Europeana DSI 2– Access to Digital Resources of European Heritage

MILESTONE

MS1.2: TECHNICAL INFRASTRUCTURE MAINTENANCE PLAN

Revision	1.0
Date of submission	9 February 2017
Author(s)	Henning Scholz (EF), Julia Welter (DIF), Maria Teresa Natale, Marzia Piccininno, Corinne Szteinsznaider (MICHAEL), Petra Böttinger, Jörg Holetschek (FUB), Dimitris Gavrilis (AthenaRC), Kate Fernie (2Culture), Marco Rendina (eFashion), Erwin Verbruggen (NISV), Eve-Marie Oesterlen (EUscreen Foundation), Wim van Dongen (APEF), Valentina Bachi (Photocons), Vassilis Tzouvaras, Arne Stabenau (NTUA)
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Description
1	08/12/2016	Henning Scholz (EF), Dimitris Gavrilis (AthenaRC)	Table of Content, Outline & template for each chapter
2	23/01/2017	Henning Scholz (EF), Julia Welter (DIF), Maria Teresa Natale, Marzia Piccininno, Corinne Szteinsznaider (MICHAEL), Petra Böttinger, Jörg Holetschek (FUB), Dimitris Gavrilis (AthenaRC), Kate Fernie (2Culture), Marco Rendina (eFashion), Erwin Verbruggen (NISV), Eve-Marie Oesterlen (EUscreen Foundation), Wim van Dongen (APEF), Valentina Bachi (Photocons), Vassilis Tzouvaras, Arne Stabenau (NTUA)	1st draft with input from all DSI aggregating partners
3	26/01/2017	Cecile Devarenne, Antoine Isaac, David Haskiya (EF)	Review & feedback
4	03/02/2017	Henning Scholz (EF) and authors listed above under revision No. 2	Revised version incorporating reviewer´s comments.
5	09/02/2017	Laura Rauscher (ONB)	Review and proofread
6	09/02/2017	Henning Scholz (EF)	Final version incl minor changes & layout edits

Statement of originality:

This milestone contains original unpublished work except where clearly indicated otherwise.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

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Table of contents

Introduction

The DSI-2 aggregating partners need to maintain their current technical infrastructure, allowing for a continuous supply chain of cultural heritage data from their data partners to the Europeana DSI. The aggregators can then increase the number of participating data partners, processing an increasing amount of data while improving the quality of data submitted to the Europeana DSI.

The purpose of this milestone document is to specify the ingestion workflow per aggregating partner and elaborate on the maintenance plan to secure efficient provision of metadata to the Europeana DSI throughout DSI-2. It is complemented by DSI-2 MS1.1 - Ingestion workflows business requirements update, which elaborates on the business requirements for the new set of ingestion tools and services (Metis). Therefore, tools and services used by Europeana Foundation are not addressed herein.

As MINT is used by a number of DSI aggregating partners, a dedicated section follows the aggregator specific sections to specify the plans for the maintenance and development of MINT, with a special focus on the migration to a unified MINT instance (MINT4ALL).

A similar plan was already produced in June 2015, to specify the development and maintenance plans in Europeana DSI-1.[1] This milestone is following up from this document. In the context of preparing for this milestone we also started the discussion of potential future developments beyond DSI-2.

Technical infrastructure maintenance plan: key actions

In this chapter the key actions of the DSI aggregating partners are summarised to give a short and concise overview of what is planned to maintain the technical infrastructure. The sections following this chapter elaborate in more detail on these actions and give a deeper insight into the ingestion workflow per aggregator.

- European Film Gateway:
    - Adjust EFG-EDM mapping to fix current bugs
    - Adjust mappings between EFG partner archives and EFG where necessary in order to simplify the ingestion workflow and increase data quality
    - Update of all transformations to EDM to take into account bug-fixes in EFG-EDM mapping as well as the newly defined rights statements
    - Gradual implementation of the newly defined rights statements in EFG
    - Adjust EFG tools (Metadata Editor, Vocabulary Editor, etc.) to perform properly under the latest version of D-NET that EFG migrated to under DSI-1.
- MUSEU:
    - Definition and implementation of a new LIDOtoEDM mapping taking into consideration the pattern defined for how to include connections to IIIF resources in LIDO
    - Operations Direct: performing a test before the end of DSI-2 through the creation of a JSON file of a small batch of records in order to draft a smooth workflow
    - Evaluation of the human and economic efforts necessary to assure that the transition toward MINT4ALL happens as smoothly as possible. This includes the evaluation of effort to keep and transfer old mappings from the previous MINT

[1] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms9-technical-infrastructure-development-plan.pdf

instances used as well as the vocabularies and the connected OAI-PMH server that assures the storage of the records.

- OpenUp!:
    - Finalizing technical improvements by adapting the OpenUp! harvesting and enrichment processes to the new BioCASe pipeline using B-HIT, with focus on working with large quantities (batch processing), to assure persistent high-level performance of the ingestion process.
    - Performing test phase for LIDO data processing, using 2 different sources within and beyond the OpenUp! domain network to enter a prototype that will provide data to Europeana; LIDO to EDM conversion, evaluation of future acquisition of high quality multimedia content.
    - Consolidate and improving metadata enrichment with vernacular names by the OpenUp! web service for Common Names as one of the value-adding services by setting up a monitoring framework to assure permanent accessibility.
- CARARE:
    - Regular maintenance and monitoring of the aggregation infrastructure and workflow;
    - Maintaining and developing metadata schema mappings from standard schemas supported by CARARE to EDM;
    - Performing testing of Europeana Shared Services.
- Europeana Fashion:
    - supervision of the migration process towards the new aggregation infrastructure MINT4ALL, including functionalities check and migration of datasets and related transformations;
    - regular maintenance and monitoring of the digital repository, the staging portal and the live portal with the related AWS services;
    - update of all the transformations to EDM, to take into account the newly defined rights statements;
    - migration of the live portal frontend to the Europeana Collection website (in collaboration with Europeana).
- APEF:
    - Evaluate current apeEAD to EDM conversion functionality from the perspective to overcome existing technical hurdles for forwarding content to Europeana
    - Start testing/piloting with the Europeana METIS system
    - Implement the new rights statements in the apeEAD to EDM conversion functionality
    - Implement forwarding information on topics/themes to Europeana via the apeEAD to EDM conversion functionality
    - Implement forwarding information on whether a finding aid contains descriptive units with links to digital objects via the API services
- Photoconsortium:
    - follow up the transition of the EP MINT (customized version of MINT according to the requirements of Europeana Photography metadata schema) to MINT4ALL, including the Multilingual mapping tool with SKOS vocabulary about photography in 16 languages

Technical infrastructure maintenance plan per partner

European Film Gateway

Ingestion workflow overview

The EFG workflow is divided into a number of steps (Fig. 1).

