DELIVERABLE

Project Acronym: **CARARE**

Grant Agreement number: **250445**

Project Title: ***Connecting ARchaeology and ARchitecture in Europeana***

D4.4.1 Live harvesting system for CARARE

Revision: final

Authors:

Dimitris Gavrilis, Costis Dallas, Panos Constantopoulos, Christos Papatheodorou, Agiatis Benardou, Stavros Angelis, DCU

Vassilis Tzouvaras, Kostas Pardalis, Arne Stabenau, Fotis Xenikoudakis, Despoina Trivela, Eleni Tsalapati, Nasos Drosopoulos NTUA

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Revision	Date	Author	Organisation	Description
0.1	28/03/2011	Dimitris Gavrilis, Costis Dallas, Panos Constantopoulos, et al.	DCU	Draft
0.2	22/4/11	Dimitris Gavrilis	DCU	Second version incorporating comments from Kate Fernie
0.3	10/5/11	Dimitris Gavrilis	DCU	Third version incorporating comments from Christian Ertmann-Christiansen, Rob Davis and Vassilis Tzouvaras

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

1. Executive Summary

The goal of the project is to harvest content regarding archaeology and architecture (mostly monuments) from a number of content providers, and deliver it to Europeana. The technical architecture has been implemented and supported by the two technical partners of the project: National Technical University of Athens (NTUA) and Digital Curation Unit of the Athena Research Centre (DCU).

This deliverable describes in detail the live harvesting system for the CARARE project, which allows content providers to provide their content to the CARARE system using a methodology which will allow for regular automated harvests. There are two main components:

a) The CARARE metadata mapping and ingestion system and

b) The CARARE repository.

Section 3 summarises the main components of the aggregation service.

Section 4 summarises the CARARE metadata mapping and ingestion system which was described in full in earlier deliverables (D2.3 and D3.4). The systems went live in April 2011 following testing.

Section 5 describes the CARARE repository and the functionality that it offers to CARARE content providers to search for and manage their content, and the OAI-PMH delivery services provided to supply metadata to Europeana. Content ingestion began in April 2011 when the system went live following testing.

Section 6 confirms the metadata formats that CARARE will deliver to Europeana.

This report presents the CARARE harvesting system and its two main components, the metadata mapping and ingestion tool and the repository service, which both went live during April 2011 to coincide with the delivery of a series of training workshops for content providers. The harvesting of content via the mapping tool to the CARARE repository has now begun.

2. Introduction

This document presents the live harvesting system for CARARE, which allows content providers to provide their content to the CARARE aggregator for harvesting by Europeana. There are two main components:

a) The CARARE metadata mapping and ingestion system and

b) The CARARE repository.

The CARARE metadata mapping and ingestion system has been described in earlier deliverables:

- D2.2.3 - Metadata Mappings
- D3.3.4 - Briefing paper on metadata mapping and the use of mapping tools

The focus of this deliverable is on the CARARE repository system.

3. CARARE aggregator

The aggregator service developed for the CARARE project consists of:

- A metadata mapping and ingestion tool
- a repository that aggregates content from the content providers
- semantic enrichment services, and
- services to deliver content to Europeana.

