCHASE metadata repository.

3.2 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. The media engine 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.

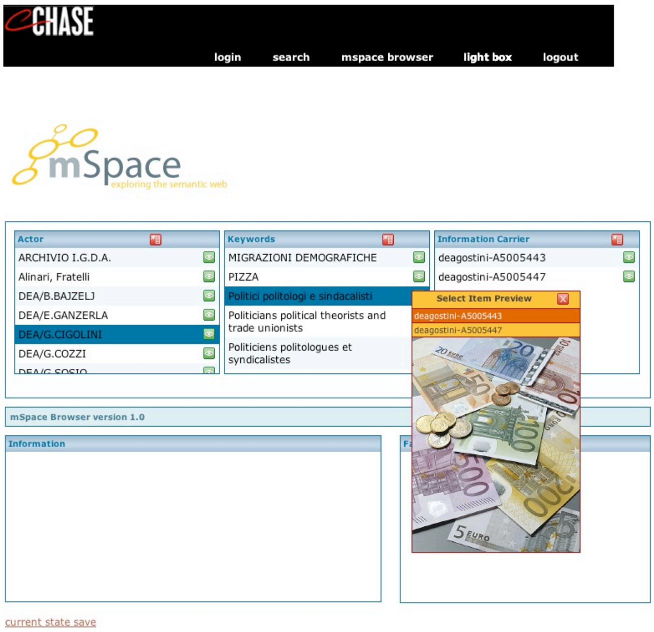
We have designed the media engine system to be flexible so that various types of content-based algorithms can be incorporated. Currently, only the 2D image Sculpteur algorithms have been implemented in the system, but we are investigating algorithms able to deal with different types of media including 3D objects, audio and video. We also intend to integrate application specific algorithms, such as a face recognition system that could attempt to find portraits in the collection.

We are also investigating the integration of ongoing work at Southampton on classification and automatic semantic annotation of media.

3.2 eCHASE Portal

The eCHASE portal provides searching and browsing of content and a facility to collect and annotate groups of objects in which users are interested. 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.



Actor
- ARCHIVIO I.G.D.A.
- Alinari, Fratelli
- DEA/B.BAJZELJ
- DEA/E.GANZERLA
- DEA/G.CIGOLINI
- DEA/G.COZZI
- DEA/G.SOSIO

Keywords
- MIGRAZIONI DEMOGRAFICHE
- PIZZA
- Politici politologi e sindacalisti
- Politicians political theorists and trade unionists
- Politiciens politologues et syndicalistes

Information Carrier
- deagostini-A5005443
- deagostini-A5005447

Select Item Preview
- deagostini-A5005443
- deagostini-A5005447

mSpace Browser version 1.0

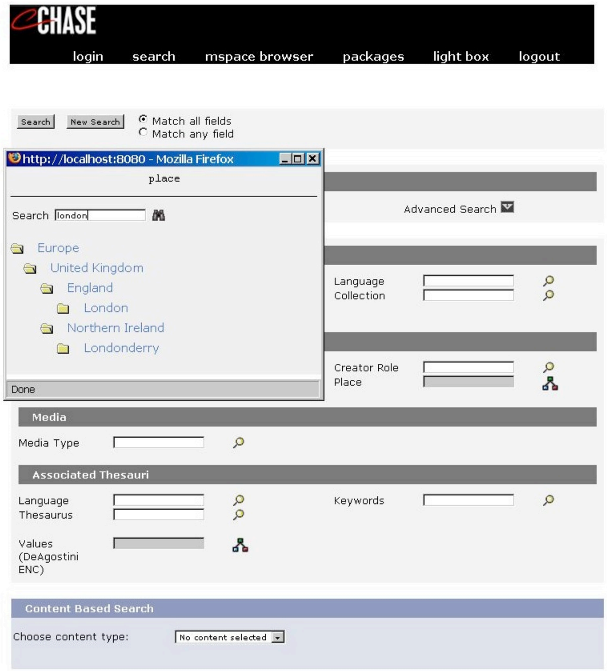
Information

current state save

Figure 3: mSpace interface.

Browsing is provided by an mSpace interface [3], an interaction model designed to allow the navigation of multi-dimensional spaces. A screen shot of the mSpace interface is shown in Figure 3. The mSpace interface uses a multi-panel display, where slices are presented as columns arranged from left to right. Selection in a slice will update the display so that the values displayed in the next slice (i.e. to the right of the current slice) are related to that value. For example, if there is a slice of artists and the next slice is painting titles, selecting an artist will display only that artist's paintings in the titles slice. Values in each slice are filtered, so that there are always results to view in the next column when a selection is made. When an item is chosen in a slice, details about that item are displayed in a detail panel; if no details are available for that item, examples of related objects are shown. Slices can be freely interchanged, removed and new slices can be added to the mSpace. Users are able to record column arrangements that they find interesting, including selected items, for quick access.

