D2.3 – Metadata Transformation Service

This deliverable is software.

Co-funded by the European Union

The project is co-funded by the European Union, through the **ICT Policy Support Programme** as part of the **Competitiveness and Innovation Framework Programme (CIP)**.

http://ec.europa.eu/information_society/activities/ict_psp/

Österreichische Nationalbibliothek

Europeana Creative is coordinated
by the Austrian National Library

Deliverable

Project Acronym: Europeana Creative

Grant Agreement Number: 325120

Project Title: Europeana Creative

D2.3 – Metadata Transformation Service

Revision: Final

Authors: Vassilis Tzouvaras (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**	

Europeana Creative is coordinated
by the Austrian National Library

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Vassilis Tzouvaras, NTUA	February 5, 2015	First draft
0.2	Draft	Vassilis Tzouvaras, NTUA	February 12, 2015	Revision
0.4	Final draft	Vassilis Tzouvaras, NTUA	February 27, 2015	Review recommendations incorporated
0.5	Final draft	Elisabeth Stricker, ONB Kristin Dill, ONB Susanne Tremml, ONB	February 27, 2015	Layout, format, editing

Distribution

Version	Date of sending	Name	Role in project
0.2	February 12, 2015	Andrew Kitchen, RAM	WP5
0.3	February 19, 2015	Valentine Charles, EF Hugo Manguinhas, EF	
0.4	February 27, 2015	Susanne Tremml, ONB	Project manager
1.0	February 27, 2015	Marcel Watelet, EC	Project Officer

Approval

Version	Date of approval	Name	Role in project
1.0	February 27, 2015	Max Kaiser, ONB	Project Coordinator

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

This deliverable reflects only the author’s/authors’ views and the European Union is not liable for any use that might be made of information contained therein.

Table of Contents

Figures

1. Executive Summary

This deliverable is about the metadata transformation service and its integration in Europeana's United Ingestion Manager (UIM)[1]. This module is enabling metadata transformation to Europeana Data Model (EDM[2]) and other formats and serialisations (i.e. Lido, IEEE LOM, CARARE). To enable data transformation between EDM and other formats, a visual mapping editor for the XSL[3] language is used. Mappings are performed through drag-and-drop and input operations which are translated to the corresponding code. In the framework of this task, NTUA integrated MINT in UIM and improved the mapping functionality. In the rest of the deliverable NTUA provides a short description of the functionality of the platform and focuses on the integration part and the parts that have been developed specifically for this task. Finally, the last section provides the requirements of the Pilot applications.

[1] http://pro.europeana.eu/blogpost/moving-to-new-europeana-data-model, accessed February 27, 2015.

[2] http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation, accessed February 27, 2015.

[3] www.w3.org/Style/XSL/, accessed February 27, 2015.

2. Metadata Transformation Service

The general architecture and function of MINT is already described and can be found here[4]. For convenience, a short description in the following section:

MINT is a web based tool that allows data providers to:

- Upload their own data to the MINT server in a selection of convenient formats and delivery ways.
- Upload in CSV, XML or JSON[5].
- Deliver from their server as HTTP or FTP download, as an OAI harvest, or as a simple HTTP upload.
- Provide a plain file, a ZIP archive or TGZ format.
- Specify inside the tool, how their data corresponds to a predefined schema (which is either agreed on or specified inside a project). This is done in a graphical manner generally accessible to non-programmers (after training).
- Preview, how their data looks if transformed into that schema. If applicable, the tool allows as well a preview on how the data would look on Europeana's portal.
- Actually execute the transformation to generate a new (derived) dataset in the target schema format.
- Let the tool check, if and how many records fulfil potential restrictions that the target schema enforces to the data.
- Let the tool handle the publication process to Europeana (if applicable).

[4] http://mint-wordpress.image.ntua.gr/mint-end-user-documentation/, accessed February 27, 2015.

[5] http://json.org/, accessed February 27, 2015.

