DELIVERABLE

Project Acronym: **CARARE**
Grant Agreement number: **250445**
Project Title: ***Connecting ARchaeology and ARchitecture in Europeana***

D3.3.1 Tested harvesting and ingestion system

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

With contributions from:

Claus Dam, KUAS

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 is being 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). The architecture specifies a three stage process, described below.

This deliverable describes in detail the testing of the CARARE harvesting and ingestion system. This system is comprised of two main components: a) the mapping tool (NTUA) and b) the CARARE repository (DCU). The first component allows the content providers to upload their metadata (or for it to be harvested by the tool) and map them to the CARARE metadata schema. The metadata is then transformed; the submission packages are created and are streamed to the repository.

The points of focus in this deliverable are the following:

- The harvesting of the content providers' metadata into the mapping tool
- The mapping to the CARARE schema
- The creation of the submission packages and their ingestion to the repository

The tests described throughout this document aim at ensuring that the CARARE system as described in the technical approach (D2.5) is capable of delivering real content to Europeana by following the entire chain of processes involved. The implementation and configuration of the services as well as their efficiency are measured. The tests focused mainly on the CARARE mapping tool (MINT) and the repository (MORE) as well as the communication protocol between them. During the testing, a small number of content records in native schema were ingested into the mapping tool, mapped into CARARE, ingested into the repository, enriched and then transformed into EDM. For these tests, both artificial and real data from the project content providers were used.

2. Introduction & Related Work

This document presents the testing of the harvesting and ingestion system that will be used in the CARARE project. There are several key components that participate in this system: an OAI repository at the content provider, the harvester, the mapping tool, the CARARE repository and the ingestion mechanism. All technologies, functionalities and protocols that will be used throughout the system are presented.

The work that was carried out in this deliverable made extensive use of the metadata used in CARARE and the mapping tool that were presented in previous deliverables of the project. These deliverables are:

- D2.2.3 - Metadata Mappings
- D2.2.5 – Technical Approach
- D3.3.4 - Briefing paper on metadata mapping and the use of mapping tools

3. Content Providers' Repository and Harvesting

Each content provider must be able to expose their metadata as XML for provision to the CARARE aggregator. A number of methods for providing their data are supported including:

- 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.

Use of the OAI-PMH protocol is preferred for sustainability and so that the- metadata can be harvested at any time by the harvester. Some of the content providers' repositories have already integrated this technology.

Training is being offered to all content partners on how to set up XML output formats and on how to install and configure an OAI-PMH service based on the REPOX repository. REPOX was been developed by the Technical University of Lisbon as part of The European Library project (TEL). It can be deployed as a standalone OAI-PMH server. It is open source and has been developed in Java (hence it can be installed and run in almost all systems) and it is capable of managing multiple internal data sources. Another key characteristic is that it provides a user interface to the content provider in order to define metadata crosswalks.

The REPOX system was presented in the Pisa workshop and content providers have been invited to three training workshops (Cologne, Jaen and Athens) where they were trained on how to install, configure and use REPOX as part of the programme.

The OAI-PMH repository harvesting method (harvester) is part of the CARARE metadata mapping and ingestion tool and it constitutes the preferred methodology of uploading content. Compliance between the harvester and OAI-PMH targets including REPOX has been tested.

4. CARARE metadata mapping and ingestion Tool

The mapping tool (http://CARARE.image.ntua.gr/CARARE) transforms the content received by the harvester using a mediating schema (the CARARE schema), creates the submission packages and sends them to the CARARE repository for ingestion. The tool also provides the content providers with the means to map their metadata schema to the mediating schema thus defining the transformation between their native schemata and the mediating schema (by producing an XSLT document).