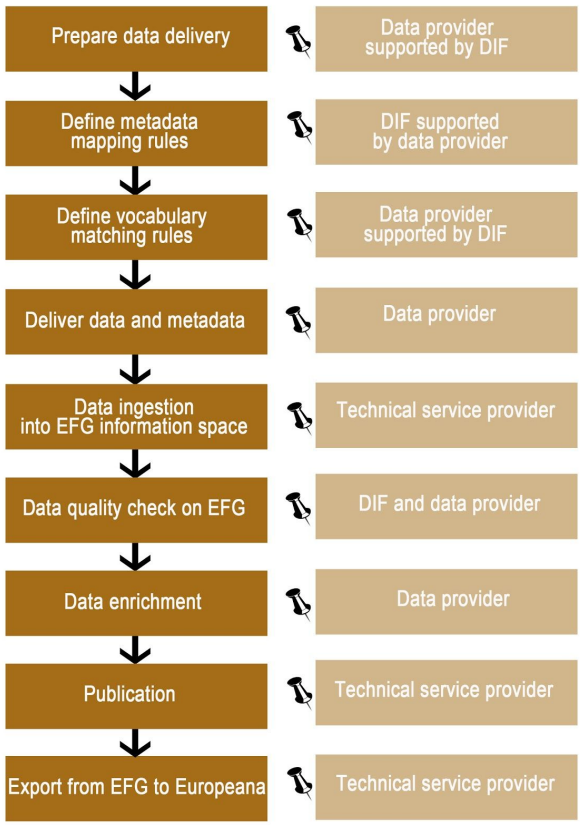

Figure 1. The EFG ingestion workflow (left: task, right: responsibility).

EFG partners provide metadata in a local XML export structure which are then mapped to the EFG schema. The mapping process consists of two steps: a) DIF and the archive establishing a conceptual mapping on which b) the technical provider builds an import filter. The same applies for the matching process. The provider and DIF match the local values to the common EFG vocabularies. Local values of the archive are then transformed into EFG values during the data ingest. Data enrichment can be carried out either on local level directly in the partners' databases (preferred) or else directly in the EFG database with the help of the Metadata Editor tool. After the ingestion of data, DIF and content provider review the data in the EFG Content Checker and as a preview in the EFG portal test environment. After the revision process, the data is published to the EFG live portal and transformed into EDM for Europeana to harvest. Further information on the EFG ingestion workflow is available in the EFG Data Provider Handbook (p. 7-10). The EFG infrastructure runs on D-NET with services adapted for the specific needs of EFG. The D-NET

Aggregator Service (AS) supports the harvesting of external repository units as well as the cleaning and enriching of data during the ingestion process.[2]

Technical infrastructure maintenance

The maintenance work of the EFG infrastructure focuses mainly on the mappings established to bring in data from EFG partner archives into the EFG repository. DIF together with the technical subcontractor are working to improve mappings (partly still originating from the early days of EFG in 2009) so that they allow for the best possible and most loss-less way to transfer data from the local databases to Europeana. These mapping revisions often help to improve not only data quality in Europeana but also make the whole ingestion process more flawless and therefore less time-consuming. Additionally, EFG will work gradually implement the new Europeana rights statements and update all transformations to EDM in order to take into account the revised EFG-EDM mapping as well as the updated rights statements. As the EFG repository was updated to a new version of D-Net under DSI-1 already, there is still some bug-fixing to be done now in order to ensure all data sets are available correctly in the updated environment. Also, EFG tools such as the Metadata Editor or the Vocabulary Editor still need to undergo some adjustments in order to perform properly. Apart from this, the EFG repository runs stable so that work of the technical service provider can focus on new data ingestions.

Towards a shared technical infrastructure for Europeana DSI

In order to better understand similarities and differences of the EFG and Europeana workflows and to identify the prerequisites for potentially moving EFG into the shared infrastructure, DIF staff met with the Europeana Data Partner Services team for two days to go over the respective workflows in detail. One of the aims of the meeting was to clarify what services Metis can offer that EFG is currently making use of. Two of the main challenges identified are related to the high number of individual mappings applied in EFG that are written in Perl rather than XSLT which will be work-intensive to transfer to Metis and the fact that Metis does not offer a vocabulary matching tool that EFG is highly dependent on if data from partner archives are supposed to be displayed in a harmonised way in Europeana. Following the discussions between DIF and EF staff, one of the next steps will be to estimate the costs and personnel that would be necessary to move the EFG infrastructure to the shared Europeana infrastructure.

[2] D-NET aggregation software source code: https://zenodo.org/record/31693#.WJNF8rbNzUK and information on licensing status: http://www.d-net.research-infrastructures.eu/?q=node/20

MUSEU

Ingestion workflow overview

The overall MUSEU ingestion workflow starts with the evaluation of the content that the providers wish to publish in Europeana. Each data partner is requested to describe their collections in terms of quantity and quality of records and of technical parameters (for instance, to check if there are obstacles to the provision of direct links to media files). New providers are also demanded to sign the DEA before starting the aggregation process.

The MUSEU data partners follow six steps to have their records published in Europeana (Fig. 2).

MUSEU data partners use a customized MINT platform to upload records. 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.

The whole MINT functionality is described below in the paragraph “Scope of MINT”.