2.1 Mapping Functionality

Metadata mapping is the most crucial step of the ingestion procedure. It formalises the notion of a metadata crosswalk, hiding the technical details and permitting semantic equivalences to emerge as the centrepiece. It involves a user-friendly graphical environment (Fig. 1 shows an example of a mapping opened in the editor) in which interoperability is achieved by guiding users in the creation of mappings between input and target elements.

The user's mapping actions are expressed through XSLT stylesheets, i.e. a well-formed XML document conforming to the namespaces in XML recommendation. XSLT stylesheets 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.

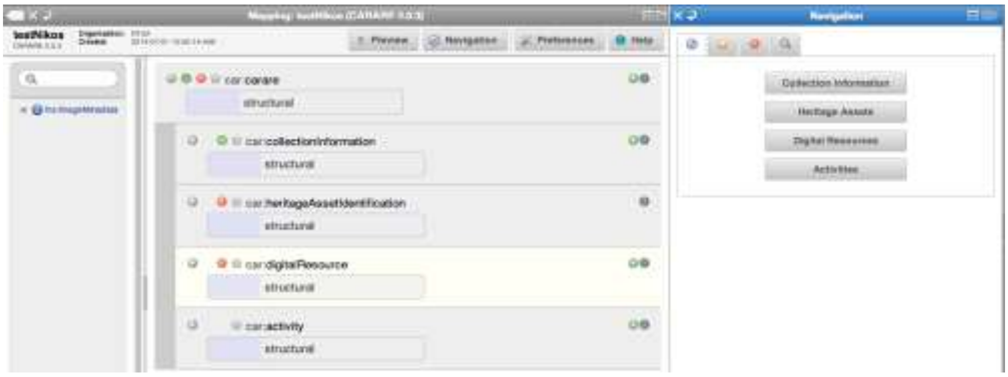


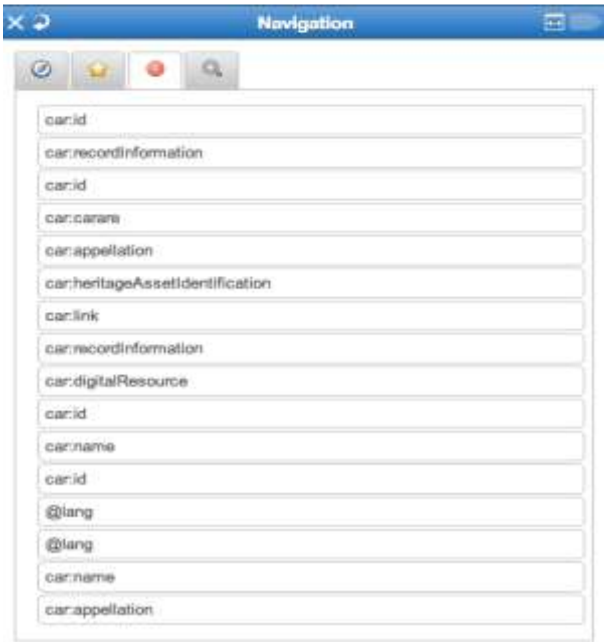
Fig. 1: Screenshot of the mapping editor

The structure that corresponds to a user's specific import is visualised 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 selected target schema and are used to access their corresponding sub-elements. These are visualised 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.
- When an element of the target schema is mandatory and it is not mapped then it appears in red and also under error panel of the navigation pane (Fig. 2).



car:id
car:recordInformation
car:id
car:carare
car:appellation
car:heritageAssetIdentification
car:link
car:recordInformation
car:digitalResource
car:id
car:name
car:id
@lang
@lang
car:name
car:appellation