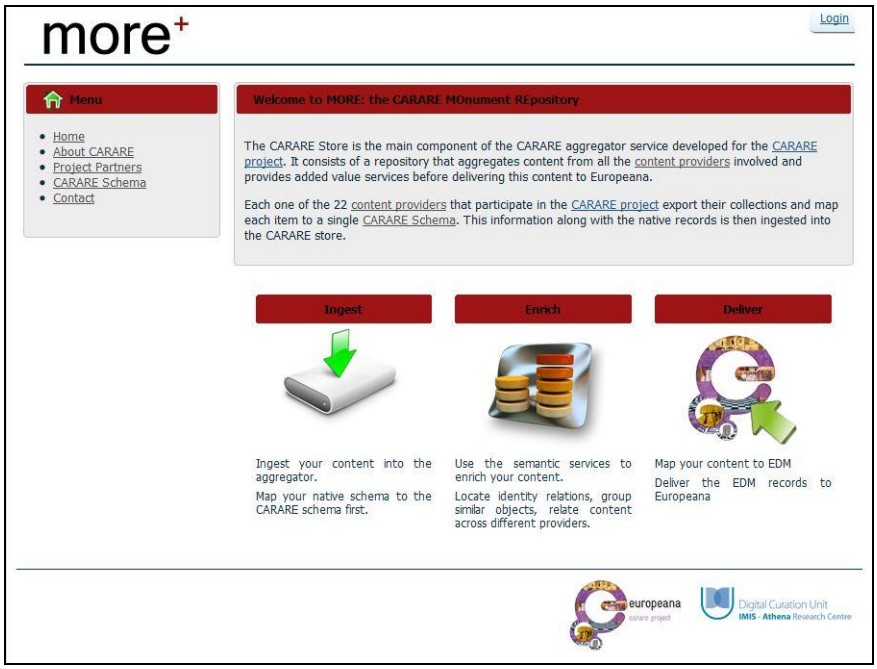

Figure 3.1: CARARE aggregator home page

The CARARE aggregator interfaces with the users (content providers) using web based services while all other services utilize international standards for information encoding and communication (such as XML, REST/SOAP web services, OAI-PMH, etc.).

4. CARARE metadata mapping and ingestion tool

A CARARE metadata mapping and ingestion tool has been provided to enable content providers to provide their content to the CARARE aggregator. The tool was released in September 2010 for testing and evaluation and went live in April 2011. It supports the harvesting/upload of native metadata from the content providers' repositories, mapping to the CARARE metadata schema and ingestion to the CARARE repository (the tool is described in detail in D2.3 and D3.4).

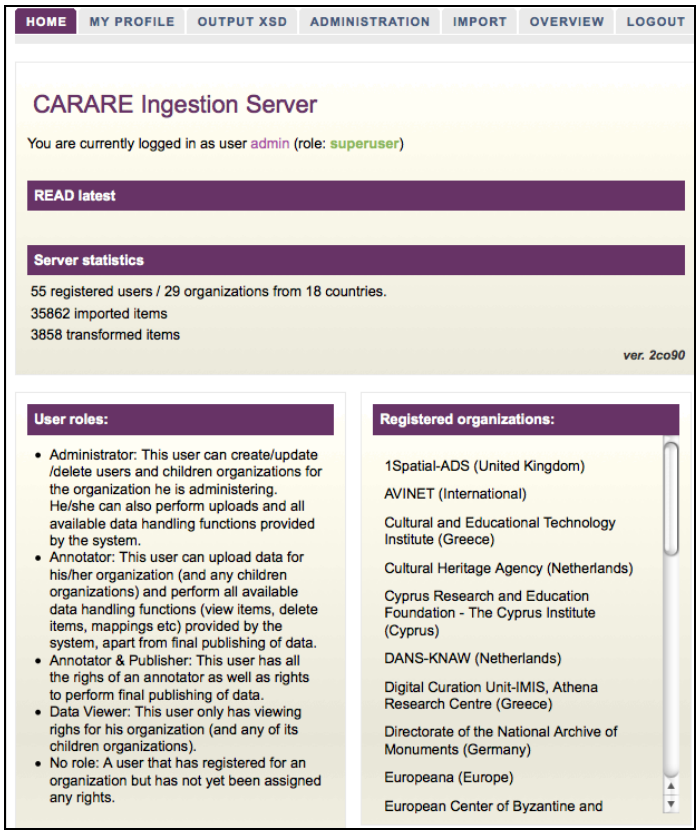

Figure 4.1: CARARE ingestion server

The mapping tool's specifications and functionalities are presented in D3.4. Briefly, the mapping tool provides the following ingest workflow:

