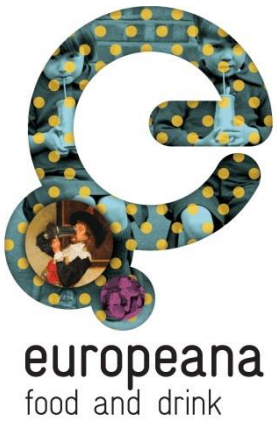


Grant Agreement 621023

Europeana Food and Drink

D2.4 Europeana Food and Drink Content Upload Report

Deliverable number	*D2.4*
Dissemination level	*PU*
Delivery date	*March 2015*
Status	*Final*
Author(s)	*Elena Lagoudi (PS)*

This project is funded by the European Commission under the
ICT Policy Support Programme part of the
Competitiveness and Innovation Framework Programme

Revision History

Revision	Date	Author	Organisation	Description
V1.0	20/03/2015	Elena Lagoudi	PS	First draft
V2.0	25/03/2015	Laura Gibson	CT	First Review
V3.0	26/03/2015	Laura Miles	CT	Second Review
V4.0	27/03/2015	Nikos Simou	NTUA	Final Review

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

Contents

Introduction

The core concept of the Europeana Food and Drink Best Practice Network is to kick-start the creative and commercial re-use of digital content relating to food and drink from the culture sector to drive new commercial applications, relationships and partnerships.

Food and drink serves the dual purpose of providing a powerful thematic focus to inspire creative re-use of digital cultural content while offering sufficient breadth to support a wide range of applications and approaches.

1.1 Background

Work Package 2 of Europeana Food and Drink will identify, describe, enhance, license and upload a body of high-quality digital assets and their associated metadata, to support the delivery of commercial applications and public engagement activity.

These objectives will be facilitated by the work under Task 2.2:

- Develop a food and drink taxonomy to support classification and resource discovery
- Map local metadata structures to the requirements of the Europeana Data Model
- Enrich the assets and their associated metadata using this classification scheme
- Apply the relevant licensing conditions drawn from the Europeana Content Re-use Framework

These objectives will be facilitated by the work under Task 2.3:

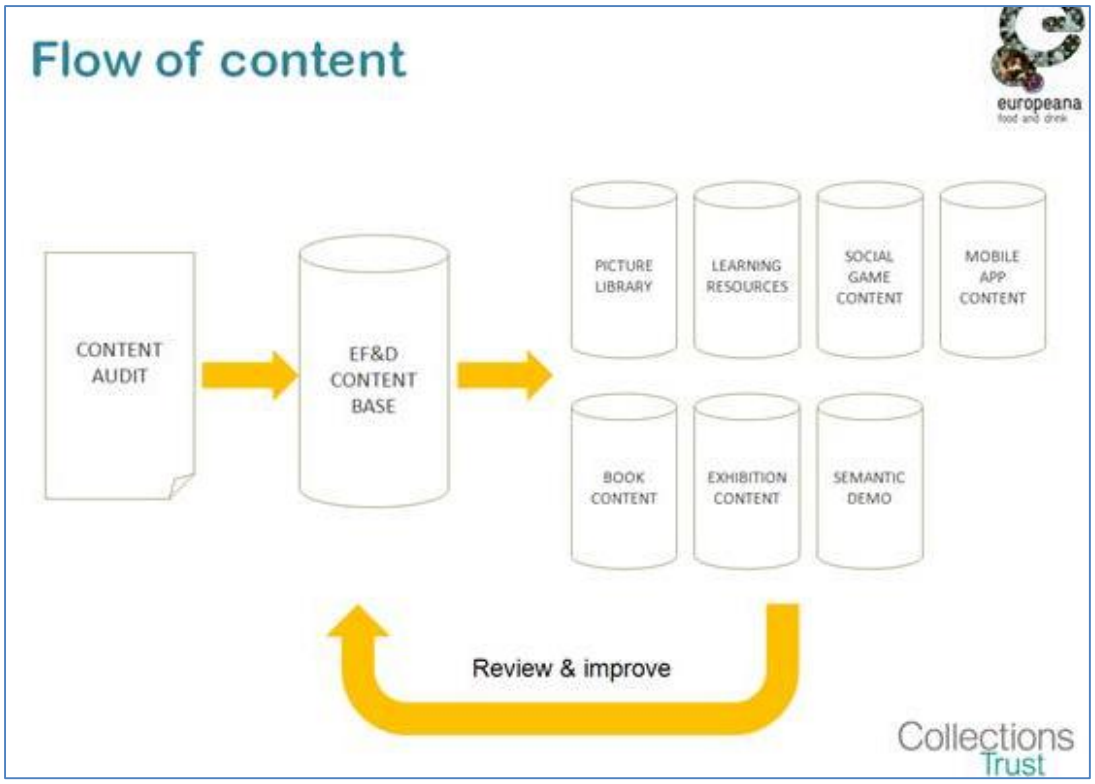
- Ingest the primary digital assets into the Europeana Cloud infrastructure
- Ingest the descriptive metadata and identifiers into Europeana

