ICT-PSP Project no. 270905

LINKED HERITAGE

Coordination of standard and technologies

for the enrichment of Europeana

Starting date: 1st April 2011

Ending date: 31st October 2013

Deliverable Number:	D 5.1
Title of the Deliverable:	Linked Heritage Technology Platform
Dissemination Level:	Public

Contractual Date of Delivery to EC:	March 2012
Actual Date of Delivery to EC:	May 2012

Project Co-ordinator

Company name :	Istituto Centrale per il Catalogo Unico (ICCU)
Name of representative :	Rosa Caffo
Address :	Viale Castro Pretorio 105, I-00185 Roma
Phone number :	+39.06.49210427
Fax number :	+39.06. 06 4959302
E-mail :	rcaffo@beniculturali.it
Project WEB site address :	http://www.linkedheritage.org

Context

WP 5	Technical Integration
WP Leader	NTUA
Task 5.1	Validation Environment
Task Leader	NTUA
Dependencies	

Author(s)	Nasos Drosopoulos, Nikolaos Simou
Contributor(s)	
Reviewers	Dov Winer, Claudio Prandoni

History

Version	Date	Author	Comments
0.1	20/03/2012	Nikolaos Simou	First Draft
0.2	09/04/2012	Nasos Drosopoulos	Final Draft
0.3	09/05/2012	Nikolaos Simou, Nasos Drosopoulos	Final
1.0	15/05/2012	A.Fresa	Submitted to EC

TABLE OF CONTENTS

1 INTRODUCTION

This document reports on deliverable D5.1 (Prototype, Public) “Linked Heritage Technology Platform”, that is made available online for validation and for the large-scale contribution of content to Europeana (in WP6) and for dissemination & training (WP7). The technology team led by NTUA integrated all the necessary components into a common technology platform, starting from the basis of the ATHENA ingestion server. The Linked Heritage Technology Platform provides content holders with the ability to perform in an efficient way the required mapping of their own metadata schemas to the project's reference metadata schema, as well as their publication to Europeana. It is based on NTUA's metadata interoperability platform MINT, that follows a typical web-based architecture offering an expanding set of services for metadata aggregation and remediation. It addresses the ingestion of metadata from multiple sources, the mapping of the imported records to a well-defined machine-understandable reference model, the transformation and storage of the metadata in a repository, and the provision of services that consume, process and remediate these metadata. Although its deployment is also guided by expediency, the system has been developed using established tools and standards, embodying best practices in order to animate familiar content provider procedures in an intuitive and transparent way.

1.1 BACKGROUND

Metadata records are critical to the documentation and maintenance of interrelationships between information resources and are being used to find, gather, and maintain resources over long periods of time. The consistent application of a descriptive metadata standard improves the user's search experience and makes information retrieval within a single collection or across multiple datasets more reliable. Descriptive, administrative, technical, and preservation metadata contribute to the management of information resources and help to ensure their intellectual integrity both now and in the future. In parallel with other domains, many researchers in the digital cultural heritage community recognized the need to lower the barriers for the management and aggregation of digital resources, by implementing some measure of interoperability among metadata standards and then with proprietary data structures. There is a wide range of proposed solutions, including crosswalks, translation algorithms, metadata registries, and specialized data dictionaries.

A crosswalk provides a mapping of metadata elements from one metadata schema to another. The prerequisite to a meaningful mapping requires a clear and precise definition of the elements in each schema. The primary difficulty is to identify the common elements in different metadata schemas and put this information to use in systems that resolve differences between incompatible records. Crosswalks are typically presented as tables of equivalent elements in two schemas and, even though the equivalences may be inexact, they represent an expert's judgment that the conceptual differences are immaterial to the successful operation of a software process that involves records encoded in the two models. A crosswalk supports the ability of a retrieval mechanism to query fields with the same or similar content in different data sources; in other words, it supports semantic interoperability.

Crosswalks are not only important for supporting the demand for single point of access or cross-domain searching; they are also instrumental for converting data from one format to another. However, aggregating metadata records from different repositories may create confusing display results, especially if some of the metadata was automatically generated or created by institutions or individuals that did not follow best practices or standard thesauri and controlled vocabularies. Mapping metadata elements from different schemas is only one step in the implementation of a crosswalk. Another level of semantic interoperability addresses datatype registration and formatting of the values that populate the metadata elements, e.g. rules for recording personal names or encoding standards for dates, and the alignment between local authority files and adopted terminologies.