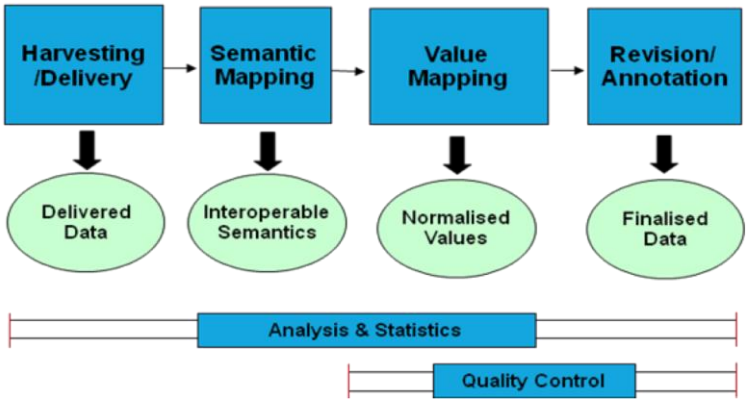

Figure 4.2: CARARE mapping tool main components and workflow

The main components are as follows:

- **Harvesting/delivery** is the process responsible for collecting the metadata from distributed repositories. It is the interface for different methods of data delivery such as HTTP or FTP upload and OAI-PMH.
- **Semantic Mapping** provides the service for assigning semantics to the harvested metadata. A mapping tool assists providers in manually aligning their local schemas to the reference data model. Providers that have metadata in supported known formats will be able to omit this step by using stored transformations from selected schemas to the reference schema based on existing crosswalks.
- **Value Mapping** ensures the correct formulation of values for controlled metadata fields. It enables providers to resolve data issues, e.g. to map their own terminology list to selected terminology lists and to automate data normalization according to selected vocabularies and best practices for values such as dates, geographical locations, nationality/language, etc.
- **Revision/Annotation** enables the revision of the transformation results as well as the editing or addition of data that is not in the original metadata (e.g. empty fields or, ones that take values from controlled vocabularies).
- Across the four phases, a set of tools for **Analysis & Statistics** provides detailed information on the metadata contributed by a provider (i.e. number of items imported, total values per field etc), while **Quality Control** procedures will automatically check and report on ingested metadata (i.e. missing values, malformed data).

Content providers can upload their content using a variety of data formats:

- XML in any schema.
- zip archives of the above
- HTTP upload; suggested only for relatively small amounts of data (<2MB)
- Upload to a dedicated FTP server.

- Remote HTTP or FTP browsing.
- OAI-PMH repository harvesting.
- SuperUser uploading from local file system (restricted).

5. CARARE repository

The CARARE repository holds the content providers' metadata and performs a number of functions that will aim at enriching these metadata. The repository consists of a set of storage spaces and of a set of services that perform various operations on the metadata.

If the repository is seen as a black box, it receives its input from the harvester and outputs to Europeana via the export module. Inside there are four main partitions where data is stored / transformed / enriched, etc. These four partitions and their interconnectivity can be seen in the Figure below. The revision / ingest / enrich and mapping services operate on these four partitions in some cases automatically or in some others (e.g. the enrichment service) triggered by the users' actions.