A user evaluation test of the CARARE was carried out to evaluate the usability of the tool from the perspective of content providers being able to

- Import their metadata
- Map their metadata to the CARARE schema
- Review their imported and mapped metadata against the CARARE schema

- Carry the user as far in the ingestion process as the tool in its present (time of test) state of development allows

A test group of content providers took part in the testing which was coordinated by KUAS between late November 2010 and early January 2011. The test users carried out a series of tasks and reported their findings to a moderator. The test results were logged and a detailed report was provided to NTUA. The general findings were:

- Import of metadata is well supported
- The general mapping procedures are clear with some additional support needed to take advantage of the full scope of possibilities offered by the tool
- The procedure for reviewing output in CARARE schema format could be made clearer
- Users asked for an overview of missed mandatory mappings
- The graphical look and feel of the tool was liked by users.

The CARARE metadata mapping and ingestion tool was updated and improved as a result of the usability testing.

4.2 CARARE Schema XSD

The CARARE metadata schema was implemented in an XSD for integration into the CARARE metadata mapping and ingestion tool. The XSD was tested by users completing mappings of their metadata using the mapping tool and reported their findings to the technical team. The general findings of the testing of the XSD were:

- The inclusion of annotations explaining the use of each element helped improve the usability of both the XSD and the mapping tool
- Specifying less mandatory elements facilitates the mapping process and transformation of native metadata to CARARE schema format
- Specifying recommended elements highlights key elements to users
- The repeating elements and repeating groups were refined.

The CARARE Schema XSD was updated and new versions were released as a result of the testing.

5. 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. A graphical representation of the repository can be seen in figure 5.1 below.

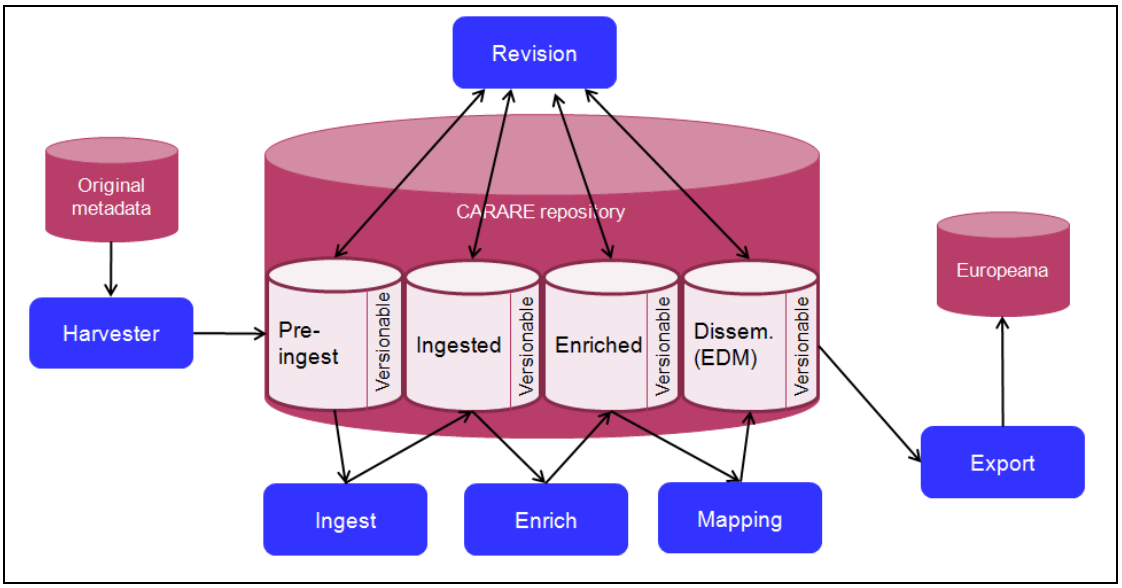

Figure 5.1 The CARARE repository main partitions.

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). In order to access the repository services, the content providers must obtain a username and password. With these credentials, they can login and search/browse through their collections. After locating a digital object, the system will be able to display the various metadata schemas:

- Native metadata
- CARARE schema
- EDM schema

Allowing the content provider to inspect the digital object and locate any errors or make improvements if necessary. Any modification 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 will keep track of the different versions of the metadata schemas (both the Native and CARARE). So for every ingest that is performed, the mapping tool will push 2 files to the repository: the Native schema and the CARARE schema.