Thesauri navigation is provided by presenting a thesaurus tree structure to the user, such as concept hierarchies or gazetteer place names, as shown in Figure 4. Users will use the thesaurus navigation and visualization to select controlled terms that are then used to perform searches on the metadata in the repository. For example, a user may want to search for items that were created in London, England. By simply performing a metadata free text search the user may find items in Londonderry, Northern Ireland as well as London, England. The thesauri browser provides a way of presenting the structure of countries, cities and so on in a way that allows users to constrain queries just on particular cities within a certain country



● Match all fields
○ Match any field

http://localhost:8080 - Mozilla Firefox

place

Search london

- Europe
  - United Kingdom
    - England
      - London
    - Northern Ireland
      - Londonderry

Done

Language
Collection

Creator Role
Place

Media

Media Type

Associated Thesauri

Language
Thesaurus

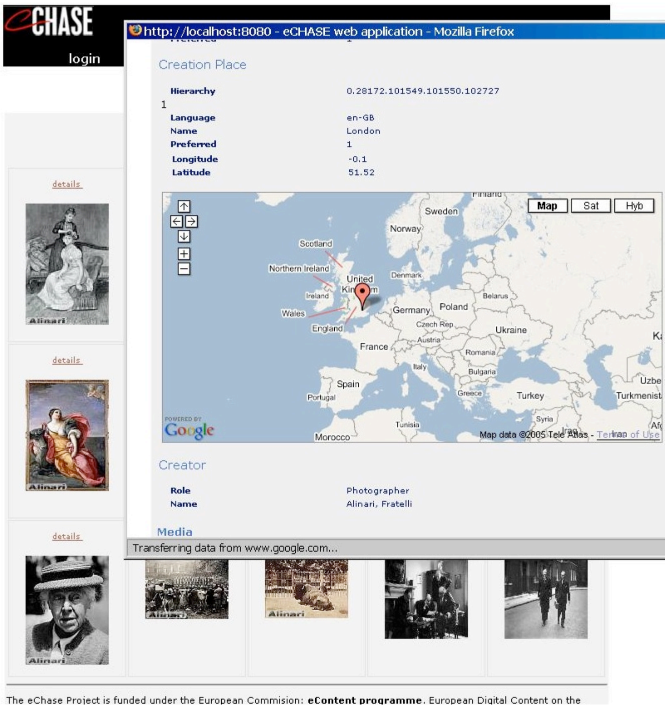
Values (DeAgostini ENC)

Keywords

Content Based Search

Choose content type: No content selected

Figure 4: Gazetteer browser.



http://localhost:8080 - eCHASE web application - Mozilla Firefox

Creation Place

Hierarchy	0.28172.101549.101550.102727
1	
Language	en-GB
Name	London
Preferred	1
Longitude	-0.1
Latitude	51.52

Creator

Role	Photographer
Name	Alinari, Fratelli

Media

Transferring data from www.google.com...

The eChase Project is funded under the European Commision: **eContent programme**. European Digital Content on the