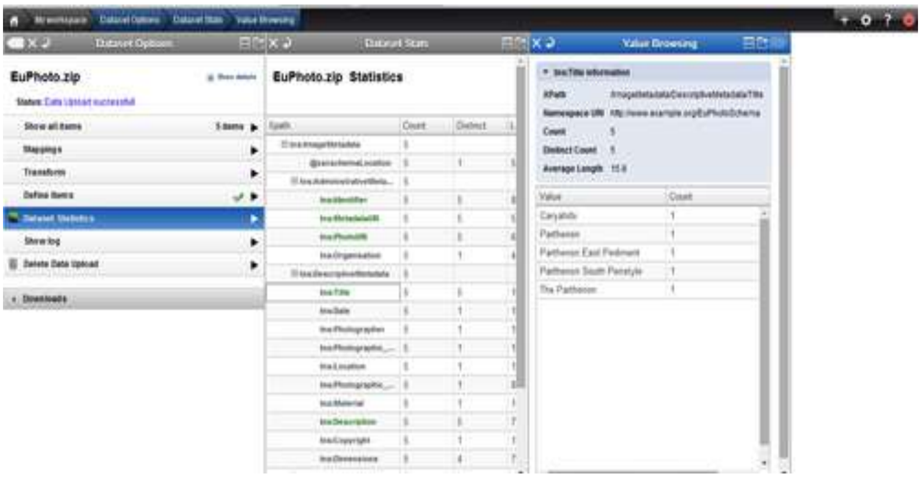
Fig. 2: MINT’s report on missing mandatory elements

Several advanced mapping features of the XSLT language are accessible to the user through actions on the interface, including:

- string manipulation functions for input elements;
- allowing providers to use part of their metadata values e.g. the first or last four characters;
- m-1 mappings with the option between concatenation and element repetition;
- structural element mappings;
- constant or controlled value assignment: this allows providers to add values to elements mandatory or recommended by the project that do not manage to upload;
- conditional mappings (with a complex condition editor);
- value mappings editor (for input and target element value lists);

2.2 Metadata Reports

Data statistics is another service of the MINT ingestion platform. It provides users with detailed information about the imported metadata. By selecting “data statistics” providers can see all the metadata elements of the imported dataset and by clicking on an element they can see its values. In the following screenshot one can see the title element. In similar manner one can view the values of all metadata elements. Moreover, additional information is provided i.e. XPath[6], namespace URI, count, distinct count, and average length. In the screenshot below count and distinct count have the same values, which means that this element has unique values.



Xpath	Count	Distinct
lsa:ImageMetadata	5	
@xsi:schemaLocation	5	1
lsa:AdministrativeMeta...	5	
lsa:Identifier	5	5
lsa:MetadataURI	5	5
lsa:PhotoURI	5	5
lsa:Organisation	5	1
lsa:DescriptiveMetadata	5	
lsa:Title	5	5
lsa:Date	5	1
lsa:Photographer	5	1
lsa:Photographic_...	5	1
lsa:Location	5	1
lsa:Photographic_...	5	1
lsa:Material	5	1
lsa:Description	5	5
lsa:Copyright	5	1
lsa:Dimensions	5	4

lsa:Title information

XPath: /ImageMetadata/DescriptiveMetadata/Title
Namespace URI: http://www.example.org/EuPhotoSchema
Count: 5
Distinct Count: 5
Average Length: 15.6

Value	Count
Caryatids	1
Parthenon	1
Parthenon East Pediment	1
Parthenon South Peristyle	1
The Parthenon	1

Fig. 3: Dataset statistics

Dataset statistics assists the providers to check the mandatory elements that are defined within the metadata target schema in order to evaluate the quality of the imported metadata, avoid duplicates and asses how rich the imported metadata is. Additionally, providers can check their diversity, for instance if the number of “distinct count” is much smaller than the number of “count”, then the imported metadata is of poor quality. Possible errors may be indicated by observing the length of the element. The length shows how the metadata has been extended or reduced. The users can perceive errors by observing the length, for example, if the length of the description is very small then the description does not correspond to the correct element.

[6] www.w3.org/TR/xpath/, accessed February 27, 2015.