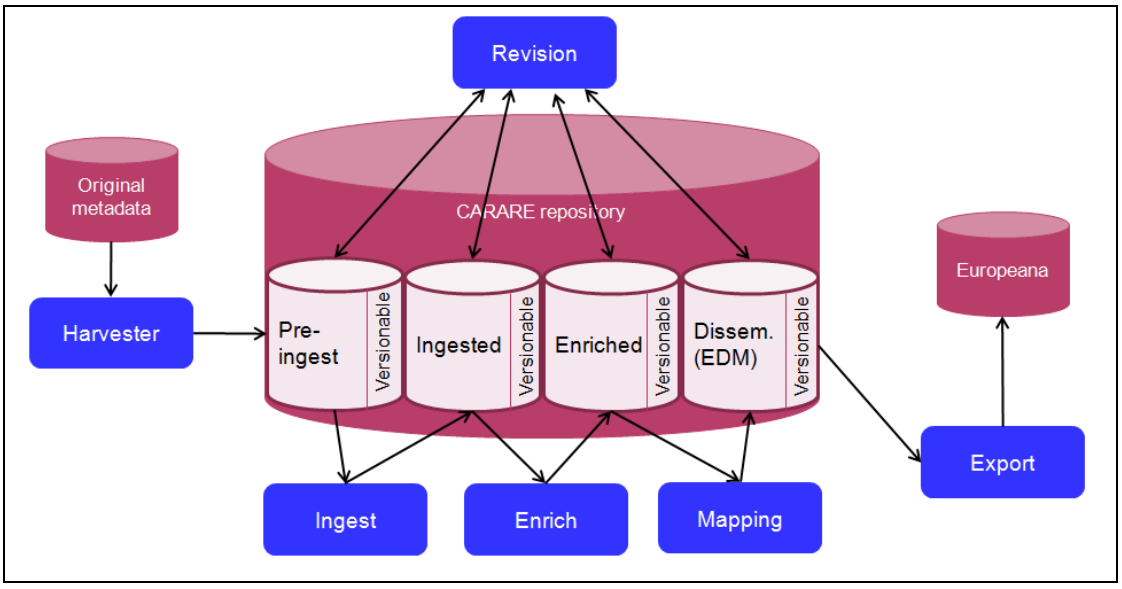

Figure 5.1 The CARARE repository main partitions.

Content is ingested into the repository using the respective ingest service in the form of submission packages. Once the content is ingested into the repository, the content providers can utilize a number of services in order to perform certain tasks (that aim to enrich the ingested content).

Content providers register with the repository to obtain a username and password to access the repository services. Users can login and search/browse through their collections. After locating a

digital object, the system is able to display the various metadata schemas:

- Native metadata
- CARARE schema
- EDM schema

This functionality allows content providers to inspect the digital object and locate any errors or make improvements if necessary. Any modifications must be made at the content provider's source repository and re-ingested into the CARARE repository following the ingest process described above.

The CARARE repository keeps track of the different versions of the metadata schemas (both the Native and CARARE). For every ingest that is performed, the mapping tool pushes 2 files to the repository: the Native schema and the CARARE schema.

The initial content is received through the mapping tool in form of submission packages.
Each package contains at least one item. Each item contains a series of distinct datastreams. These are:

- The administrative metadata for the item (e.g. content provider and user information, basic record information, etc.).
- The Native record in XML format
- The CARARE record in XML format.
- The mapping XSLT document used for transforming the Native record to CARARE.

These packages are received from the repository's ingest service, are checked and if they pass all checks, an archival package is created and passed on to the versioning and indexes services. The versioning service determines where to ingest each item (contained in the submission package). The indexing service extracts all index information that will be used from the rest of the repository's services.

Inside the repository, a set of services such as enrichment and geo-related services operate automatically or based on users' interaction. These services mainly operate on CARARE XML records and produce dissemination packages that are presented to the users. Dissemination packages will be sent to the Europeana in form of EDM XML records.