The Linked Heritage Technology Platform implements an aggregation infrastructure offering a crosswalk mechanism to support subsequent critical activities:

- harvesting and aggregating metadata records that were created using shared community standards or proprietary metadata schemas,
- migrating from providers' models (whether standard or local) to a reference model,
- transforming records from the Linked Heritage model to the Europeana Semantic Elements and the Europeana Data Model.

1.2 ROLE OF THIS DELIVERABLE IN THE PROJECT

This deliverable corresponds to the technology platform that implements Task 5.1 "Validation environment". The software services are employed for the aggregation of metadata in order to realize the content delivery to Europeana as it is planned and executed within WP6: "Coordination of Content". The validation environment enables users to

- Provide metadata records in a range of "source" formats
- Convert metadata to the Linked Heritage metadata model
- Map local terminologies to the adopted reference terminologies
- Submit the records to Europeana via the Linked Heritage gateway

Specifications are informed by the results of WP2 and WP4 concerning metadata modeling for the public and private sector, as well as by the adoption of terminologies and development of the terminology system designed in WP3. The adoption of LIDO as the metadata model of reference for the project is one of the important decisions of WP2 that guided deployment. The platform is used to realize Task 6.1: "Content delivery to Europeana" and is assessed and refined within the same WP and specifically Task 6.2: "Feedback". The delivery of the system is a project milestone (MS7 " Ingester ready for delivery of content to Europeana'', Month 12) and constitutes the starting point for mapping and delivery of metadata records to Europeana.

1.3 APPROACH

Work for this deliverable was carried out by the technology team led by NTUA, that integrated all the necessary components into a common technology platform starting from the basis of the ATHENA ingestion server. The ingestion tool is based on the metadata interoperability services suite (MINT) that is developed and maintained by NTUA. MINT services compose a web based platform that was designed to facilitate aggregation initiatives for cultural heritage content and metadata in Europe. It is employed from the first steps of such workflows, corresponding to the ingestion, semantic alignment and aggregation of metadata records, and proceeds to implement a variety of remediation approaches.

The platform has been deployed for a variety of aggregation workflows corresponding to the whole or parts of the backend services. Specifically, it has served the aggregator of museum content for Europeana (and one of the largest in volume and significance), the ATHENA project, that has ingested from 135 organizations over 4 million items, which users aligned to the LIDO format. The resulting repository offers an OAI-PMH interface exposing the records in the Europeana Semantic Elements schema. The use of a reference model allowed the rapid support of updated ESE versions that were introduced in the duration of the project (2008-2011), with minimal input from providers. User efforts to align their data to an adopted domain model motivated them to update their collection management systems and improve the quality of their annotations in order to take advantage of a well defined, machine understandable model and, subsequently, control and enrich their organization's contribution and visibility through Europeana.

The EUscreen project also follows the same aggregation workflow for Europeana while, in addition, it provides a portal for Europe's television heritage where both the video content and metadata records are offered to users. The metadata records for the portal are based on the selected reference models (EUscreen and EBUcore) for which an item annotator was introduced. MINT serves the aggregation and remediation of records both for the portal (also offering the Lucene indexes for the search engine) and Europeana (OAI-PMH for ESE).

Similarly, the CARARE, JUDAICA Europeana, ECLAP, DCA and LinkedHeritage projects utilize MINT to accommodate their aggregation and remediation requirements for their specific domain and project, and for Europeana. The group is actively participating in several metadata modeling activities such as LIDO, where the tool is used for the presentation and revision of the LIDO schema by the corresponding CIDOC working group, EDM, for the prototyping of the Europeana Data Model harvesting XSD and RDFS ontology, and EBU Tech. It is also involved in the development of Europeana and the Digital Public Library of America. The growing user base (more than 350 cultural heritage organizations and 500 users) contributes to its ongoing development, improvement and support, while the first version, MINT-Athena, was released under an open source license in July 2011.

