



**Project Acronym:** Europeana Sounds  
**Grant Agreement no:** 620591  
**Project Title:** Europeana Sounds

## D5.2 Deployment of fully functional updated aggregation system deployed by UIM

**Revision:** Final  
**Date:** 07/06/2015  
**Authors:** Natasa Sofou (NTUA) [natasa@image.ntua.gr](mailto:natasa@image.ntua.gr)  
Vassilis Tzouvaras (NTUA)  
Cecile Devarenne (EF)

**Abstract:** We report on the deployment of the updated aggregation system for Europeana Sounds, contributing to task T5.3 *Aggregator deployment and maintenance* and using the outcomes of T5.1 *Aggregation infrastructure design* as well as experience obtained in T5.2 *Aggregation infrastructure evaluation*. The aggregation is a two-phase process. The first phase, handled by the MINT ingestion platform, involves data providers in the aggregation process. In the second phase, transformed metadata are delivered to Europeana in EDM via the OAI repository and then published on the Europeana website. Details on the aggregation phases and aggregation functionalities are presented.

Dissemination level	
Public	X
Confidential, only for the members of the Consortium and Commission Services	



## Revision history

Version	Status	Name, organisation	Date	Changes
0.1	ToC	Vassilis Tzouvaras, NTUA	27/04/2015	ToC
0.2	1st draft	Vassilis Tzouvaras, NTUA Natasa Sofou, NTUA	10/05/2015	First round of contributions
0.3	2nd draft	Natasa Sofou, NTUA Cecile Devarenne, EF	21/05/2015	Second round of contributions
0.4	Final draft	Natasa Sofou, NTUA	04/06/2015	Updated after internal review
1.0	Final	Richard Ranft, BL	07/06/2015	Layout, minor changes

## Review and approval

Action	Name, organisation	Date
Reviewed by	Sergiu Gordea, AIT Alessio Piccioli, NET7	04/06/2015
Approved by	Coordinator and PMB	04/06/2015

## Distribution

No.	Date	Comment	Partner / WP
1	07/06/2015	Submitted to the European Commission	BL/WP7
2	07/06/2015	Posted on Europeana Pro website	BL/WP7
3	07/06/2015	Distributed to project consortium	BL/WP7

## Application area

This document is a formal output for the European Commission, applicable to all members of the Europeana Sounds project and beneficiaries. This document reflects only the author's views and the European Union is not liable for any use that might be made of information contained therein.

## Statement of originality

This document contains original unpublished work except where clearly indicated otherwise.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

## Project summary

Europeana Sounds is Europeana's 'missing' fifth domain aggregator, joining APEX (Archives), EUscreen (television), the Europeana film Gateway (film) and TEL (libraries). It will increase the opportunities for access to and creative re-use of Europeana's audio and audio-related content and will build a sustainable best practice network of stakeholders in the content value chain to aggregate, enrich and share a critical mass of audio that meets the needs of public audiences, the creative industries (notably publishers) and researchers. The consortium of 24 partners will:

- Double the number of audio items accessible through Europeana to over 1 million and improve geographical and thematic coverage by aggregating items with widespread popular appeal such as contemporary and classical music, traditional and folk music, the natural world, oral memory and languages and dialects.
- Add meaningful contextual knowledge and medium-specific metadata to 2 million items in Europeana's audio and audio-related collections, developing techniques for cross-media and cross-collection linking.
- Develop and validate audience specific sound channels and a distributed crowd-sourcing infrastructure for end-users that will improve Europeana's search facility, navigation and user experience. These can then be used for other communities and other media.
- Engage music publishers and rights holders in efforts to make more material accessible online through Europeana by resolving domain constraints and lack of access to commercially unviable (i.e. out-of-commerce) content.

These outcomes will be achieved through a network of leading sound archives working with specialists in audiovisual technology, rights issues, and software development. The network will expand to include other data-providers and mainstream distribution platforms (Historypin, Spotify, SoundCloud) to ensure the widest possible availability of their content.

For more information, visit <http://pro.europeana.eu/web/europeana-sounds> and <http://www.europeanasounds.eu>

## Copyright notice

Copyright © Members of the Europeana Sounds Consortium, 2014-2017. This work is licensed under the Creative Commons CC-BY License: <https://creativecommons.org/licenses/by/4.0/>.

## Contents

Executive summary: D5.2 Deployment of fully functional updated aggregation system deployed by UIM	6
1 Introduction	7
1.1 Background	7
1.2 Scope	7
1.3 Related documents	7
2 Metadata ingestion workflow in Europeana Sounds	8
2.1 Workflow	8
3 MINT services	10
3.1 Specific updates of MINT for Europeana Sounds	10
3.1.1 Backend reconstruction	11
3.1.2 Frontend reconstruction - User Interface	11
3.1.3 Mapping functionalities	13
3.1.4 EDM Sounds Profile implementation and deployment	14
3.1.5 SKOS vocabularies support	14
3.1.6 OAI publication	14
3.2 Description of basic aggregation functionalities in MINT	19
3.2.1 Metadata upload and preparation	19
3.2.2 Metadata mapping	20
3.2.3 Transformation, quality check and publication	23
3.3 Technical Details	24
3.3.1 Platform	24
3.3.2 Ingestion	24
3.3.3 Processing	25
3.3.4 Normalization and vocabularies	25
4 Europeana data processes	25
4.1 Ingestion	25
4.2 Data transformations	25
4.2.1 Cleaning and quality checks (MINT)	25
4.2.2 Mapping, transformation, validation (MINT)	26
4.2.3 Itemization and Unique Identifiers Generation (UIM)	26
4.3 Preview caching (Media Harvester)	26
4.4 Enrichment (UIM)	26
4.4.1 Dereferencing and vocabulary mapping	26
4.4.2 Semantic enrichment	26
4.5 Publication	27
5 Summary	27

6	References	28
	Appendix A: Terminology	28

## Executive summary: D5.2 Deployment of fully functional updated aggregation system deployed by UIM

Europeana Sounds focuses specifically on audio and audio-related content, primarily music and speech audio, including out-of-commerce recordings and a large number of unpublished works from Europe's major sound archives that are not widely available. Efficient ingest of this audio and audio-related metadata is one of the main objectives of the project and requires an aggregation platform as well as a content selection policy.

Europeana Sounds' aggregation workflow is implemented using the MINT platform, the Europeana Data Model - Sounds Profile specification (D1.4), and the Europeana Sounds Ontologies as introduced in D1.3. The operation of the aggregation is planned and monitored by WP1 in order to coordinate content provision (workflow guide, publication cycles, contingency planning etc.) The technical infrastructure is implemented, monitored and maintained by WP5 in order to support the activities of WP1, implement the produced specifications and enable emerging ingestion and publication requirements. WP5 also continues the work on activities that enable several workflow steps, and introduce tools and services for enrichment & linking.

This document contributes to T5.3 - *Aggregator deployment and maintenance* and uses the results from T5.1 - *Aggregation infrastructure design* as well as experience obtained in T5.2 - *Aggregation infrastructure evaluation*.

The main focus of this document is to present the aggregation workflow of Europeana Sounds as a two-phase process, where the first phase is handled by MINT ingestion platform and is the part where providers get involved in the aggregation process and the second phase takes part when the transformed metadata are delivered to Europeana in EDM via OAI repository and are then handled and published on Europeana website.

We present in the first section the specializations and updates that were applied to the MINT platform for the Europeana Sounds project, such as backend and frontend reconstruction, additional mapping functionalities offered, EDM sounds profile implementation and deployment, Sounds vocabularies support and OAI publication. Then there is a description of all aggregation functionalities offered by MINT such as metadata upload and preparation, mapping, transformation, quality check and publication.

Since the second phase of aggregation is handled by Europeana, there is a description of the Europeana data processes that take place such as ingestion, data transformations, preview caching, enrichment and publication.

# 1 Introduction

## 1.1 Background

Europeana Sounds focuses specifically on audio and audio-related content, primarily music and speech audio, including out-of-commerce recordings and a large number of unpublished works from Europe's major sound archives that are not widely available. Efficient ingest of this audio and audio-related metadata is one of the main objectives of the project and requires an aggregation platform as well as a content selection policy.

Within the Europeana Sounds project, the aim of WP5 Technical Infrastructure is primarily to enable metadata aggregation by providing an appropriate mechanism that extends and enhances the existing Europeana infrastructure. Metadata aggregation services are handled by the MINT (Metadata INTERoperability) platform and cover registration and metadata import, mapping and transformation, cleaning and normalisation, link checking, thumbnail caching and quality checking prior to publication

## 1.2 Scope

Europeana Sounds' aggregation workflow is implemented using MINT, the Europeana Data Model - Sounds Profile specification (D1.4), and the Europeana Sounds Ontologies as introduced in D1.3. The operation of the aggregation is planned and monitored by WP1 *Aggregation* in order to coordinate content provision (workflow guide, publication cycles, contingency planning etc. The technical infrastructure is implemented, monitored and maintained by WP5 in order to support the activities of WP1, implement the produced specifications and enable emerging ingestion and publication requirements. WP5 also continues the work on activities that enable several workflow steps, and introduce tools and services for enrichment & linking.

This document contributes to task T5.3 - *Aggregator deployment and maintenance* and uses the results from T5.1 - *Aggregation infrastructure design* as well as experience obtained in T5.2 - *Aggregation infrastructure evaluation*.