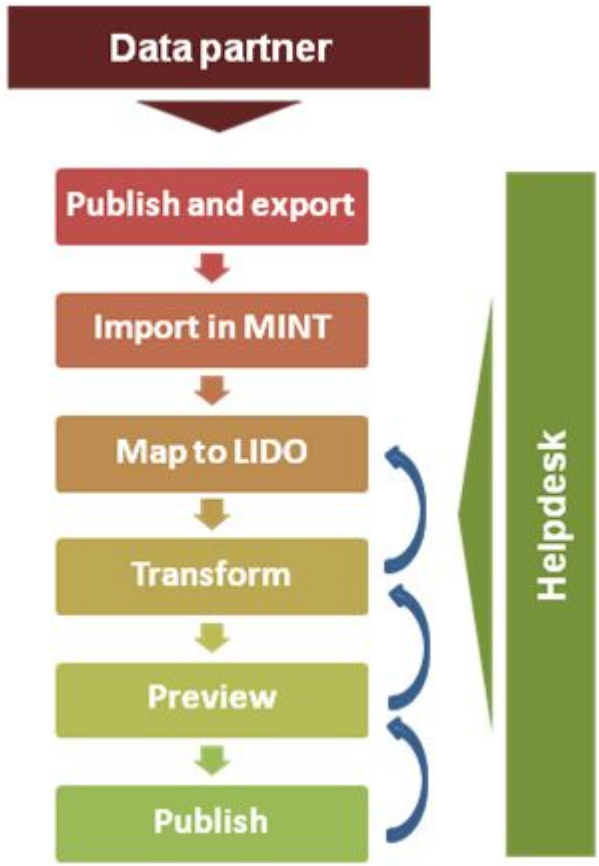

Figure 2. The MUSEU ingestion workflow.

A data partner performs an export of his metadata from his own content management system. Then this set of records is imported into MINT/MUSEU in either CSV and XML formats or using an Excel file. MINT supports every metadata source formats. After that using the user-friendly interface of MINT the provider performs a mapping to LIDO that acts as the intermediate schema towards EDM. LIDO preserves the complex richness of the museum records.
Once this is done the data partner can check the validity of the mapping through a set of preview interfaces that give him inline schema validation and also a view of what his records would look like when published on Europeana. Once these checks are performed he transforms his metadata to LIDO creating a new dataset. After that the MINT tool performs the transformation of LIDO records into EDM. Again this process results into a new EDM dataset and once this is ready the data partner can send his metadata to NTUA's OAI-PMH server.

This process ends with the communication to Europeana from the MUSEU helpdesk to Europeana of what's ready for the harvesting and, in case of updating, which old datasets must be deleted from Europeana to avoid duplicates (Fig. 3). Mapping is the activity that requires the major support.

Figure 3. Relation diagram of the MUSEU stakeholders.

A helpdesk service supports the data partners in each step of the aggregation workflow.

MINT supports SKOS vocabularies by allowing specific target schema elements to take values only from them. The vocabularies used are: ISO 639-1:2002 standard for representation of names of language and metadata language; AAT for object work type, cultural heritage type, cultural context, period, and materials/technique; PICO for cultural context and period; UDC summary for thematic context; TGN for location and place; ULAN for actor.

The documentation about the workflow is being collected at this address: http://www.museuhub.eu/ (website under construction).

Technical infrastructure maintenance

The MINT/MUSEU infrastructure was built during the project Europeana DSI. Therefore, during DSI-2 the level of maintenance is low because the necessary updates were already performed, including the alignment of LIDO to the latest version of EDM and the 'Requirements for Visualization and Indexing of Museum Content' issued during the AthenaPlus project.

The partner NTUA assures, as for all other DSI-2 MINT users, the maintenance of the technical infrastructure of MUSEU and the OAI-PMH server. Currently NTUA is taking care of the alignment of this MINT instance to the new rights values Europeana promoted, that the data partners find in a drop-down menu in the tool.

One part of the maintenance work of the MUSEU infrastructure is the continued updating of the LIDO to EDM mapping together with the correspondent implementation into the stylesheet used in MINT to transform LIDO into EDM records in order to reflect changes and developments of Europeana's EDM implementation. An important step is to support the ingestion of connections

with images published according to the International Image Interoperability Framework (IIIF). Therefore, a pattern was defined for how to include such connections to IIIF resources in LIDO, which relates to Europeana's EDM profile for IIIF[3]. It is now easily possible for data partners to submit IIIF resources from their own services through LIDO similarly than other web resources, just by qualifying them in a standardized way. Having integrated this pattern into the MINT customization for MUSEU, the required representation of IIIF resources in EDM is generated from the LIDO records. A first dataset including IIIF resources was already harvested by Europeana.

MUSEU will switch to the use of the MINT4ALL shared infrastructure along next summer. The integration with the Metis services will happen then in this new environment as it is described below in the paragraph dedicated to MINT4ALL. What will be done before the end of DSI-2 is the evaluation of the human and economic efforts necessary to assure that this transition will happen as smoothly as possible. This includes the evaluation of effort to keep and transfer old mappings from the previous MINT instances used (ATHENA, Linked Heritage, AthenaPlus, MUSEU) as well as the vocabularies and the connected OAI-PMH server that assures the storage of the records.

As to Operations direct, MUSEU has planned to perform a test before the end of DSI-2 through the creation of a JSON file of a small batch of records in order to draft a smooth workflow and start training the partners.

[3] http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_profiles/IIIFtoEDM_profile_042016.pdf.

OpenUp!

Ingestion workflow overview

The OpenUp! project has created a workflow for harvesting community standard data, their conversion, enrichment and provision as EDM to the Europeana DSI (Fig. 4). The data provision process has been continuously improved to supply metadata offering the highest quality possible. Since 2016, this aggregation workflow of OpenUp! has been adapted and opened up to new standards and protocols. The heterogeneous databases of various providers can now be imported not only via BioCASe, the Biological Collections Access Service for Europe, but also by using the Integrated Publishing Toolkit (IPT) of the Global Biodiversity Information Facility. (BioCASe see: http://www.biocase.org/products/; GBIF see: http://www.gbif.org/ipt).