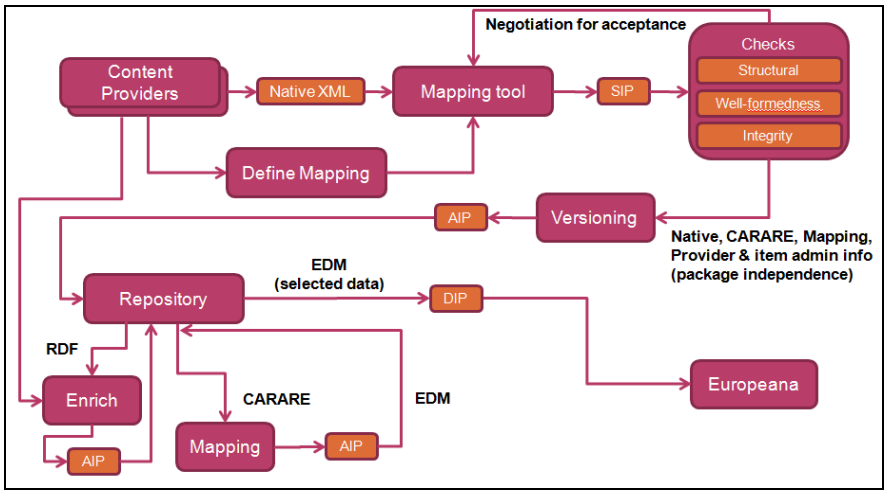

Figure 5.2. The information flow from the CARARE aggregator

Search functionality

Once ingested into the repository, the items are able to be retrieved by the users and enriched using the repository's added value services. The retrieval process utilizes the indexes extracted for each record. Since all CARARE records contain four top level elements, the search query forms allow the user to search specifically for a top level element. So the user can search for heritage assets, digital resources, activities.

This provides greater flexibility and allows the users to perform complex searches and get accurate results. The main functionalities of the search process are presented below:

- **Type of search**
  Simple search. This allows the user to search across all indexes for the specific top level element.
  Advanced search. This allows the user to define separately multiple terms for each index.
- **Scope**
  The scope of the search allows the user to search within a specific (or all) content provider's records and also to specify in which collection to search.
- **Labels**
  User defined labels provide a flexible way of categorizing objects. This provides the users with the means to quickly group objects together.