These objectives will be facilitated by the work under Task 2.4:

- Ingest new content contributions to the Dark Aggregator (Europeana Inside)
- Ingest user-generated content alongside existing assets
- Apply appropriate licensing and metadata schemata to new content

As seen above, the content (digital assets) was to be ingested into the Europeana Cloud infrastructure and the metadata into Europeana though the MINT instance [1]that has been set up for this project. Since Europeana Cloud will not be finished in time for the first iteration of Food and Drink, the Content Base for the project is an umbrella term. Content will be stored on institutions' servers if their content is already available online through Europeana, or on the Europeana Inside Dark Aggregator if their content is not yet available online. Access to the metadata for creative re-use will be given thought the EuropeanaAPI once they are published the API in both cases.

[1] http://guinness.image.ntua.gr:9090/foodanddrink/



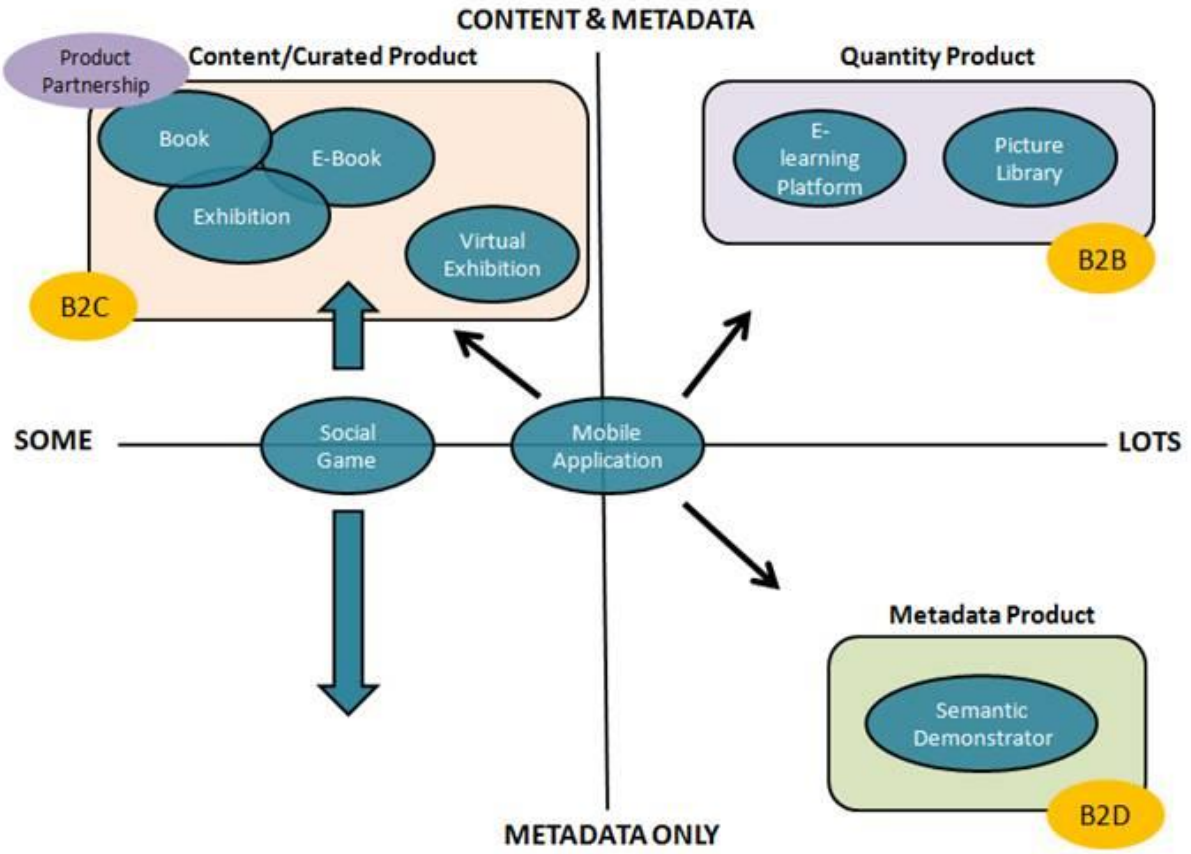
1.2 Role of deliverable in the project

WP2 as a whole will support the objective of identifying, describing, enhancing, licensing and uploading a body of high-quality digital assets and their associated metadata, to support the delivery of commercial applications and public engagement activity.

The Content Base developed throughout the project will be available for cultural institutions, creative industries, professional users and third party developers in order to easily search for the cultural resources that meet their retrieval criteria so as to use and re-use them for the development of applications, products and services.

The result will be a body of high quality digital assets and semantically-enriched metadata that can support a wider range of (multi-lingual) natural language applications such as search, discovery and browse.