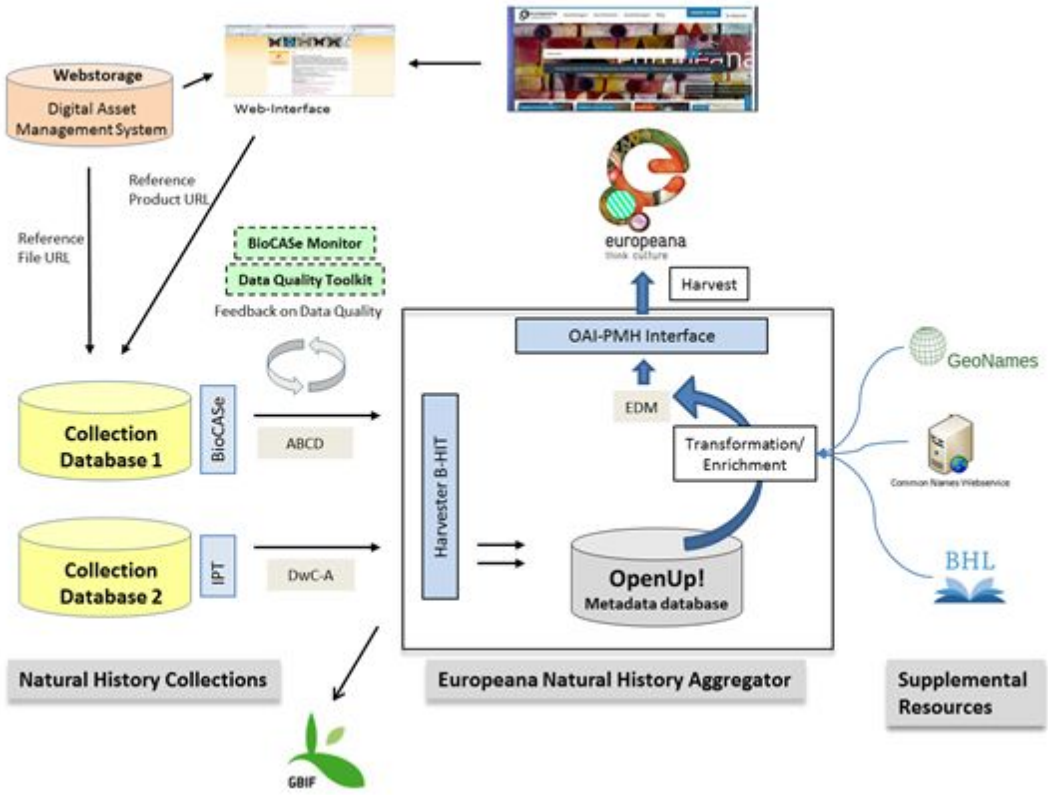

Figure 4. Aggregation Workflow of OpenUp!.

The source data are harvested from the BioCASe or IPT providers using the Berlin Harvesting and Indexing Toolkit (B-HIT). B-HIT stores metadata as XML files (extensible mark-up language) for BioCASe data sources and CSV files (comma separated values) for IPT providers or custom-made DarwinCore archives. In a first step, these files are parsed into a relational metadata database. This database is not bound to a specific data standard; it can accommodate both ABCD and Darwin Core data. In particular, it contains the URLs for the multimedia objects that Europeana is centred around.

For data transformation, OpenUp! uses the open source business intelligence tool Pentaho that offers extraction, transformation and loading (ETL) capabilities. The Mapping Tool (Pentaho Kettle – Transformation) receives ABCD and DarwinCore records and processes them to EDM

(Europeana Data Model). In a subsequent step, the natural history metadata is enriched with information drawn from supplemental resources. By now, the Natural History Aggregator uses information of three kinds to augment the original collection data:

1. GeoNames, a geographic thesaurus database available through web services,
2. The OpenUp! vocabulary webservice for common names, currently offering 635,020 common names for 259,300 species from 30 different taxonomic thesauri, and
3. The Biodiversity Heritage Library (BHL), providing open access to 183,385 volumes of biodiversity literature.

After processing, both the original specimen metadata and the enrichments are stored as EDM documents and are published through an OAI-PMH server (Open Archive Initiative, Protocol for Metadata Harvesting), ready for the periodic harvest by Europeana. In addition, all records available on the portal can be retrieved through the web services defined in the Europeana API. This allows a wider dissemination of the data; they can be used by other portals or by apps on mobile devices.

Technical infrastructure maintenance

FUB as coordinator, together with the subcontractors, maintain the OpenUp! Network (tools, community, business model). The OpenUp! project created a software suite and workflow for the harvesting of community standard data, providing it to the Europeana DSI. The *OpenUp! Natural History Aggregator* is contained in a Virtual Machine (VM) that can be operated in different places. A first mirror has already been installed at the BGBM. The aggregation platform is operated by a subcontractor; outsourcing has proven to be the most cost-effective option. The aggregation platform under Europeana DSI is continued using the OpenUp! sustainability model, i.e. the maintenance is financed by the data partners under an SLA with the operator of the platform as a partner.

To keep up with increasing amounts of data as well as increasing complexity, OpenUp! works on putting in place automated and simple workflows to steadily update transformation of data. The subcontractor tests switching from harvesting the data partner's BioCASe web service record-by-record to harvesting the aggregated standard dump now provided by the BioCASe Provider Software, and further automates the OpenUp! Natural History Aggregator with respect to metadata ingestion, transformation and enrichment workflow.

OpenUp! is investigating the consequences of upscaling the workflow to face and handle a steadily growing numbers of data partners as well as the number of datasets to be transformed. In collaboration with subcontractor A the OpenUp! Natural History Aggregator will be further automated to allow batch processing of multiple sources with B-HIT and increase the efficiency of metadata enhancements in collaboration with subcontractor B.

The data quality is significantly improved by adding additional data sources to the metadata enrichment process, improving public accessibility and ensuring re-usability of data (e.g. by checking for valid licenses).