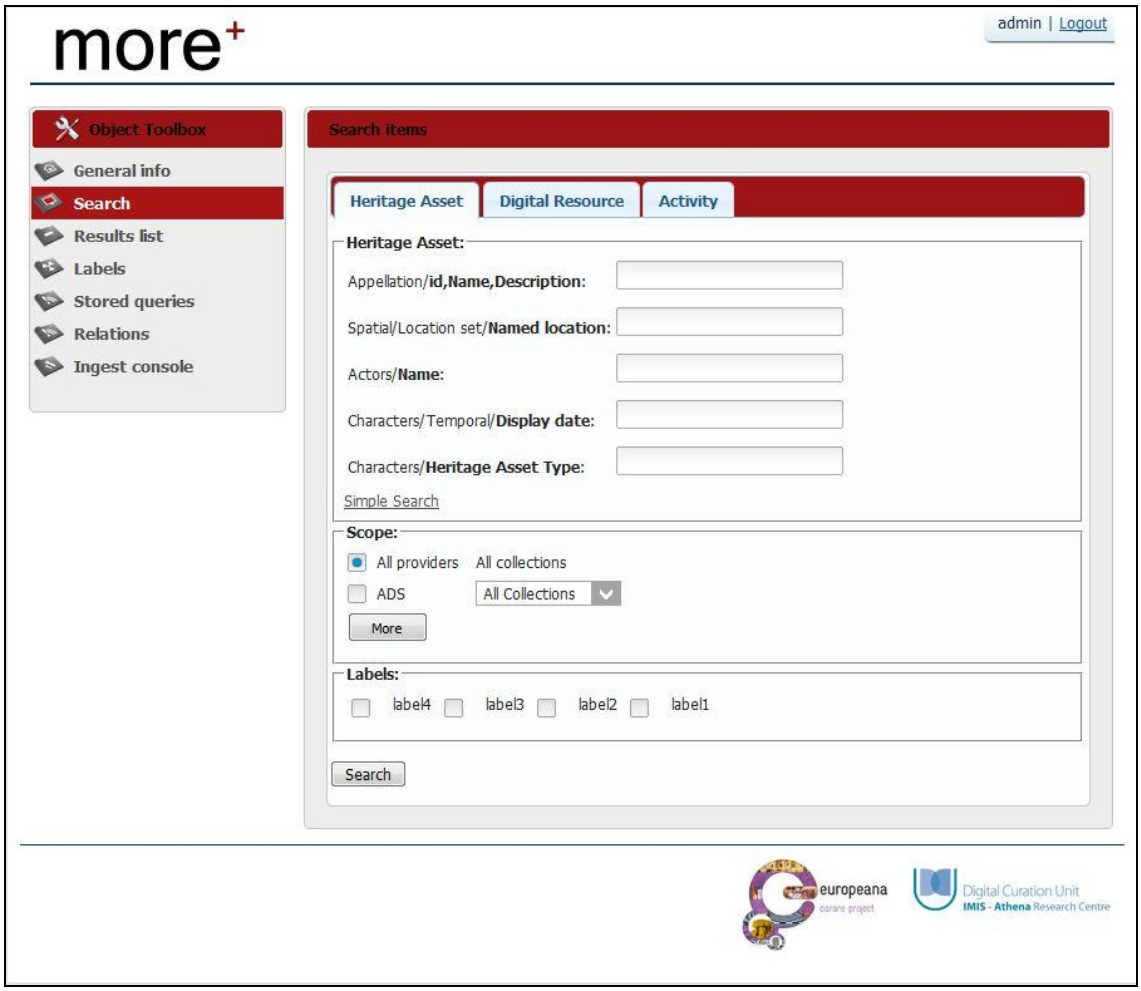

Figure 5.3 Search functionality in the CARARE repository

Object information

Once the search is performed, the search results page is presented and the user can select the object he/she wishes to display. Within the search results screen, the user can store the search query so that he/she can re-submit later. This is a convenient way of saving complex or frequently used searches.

Once inside an object, the user is presented with the brief information about the object. A set of object specific actions appear on the left navigation menu:

- Object information
  This option displays the available datastreams in human readable format (html).
- Object information (XML)
  This option displays the available datastreams in XML format.
- Relations
  This option presents the user with the enriched relations for this object. The user can call the relation editor to add more relations or he/she can delete existing ones.

Permissions

The permissions action allows the user to activate/withdraw or enable/disable an object.