The Europeana Food and Drink Content Base will feed content into the applications and products the Consortium will develop which includes a social game and exhibition. Based on the quantity of content and metadata needed, the applications were split into the following diagram:



D2.4 is one of a series of Content Upload reports (approximately every 6 months) aimed at monitoring the progress of uploading digital assets to the Content Base whereby they can be ingested to Europeana. New partners who do not already have their digital assets available online can store their content on the Dark Aggregator.

2. Approach

2.1 Methodology

First contact with Europeana Cloud was made in February 2015. Until then, in order to move things forward until the launch of Europeana Cloud, we investigated various solutions for hosting the Content Base and researched their functionalities and specifications.

One solution we investigated was that of “merging” the Picture Library product with the Content Base. The Consortium felt that, for the project’s content to be findable and exploitable, a more commercial focused application had to be used. It was decided, since the Content Base needs to provide the underlying infrastructure from which the other products will draw their content from, it has to be merged with the Picture Library, so that the feasibility of this entrepreneurial model that the Consortium is developing for GLAMS can be piloted.

The requirements and functionalities of this Content Base/Picture Library were summarized as:

–Storing high-quality images & media
–Storing associated metadata
–Managing rights associated with content & metadata
–Easy workflows for ingestion, mapping, data management and export
–Export to LIDO/EDM compatible format
–Uses existing technology/infrastructure
–Scalable to meet future demand

In this direction, we researched using some of the partners' infrastructures or products.

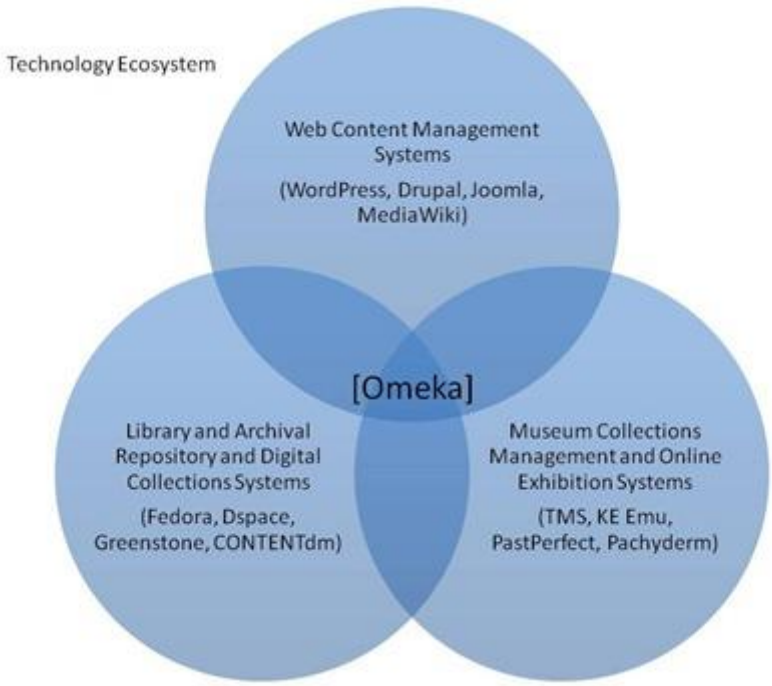
Initially, we investigated ALINARI's infrastructure. This option unfortunately was not deemed viable and the infrastructure could not support the scale and volume of content it needed to host.