CARARE

Overview

The CARARE technical maintenance involves three main tasks: the maintenance of the aggregation technical infrastructure, the mappings, the operations & communication. These three types of tasks require different and specialized skills such as: systems administration, software development, metadata design & mappings, operations & support.

Ingest workflows

The CARARE ingest workflow is divided into a series of activities which begin with preparation of the data and a series of decisions taken by the data partners. The major decisions concern the method by which metadata will be delivered for harvesting, which influences the format in which it is exported and whether MINT is used for metadata mapping. Figure 5 illustrates the initial steps and the responsibilities.

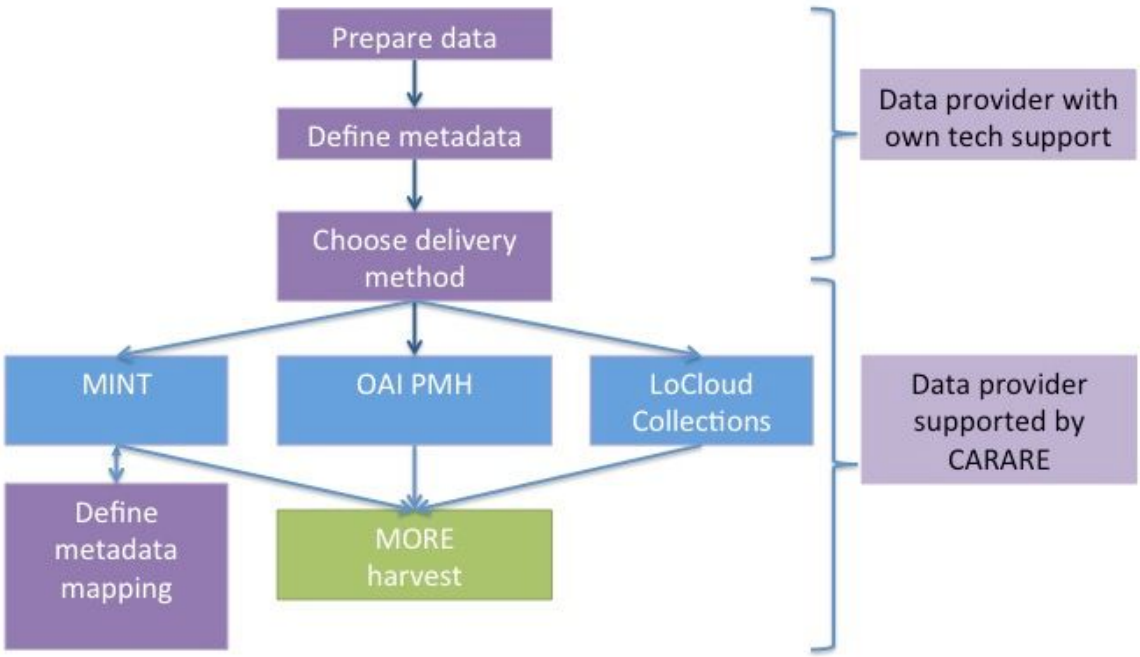

Figure 5. Initial workflow for CARARE data partners

Once the metadata has been prepared for harvesting (via OAI-PMH, MINT or LoCloud Collections) the main CARARE aggregation workflow seen in figure 2 below which involves the following backbone: Harvest → Ingest → Transform → Enrich → Publish.

Of course it is possible to delete, reject and withdraw content at specific steps (shown in Figure 6).

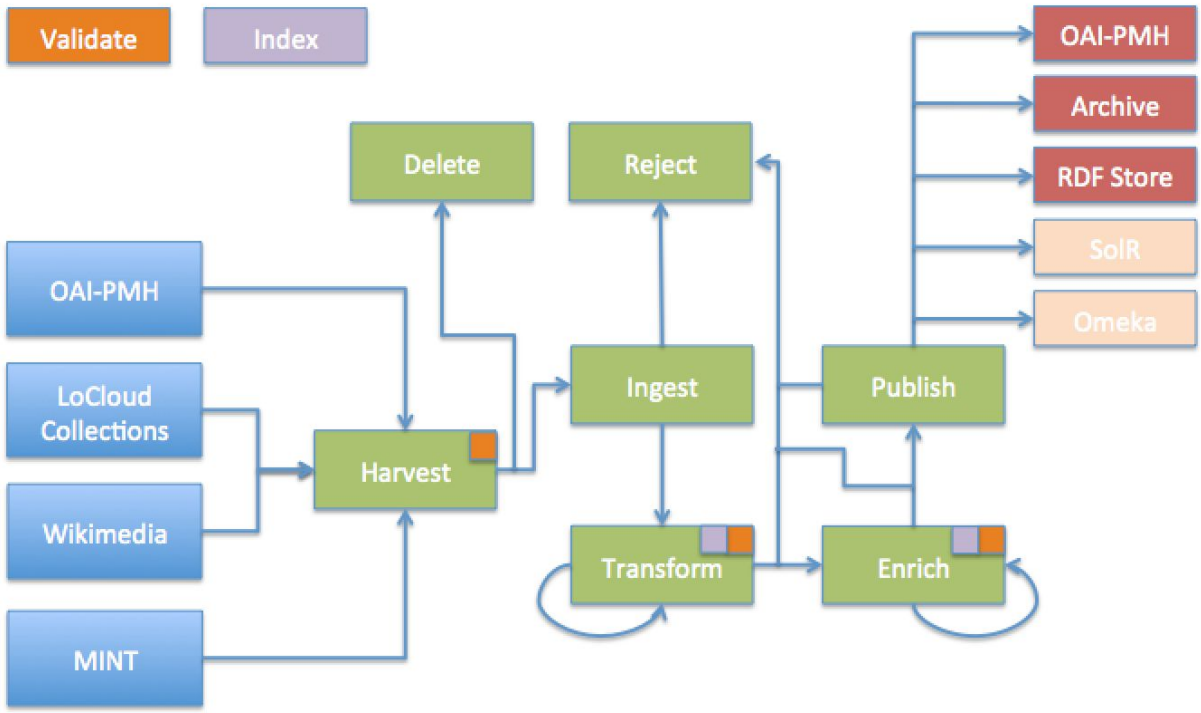

Figure 6. Main workflow used in CARARE

The workflow is not static as it is possible to perform multiple iterations on the main tasks, especially in the transform and enrich. Also, there are tasks that are auto-triggered and are not controlled by the user. These are the validation, indexing and quality measurement and are there to ensure integrity, quality and robustness. The data flow through the system through actions that are triggered and managed by the aggregation manager (a non-IT person).