5.1 Ingest protocols

In the present architecture, there are two points where an ingest action is identified:

1. The ingest between the content provider and the MINT tool
2. The ingest between the MINT tool and the repository

In the first case, the OAI-PMH protocol is used (http://www.openarchives.org/OAI/openarchivesprotocol.html). OAI is a well established and widely known protocol, many open source and proprietary repositories support it. Since content providers only have to send metadata to the MINT tool, OAI is sufficient and simple.

The second case is somewhat more complicated since a number of distinct objects have to be ingested into the repository. This is because the repository offers two key services: a) preservation of objects and b) version control. For this reason, and following the OAIS standard, a submission package is created and is ingested into the repository using a REST a protocol. The ingest service negotiates the submission and in case where an structural or integrity checks fails, an error is returned. The submission package consists of the following files:

- The native metadata record of the object (an XML document)
- The CARARE metadata record of the object (an XML document)
- The mapping mechanism between them (the XSLT document)
- An XML document containing administrative information on the object (content provider info, object native identifier and name, user that performed the transformation, etc).

6. Testing scope and methodology

The purpose of the testing that was carried out and described in this deliverable was to ensure that the CARARE system's main components (the mapping tool and the repository) are functioning as they are described in the technical approach of the project and to ensure that the CARARE system as a whole is also functioning within the defined specifications.

The tests that were performed aimed mainly at these two components and as long as their interconnectivity (the communication protocol between them). More specifically, the most important functionalities that were tested are:

- The mapping process (usability, mandatory elements checks, etc.)
- The ingest process from the content providers to the mapping tool
- The search functionality and usability of the repository
- The enrichment functionality of the repository
- The mapping to EDM

The testing methodology involved the use of both artificial and real content (taken from some of the content providers). The content was ingested into the mapping tool, traversed the entire chain of processes involved,

was delivered to the repository, enriched and mapped to EDM. A group of archaeologists and information scientists were involved in the testing and evaluation process. The feedback from these experts was used to improve on the implementation of the various services.

7. Testing of the repository services

The testing of the services described above has been done in many levels:

- Creation of the submission packages using real content from a few content providers
- Testing of the repository ingest service
- Testing the repository services

The submission packages provide a robust and efficient method of transferring information from the MINT tool to the repository. The submission packages use a submission ingest protocol (SIP) which has been designed specifically for CARARE and has been implemented by the MINT tool and the repository. The benefits of using SIP are that:

- It can transfer large amounts of data since it uses compression for encoding this data.
- It can contain different datastreams for each item.
- It can implement a negotiation protocol before accepting a package.
- It can transfer both binary and XML datastreams

As it can be seen from Fig. 7.1 below, the content providers' native metadata (in XML format) are harvested by the MINT tool and are mapped into CARARE. Then, a SIP package is created (containing one or multiple items) which is harvested by the repository. Each package goes through a series of checks before it is accepted. These checks can be split into three levels:

- Structural. The contents and structure of each item is verifired.
- Well-formedness. The XML datastreams are checked.
- Integrity. For each package all necessary information for its ingestion is located and verified (native identifier, item label, provider identifier).

After it has been accepted, an archival package is created and each item is ingested into the repository as a new item or as a new version of an already existing item.

Regarding the test of SIP, the following tests have been carried out:

- A series of tests SIP packages have been created along with testing scenarios that check altogether all cases where something can go wrong. The implementation of SIP has been verified based on these tests.
- A live test using dummy data was carried out with packages from the MINT tool to the repository.
- A real test using actual data from content providers was carried out.

The purpose in all tests was to locate possible errors and return the correct XML error response. For the packages that were verified, all the items they contained were ingested into the repository. In cases where the items already existed, a new version had to be created.