## 1.3 Related documents

- D1.3 – *Ontologies for sound*: the recommendations from work on ontologies and language.
- D1.4 – *EDM profile for sound*: the result from the working group defining audio extensions to EDM.
- MS23 – *Revised aggregation design available* – outlines the basic components of aggregation infrastructure.

- MS24 – *Aggregation infrastructure prototype available* – reports on the delivered prototype that offers aggregation services such as registration and metadata import, mapping and transformation, and publication
- MS25 – *Sounds SKOS ontology normalization and cleaning module beta* – outlines advance functionalities of aggregation mechanism.
- MS26 – *Aggregation mechanism ready* – reports on the delivery of the aggregation mechanism.
- D5.1 – *Report on the evaluation of the aggregation mechanism*: Report with recommendations on evaluation of the aggregation toolset and pilot phase for data provider familiarisation with the technology.

## 2 Metadata ingestion workflow in Europeana Sounds

### 2.1 Workflow

Europeana's toolset consists of the United Ingestion Manager (UIM) that controls the harvesting process, which is done through Repox (repo.ist.utl.pt) and the mapping and transformation process, which is done through MINT.

The aggregation workflow can be considered as a two-phase process, illustrated in Figure 2-1. Overall workflow as a 2-step process. The first phase is handled by MINT and is the part where providers get involved in the aggregation process. In the second phase, transformed metadata are delivered to Europeana in EDM via the OAI repository and are then handled and published on Europeana website.

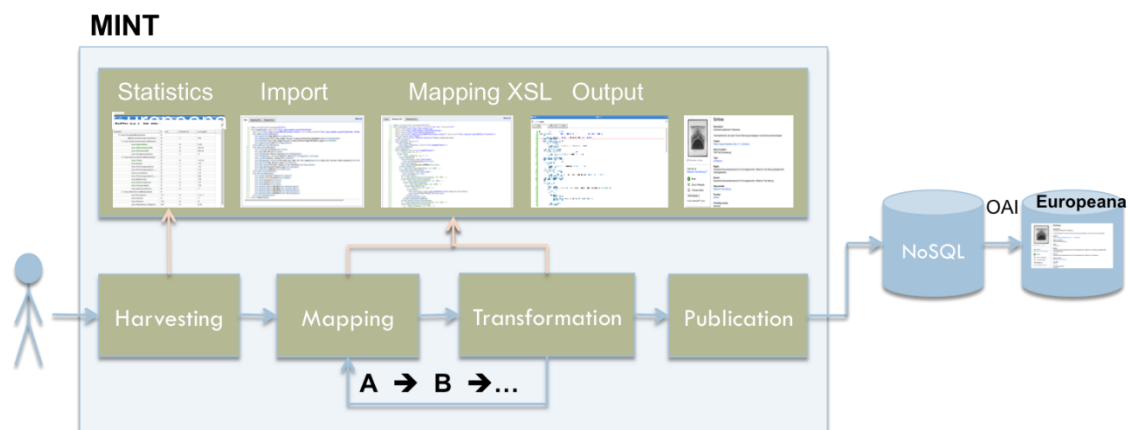


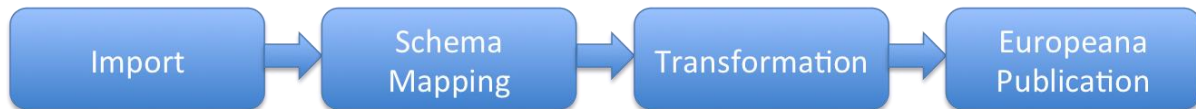
Figure 2-1. Overall workflow as a 2-step process

The MINT aggregation platform facilitates the ingestion of semi-structured data and offers the ability to establish crosswalks to the reference schemas (EDM Sounds profile, EDM) in order to take advantage of a well-defined, machine understandable model. The underlying data serialization is in XML, while the user's mapping actions are registered as XSL transformations.



The main role of the MINT ingestion platform in the Europeana Sounds project is to enable users to:

- Provide metadata records in a range of “source” formats
- Convert metadata to selected target schema (EDM and EDM Sounds profile - used as an intermediate standard before publishing to Europeana)
- Monitor the progresses of content provision



**Figure 2-2. MINT Ingestion workflow**

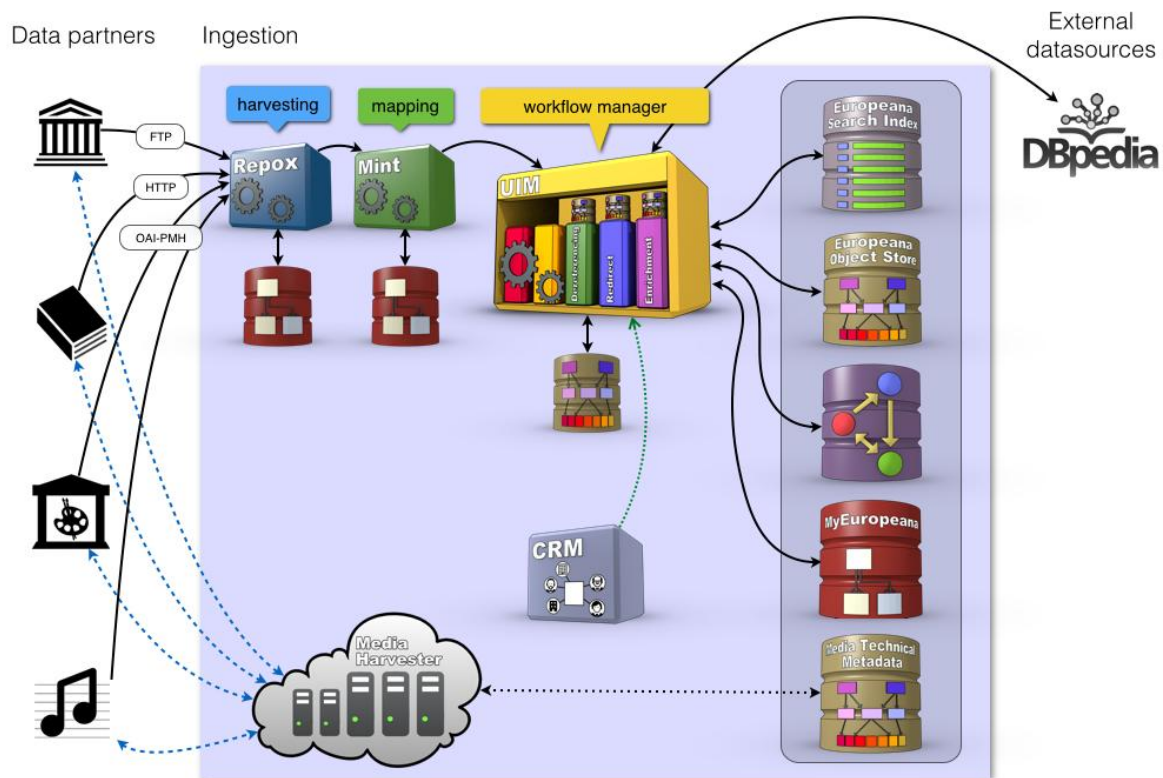
The metadata ingestion workflow handled by MINT, as illustrated in Figure 2-2. MINT Ingestion workflow, consists of four main discrete stages. First is the *Import* of provider’s metadata using common data delivery protocols, such as OAI-PMH, HTTP and FTP. Next is the *Schema Mapping* procedure, during which the imported metadata are mapped to EDM Sounds profile. A graphical user interface assists data providers in mapping their metadata to the target schema, using an underlying machine-understandable mapping language. Furthermore, it provides useful statistics about the provider’s metadata and supports the share and reuse of metadata crosswalks and the establishment of template transformations. The third step is the *Transformation* procedure, during which providers’ metadata are transformed to the selected schema by using the mapping they made in the previous step. The last step is the *Europeana Publication* procedure, during which metadata are transformed from EDM Sounds profile to EDM - according to the project requirements - and stored at NTUA’s OAI-PMH server. Publication to Europeana is then performed by informing Europeana’s Ingestion office to harvest metadata from the NTUA’s server.

Regarding the technical details of the ingestion workflow, the underlying data serialization is in XML while the user's mapping actions are translated into XSL transformations. The EDM Sounds profile functions as an anchor, to which various data providers can be attached and become, at least partly, interoperable.

Some of the key functionalities included in MINT aggregation workflow are:

- Organization and user level access and role assignment;
- Upload structured and semi-structured data;
- XML collection and record management;
- Direct importing and validation according to a standard schema (XSD);
- OAI-PMH harvesting and publishing;
- Visual mapping editing for the XSLT language;
- Transformation and publication ;

The proposed approach delivers a sustainable core part of the Europeana infrastructure, efficiently and transparently integrating aggregation functions into an aggregation workflow. More details about the aggregation functionalities are presented in following section.



**Figure 2-3 Europeana Ingestion Workflow**

Once data providers have exposed their metadata through the Europeana publication procedure in MINT, Europeana can harvest them from NTUA's OAI-PMH server into its system. Data are then transferred in Europeana's MINT where:

- Quality assurance is performed;
- Data are mapped and transformed from EDM External to the EDM Internal standard;
- Validation is run according to the EDM Internal standard schema (XSD).

The transformed data are re-imported in UIM's Mongo database and a series of operations is triggered to enrich the data prior to their publication onto Europeana's portal and API: link caching, identifiers generation, redirects creation, enrichment. More details about the Europeana part of the workflow are presented in section **Error! Reference source not found.**

## 3 MINT services

### 3.1 Specific updates of MINT for Europeana Sounds

The MINT ingestion platform has been previously used in Europeana related projects, making it by now one of the mature solutions for metadata handling and reuse. Even if the main aggregation functionalities of the MINT platform – that is the transformation of the metadata extracted from the provider's metadata management systems in various standards to a common metadata standard for the project – remains the same, necessary adjustments for the implementation within a given data-domain,

and continued development from a service provider's perspective are needed to meet the requirements and objectives of Europeana Sounds project.