Ingest tools

MINT is used by some CARARE data partners for metadata mapping. The ingest tool used in CARARE is actually the MORe aggregator, a scalable micro-service oriented system that employs a series of core-services and micro-services in order to accomplish the task in hand. The core-services are the following:

- Harvest
- Ingest
- Transform - this service uses metadata mappings from the CARARE, LIDO, DC and Omeka XML schemas to EDM which are produced by CARARE (Athena RC with 2Culture)
- Enrich
- Publish
- Withdraw
- Validate
- Index
- Delete
- Reject

The above core-services are orchestrated by MORe itself, are accessible through the MORE UI and their operation is mainly dictated by the main workflow.

Furthermore, some of the above core-services are responsible for managing a series of micro-services. These can be seen in the table below:

Core-service	Micro-services
Harvest	OAI-PMH
	File Upload
	Omeka
	MINT
	Wikimedia
Validation	Structural
	Well formedness
	Broken links
	Schematron rules
Enrichment	Language identification
	Geo-coding
	Vocabulary enrichment
	Reverse geo-coding
	Historic place names
	Thesauri mappings
	Perio.do
	Geo-normalization
	Date-normalization
	...
Publish	OAI-PMH
	Elastic search
	Archive
	Virtuoso RDF store
	Sesame RDF store

Operations

In terms of day to day operations, the MORe aggregator requires an aggregation manager (usually a non-IT person) that is responsible for aggregating the data, inspecting, enriching and monitoring quality.

Apart from operating MORe, there are cases where data are encoded in a new format or in an existing format that has changed, XSLT based mappings need to be designed and developed or revised. This requires either an information scientist or an IT specialist that will implement the XSLT mappings. These cases are not rare, instead they occur at least on a monthly basis.

Even when a system is stable for years (like with the case of MORe), the different services themselves require maintenance. This maintenance involves the following tasks:

Maintenance category	Description	Frequency
Debugging	This is quite common and with the size of MORe (in terms of core and micro services).	Monthly
Service improvement	This is also quite common and with the number of MORe services. The improvements mostly have to do with bugs and modifications to the external services MORe uses.	Weekly
New services	Development of new services.	twice a year
Schema aware development	Making services to schema aware (for new schemas)	twice a year
Schema aware maintenance	Adapting services to changes to existing schemas	every three months

Europeana Fashion

Ingestion workflow overview

The actual aggregator infrastructure of Europeana Fashion is based on MINT, a metadata aggregation platform developed by NTUA.

Figure 7 summarises the Europeana Fashion ingestion and publication workflow.

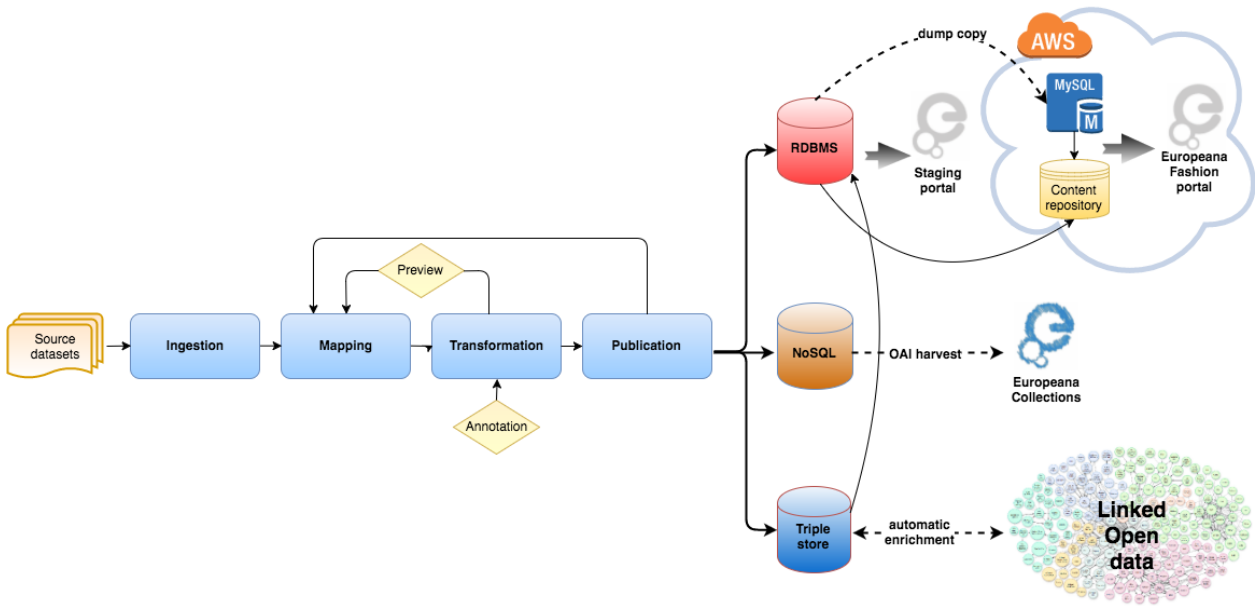

Figure 7. The Europeana Fashion ingestion and publication workflow.

As shown in Figure 7, a Europeana Fashion data partner can import his metadata into MINT that allows him to perform a mapping to the EDM-Fashion profile (EDM-FP) through a user-friendly mapping interface. In addition, different modules are implemented into MINT providing statistics about the imported metadata, the resolving XSLT and also previews of how the records would look like when published on the Europeana Fashion and the Europeana Collections portals.