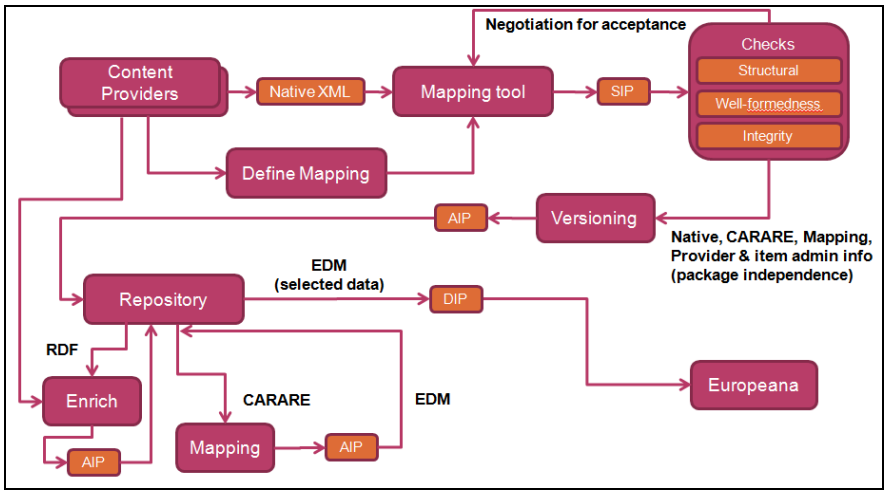

Figure 7.1. The information flow from the CARARE aggregator.

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. item identifier and label, provider identifier information, user 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.

The SIP protocol negotiates the acceptance of each package (see Fig. 7.2 below) using a set of REST based web services that return an XML response for each case. When a package is accepted, its containing items processed by the version control and indexing services of the repository in order to be correctly indexed and versioned. A series of dissemination packages are also created each case an item has to be sent to the previewing service (HTML format) or to Europeana (in EDM format).

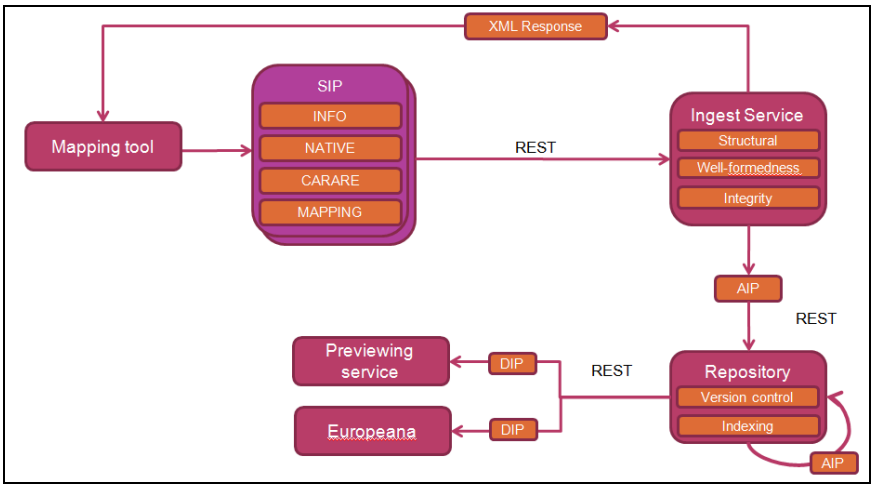

Figure 7.2.Handling of the SIP packages.

Once ingested into the repository, the items must be able to be retrieved by the users and enriched using the repository's added value services. For this case, a number of usage scenarios have been created in order to simulate the users' search operations. Based on these scenarios, a number of three working prototypes were created. Each prototype followed a different concept. Those prototypes were evaluated and a final solution was adopted.

The prevailing approach contained a different search form for each one of the CARARE schema top-level elements (heritage assets, digital resources, activities). Each search form had the following general characteristics:

- A simple and advanced search form
- Allowed the user to define the scope of the search (search across all content providers or a specific one, search inside all collections all a specific one).
- Ability to search records that were labeled using one or more user defined labels.

A screenshot of the heritage asset identification advanced search form can be seen in Figure 7.3 below.