TOPFOTO's Picture Kitchen has the obvious advantage that it is already a commercially set platform for licensing food and drink related content, so the business model is already tested and working. The disadvantages are that the documentation and search functionalities are quite basic, TOPFOTO has probably not had the experience of many different providers ingesting content into their Picture Library so various issues with mapping from the content providers' schemata to TOPFOTO's schema and work flow and technical support during ingestion. However, they can export to XML, as they did for the Europeana Photography project by importing the XML into NTUA's MINT server which mapped to EDM via LIDO.
They use Fotostation to annotate their metadata into the IPTC headers.

KEEPTHINKING's QI, on the other hand, seems much more suited to support such a demanding process, although it remains to be investigated how it can support content uploaded by many different partners. Its technical features though better cover the need for discoverability, faceted search, browsing functionalities, versioning, licensing etc. However, since it is a proprietary software with a set-up and maintenance cost it does pose an obstacle in terms of the Business Model for using it for similar joint sector projects.

Another path was investigated with the use open source software. We investigated using Low Cloud, which is built using the open source OMEKA platform.

OMEKA is a free, flexible, and open source web-publishing platform for the display of library, museum, archives, and scholarly collections and exhibitions. Its "five-minute setup" makes launching an online exhibition as easy as launching a blog. It accepts and stores all types of files, including images, video, audio, multi-page documents and PDFs, Power Point presentations, et al. Individual items may contain multiple files. It falls at a crossroads of Web Content Management, Collections Management, and Archival Digital Collections Systems:



In terms of documentation, it is interoperable with o Unqualified Dublin Core data, combined with Omeka-generated feeds and OAI-PMH harvestable data, which gives Omeka sites the ability to share data among different systems and with other Omeka sites. Its Data Output Formats includes Atom, DCMES-XML, JSON, and RSS2. In terms of data migration, items can be added individually or batch added using data migration tools, such as the OAI-PMH harvester, and CSV, EAD, and Zotero importer plugins.

We made contact with Katie Fernie and Marcin Werla who lead the development of LoCloud Collections, which is a a Best Practice Network co-funded under the CIP ICT-PSP programme of the European Commission, as well as a hosting service and publishing platform based on Omeka, designed to be used by smaller institutions.

Calculating our storage requirements for the volume of content of Europeana Food and Drink, they suggested that we could not store our content in a single Omeka collection, but had to create multiple instances and configuring each instance with the specific vocabularies necessary. This would pose an issue with findability, as well as separating our content into sections, which we deemed as not a viable solution. Also, since we wanted to use LIDO as a schema, LoCloud uses Dublin Core, which would need added mapping. Finally, there was a budgetary issue with using LoCloud, which also contributed to the factors that led us to cross out LoCloud from our options for hosting Europeana Food and Drink content.

We also looked at DSpace as an option. DSpace is also an open source solution and it is the software of choice for academic, non-profit, and commercial organizations building open digital repositories. It is free and easy to install "out of the box" and completely customizable to fit the needs of any organization.
DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets.

The main issue with DSpace is that it is best suited to academic content, focusing on bibliographical documentation and open access to research publications. Moreover, it

requires extensive modifications to be able to incorporate custom vocabularies and thesauri. Also, DSpace has not been successfully used for GLAMS collections, except with some modifications and plug ins, such as the one developed by the University of Texas called Manakin, to facilitate the showcasing of collections of visual interest, such as our Content.

Pavel Kats, Europeana's Chief Technology Officer from the Europeana Foundation, also responsible for WP2 - **Developing the Infrastructure for Europeana Cloud** made contact with us recently (February 2015). We have been in discussions since then, trying to establish what Europeana Cloud can offer suits our needs and to establish a work-flow.

Europeana Cloud is a cloud-based infrastructure which will:

- Source, prepare and add new data (2.4 million new metadata records and 5 million digital items) to the Cloud infrastructure;
- Allow content providers and aggregators to efficiently store, share and provide access to digital cultural heritage;
- Give researchers new services and tools, with which they can access, work on and share the content stored in the Cloud.
- Enable everyone to feedback enriched data back into the cloud, for use by institutions and researchers alike.