An important module that has been implemented and used within the Europeana Fashion aggregator is the annotator. This module gives the opportunity to either annotate/curate elements with information that was missing from providers' original datasets (single item annotator) or to group records based on specific values and to annotate them accordingly (group annotator) - a process that can be repeated more than once.

The module that mostly differentiated Europeana Fashion from all the other aggregation projects is its publication module that publishes the EDM-FP metadata to three different back ends that are used for different purposes. In detail the EDM-FP metadata are first published to a Relational Database Management System (RDBMS) that acts as a backend for the Fashion portal. The records are published in a staging portal, where providers can check how all their published datasets look like. On a monthly basis the database and the index on the staging portal are copied to the Europeana Fashion live server (www.europeanafashion.eu), hosted on an Amazon cloud infrastructure, where also the content repository is hosted. Once this is done, each of the

published records achieves a unique identifier on the Europeana Fashion portal that is used as the edm:isShownAt value in the EDM-FP record.

After this step, the publication to the NoSQL backend - on top of which the OAI server has been set-up - begins, enabling the Europeana harvesting and publication in EDM format.

Because Europeana Fashion applies automatic enrichment to its metadata, the third step in the publication process is performed to a triplestore from where semantic algorithms are employed, normalizing given values and also enriching metadata by adding entirely new vocabulary values. Apart from the semantic enrichment, the actual content (i.e. the images) has been processed by image analysis algorithms that automatically extracted colour characteristics. These enrichments are currently saved to the RBDMS and displayed on the Europeana Fashion portal.

Technical infrastructure maintenance

The maintenance of the MINT infrastructure is performed by NTUA (see chapter below), while the Europeana Fashion International Association is taking care of the monitoring and maintenance of the staging portal, and of the content repository and live portal that are hosted on Amazon cloud services (AWS).

The staging version of the Europeana Fashion Portal remains hosted together with the MINT platform in the NTUA server farm, to ensure that the network will have a minimal effect in the interoperation of the two systems, while the live server and the content repository are hosted on Amazon AWS using the following services:

- **Amazon EC2**: Amazon Elastic Compute Cloud is a web service that provides resizable computing capacity in the cloud. It is designed to make web-scale computing easier for developers.
- **Amazon S3**: Amazon S3 provides a simple web-services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, secure, fast, inexpensive infrastructure that Amazon uses to run its own global network of web sites.
- **Amazon RDS**: Amazon Relational Database Service is a web service that makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while managing time-consuming database administration tasks.
- **Amazon CloudFront**: CloudFront is a content delivery web service. It integrates with other Amazon Web Services to provide an easy way to distribute content to end users with low latency, high data transfer speeds, and no commitments.
- **Amazon ElasticCache**: ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud. The service improves the performance of web applications by allowing you to retrieve information from fast, managed, in-memory caches, instead of relying entirely on slower disk-based databases.

The eFashion team is constantly monitoring the two versions of the portal to ensure their smooth operation. Updates are being applied for bug fixes, and both the software and the hardware are monitored using various tools. When bugs on the portal are reported, they are fixed and deployed initially on the staging server for validation, subsequently leading to the update of the live server with the successful fixes.

Periodically, snapshots are taken from the database of the staging server and replicated on the live server. This is scheduled to happen once per month in the condition that new content is

delivered but in some cases it had to be scheduled more frequently depending on the circumstances. These snapshots are related to both the database and the search index.

The caching mechanism is executed and monitored to ensure that the images are cached on amazon S3 and thumbnails are generated. In case that there are broken links or problematic images, they are removed from the database in order to increase the quality of the delivered experience to the end user, and these broken links are reported to the content ingestion team that try to fix them in collaboration with data partners.

Finally, data are exported from the portal, transformed into EDM and delivered to Europeana through the OAI-PMH repository that is hosted by NTUA. This process happens upon request depending on the publication schedule of Europeana.

Beside all of this, eFashion is also collaborating with the Europeana development team in order to migrate the Europeana Fashion portal to the Europeana Thematic Collections infrastructure. The migration will be completed by March 2017, as far as concern the portal front-end. The migration of the content repository will happen when the Europeana DSI will offer such a service.

EUscreen

Ingestion workflow overview

EUscreen uses NTUA's MINT aggregation platform to ingest its comprehensive body of audiovisual metadata and a specific AV toolset for managing media content, effectively making its audiovisual collections accessible through its own portal (euscreen.eu) and Europeana Collections.

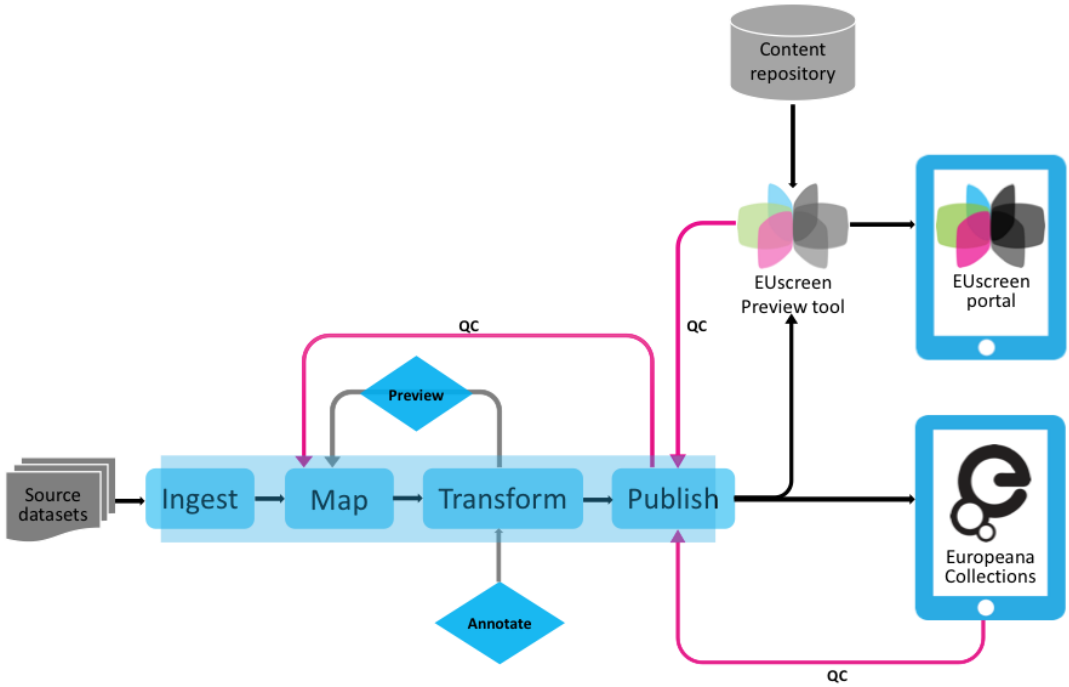

Figure 8. EUscreen's twin-track aggregation workflow.