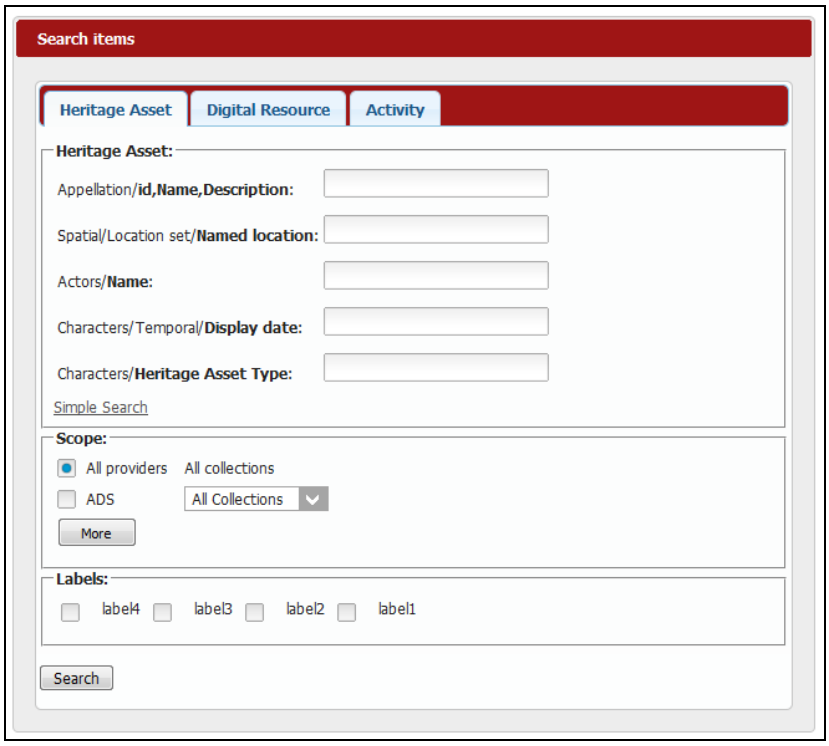

Figure 7.3. A screenshot of the Heritage Asset advanced search form of the repository.

Another crucial service of the repository is to be able to enrich each item using EDM supported relations. These relations would be among items of the repository. The relation service should allow the user to quickly locate related objects from within an object, and set the relation type. In order to achieve this, a usable AJAX based service was used and a number of tests were performed by non-IT users.

The information architecture and presentation was also evaluated and refined. This included the type of information displayed to the user and how this information was organized. This information included the different schemas (NATIVE, CARARE, CARARE Enriched, EDM) along with some other information such as the spatial information that was extracted, the completeness check for the object, possible user-defined labels associated with this record, etc. A example can be seen on Figure 7.4 below. The evaluation of the usability of these services resulted in improvements both in the way the services worked and in their user interface.

The services of the repository that were exhaustively tested were:

- Search service
- Display of search results
- Display of object information
- Object labeling service
- Stored query service
- Relations services
- Completeness check service