However, Europeana Cloud is currently very much at the alpha stage and will not be ready in time for our project. They are finalizing their roadmap for 2015 in March 2015 and it will contain a milestone in summer which will be internal beta and by the end of the year they'll have the external beta. They are happy to start working with us much earlier, but made it clear that it will be an experimental phase. They will be ready to commit to full service with SLA by the end of the year.

In preparation for this, we opened a Basecamp space to investigate together how the Food and Drink content partners can efficiently share content for further use by the project. Since one of the goals of the project is to experiment with how high-quality digital assets can be used by commercial applications, our first step should be making these digital assets available in a standard way for further re-use by these applications. At the moment, we are examining if Europeana Cloud can provide this standard way to store and share content.

To start doing this we need to establish how our Food and Drink content can reach Europeana Cloud. This is a function of the systems used for managing content, external interfaces they allow, data format and volume but also of the intended manner of usage (e.g. if it is going to be an one-time upload of data, a continuous integration of it and other related questions).

A set of three groups of questions was shared with all Content Partners on Basecamp: the first one about the content itself, the second one about the systems hosting it now and the third one about the work against Europeana Cloud. If we can have the partners covering all these topics, we will already have a very good idea about the effort we need to undertake in Europeana Cloud to support Food and Drink.

<u>Content Specification</u>

- Amount of digital objects

- Mime types
- Total volume (in GB)
- Internal organisation - are digital objects organised internally into thematic or other kind of collections?
- Are there any links between different objects?

System Specification

- What system stores the content today?
- What external interfaces for content it has?
- What is the preferred (technical) way of the provider to contribute content? (can be using the system's interfaces or not)

Aggregation Flow

- How does content make it to Europeana Cloud and how direct links to it are planted to metadata records?
- Does the partner envision a one-off upload of content or it will be done several times?

We are now discussing our functional and non-functional requirements.

The non-functional requirements are:

1. Performance (response time, throughput, efficient resource usage for specific performance requirements)
2. Scalability (number of organizations and users, ease of resource allocation to accommodate changing load)
3. Availability and Recoverability (ability to maintain an accepted level of performance over time, recovery from errors)
4. Data Security and Integrity
5. Usability (efficiency, documentation, ease to learn, satisfying for a target user community)
6. Interoperability (use services from and provide services to other systems for Digital Cultural Heritage)

During the March plenary meeting in Athens, Content Partners had the opportunity to ask questions and there was a discussion about the work flow for metadata and content.

During his presentation, Nikos Simou (NTUA) also presented us with an alternative for content hosting, a new platform called WITH that NTUA are developing, for storing content and metadata.

The Technical Workshop agenda covered the following:

1. Nikos Simou from NTUA gave us a live demonstration of the mapping-to-EDM tool "MINT" which we will use to deliver our metadata to Europeana.
2. Vladimir Alexiev from ONTOTEXT will take us to a journey in semantic technologies, showcasing the EFD Classification Scheme and talk about semantic enrichment.
3. Pavel Kats gave an overview of Europeana Cloud

Favoured solution:

Following on from our research and All Partner meeting in Athens, 19 – 20 March 2015, the project agreed that Europeana Cloud will not suit our purposes. Instead, we propose the solution below for uploading content and metadata for this project that can be ingested into Europeana. The proposed solution offers two routes, depending on whether the providers have their content available online[2] or not, that are highlighted in the following figures:

Route 1 is for content providers who have their content available online. These providers will either have to use the Europeana Connection Kit (ECK) or the MINT mapping tool for transforming their metadata according to the project and the Europeana requirements and publish them on Europeana.

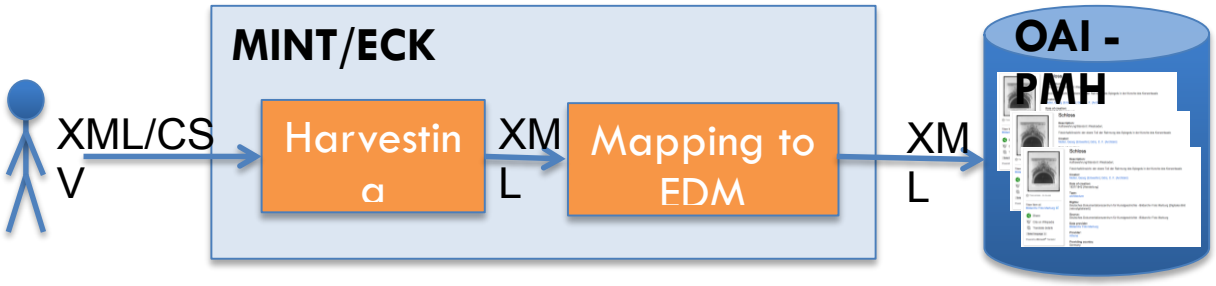


Figure 1: Providers that do not have the content available online

ECK was developed during Europeana Inside and any partners who use a CMS that has been adapted to incorporate the ECK can use this route. Partners who were also part of the Europeana Inside consortium (e.g KMKG and PS) may find this is the easiest route for them.