In the remaining of this section, there is a presentation of the specializations that were implemented for the Europeana Sounds project.

### 3.1.1 Backend reconstruction

One of the most important developments that are implemented for the Europeana Sounds project is the major reconstruction of the MINT's backend platform. More specifically a new module for processing the XML imports has been implemented considering the experience gathered from previous usage of MINT as well as feedback from data providers. The main problem that data providers had experienced in previous MINT releases was its non-scalable behaviour, especially in operations like item previewing and dataset statistics.

This new processing module is responsible for the itemization of the imported metadata. The overall backend scalability is improved since:

- Items of an import are pre-calculated and computational time is saved by preventing the dynamic fetch of items.
- Items statistics are calculated during import thus saving computational time by preventing its dynamic calculation.
- Indexing of the imported metadata is not performed anymore cause it proved to be time consuming and without any useful functionality.

### 3.1.2 Frontend reconstruction - User Interface

Another update in the MINT release used for the Europeana Sounds project was the redesign of the user interface. The main objective was to redesign the user interface of the mapping tool so as to allow easy access and to better understanding of the overall workflow towards Europeana.

**Aggregation workflow guided interface.** In order to enable users' guidance thought the aggregation process, the jQuery library Kaiten<sup>1</sup> has been used for the new user interface of the MINT mapping tool. Interactions between the user and the application are a stack of contiguous screens as seen in Figure 3-1, where each step of the workflow is presented as a contiguous column allowing the user to perform specific actions depending on their starting point.

---

<sup>1</sup> <http://www.officity.com/kaiten/>

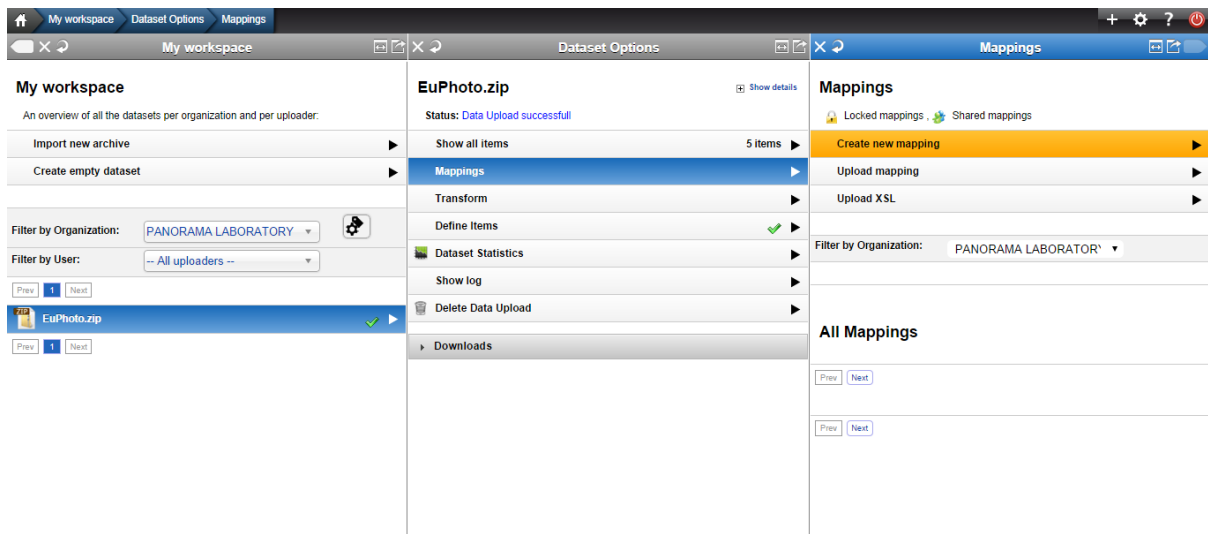


Figure 3-1 Workflow guided interface

**Improved browsing schema.** The new MINT interface, as seen in Figure 3-2 enables providers to search elements in both input and target schema. Additionally, in an effort to further assist providers to meet the project requirements and get familiarized with the target schema, bookmarks have been set up for the Europeana Sounds recommended elements. Bookmarks can be viewed by selecting the star icon in the navigation area (see Figure 3-3), enabling the user to see the respective elements and easily map a value from his/her metadata.

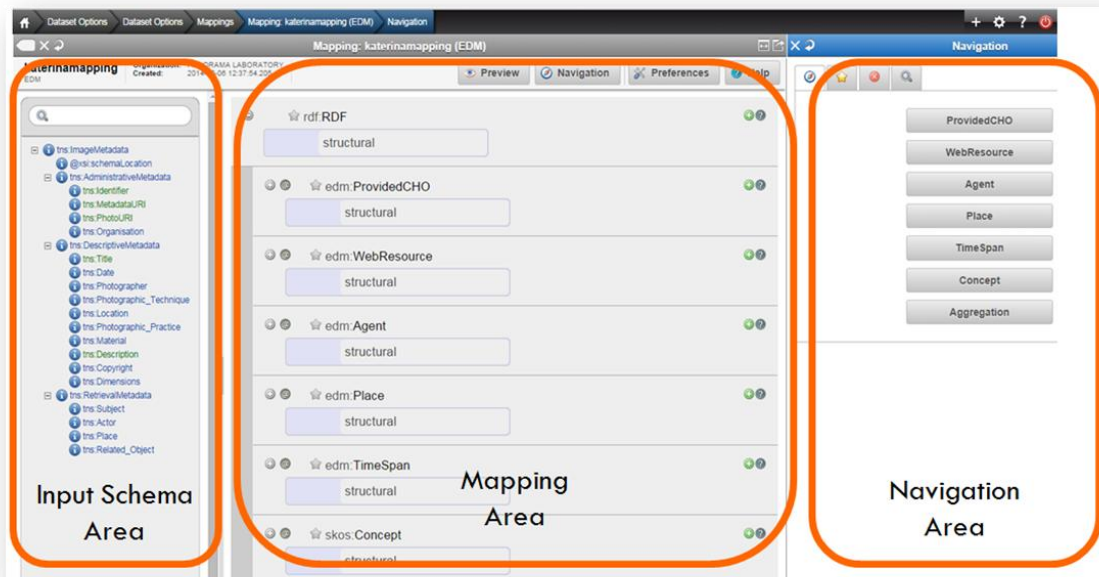
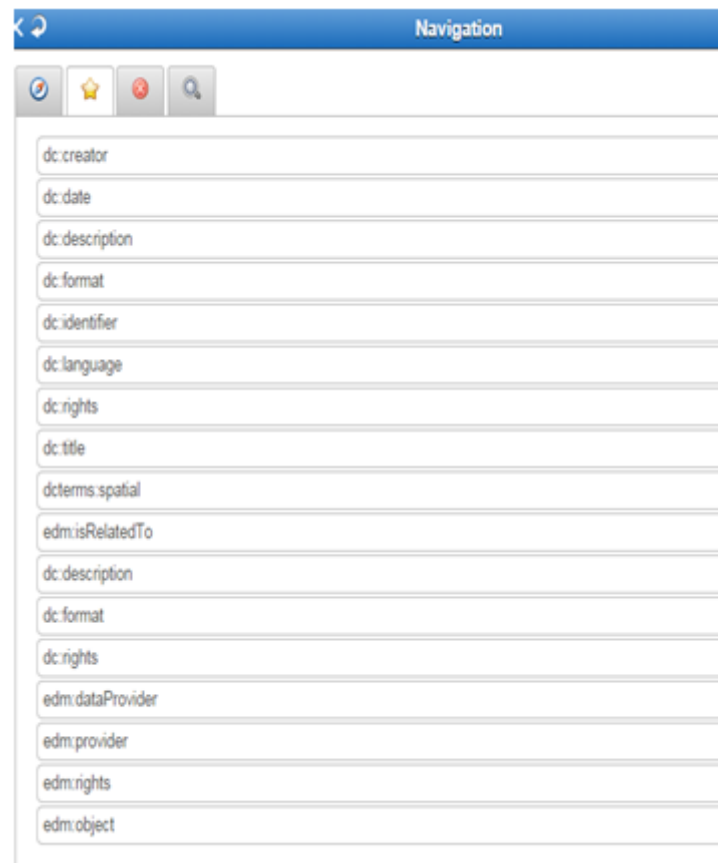


Figure 3-2 MINT browsing schema



**Figure 3-3 Bookmarks**

### 3.1.3 Mapping functionalities

**Advanced mapping functionalities** have been developed in order to assist the data providers to enhance their mapped metadata. Such functionalities are:

- The implementation of the negation for all the conditions used in conditional mappings. By using this functionality a data provider can perform mappings by separating his metadata into two sets according to its values (e.g. those that start with, and those that don't start with) and treating each of them in a different way.
- The implementation of the 'else' statement for further completing the conditional mapping functionality of the tool.

“Group by” functionality. Some of the providers have metadata exports in schemas that do not have any hierarchical structure like MARC. In detail the different types of information that describe their collections appear as different occurrences of the same element but with different attributes. The problem with data providers that use input schemas like MARC through MINT is that only one element appears on the mapping editor – and therefore providers cannot distinguish between their input values and the desired mappings. The “group by” mapping functionality solves that problem by permitting data

providers of such schemas to treat each element occurrence independently considering its different attributes.