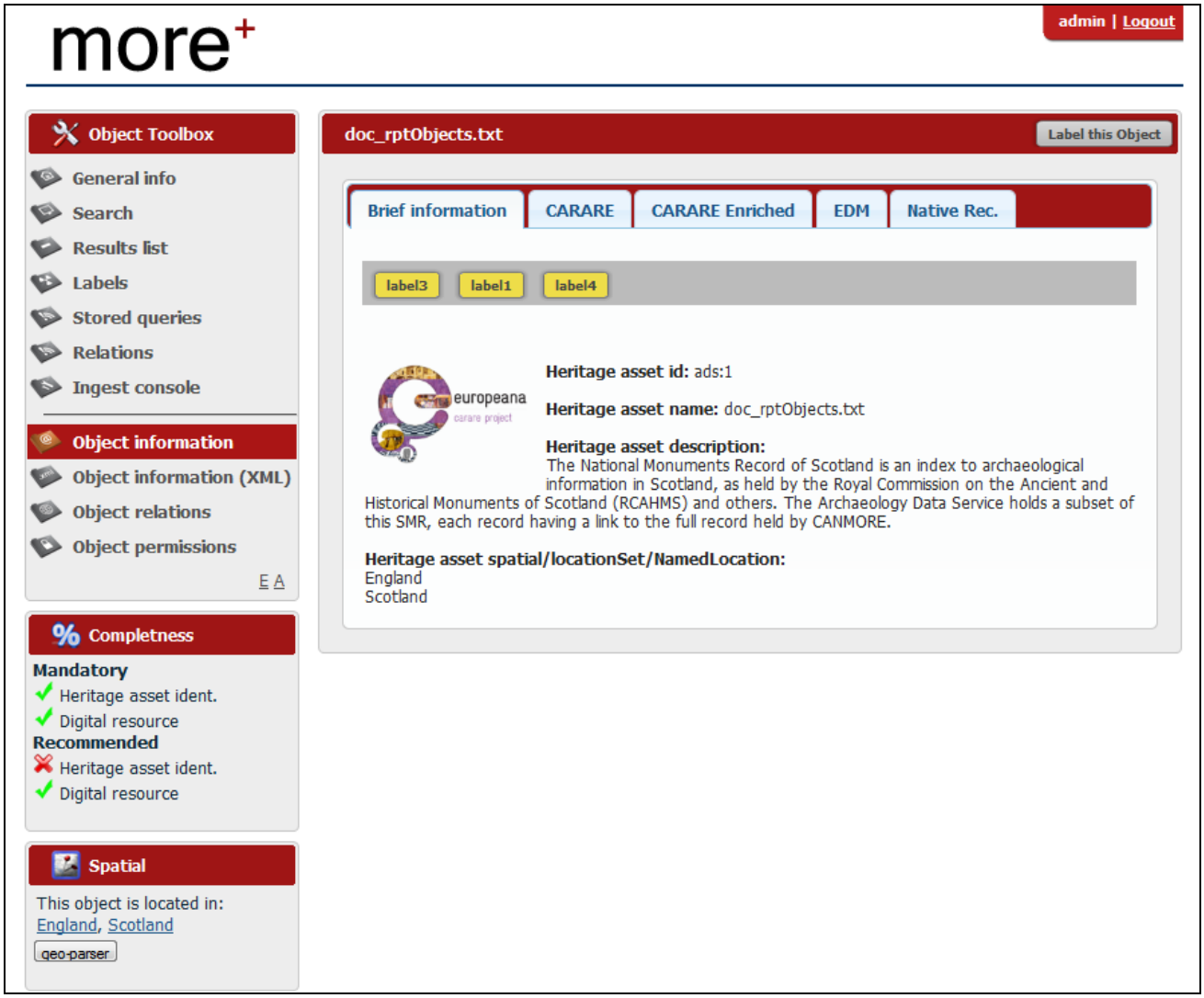

Figure 5.4. A screenshot of the Brief information display of an object.

Permissions

The permissions in the CARARE repository allow the user to disable an object or to withdraw it. If an object is disabled, it cannot be retrieved in search results and thus it cannot be enriched.

If an object is withdrawn, it can be retrieved and enriched but it is not exported to Europeana.

Relations editor

The relations editor is a powerful mini-sized component that is called from within an object and allows to locate related objects and relate them to each other with EDM supported relations. The relations editor search forms are identical with the search forms found in the repository. However, the search results and object preview are displayed in one screen and allow for fast browsing of objects. Once the user locates a related object he/she can relate it and quickly go back to continue from where he/she left of.

The whole process is implemented using AJAX so that the user does not need to refresh the page and re-submit any search forms, etc.