3. Integration with Europeana

This section presents the integration of MINT in UIM. Before the actual integration work, NTUA needed to investigate which technologies could be used to realise this integration. Fig. 4 shows Europeana's high-level aggregation infrastructure workflow, including MINT.

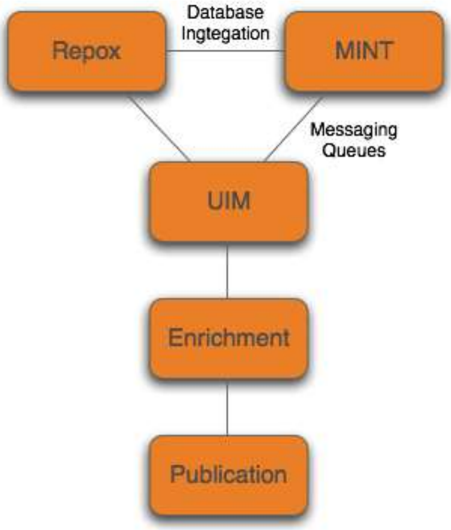


Fig. 4: Aggregation infrastructure

MINT is integrated between REPOX[7] and UIM. REPOX and MINT are integrated on the database schema level. The use of the same database technology enables easy integration of the two modules. MINT and UIM are integrated using a messaging protocol system that is described in section 3.

The overall architecture of MINT in terms of communication with internal and external modules is depicted in Fig. 5.

[7] http://repox.ist.utl.pt/, accessed February 27, 2015.

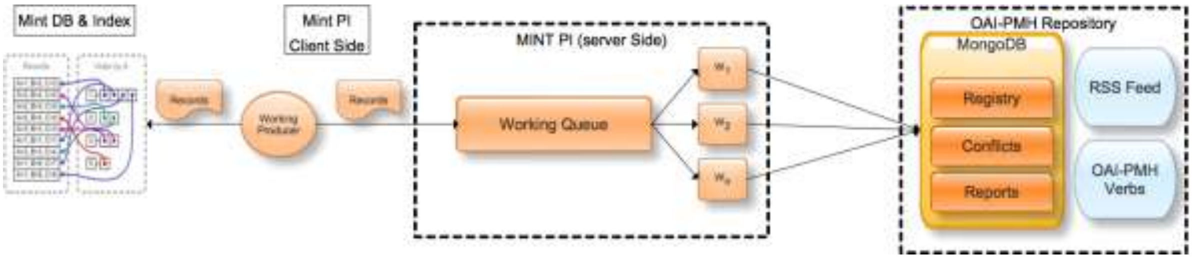


Fig. 5: MINT communication architecture

3.1 MINT Processing Infrastructure (PI)

MINT PI is part of the metadata interoperability services suite and offers a scalable mechanism for structured data processing. It is built using RabbitMQ1[8] acting as a message broker in its core and utilises a number of message queue patterns based on the AMQP2[9], a standard wire-level protocol and semantic framework for high performance enterprise messaging. AMQP is an Open Standard for Messaging Middleware.

By complying with the AMQP standard, middleware products written for different platforms and in different languages can send messages to one another. AMQP addresses the problem of transporting value-bearing messages across and between organisations in a timely manner. The approach of using message queues was decided based on the requirement to scale on both large and small datasets without reducing the efficiency of the overall system. The overall architecture of MINT PI is depicted in Fig. 6.

[8] www.rabbitmq.com/, accessed February 27, 2015.

[9] http://psnprofiles.com/amqp2, accessed February 27, 2015.