Figure 5: Item details including place visualisation

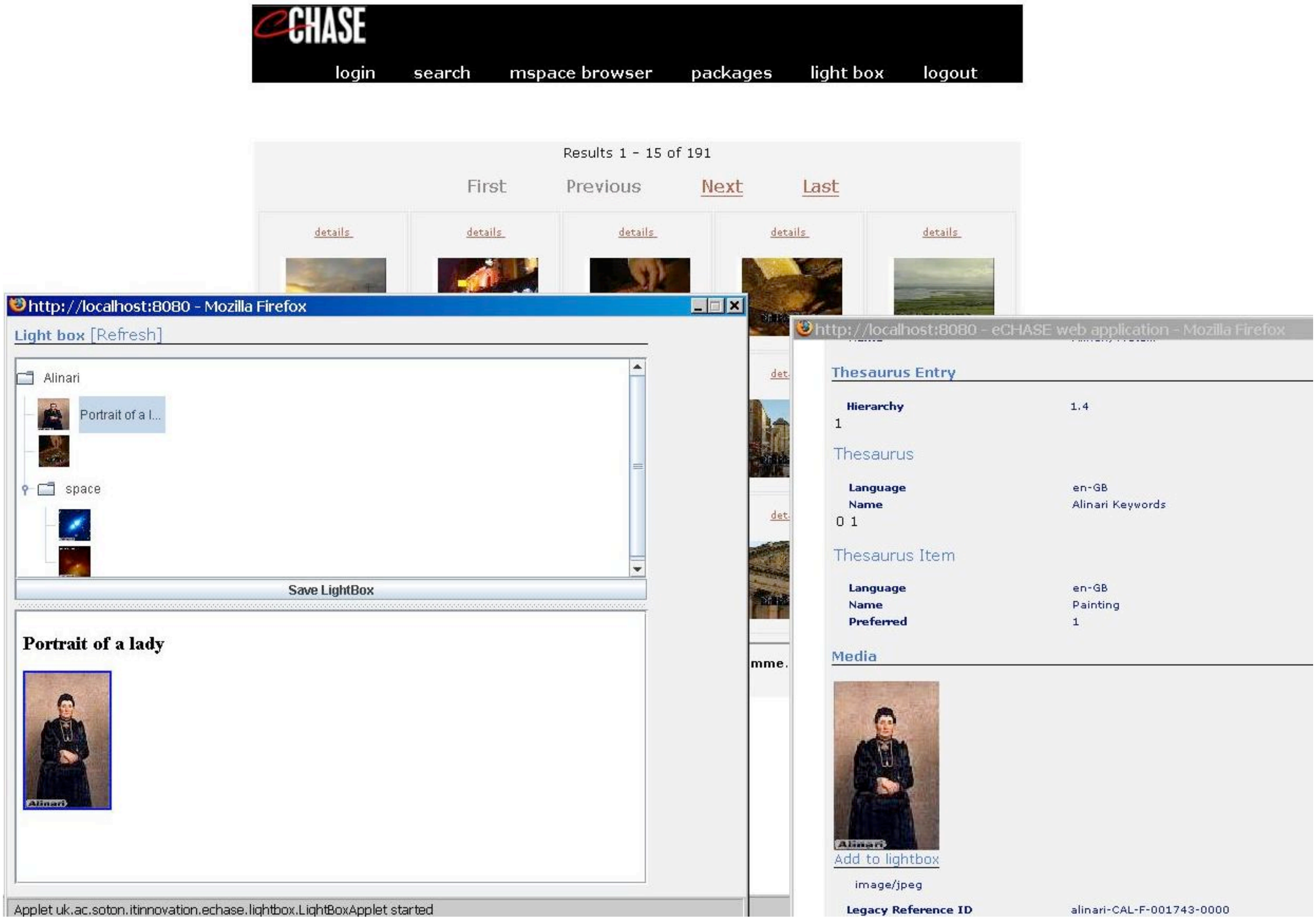
Gazetteer latitude and longitude data is also used for visualizing place information for each object. Figure 5 shows the details page for an object, and displays a map of the place it was created. The map visualization interface is provided by the Google Maps API [5]. Therefore, even if the place associated with a search result is obscure (e.g. the name of a small village in a remote part of Poland), then it is still very easy for the user to get a visual impression of where that place is in the context of Eastern Europe. The Google Maps interface through eCHASE is fully navigable, e.g. the user can zoom in to see details of the surrounding area, e.g. roads and local places.

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.

User content is managed by a personal light box tool, which gives authors and experts a mechanism to collect, organize and annotate information for later use. Items found through the search engine with the web interface can be added to one of the users light box collections to either be used for future searches or to be exported as an item in the content package the user is building. For example, the user finds a fine art painting and would like to find and create a content package containing other paintings by the same artist. The user can save the item in their light box and then use the saved item to search for the other paintings by the artist. The paintings found through subsequent searches can be placed in the user's light box and the collection can be arranged and annotated. Once a collection has been completed it can be exported as a complete package for further use in the content creation process. These bundles will contain the media and associated metadata in XML format. A screen shot of the light box is shown in Figure 5.

4 Conclusions

The eCHASE project is developing sustainable models for accessing and using public sector cultural heritage content.
This paper has described how value is amplified by aggregation and contextualisation of cultural heritage from multiple sources, and presented a scenario that highlights the vision of the eCHASE project. The eCHASE demonstrator system that is being developed to showcase the business models being created within the project is introduced. The demonstrator is supported by the eCHASE framework for importing and processing the media and metadata from our partners' collections.



Results 1 - 15 of 191

Portrait of a lady

Applet uk.ac.soton.itinnovation.echase.lightbox.LightBoxApplet started

Thesaurus Entry

Hierarchy 1.4
1

Thesaurus

Language en-GB
Name Alinari Keywords
0 1

Thesaurus Item

Language en-GB
Name Painting
Preferred 1

Media

image/jpeg

Legacy Reference ID alinari-CAL-F-001743-0000

Figure 5: Screen shot of the light box.

Acknowledgements

eCHASE is a 3.5 MEURO eContent project that started in January 2005 and runs for 2 years. eCHASE 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. 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.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] Google Maps API, http://www.google.com/apis/maps/, 2005

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

[7] Taverna. http://taverna.sourceforge.net, 2005.

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

[9] World Gazetteer, http://www.world-gazetteer.com/, 2005