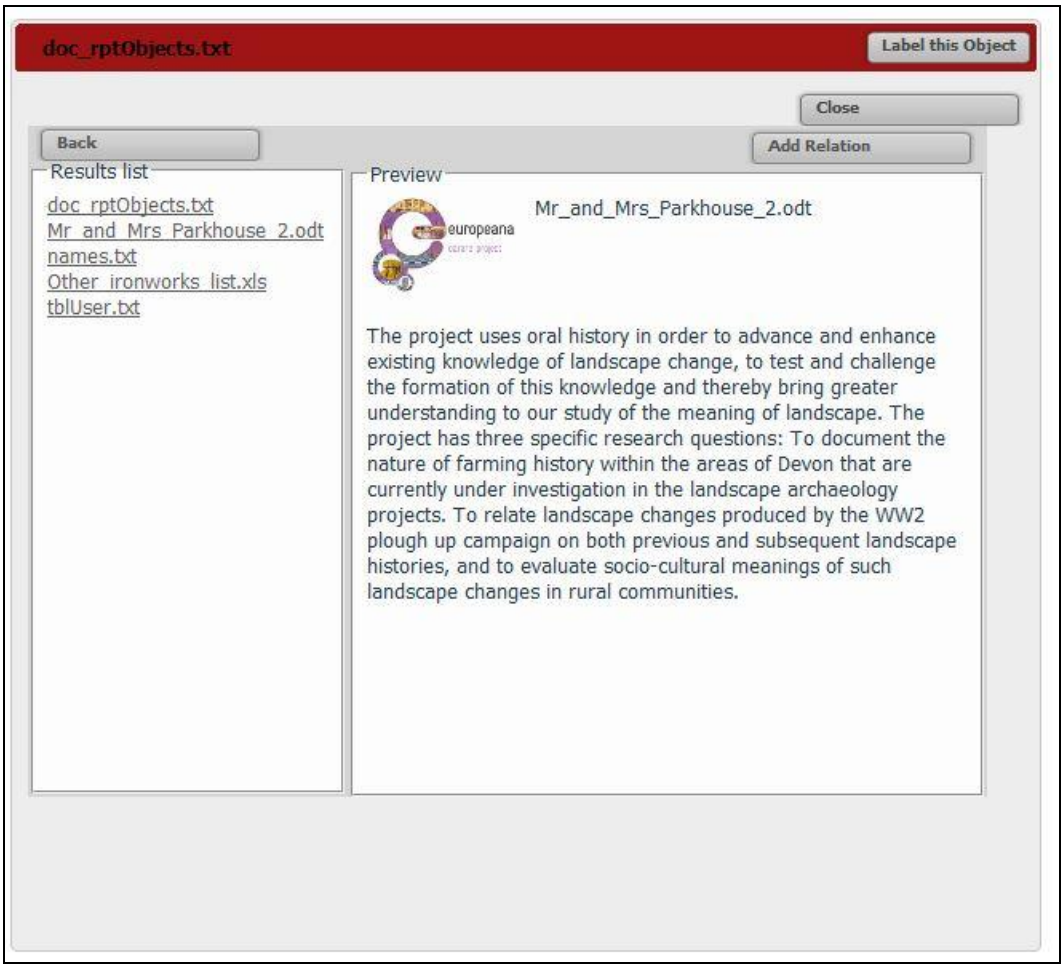

Figure 5.5: Previewing an object in the CARARE repository

Metadata completeness check

The CARARE repository presents the user with additional services such as a completeness report for each object. The completeness check is presented to the user via an information block on the left and the check is performed for mandatory and recommended elements and for each top-level element.

Figure 5.6: Completeness check and spatial information blocks.

Spatial information

The spatial information block presents the user with the spatial information extracted from the record. This information can be in form of coordinates or place names. Europeana geo-parser services are also offered to enable users to cross reference any named locations.

Temporal information

The temporal information block presents the user with the temporal information extracted from the record.

6. Delivery to Europeana

CARARE will deliver metadata to Europeana in EDM using the OAI-PMH protocol and in XML format. Delivery of metadata will commence following the implementation by Europeana of ingestion services to support metadata which are due to be implemented by the end of June 2011.

7. Conclusions

This report presents the live CARARE harvesting system. It illustrates the two distinct main components: the repository and the mapping tool and how they interoperate to deliver content from the content providers to Europeana. The system has been implemented according to the specifications described in the technical approach of the project (D2.5) and is now live and is accepting content.

8. References

D2.2.3 - Metadata Mappings: http://www.CARARE.eu/eng/Resources

D2.2.5 – White paper on the CARARE technical approach: http://www.carare.eu/eng/Media/Files/White-paper-on-CARARE-technical-approach

D3.3.4 - Briefing paper on metadata mapping and the use of mapping tools: http://www.CARARE.eu/eng/Media/Files/D3.4-Briefing-paper-on-metadata-mapping-and-the-use-of-mapping-tools

CARARE Metadata Ingestion and Mapping Tool. http://carare.image.ntua.gr/carare/

CARARE Repository. http://store.carare.eu/

In this section you may find links to the services and technologies that are used or described throughout this document.

- Extensible Markup Language (XML). http://www.w3.org/XML/
- Java. http://www.java.com/
- OAI-PMH. http://www.openarchives.org/
- REPOX. http://repox.ist.utl.pt
- Web services. http://www.w3.org/TR/ws-arch/
- XSLT. http://www.w3.org/TR/xslt
- XML Schema. http://www.w3.org/XML/Schema