### 3.1.4 EDM Sounds Profile implementation and deployment

The EDM sounds profile specification was designed and released in D1.4 –*EDM profile for sound* [REF 1] as part of task T1.3 - *EDM profile*.

Using the aforementioned specification, NTUA implemented the actual schema in XSD, which was then deployed in MINT, in order to enable users to map their metadata into the EDM sounds profile schema, which is the sound enriched version of EDM used as intermediate schema before publishing to Europeana. The EDM sounds profile XSLT is provided as complementary material.

### 3.1.5 SKOS vocabularies support

SKOS vocabularies specific to the sound content have been designed and developed within WP1, T1.2-*Vocabularies* and reported in D1.3 – *Ontologies for Sound*, in order to aid and enrich the metadata production and delivery to Europeana by the data providers. Since basic aggregation services are performed through MINT, the integration of SKOS vocabularies into the MINT mapping tool was vital. In order to extend MINT’s functionality to support SKOS vocabularies an additional module has been developed. More specifically, a semantic repository<sup>2</sup> has been set up to which the SKOS vocabularies are stored. The communication of the MINT mapping tool with the SKOS repository is established using SPAQRL 1.1 to retrieve the vocabularies’ terms based on the SKOS specification<sup>3</sup>. Additional semantic properties can be supported - if necessary - by the vocabularies for controlling selectable and non-selectable terms through the mapping tool (skos:member, skos:Collection) and also for selecting to display only subcategories of them(skos:inScheme, skos:ConceptScheme).

### 3.1.6 OAI publication

The main objective of the publication task is the delivery of high quality metadata to Europeana. MINT enables crosswalks between EDM Sounds profile and EDM schema through a user-friendly interface. In addition, it provides validation services for both EDM Sounds and EDM metadata models together with preview interfaces, through which providers can a priori check how their metadata would look like when published on Europeana, thus ensuring high quality metadata.

#### 3.1.6.1 Mapping (XSLT) from EDM Sounds to EDM

As mentioned earlier the metadata are delivered to Europeana in EDM format. Mapping metadata from EDM Sounds to EDM is performed when users perform the “publish” functionality from MINT. This mapping is completed automatically using the corresponding XSLT that was specifically implemented and deployed within MINT platform for this purpose. The transformation is 1-1, meaning that each EDM Sounds element is mapped to one EDM element. For fields that exist in both schemas the assignment is obvious. For the rest of the fields the assignment is performed according to following table:

---

<sup>2</sup> [https://jena.apache.org/documentation/serving\\_data/](https://jena.apache.org/documentation/serving_data/)

<sup>3</sup> <http://www.w3.org/TR/skos-reference/>

**Table 1. EDM Sounds to EDM term correspondence**

EDM Sounds		EDM	
edm:ProvidedCHO		edm:ProvidedCHO	
	dcterms:dateCopyrighted		dc:date or dc:rights
	dcterms:modified		dc:date
	edm:isGatheredInto		dcterms:isPartOf
	skos:note		dc:coverage
	mo:remaster_of		edm:isDerivativeOf
	mo:record_side		dcterms:extent
	mo:track_number		dcterms:extent
	mo:track_count		dcterms:extent
	ebucore:hasGenre		dc:subject
	ebucore:dateDigitised		dc:date
	ebucore:duration		dcterms:extent
	ebucore:audioChannelNumber		dcterms:extent
	ebucore:audioTrackConfiguration		dc:format
	ebucore:bitRate		dc:format
	ebucore:fileSize		dc:format
	ebucore:hasAudioEncodingFormat		dc:format
	ebucore:hasMimeType		dc:format
	ebucore:sampleRate		dcterms:extent
	ebucore:sampleSize		dc:format
	schema:version		edm:isDerivativeOf
<b>edm:Collection</b>		<b>edm:ProvidedCHO</b>	
	dc:creator		dc:creator
	dc:identifier		dc:identifier
	dc:language		dc:language
	dc:relation		dc:relation
	dc:rights		dc:rights
	dc:subject		dc:subject
	dc:title		dc:title
	dcterms:accrualPeriodicity		dcterms:extent
	dcterms:alternative		dcterms:alternative
	dcterms:audience		dc:description
	dcterms:description		dc:description
	dcterms:extent		dcterms:extent
	dcterms:hasPart		dcterms:hasPart
	dcterms:isPartOf		dcterms:isPartOf
	dcterms:isReferencedBy		dcterms:isReferencedBy
	dcterms:provenance		dcterms:provenance
	dcterms:spatial		dcterms:spatial
	dcterms:temporal		dcterms:temporal
	edm:highlight		
	edm:isRelatedTo		edm:isRelatedTo

	edm:itemGenre		dc:subject
	cld:dateItemsCreated		dcterms:temporal
	cld:itemFormat		dc:format
	cld:itemType		dc:type and edm:type
<b>edm:WebResource</b>		<b>edm:WebResource</b>	
	ebucore:dateDigitised		dcterms:created
	ebucore:duration		dcterms:extent
	ebucore:audioChannelNumber		dcterms:extent
	ebucore:audioTrackConfiguration		dc:format
	ebucore:bitRate		dc:format
	ebucore:fileSize		dc:format
	ebucore:hasAudioEncodingFormat		dc:format
	ebucore:hasMimeType		dc:format
	ebucore:sampleRate		dcterms:extent
	ebucore:sampleSize		dcterms:extent
	mo:record_side		dcterms:extent
	mo:remaster_of		edm:isDerivativeOf
	mo:track_number		dcterms:extent
	mo:track_count		dcterms:extent
	schema:version		edm:isDerivativeOf
	owl:sameAs		owl:sameAs
<b>edm:Agent</b>		<b>edm:Agent</b>	
	skos:hiddenLabel		skos:altLabel
	rdaGr2:placeOfBirth		rdaGr2:placeofBirth
	rdaGr2:placeOfDeath		rdaGr2:placeofDeath
<b>edm:Place</b>		<b>edm:Place</b>	
	skos:hiddenLabel		skos:altLabel
<b>edm:TimeSpan</b>		<b>edm:TimeSpan</b>	
	skos:hiddenLabel		skos:altLabel
<b>skos:Concept</b>		<b>skos:Concept</b>	
	skos:hiddenLabel		skos:altLabel
<b>mo:MusicGroup</b>		<b>edm:Agent</b>	
	skos:prefLabel		skos:prefLabel
	skos:altLabel		skos:altLabel
	skos:hiddenLabel		skos:altLabel
	skos:note		skos:note
	dc:date		dc:date
	dc:identifier		dc:identifier
	dcterms:hasPart		dcterms:hasPart
	dcterms:isPartOf		dcterms:isPartOf
	edm:begin		edm:begin
	edm:end		edm:end
	edm:hasMet		edm:hasMet
	edm:isRelatedTo		edm:isRelatedTo
	foaf:name		foaf:name
	rdaGr2:biographicalInformation		rdaGr2:biographicalInf



	on		ormation
	rdaGr2:dateOfBirth		rdaGr2:dateOfBirth
	rdaGr2:dateOfDeath		rdaGr2:dateOfDeath
	rdaGr2:dateOfEstablishment		rdaGr2:dateOfEstablishment
	rdaGr2:dateOfTermination		rdaGr2:dateOfTermination
	rdaGr2:gender		rdaGr2:gender
	rdaGr2:professionOrOccupation		rdaGr2:professionOrOccupation
	rdaGr2:placeOfBirth		rdaGr2:placeOfBirth
	rdaGr2:placeOfDeath		rdaGr2:placeOfDeath
	owl:sameAs		owl:sameAs
	mo:collaborated_with		edm:hasMet
	mo:member_of		edm:hasMet

### 3.1.6.2 OAI

A NoSQL (MongoDB) backend has been set up for storing the successfully transformed metadata to EDM and serving them using the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH is a low barrier mechanism for repository interoperability. Data providers set up repositories that expose structured metadata, to which service providers make OAI-PMH service requests to harvest that metadata. The protocol consists of a set of six verbs or services invoked within HTTP. In the context of an aggregation, OAI-PMH provides a mechanism for interoperability between the ingestion platform and other modules or platforms (e.g. Europeana's United Ingestion Mechanism).

MINT's OAI repository is capable of managing heterogeneous collections of metadata records while exposing services for mapping and transforming from one metadata schema to another. In order to extend these functionalities with the OAI-PMH protocol and thus to expose metadata through an interoperable mechanism, MINT implements the defined OAI-PMH verbs on top of the underlying, domain-specific data layer. An issue that arises in the case of aggregations is that while being able to manage collections of metadata records, the OAI-PMH verbs operate on an item level, which makes the implementation of the appropriate verbs, directly on top of a collection-based data layer, a challenging task. For this reason, and also to design and introduce a set of robust, enhanced metadata processing services, an export mechanism is implemented in the MINT platform, facilitating scalable and reliable data delivery and exchange between different data layers and repositories.