During the first year of the project the workflow for the project's content delivery was defined and modeling requirements concerning the Linked Heritage aggregation schema (LIDO) and related terminologies were put in place by the respective work packages. WP5 implemented these specifications in the technology platform that is evaluated and validated within WP6. Integration tasks also involved the inclusion of an online process for review and acceptance of the Data Exchange Agreement that providers sign with Europeana as well as the design and implementation of APIs to interface with services developed for the terminology management (WP3).

The platform is hosted by NTUA and administered in collaboration with WP6. Users can register and access the services at http://mint-projects.image.ntua.gr/linkedheritage

1.4 STRUCTURE OF THE DOCUMENT

The rest of the document presents more details on the aggregation workflow that is implemented in the Linked Heritage Technology Platform (Section 2), emphasising on the metadata mapping tool (Section 3) and, closing with detailed technical specifications for the implementation (Section 4).

2 METADATA AGGREGATION PLATFORM

The key concept behind the introduced aggregation workflow is that, although simpler models lower the barrier to entry and initially reduce the required effort and cost (an approach followed in the first stages of Europeana through the introduction of the Dublin Core-based ESE data model), a more expressive, domain-specific model reinforces the homogeneity and richness of the aggregation. Moreover, since the technological evolution of consuming services for cultural heritage is greater than that of most individual organizations, a richer schema, like in this case LIDO, allows harvesting and registering of all annotation data regardless of the current technological state of the repositories or their intended (re-) use.

The developed system facilitates the ingestion of semi-structured data and offers the ability to establish crosswalks to the reference schema in order to take advantage of a well-defined, machine understandable model. The underlying data serialisation is in XML, while the user's mapping actions are registered as XSL transformations. The common model functions as an anchor, to which various data providers can be attached and become, at least partly, interoperable.

Key functionalities include:

- Organization and user level access rights and role assignment.
- Collection and record management (XML serialisation).
- Direct import and validation according to registered schemas (XSD).
- OAI-PMH based harvesting and publishing.
- Visual mapping editor for the XSLT language.
- Transformation and previewing (XML and HTML).
- Repository deployment and remediation interfaces.

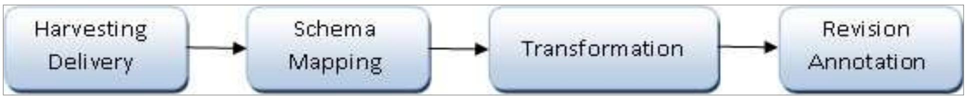


Figure 1. Ingestion workflow

The metadata ingestion workflow, as illustrated in *Figure 1*, consists of four main steps. First is the Harvesting/Delivery procedure, which refers to the collection of metadata from content providers through common data delivery protocols, such as OAI-PMH, HTTP and FTP. Following is the Schema Mapping procedure, during which the harvested metadata are mapped to the common reference model. A graphical user interface assists content providers in mapping their metadata structures and instances to a rich, well defined schema (e.g. LIDO), using an underlying machine-understandable mapping language. Furthermore, it provides useful statistics about the provider’s metadata while also supporting the share and reuse of metadata crosswalks and the establishment of template transformations. The third step is the Transformation procedure, which also aims at the transformation of the content provider's list of terms to the vocabularies and terminologies introduced by the reference model. The last step is the Revision/Annotation procedure that enables the addition and correction of annotations, group editing of items in order to assign metadata not available in the original context and, further transformations and quality control checks according to the aggregation guidelines and scope (e.g. for URI/URLs).

3 MAPPING EDITOR

Metadata mapping is the crucial step of the ingestion procedure. It formalizes the notion of a metadata crosswalk, hiding the technical details and permitting semantic equivalences to emerge as the centerpiece. It involves a user-friendly graphical environment (*Figure 2* shows an example mapping opened in the editor) where interoperability is achieved by guiding users in the creation of mappings between input and target elements. User imports are not required to include the respective schema declaration, while the records can be uploaded as XML or CSV files. User's mapping actions are expressed through XSLT style sheets, i.e. a well-formed XML document conforming to the namespaces in XML recommendation. XSLT style sheets are stored and can be applied to any user data, exported and published as a well-defined, machine understandable crosswalk and, shared with other users to act as template for their mapping needs.