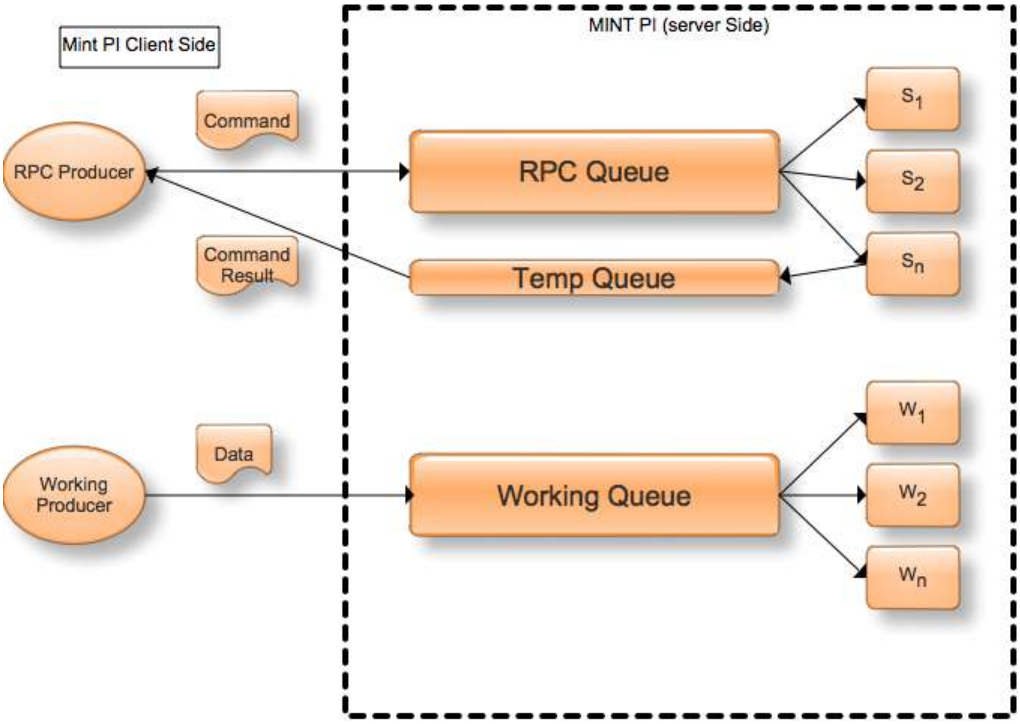


Fig. 6: The MINT PI Architecture

MINT PI uses two distinct queue patterns. One is an RPC Queue pattern, which is used for cases in which the client desires to block while the processing is executed and also awaits for a response in a pre-defined format. The second is a Working Queue pattern, which is used for non-blocking processing in which the client submits the data for processing and does not wait for a response. The first case is mainly used for the implementation of specific commands, e.g. for implementing the cleaning or deletion of a repository, while the second case is used for bulk processing of raw data, e.g. data transformation and enrichment of records.

Scalability is achieved by the parallel processing of many workers for the case of the Working Queue and the existence of number of RPC Consumers that are running concurrently for the case of the RPC Queue. In both cases the workers and the RPC consumers might be running on different nodes of the cluster that is materialised while more workers and consumers can be added to the system at any time and thus increasing the processing power of the system. At the same time, the RabbitMQ broker is also scalable, new nodes can be added to the broker and thus increasing the message per second ratio that can be handled by the system. In this way an overall scalable architecture for message delivery is defined which can be extended with minimal administrative effort.

MINT PI is not limited in a certain type of processing or data schema that is delivered using the messages. This is achieved by utilising a software design pattern named the Strategy or Policy Pattern, using this pattern it is possible to select different algorithms for execution on runtime. An additional benefit of this design pattern is the abstraction introduced between the messaging layer of the system and the implementation of various algorithms for processing. A developer does not have to know about the intrinsic details of the messaging system in order to implement another algorithm for data processing by MINT PI.

4. Description of Software Developed within Europeana Creative

Link to software	http://sandbox36.isti.cnr.it:8080/europeanamint
Log-in information	Log-in is required. Anyone can register and then log into the system
Development environment	Java
Programming language used	Java 7
Application server used	Apache Tomcat
Database requirements	Postgress
Operating system requirements	Deployed on Linux. Solaris
Port requirements / default ports used	8080
Interface	XML, OWL/RDF, JSON, CSV, OAI-PMH, FTP, HTTP
Licensing conditions	EUPL – Open Source

5. Pilots Requirements

This section summarises the requirements gathered from the Pilots for all services of WP2. It is included within this deliverable based on a recommendation of the first Technical Review of Europeana Creative.