- Spatial information extraction service
- Object permissions set service

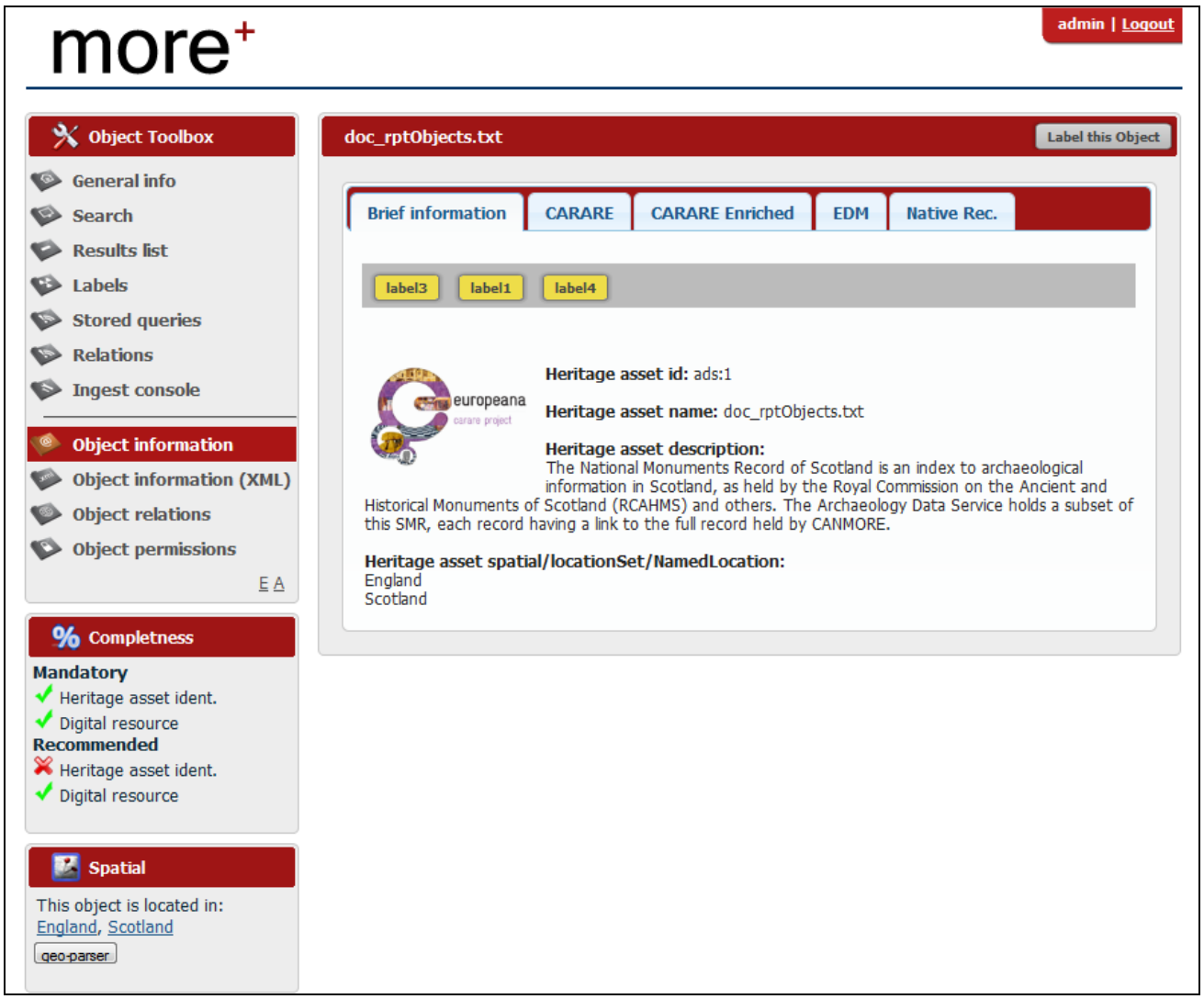

Figure 7.4.A screenshot of the Brief information display of an object.

Finally, since the repository is fully web based a set of browser compatibility tests were performed. The repository was verified to be compliant with the following browsers (latest 2 major versions):

- Mozilla Firefox
- Internet Explorer
- Safari

Also, the repository can be accessed from different screen sizes and resolutions including:

- Normal/Wide screens (resolution 1024x768 or greater)
- Tablet devices (iPad verified)

8. Conclusions

During the testing phase, a number of issues were raised from the evaluation team. These issues focused mainly on the usability of the system and led to the improvement of certain services. Specifically, the retrieval process in the repository was greatly improved throughout the testing process as long as the tool that is used to perform the enrichment of the records.

Apart from the usability issues, certain light modifications were made to the communication protocol and its implementation (the communication protocol is used for the transfer of content from the mapping tool to the repository).

9. References

This section includes links to the services and technologies that are used or described throughout this document.

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

Claus Dam, KUAS, User evaluation (usability test) of the CARARE ingestion tool, internal report, 30th January 2011

CARARE, 2010, Papatheodorou, C., Carlisle, P., Ertmann-Christiansen, C. and Fernie, K., CARARE metadata schema outline, v1.0: http://www.CARARE.eu/eng/Resources/CARARE-metadata-schema-outline-v1.0

CARARE Metadata Interoperability Tool: http://CARARE.image.ntua.gr/CARARE/

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

Gavrillis, D., Angelis, S. and Papatheodorou, C., 2010, Mopseus – a digital repository system with semantically enhanced preservation services, http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/gavrilis-34.pdf

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