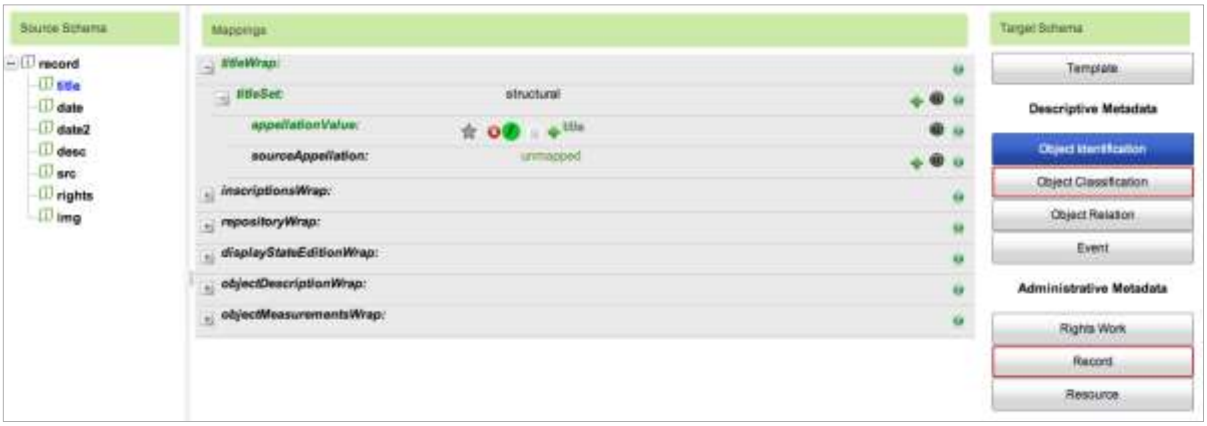


Figure 2: Screenshot of the mapping editor

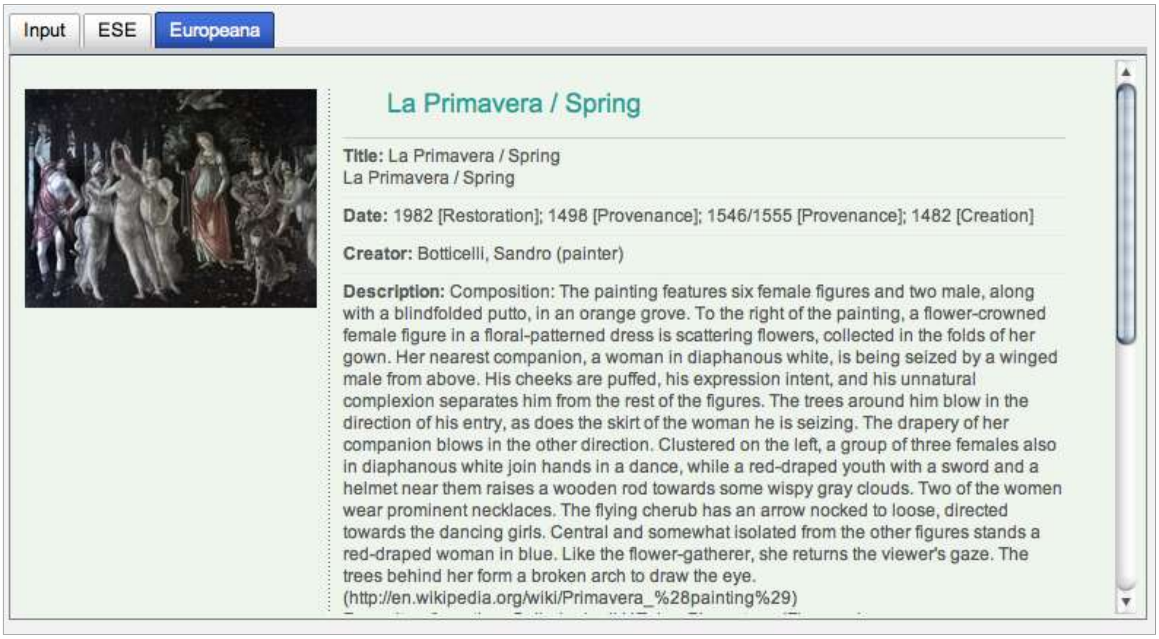
The structure that corresponds to a user's specific import is visualized in the mapping interface as an interactive tree that appears on the left hand side of the editor. The tree represents the snapshot of the XML schema that is used as input for the mapping process. The user is able to navigate and access element statistics for the specific import while the set of elements that have to be mapped can be limited to those that are actually populated. The aim is to accelerate the actual work, especially for the non-expert user, and to help overcome expected inconsistencies between schema declaration and actual usage.