5.1 History Education Pilot

The objective of the **History Education Pilot** is to stimulate the re-use of digitised heritage objects, specifically those made available through Europeana in history education. In order to achieve this, EUROCLIO has joined together with Webtic and an international community of history educators who work as professional volunteers on the creation of (online) learning activities. Within the learning section, history educators can search a set of sources that are pre-selected by their relevance for history education, their quality and their license (that should allow for re-use in education). These sources can be searched by source type, people, locations and time. This set of sources includes newspapers, postcards, posters, diaries, music, monuments, official documents and newsreels, and they come from a range of digital collections from across Europe. The learning section includes the first tools that educators can use to make their own online learning activities to make best use of the digitised heritage. The design focused on the critical analysis of visual sources contributes to the innovation of history education and use new technologies to promote historical and critical thinking skills. The learning section also contains learning activities that will be featured on the Historiana website as an example of how to implement certain teaching methods, overcome teaching challenges and/or how to promote the students' acquisition of historical thinking skills. The learning activities provide all the information needed for the students and their teachers to implement the activity in practice. The exemplar educational resources focus on the First World War, but EUROCLIO is committed to developing resources on other key moments and developments in history, as well as new tools. The combination of free access and quality control will provide a unique social value for all stakeholders.

The type of content used by the History Education Pilot is text and image from Europeana, Wikipedia and the Library of Congress. The metadata schemas used are EDM, Custom. The Pilot is expected to support multiple licensing for content and the licensing scheme features are non-commercial, educational. Some potential data sources that can be used for linking data are: DBpedia, Freebase, Geonames.

The History Education Pilot is a web-based pilot that allows for metadata ingestion, pilot creation (once), content curation (periodical updates by curator), new users (once per user), application start. It supports constant access to a repository that allows live search and retrieval. The users are allowed to read, edit and add information.

The WP2 services that are integrated into the Pilot are described by the following:

- locate content from the Europeana portal: Historiana Database, europeana.eu, Europeana project websites (EDF, 1914-1918, 1989, EUScreen, Europeana Newspapers),

- request for copyright clearance: Copyright Clearance Letter + Form, overview of content,
- import of records with metadata for re-use in the Historiana website: search and select,
- the Analysis Tool, Learning Section and Learning Activities of the Historiana website.

The planning concerning the applications designed within WP2 is described by the following: Continued UX-testing, development on (at least 4) new tools in a project, development of a teacher’s guide, continued trainings on the use of the tools, searching/selecting/contextualising additional sources, making a better Europeana API integration in the Search and Select tool, traineeships linked to Historiana, set up of a newsletter for Historiana.

Other requirements of the Pilot include the ability to filter content that is retrievable (e.g. directly downloadable via a URL), to filter on the resolution of the images, to add a multilingual search, to filter further down per tag, and finally, to look for similar sources (e.g. Google image search).

5.2 Natural History Education Pilot

The objective of the Natural History Education Pilot is to facilitate creative industry access to Europeana metadata and digital content from the Natural History domain to deliver education applications. Its purpose is to stimulate the interest of learners, educational publishers and other stakeholders.

The Pilot supports an interactive application for the creation of education material and an adventure game for education in natural history. The game can be seen as a virtual exhibition in a multimedia environment, where a visitor is given all the necessary information to solve puzzles and questions. While answering questions the visitor is engaged in a learning procedure concerning animals and plants, as well as the history of research and the context of the discoveries.

The content source of the Pilot is Europeana and relevant projects. It makes use of Europeana API extensions and Content services.

5.3 Social Networks Pilot

The Social Networks Pilot is called 'Sound Connections' and encourages and enables visitors to actively enrich the geo-pinned sounds with supplementary media from various sources. Sound collections from BL and NISV are used in this Pilot.