MINT development has started by NTUA from the ATHENA[3] ingestion server and evolved through other projects like Linked Heritage[4], EuScreen[5] and ECLAP[6]. It follows a typical web-based architecture offering an expanding set of services for metadata aggregation. It addresses the ingestion of metadata from multiple sources, the mapping of the imported records to the intermediate metadata schema and the transformation and storage of the metadata in a repository.

The main role of the MINT mapping tool in the Food and Drink project is to enable users to

- Provide metadata records in a range of “source” formats
- Convert metadata to EDM or to an EDM food and drink profile that may be developed within the project.
- Submit the records to Europeana

[2] Having the content available online is one of the main Europeana requirements for publishing on its portal.

[3] http://www.athenaeurope.org/

[4] http://www.linkedheritage.eu/

[5] http://euscreen.eu/

[6] http://www.eclap.eu/drupal/?q=en-US

Its 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.

MINT allows providers to perform mappings from their schemas to any target schema (EDM or EDM-FD) through a very user-friendly interface (see figure below).

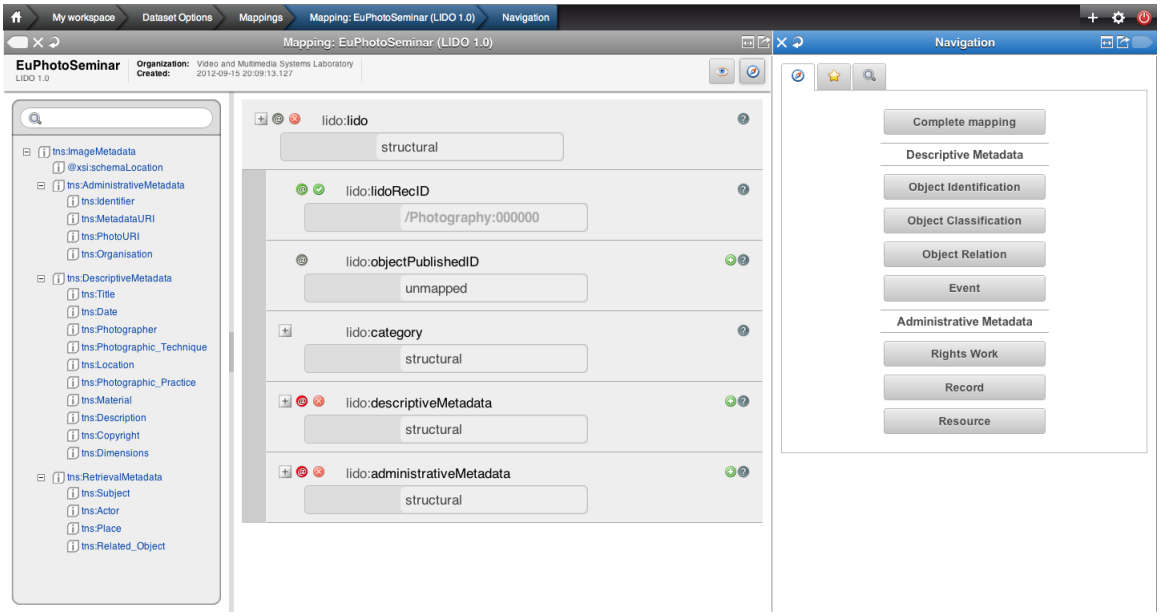


Figure 2:MINT's mapping editor

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 and also to search the tree by using the text field on the top.

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 resulting repository offers an OAI-PMH interface currently exposing the records in the Europeana Data Model. Publication to Europeana is performed by informing Europeana’s Ingestion office to harvest metadata from the NTUA’s server.

Route 2 is for content providers who do not have their content available online.