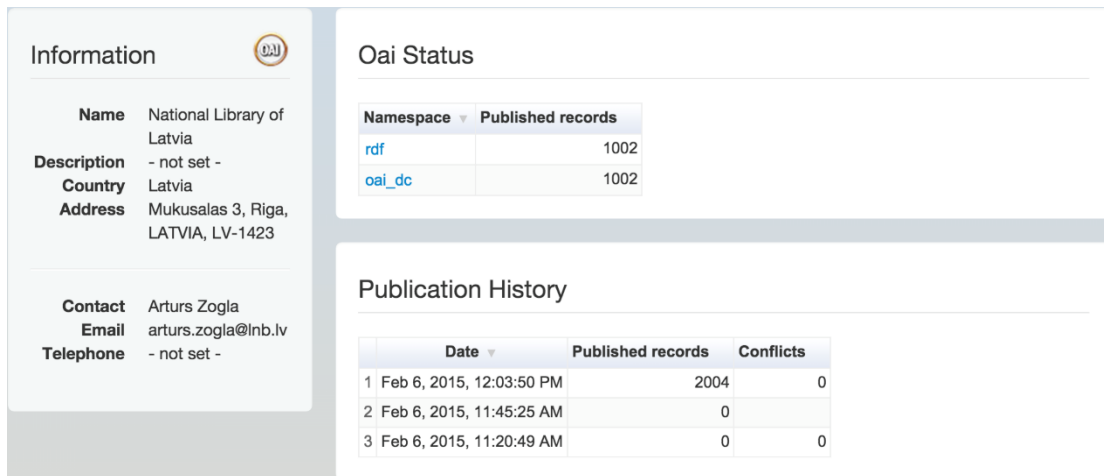
For the needs of Europeana Sounds, the insertion mechanism has been re-implemented allowing the use of more than one schemas through MINT resulting in multiple OAI namespaces for the same organization in the repository. In short, the previous insertion mechanism allowed the use of only one schema, which was EDM. The transformed EDM records from MINT were transferred to the NoSQL database while the OAI\_DC records were constructed from them and were served under two different OAI namespaces that were predefined. The new insertion mechanism is more sophisticated and permits the use of arbitrary schemas - in our case EDM Sounds and EDM metadata - and the OAI exposure under different or the same namespace - in our case EDM records ends up in rdf namespace while EDM Sounds records end up in EDM Sounds namespace .

The old version can be found at:

<http://panic.image.ntua.gr:9000/manager/projects/sounds>

The old version of the OAI repository illustration information for an organisation can be viewed at:

<http://panic.image.ntua.gr:9000/manager/projects/sounds/organizations/1003>



**Information**

**Name** National Library of Latvia  
**Description** - not set -  
**Country** Latvia  
**Address** Mukusalas 3, Riga, LATVIA, LV-1423

**Contact** Arturs Zogla  
**Email** arturs.zogla@lnb.lv  
**Telephone** - not set -

**Oai Status**

Namespace	Published records
rdf	1002
oai_dc	1002

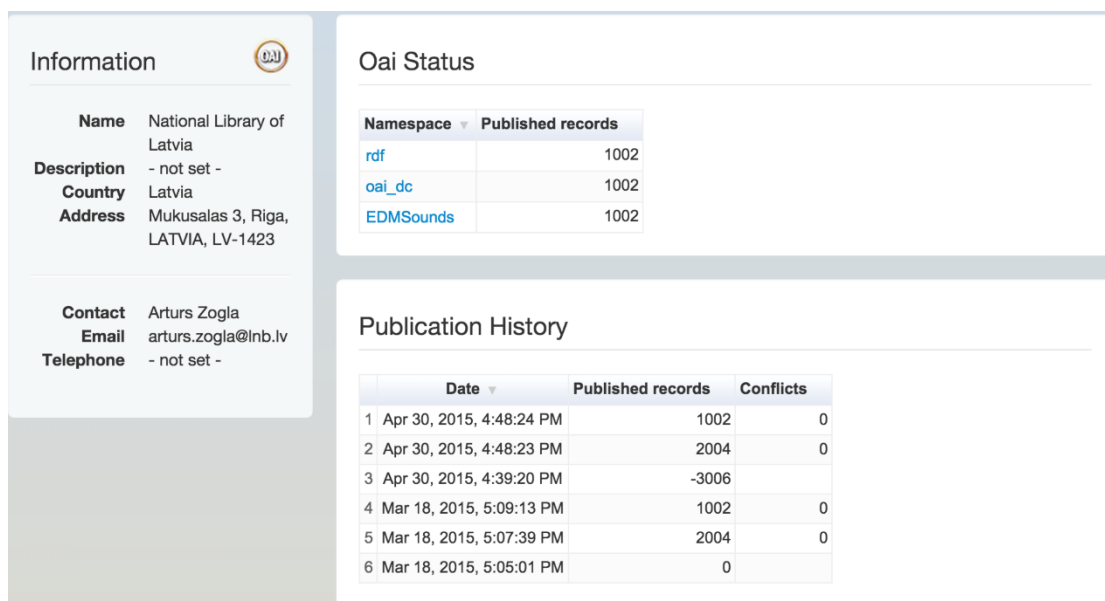
**Publication History**

Date	Published records	Conflicts
1 Feb 6, 2015, 12:03:50 PM	2004	0
2 Feb 6, 2015, 11:45:25 AM	0	
3 Feb 6, 2015, 11:20:49 AM	0	0

**Figure 3-4. Publication history for one organization using the old version of OAI**

The updated version of the OAI repository is available at:

<http://panic.image.ntua.gr:9876/manager/projects/sounds/organizations/1003>



**Information**

**Name** National Library of Latvia  
**Description** - not set -  
**Country** Latvia  
**Address** Mukusalas 3, Riga, LATVIA, LV-1423

**Contact** Arturs Zogla  
**Email** arturs.zogla@lnb.lv  
**Telephone** - not set -

**Oai Status**

Namespace	Published records
rdf	1002
oai_dc	1002
EDMSounds	1002

**Publication History**

Date	Published records	Conflicts
1 Apr 30, 2015, 4:48:24 PM	1002	0
2 Apr 30, 2015, 4:48:23 PM	2004	0
3 Apr 30, 2015, 4:39:20 PM	-3006	
4 Mar 18, 2015, 5:09:13 PM	1002	0
5 Mar 18, 2015, 5:07:39 PM	2004	0
6 Mar 18, 2015, 5:05:01 PM	0	

**Figure 3-5. Publication history of one organization using the updated OAI**

## 3.2 Description of basic aggregation functionalities in MINT

Up to now we provided a detailed description of MINT updates specialized for Europeana Sounds project. The remaining of this section refers to the basic aggregation functionalities serviced by MINT, accompanied by a brief description.

### 3.2.1 Metadata upload and preparation

Registered users are able to upload their metadata records in XML or CSV serialisation, using the HTTP, FTP and OAI-PMH protocols. XML records will be stored and indexed for statistics, previews, access from the mapping tool and subsequent services.

When an upload is successfully performed, providers are prompted to prepare their XML records for mapping by defining items, as illustrated in Figure 3-6 Dataset options and Figure 3-7 Define items. When items in the uploaded dataset are successfully defined, providers can preview their dataset statistics as shown in Figure 3-8 Dataset statistics.