On the right hand side, buttons correspond to high-level elements of the target schema and are used to access their corresponding sub-elements. These are visualized on the middle part of the screen as a tree structure of embedded boxes, representing the internal structure of the complex element. The user is able to interact with this structure by clicking to collapse and expand every embedded box that represents an element, along with all relevant information (attributes, annotations) defined in the XML schema document. To perform an actual (one to one) mapping between the input and the target schema, a user has to simply drag a source element from the left and drop it on the respective target in the middle.

The user interface of the mapping editor is schema aware regarding the target data model and enables or restricts certain operations accordingly, based on constraints for elements in the target XSD. For example, when an element can be repeated then an appropriate button appears to indicate and implement its duplication. Several advanced mapping features of the language are accessible to the user through actions on the interface, including:

- String manipulation functions for input elements.
- m-1 mappings with the option between concatenation and element repetition.
- Structural element mappings.
- Constant or controlled value assignment.
- Conditional mappings (with a complex condition editor).
- Value mappings editor (for input and target element value lists)

Mappings can be applied to ingested records, edited, downloaded and shared as templates. Preview interfaces (Figure 3) present the steps of the aggregation such as the current input xml record, the XSLT code of mappings, and the transformed record in the target schema, subsequent transformations from the target schema to other models of interest, and available html renderings of each xml record. Users can transform their selected collections using complete and validated mappings in order to publish them in available target schemas for the required aggregation and remediation steps.



La Primavera / Spring

Title: La Primavera / Spring
La Primavera / Spring

Date: 1982 [Restoration]; 1498 [Provenance]; 1546/1555 [Provenance]; 1482 [Creation]

Creator: Botticelli, Sandro (painter)

Description: Composition: The painting features six female figures and two male, along with a blindfolded putto, in an orange grove. To the right of the painting, a flower-crowned female figure in a floral-patterned dress is scattering flowers, collected in the folds of her gown. Her nearest companion, a woman in diaphanous white, is being seized by a winged male from above. His cheeks are puffed, his expression intent, and his unnatural complexion separates him from the rest of the figures. The trees around him blow in the direction of his entry, as does the skirt of the woman he is seizing. The drapery of her companion blows in the other direction. Clustered on the left, a group of three females also in diaphanous white join hands in a dance, while a red-draped youth with a sword and a helmet near them raises a wooden rod towards some wispy gray clouds. Two of the women wear prominent necklaces. The flying cherub has an arrow nocked to loose, directed towards the dancing girls. Central and somewhat isolated from the other figures stands a red-draped woman in blue. Like the flower-gatherer, she returns the viewer's gaze. The trees behind her form a broken arch to draw the eye. (http://en.wikipedia.org/wiki/Primavera_%28painting%29)

Figure 3. Transformation and HTML rendering preview

4 TECHNICAL SPECIFICATIONS

4.1 PLATFORM

It is written in JAVA, JSP, HTML and Javascript. It uses PostgreSQL as an object-relational database with Hibernate as the data persistence framework, and mongoDB as a document-oriented database. MINT is also reusing other open source development frameworks and libraries according to specific deployments and customizations. Mint source code versions are released under a free software license (GNU Affero GPL).

The platform offers a user and organisation management system that allows the deployment and operation of different aggregation schemes with corresponding user roles and access rights. A Restful web service is available for user management and authentication. Specifically in Linked Heritage, the authentication API is being reused by the terminology management system developed in WP3.

4.2 INGESTION

Registered users can upload their metadata records in XML or CSV serialization, using the HTTP, FTP and OAI-PMH protocols. Users can also directly upload and validate records in a range of supported metadata standards (XSD). XML records are stored and indexed for statistics, previews, access from the mapping tool and subsequent services.

Current developments aim to support relational database schemata and OWL/RDFS ontologies as input.