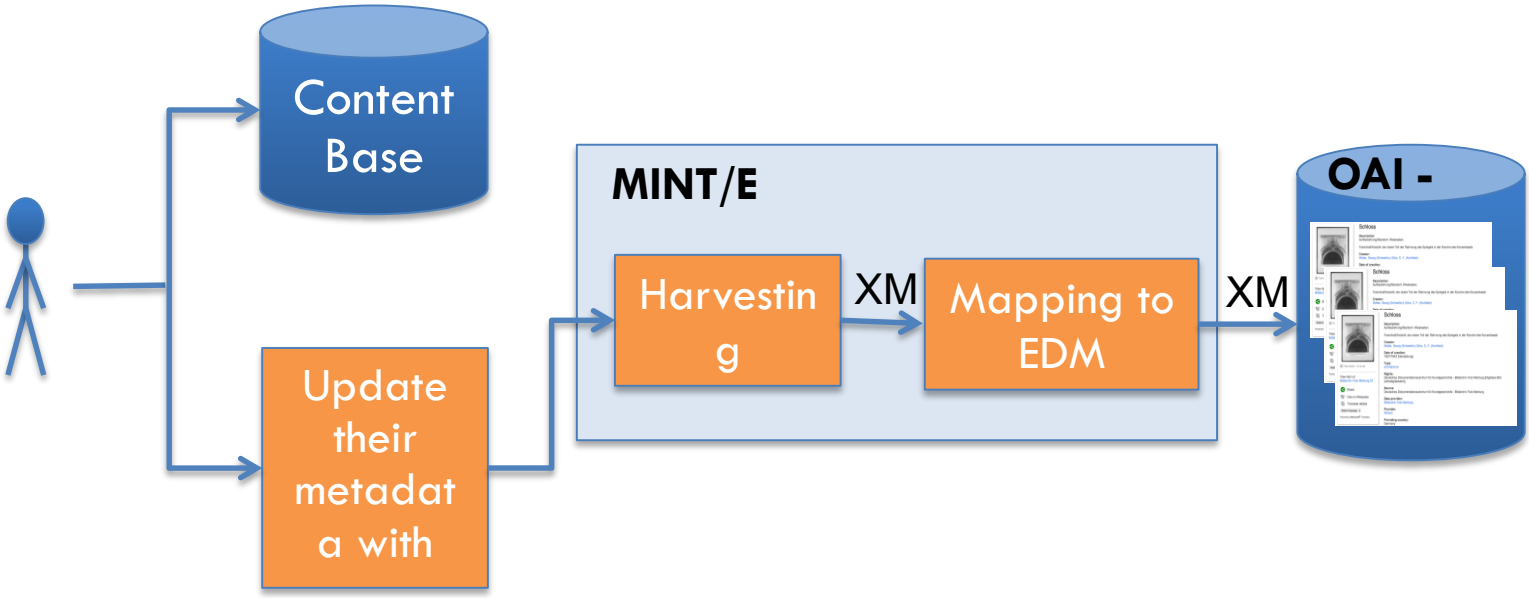


Figure 3: Providers who do not have their content available online

The main difference of this route - as it can be observed in Figure 3- is that providers will need to make an extra step before reaching the MINT or ECK tools for manipulating their metadata. This step will include the upload of their content to the Content Base repository that will be used for hosting the content. After that, they will have to appropriately modify their metadata by including the resulting links. Finally the updated metadata will be imported to one of the metadata tools for their Europeana publication as done in route 1.

Europeana Food and Drink Technical Workshop

Collections Trust is currently organizing a one day technical workshop at the beginning of May for content providers using the first route. An important stage in planning this meeting is identifying content providers who need to use MINT or need to upload their metadata to ECK. Content providing partners will bring their metadata with them and learn as they work in hands-on sessions. The intended outcomes of the workshop will be

- the familiarization of the users with the MINT enviroment,
- the correct mapping – in terms of semantic alignment - of their in-house metadata to the Europeana Data Model, and the
- the publication of their metadata to Europeana.

Next steps

During the next few weeks the metadata ingestion through MINT and ECK will start. NTUA are going to circulate instructions and the MINT instance link and partners will embark on mapping their metadata to EDM.

Partners who have previously used the ECK during other projects will be able to start work instantly and content should begin to flow.

Additionally, content providers will attend the one day technical workshop organized by CT and outlined above.