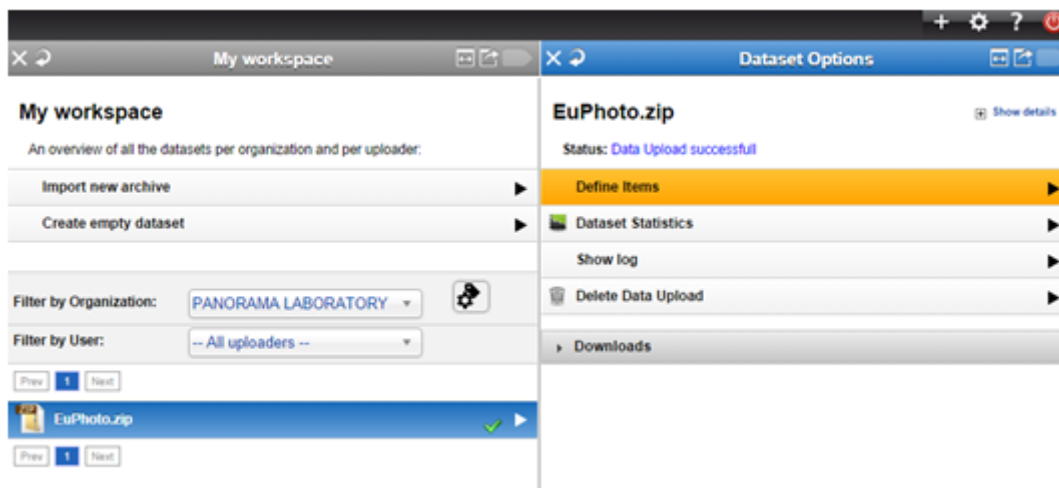


Figure 3-6 Dataset options

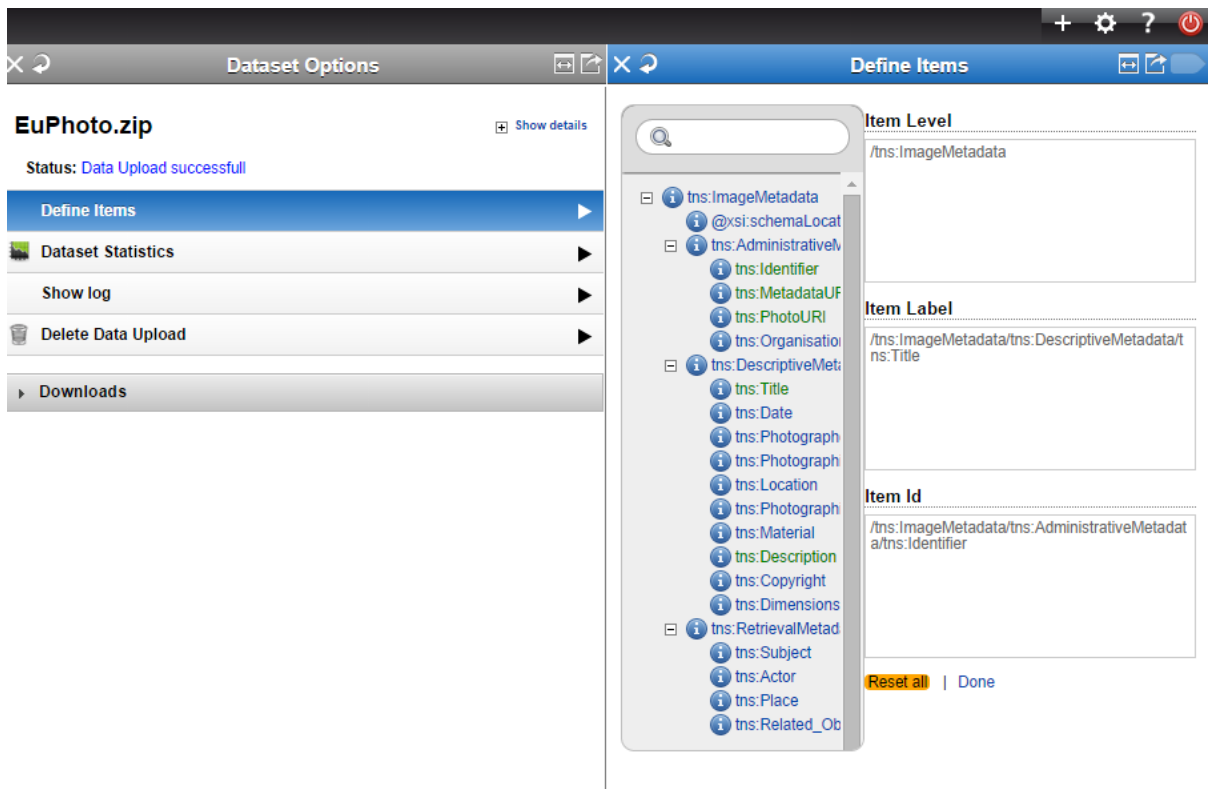
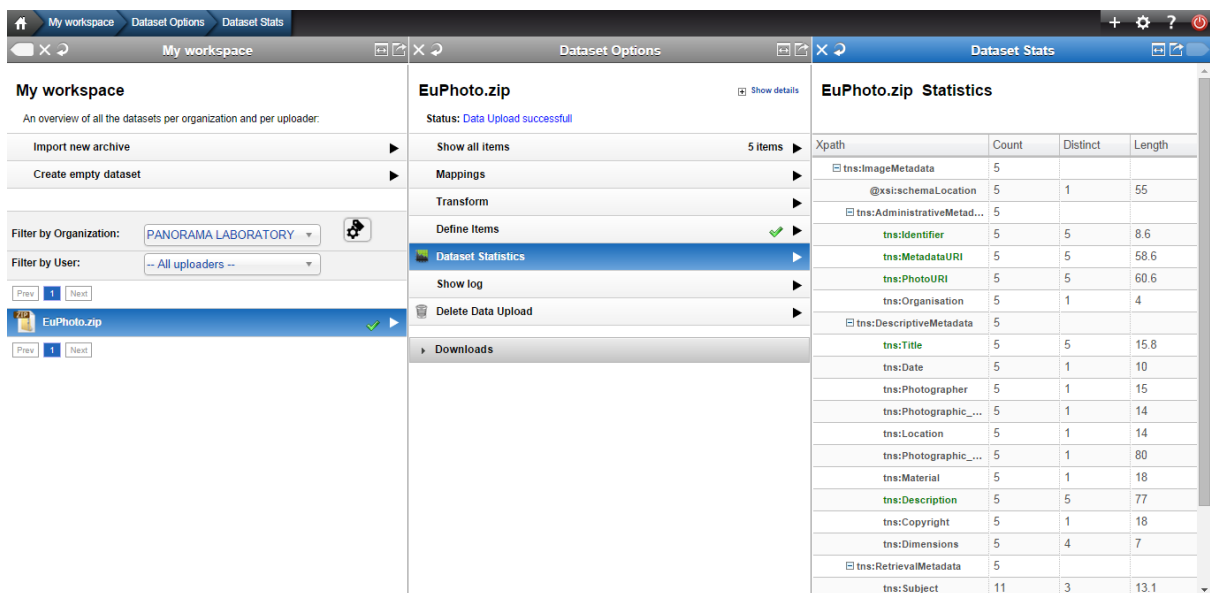


Figure 3-7 Define items



The screenshot shows the 'Dataset Statistics' interface for the 'EuPhoto.zip' dataset. The status is 'Data Upload successful'. The interface is divided into several sections:

- My workspace:** A sidebar with buttons for 'Import new archive', 'Create empty dataset', 'Filter by Organization: PANORAMA LABORATORY', 'Filter by User: -- All uploaders --', and 'EuPhoto.zip'.
- Dataset Options:** A sidebar with buttons for 'Show all items', 'Mappings', 'Transform', 'Define Items', 'Dataset Statistics', 'Show log', 'Delete Data Upload', and 'Downloads'.
- EuPhoto.zip Statistics:** A table showing the following data:

Xpath	Count	Distinct	Length
tns:ImageMetadata	5		
@xsi:schemaLocation	5	1	55
tns:AdministrativeMetad...	5		
tns:Identifier	5	5	8.6
tns:MetadataURI	5	5	58.6
tns:PhotoURI	5	5	60.6
tns:Organisation	5	1	4
tns:DescriptiveMetadata	5		
tns>Title	5	5	15.8
tns>Date	5	1	10
tns:Photographer	5	1	15
tns:Photographic_...	5	1	14
tns:Location	5	1	14
tns:Photographic_...	5	1	80
tns:Material	5	1	18
tns:Description	5	5	77
tns:Copyright	5	1	18
tns:Dimensions	5	4	7
tns:RetrievalMetadata	5		
tns:Subject	11	3	13.1

Figure 3-8 Dataset statistics

### 3.2.2 Metadata mapping

Mapping of metadata into the desired schema is the basic functionality of MINT ingestion platform. A visual mapping editor for the XSL language is used, as illustrated in Figure 3-10. Mapping is performed following a drag-and-drop procedure, selecting items from the input schema area and dropping them in

the mapping area. Input operations are translated into the corresponding XSLT code. The mapping editor supports string manipulation functions for input elements, structural element mappings, constant or controlled value (target schema enumerations) assignment, conditional mappings and value mappings between input and target value lists. All different mapping options are illustrated graphically in Figure 3-11. By setting preview options users have the ability to preview XML code of import and transformed items, as in Figure 3-12.

Handling of metadata records includes indexing, retrieval, update and transformation of XML files and records. XML processors 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 services are multi-threaded and queue processing mechanisms are implemented.