The type of content used by the Pilot is audio in wav format. The content source of the Pilot is Europeana. The Pilot supports multiple licensing for content and the creative commons licenses scheme features are used. The metadata schemas that are used by the Pilot are the EDM –

Europeana and the European Library schema. Potential data sources for linking are DBpedia, Geonames, Wikipedia, Flickr. The type of enrichment is URL and data (resource retrieval from external data source).

The Social Networks Pilot is a web-based application that supports metadata ingestion for content curation (periodical updates by curator). Its users are allowed to read, edit and add. The Pilot integrates audio files from the BL and NISV via the Europeana API into the 4 thematic Pilot pages (aviation, birds, Amsterdam sounds, London sounds). Within WP2 Ontotext contributed to the enrichment of geo-information for the datasets, as an additional requirement regarding the use of content in the Pilot is that the sounds need to have a date and a geo-location.

5.4 Tourism Pilot

The Tourism Pilot supports the integration of Europeana metadata and public domain content into touristic services and travel activities based on cultural itineraries. Towards this direction PLURIO.NET works with Culture24 to develop applications that serve the needs of the online tourist.

The Pilot holds a platform for cultural tourism that provides access to content from Europeana and other databases (Wikipedia, Flickr, Openstreetmap) to touristic operators and individual travellers. The purpose is to allow users to select, publish and share their travel itineraries and curate content from Europeana and other public domain sources in order to allow for creative interaction with cultural heritage and objects.

The content is provided by Europeana, as well as directly by museums. The Pilot supports multiple licensing for content and uses the licensing scheme features: PD, CC0, CC-BY, CC-BY-SA. It provides a platform where users can manually download text and image files. A special field that links to the relevant data in Europeana is available.

5.5 Design Pilot

The purpose of the Design Pilot is to connect the digital cultural heritage provided by Europeana to open design communities of craft and media designers who wish to use cultural heritage objects as sources for new, derivative designs, such as embroidery, textile patterns, 3D printed objects, media art, etc., that also wish to share with the community their designs back in re-usable forms. The Design Pilot serves to capture the interest of designers, artists, practitioners as well as creative industries towards the re-usage of Europeana content. The Pilot is carried out by Aalto University (Aalto), Spild af Tid ApS (SAT) and Austrian Institute of Technology (AIT).

The requirements of the Design Pilot within the scope of WP2 were to design and develop a new advanced search mechanism to support various types of visual search among its visual contents (e.g. colour and shape). The tool, Culture Cam, is a digital “live” similarity tool developed by SAT that helps recognise a colour, a shape or a pattern by using a web camera.

Once an object is scanned, search results of objects with a similar colour, shape or pattern are shown directly from the Europeana archive.

A relevant work was developed within WP2 by the Austrian Institute of Technology (AIT) concerning an Image Similarity Tool for Europeana. Against this background SAT found an ideal possibility of making use of AIT's knowledge for the development of Culture Cam, both in regards of similarity search and the Europeana API. Finally, the services and messaging APIs software developed within WP2 have been integrated into the project.

6. Conclusion

Deliverable 2.3 presented the metadata transformation service as part of the content re-use infrastructure of WP2. This module enables the transformation of metadata records in EDM formats or to other metadata standards (i.e. Lido, CARARE, IEEE LOM) and data serialisations (i.e. XML, RDF). The basic functionality of the MINT aggregation platform was presented, which is used for metadata mapping and transformation, as well as its integration in the Europeana aggregation infrastructure. Special attention was given that the technologies used enable deep and easy integration. Using the metadata transformation module, external and Pilot applications wanting to re-use Europeana content have the ability to transform their data in appropriate formats and metadata standards to better fit better their use cases (e.g. educational). Finally, the requirements gathered from the Pilots for the development of WP2 services were included, to follow the recommendations of the first Technical Review of Europeana Creative.