4.3 PROCESSING

Handling of metadata records includes indexing, retrieval, update and transformation of XML files and records. XML processors (Apache Xerces, SAXON, Nux) are used for validation and transformation tasks as well as for the visualization of XML and XSLT. For issues of scalability with respect to the amount of data and concurrent heavy processing tasks, parts of the services are multi-threaded or use specific queue processing mechanisms.

4.4 NORMALIZATION & VOCABULARIES

Various additional resources such as terminologies, vocabularies, authority files and dictionaries are used to reinforce an aggregation's homogeneity and interoperability with external data sources. A typical usage scenario is the connection of a local (server) or online resource with a metadata element in order to be used during mapping/normalization. These resources can be XML, RDFS/OWL, SKOS or even proprietary systems accessed through APIs.

Normalization services such as group editing and value mapping are currently being implemented as standalone tasks for direct imports.

4.5 LINKING & ENRICHMENT

MINT uses 4Store and Sesame for RDFS/OWL storage and processing, and links data sources to external SPARQL endpoints using string-based and knowledge-assisted matching strategies. Entity/term extraction and/or natural language processing frameworks are evaluated to expand the number of suggested links. RabbitMQ is used to allow for a reliable, scalable and portable messaging and processing system, used in and between different services.

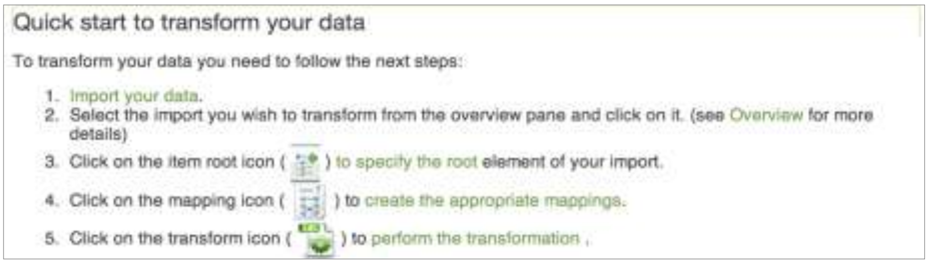
4.6 REMEDIATION

MINT is being used to publish metadata in XML, JSON or RDFS/OWL according to the mechanism and usage. Typical scenarios include an OAI-PMH repository for XML records, SPARQL endpoints for triple stores, Lucene-based indexes for search engines and Restful or Restless APIs for third party services.

4.7 DOCUMENTATION

Online documentation is available within the tool and on a dedicated wiki at

http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/User_manual



Quick start to transform your data

To transform your data you need to follow the next steps:

1. Import your data.
2. Select the import you wish to transform from the overview pane and click on it. (see Overview for more details)
3. Click on the item root icon ( ) to specify the root element of your import.
4. Click on the mapping icon ( ) to create the appropriate mappings.
5. Click on the transform icon ( ) to perform the transformation .

Figure 4. Quick start guide

In collaboration with WP7, Linked Heritage-specific training material will be produced in order to assist content providers in the use of the validation environment, tools and services.

5 CONCLUSION

The present document constitutes the report of deliverable D5.1 (Prototype, Public), the Linked Heritage Technology Platform, that is made available online for validation and for the large-scale contribution of content to Europeana (in WP6) and for dissemination & training (WP7). The platform implements an aggregation infrastructure offering a crosswalk mechanism to support subsequent critical activities:

- harvesting and aggregating metadata records that were created using shared community standards or proprietary metadata schemas,
- migrating from providers' models (whether standard or local) to a reference model,
- transforming records from the Linked Heritage model to the Europeana Semantic Elements and the Europeana Data Model.

5.1 RESULTS

The objective of the deliverable is the deployment of the prototype validation platform (Task 5.1), that is available online to the project partners at http://mint-projects.image.ntua.gr/linkedheritage

5.2 IMPACT

The delivery of the platform and the present report achieve a project milestone, MS7 “Ingester ready for delivery of content to Europeana”. This enables WP6 to start processing metadata and together with D5.2 “Metadata Gateway” work towards milestone MS8 “First lot of 500,000 content successfully delivered to Europeana”. In parallel, WP6 will assess the platform and provide a validation report as well as feedback from users. WP3 reuses the authentication API to integrate the introduced terminology management system , while WP2 can start experimenting towards Task 2.4 “Enabling linked cultural heritage data”.