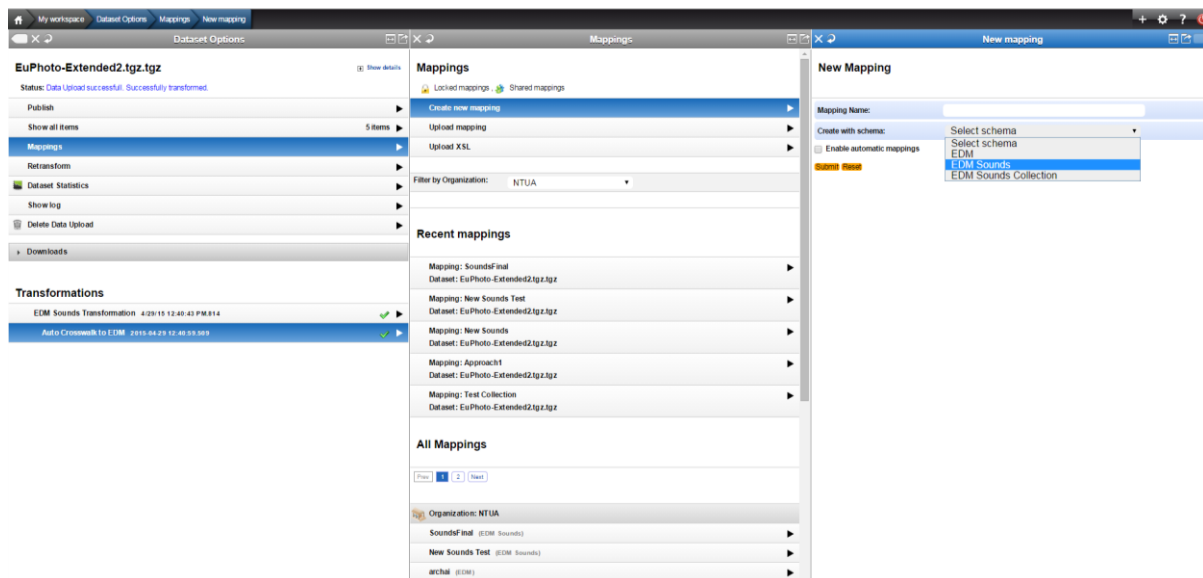


Figure 3-9 Create a new mapping

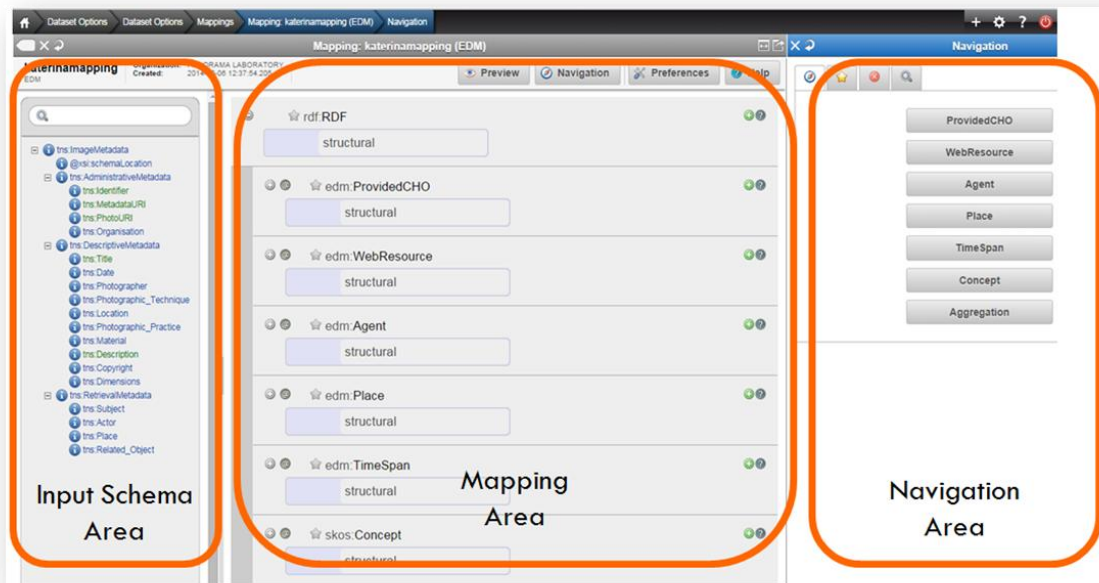
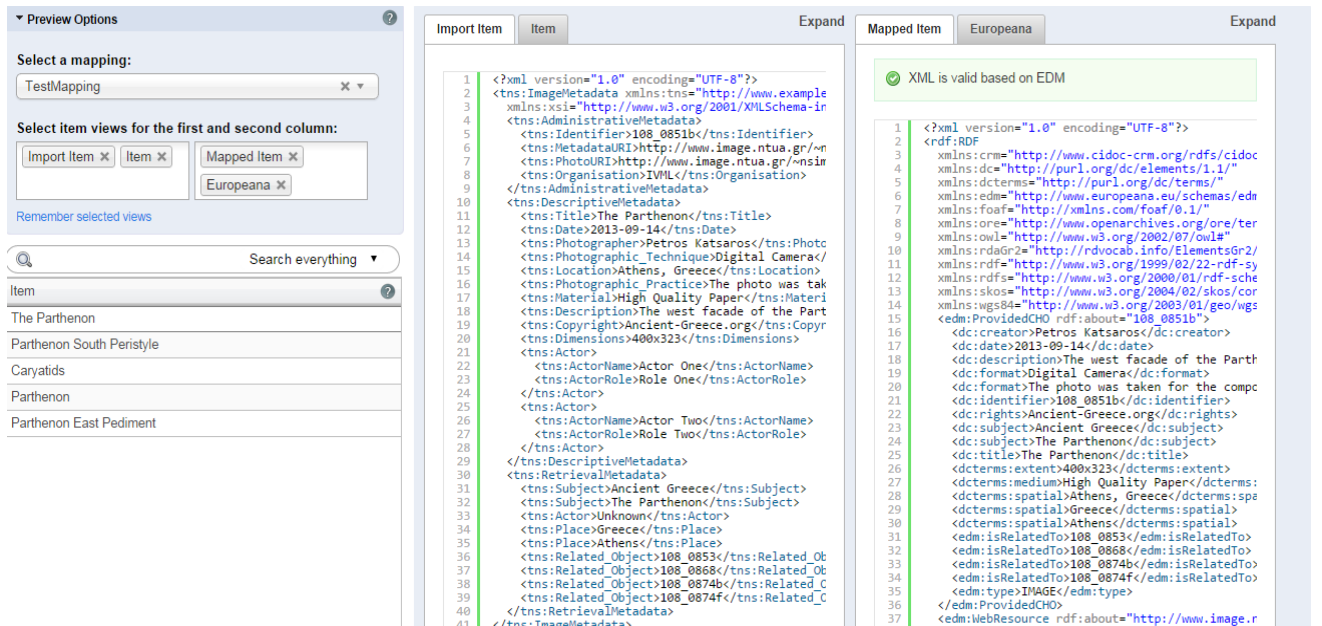


Figure 3-10 The mapping editor



Figure 3-11 Mapping options



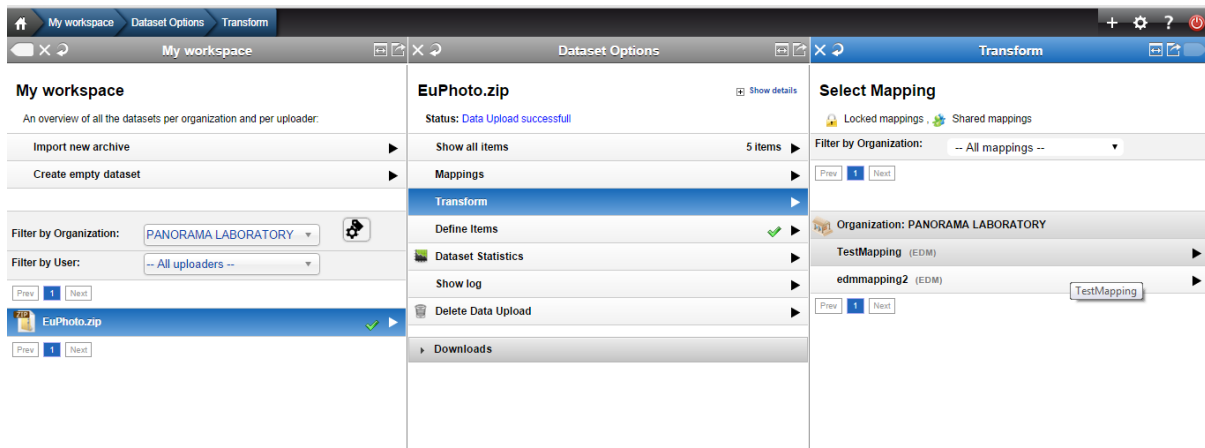
The screenshot shows the MINT interface with three main panels:

- Preview Options:** A sidebar on the left with a search bar and a list of items including 'The Parthenon', 'Parthenon South Peristyle', 'Caryatids', 'Parthenon', and 'Parthenon East Pediment'. It includes 'Import Item', 'Mapped Item', and 'Europeana' buttons.
- Import Item / Item:** A central panel displaying XML metadata for 'The Parthenon', including fields like <Identifier>108\_0851b</Identifier>, <Title>The Parthenon</Title>, and <Subject>Ancient Greece</Subject>.
- Mapped Item / Europeana:** A right panel showing a green message: 'XML is valid based on EDM'. Below it, the mapped XML structure is visible, including <dc:format>Digital Camera</dc:format> and <dc:subject>Ancient Greece</dc:subject>.

Figure 3-12 Preview options

### 3.2.3 Transformation, quality check and publication

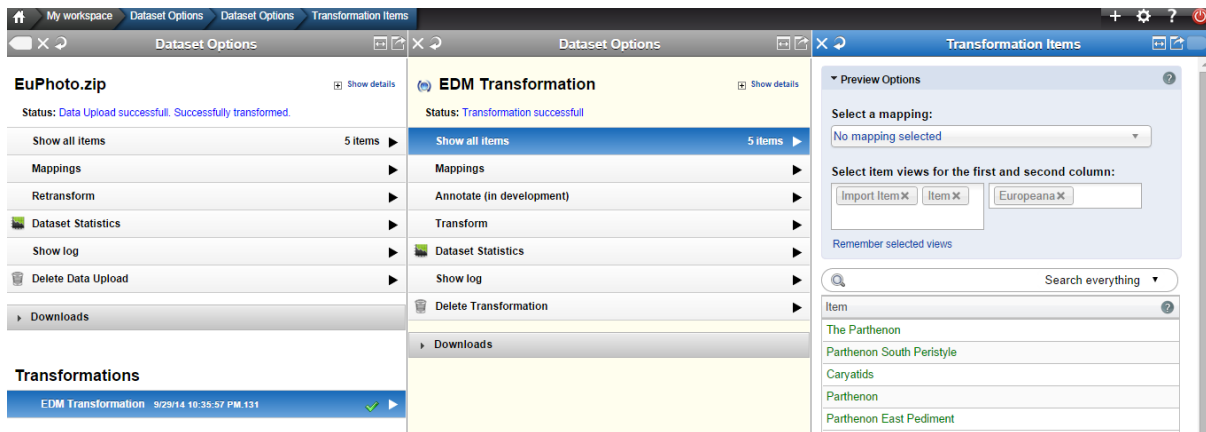
After completing a successful mapping, data providers can use MINT to transform their metadata to the EDM sounds profile target schema.



The screenshot shows the MINT interface with three main panels:

- My workspace:** A sidebar on the left showing a list of datasets, including 'EuPhoto.zip' with a status of 'Data Upload successful'.
- Dataset Options:** A central panel showing a workflow for 'EuPhoto.zip' with steps: 'Show all items', 'Mappings', 'Transform' (highlighted), 'Define Items', 'Dataset Statistics', 'Show log', and 'Delete Data Upload'.
- Select Mapping:** A right panel showing a list of mappings for 'Organization: PANORAMA LABORATORY', including 'TestMapping (EDM)' and 'edmmapping2 (EDM)'.

Figure 3-13 Data transformation



**Figure 3-14 Transformation - Preview items**

In order to publish metadata to Europeana, providers have to pass a number of quality checks. Structural validity, completeness of metadata elements and use of controlled vocabularies are among the checks that can be applied. State of the art technologies are used to allow for a reliable, scalable and portable messaging and processing system, used in and between different services. Metadata will be published in XML, JSON or RDF according to the mechanism and usage.

After quality check, the transformed dataset can be prepared for publication by consenting to the Europeana Data Exchange Agreement and setting appropriately the Europeana Rights.

### 3.3 Technical Details

#### 3.3.1 Platform

The platform is developed using 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 also reuses 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 MINT platform offers the user an 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.

#### 3.3.2 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 ingestion platform and subsequent services. Current developments aim to support relational database schemata and OWL/RDFS ontologies as input.



### 3.3.3 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.

### 3.3.4 Normalization and 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. The vocabularies have to be represented in SKOS.

## 4 Europeana data processes

### 4.1 Ingestion

Data are imported to Europeana's servers using the open-source REPOX solution. REPOX can communicate with remote data repositories via a number of protocols, the ones primarily used being OAI-PMH, FTP, and HTTP. EDM data for the Europeana Sounds project are harvested in EDM using OAI-PMH and exposed on the Europeana OAI repository:

[http://uim-external-apps.isti.cnr.it:8080/repoXUI\\_Europeana/OAIHandler?verb=Identify](http://uim-external-apps.isti.cnr.it:8080/repoXUI_Europeana/OAIHandler?verb=Identify)

The harvested data are stored on REPOX's PostgreSQL database.

### 4.2 Data transformations

From REPOX's database, data are copied to Europeana Mint's PostgreSQL database. Operations officers perform in MINT cleaning, mapping and validation operations.

#### 4.2.1 Cleaning and quality checks (MINT)

A number of quality checks are performed using the preview and statistics functionalities of MINT: general structure of the provided EDM data, existence of unique and well-formed identifiers, and richness of the literal values. For the Europeana Sounds project, the quality assurance is also done according to the data requirements specified by WP1 and reproduced here in Table 1. Extra attention is paid to the use of SKOS ontologies (See 3.1.5). When necessary, cleaning is performed using advanced mapping functionalities.

## 4.2.2 Mapping, transformation, validation (MINT)

From EDM External, data are transformed to the EDM Internal variation of the schema. XSD and Schematron validation is performed and records that do not meet the validation are marked as invalid.

## 4.2.3 Itemization and Unique Identifiers Generation (UIM)

The transformed valid data produced are (re-)imported and stored into UIM's Mongo database. The process of re-importing triggers additional transformations on the data:

- Itemization: the Europeana Sounds data make use of new EDM entities such as edm:Collection and in order to properly serve these data on the Europeana portal and API, a splitting process is needed to extract sets of interlinked EDM entities according to their semantic relationships and generate one item for each edm:ProvidedCHO entity.
- Deduplication and unique identifiers generation: the uniqueness of each obtained item is checked within each collection and Europeana identifiers are created for generation of permalinks to access the published records.

## 4.3 Preview caching (Media Harvester)

Most Europeana sounds objects are submitted to Europeana including a link to a visual preview. Provided links are checked and analysed, previews are generated and stored to be made retrievable on both portal and API.

## 4.4 Enrichment (UIM)

### 4.4.1 Dereferencing and vocabulary mapping

Specifically for the Europeana Sounds project and as described in section 3.1.5, links to concepts from SKOS sounds related ontologies are included in the provided EDM data. Three ontologies used by the project were mapped from SKOS to EDM: Sounds genres, DISMARC genres and DISMARC formats.

The Europeana dereferencing plugin is triggered on the EDM sounds data, enabling:

- Queries using SPARQL on these three external datasets;
- Transformation of the SKOS relevant data into EDM and download of the result in a Mongo database;

Multilingual enrichment of the EDM Sounds data: values from the contextual resource are added to the original object description.

### 4.4.2 Semantic enrichment

Europeana's current enrichment process is based on the Annocultor tool. Europeana enriches all provided data by creating links to contextual resources - places, concepts, agents and time periods.

Named entities are found in records using a set of heuristics. Contextual information, known about these entities from external data sources, is appended to the original record in the form of EDM contextual classes. This information currently includes multilingual representation of the identified contextual entity, as well as links to similar contextual resources. For the moment, contextual information is gathered from such external data sources as: *Geonames*, for geographical places; *DBPedia* for people and concepts; *Gemet* for concepts; and *Semium* for time periods. The processed output is stored locally in a MongoDB database. The enrichment plugin, which implements the described process, is also responsible for creating the version of the record database and search index to be published later on the production environment.

## 4.5 Publication

The process of transferring the up-to-date content produced by the ingestion process to the production environment happens currently on a monthly basis. During the process, the content is finalised and transferred manually to the production environment. In the future, Europeana is planning to introduce a continuous publication process which will not require rigidly defined regular cycles but instead allow for more frequent updates of published content.

In May 2015, the first 26,620 Sounds records from nine data providers were successfully aggregated and published in Europeana. Monthly deliveries for additional data as well as updates will ensure that the project targets are met gradually.

## 5 Summary

Europeana Sounds' aggregation workflow is implemented using the MINT platform, the Europeana Data Model - Sounds Profile specification (D1.4), and the Europeana Sounds Ontologies (D1.3). The operation of the aggregation is planned and monitored by WP1 in order to coordinate content provision (workflow guide, publication cycles, contingency planning etc. The technical infrastructure is implemented, monitored and maintained by WP5 in order to support the activities of WP1, implement the produced specifications and enable emerging ingestion and publication requirements. This document reports on the deployment of fully functional updated aggregation system, contributes to task T5.3 - *Aggregator deployment and maintenance* and uses the outcome of T5.1 - *Aggregation infrastructure design* as well as experience obtained in T5.2 - *Aggregation infrastructure evaluation*.

The aggregation as described throughout the document is a two-phase process. The first phase is handled by MINT ingestion platform and is the part where providers get involved in the aggregation process. In the second phase, transformed metadata are delivered to Europeana in EDM via OAI repository and are then handled and published on Europeana website.

The MINT ingestion platform has been previously applied in Europeana related projects, making it by now one of the mature solutions for metadata handling and reuse. While the base of the main aggregation functionalities of the MINT platform remains the same, necessary adjustments for the implementation within the given data-domain, and continued development from a service provider's perspective have been applied to meet the requirements and objectives of Europeana Sounds project.

All specializations applied to MINT platform with respect to Europeana sounds project such as frontend and backend reconstruction, advance mapping functionalities, EDM sounds profile implementation and deployment, SKOS vocabulary support and OAI publication, are described in this document. Additionally the basic aggregation functionalities offered by MINT - metadata upload and preparation, mapping, transformation, quality check and publication – are presented.

After the completion of MINT aggregation services, a set of data processes takes place on European’s site before data publication on the Europeana portal. These processes, as further explained in this report, include ingestion, transformation, and preview caching, enrichment and publication. In May 2015, 26,620 Sounds objects from nine data providers were successfully aggregated and published in Europeana. Monthly deliveries for additional data as well as updates will ensure that the KPIs are met gradually.

## 6 References

Ref 1	D1.3 – Ontologies for sound <a href="http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Deliverables/EuropeanaSounds-D1.3-Ontologies-for-sound-v1.2.pdf">http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Deliverables/EuropeanaSounds-D1.3-Ontologies-for-sound-v1.2.pdf</a>
Ref 2	D1.4 – EDM profile for sound <a href="http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Deliverables/EuropeanaSounds-D1.4-EDM-profile-for-sound.pdf">http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Deliverables/EuropeanaSounds-D1.4-EDM-profile-for-sound.pdf</a>
Ref 3	D5.1 – Report on the evaluation of the aggregation mechanism <a href="http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Deliverables/EuropeanaSounds-D5.1-Evaluation-of-the-aggregation-mechanism-v1.0.pdf">http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Sounds/Deliverables/EuropeanaSounds-D5.1-Evaluation-of-the-aggregation-mechanism-v1.0.pdf</a>
Ref 4	MS23 – Revised aggregation design available
Ref 5	MS24 – Aggregation infrastructure prototype available
Ref 6	MS25 – Sounds SKOS ontology normalization and cleaning module beta
Ref 7	MS26 – Aggregation mechanism ready
Ref 8	EDM profile for Sound <a href="http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_task_forces/EDMSound//TF_Report_EDM_Profile_Sound_301214.pdf">http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_task_forces/EDMSound//TF_Report_EDM_Profile_Sound_301214.pdf</a>

## Appendix A: Terminology

A project glossary is provided at: <http://pro.europeana.eu/web/guest/glossary>.

Additional terms are defined below:

Term	Definition
APEX	Archives Portal Europe network of excellence
CSV	Comma-separated values

EC-GA	Grant Agreement (including Annex I, the Description of Work) signed with the European Commission
EDM	Europeana Data Model
FTP	File Transfer Protocol
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
MARC	MAchine-Readable Cataloguing standard
MINT	Metadata INTeroperability platform
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System data model
TEL	The European Library
UIM	United Ingestion Manager, part of Europeana's ingestion toolset
WP	Work Package
XLST	EXtensible Stylesheet Language Transformations
XML	Extensible Markup Language
XSD	XML Schema Definition