As Figure 8 illustrates, EUscreen's technical infrastructure has been set up to enable a twin-track approach to metadata aggregation and ingestion of media content. A customised version of MINT [4] allows data partners with sufficient metadata and digitised audiovisual content to ingest, map, preview and transform their data for direct publication on Europeana; it also gives partners the possibility to manually enrich the quality of their source metadata using the MINT annotator functionality. The AV toolset devised by Noterik B.V. enables partners publishing content on the EUscreen portal to upload and safely stream their media files via EUscreen's streaming server and to preview and manage their content via the EUscreen Preview tool (see also Fig. 9).[5]

[4] See http://mint-projects.image.ntua.gr/euscreenxl/Login_input.action
[5] See http://preview.euscreen.eu

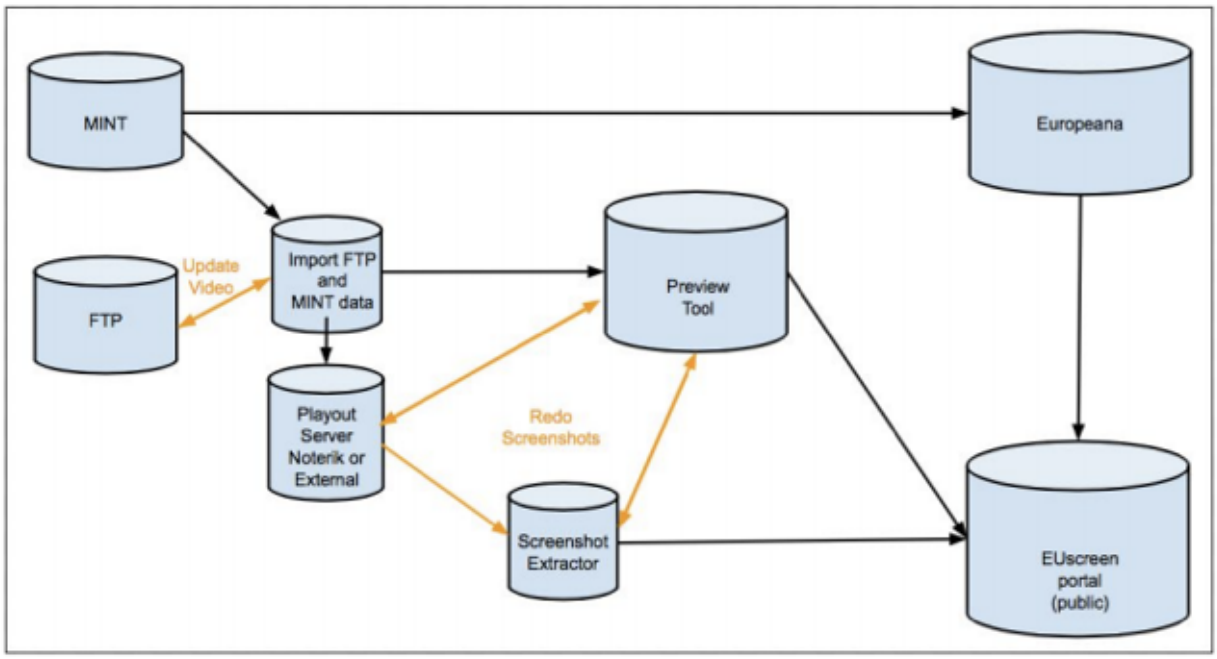

Figure 9. Expanded workflow via EUscreen's preview tool.

A data manager has been subcontracted to work closely with data partners throughout the aggregation process: this ranges from support with the ingestion of new/updated content and openly accessible data, to the provision of closed feedback loops throughout the mapping and publishing process, something that not only encourages partners to provide curated and enriched high-quality data but also fosters active involvement in the AV expert hub community. A variety of resources (handbooks and online tutorials) has been created and are constantly being updated to respond to changes in standards and to reflect data partners' needs. A training and networking event has been set up to consolidate sharing of best practice and expertise in the wider AV domain.

Technical infrastructure maintenance

The EUscreen infrastructure maintenance and upkeep are continuously monitored to keep up the levels of security desired by the content owners, the playback of video files over a wide range of browsers and platforms, and the correctness of data across the entire ecosystem. Bug tracking takes place in Mantis, a bug tracking platform provided by Noterik and shared with NTUA. In bi-weekly meetings between the technical teams at Noterik and NTUA, the data manager and DSI project lead for EUscreen, technical improvements are tackled in an agile manner.

APEF

Ingestion workflow overview

The Archives Portal Europe offers its data partners an account on the back-end of the portal – the Dashboard – via which they can upload, process and publish their data themselves and in 'real time'. The Dashboard offers a variety of possibilities to perform the necessary steps to get the data online.[6]

The most basic approach is to upload XML files (one XML file or a zipfile containing more than one XML file) manually from a computer or network hard drive into the Dashboard (see Fig. 10) and then subsequently perform all basic processing steps manually: converting the file(s) to the APE formats (i.e. apeEAD/XML, a subset of the official EAD2002/XML), validating them against the schema's of the APE formats and indexing and publishing them. These actions can be activated and monitored in the most important screen of the Dashboard, the Content manager screen (see Fig. 11).

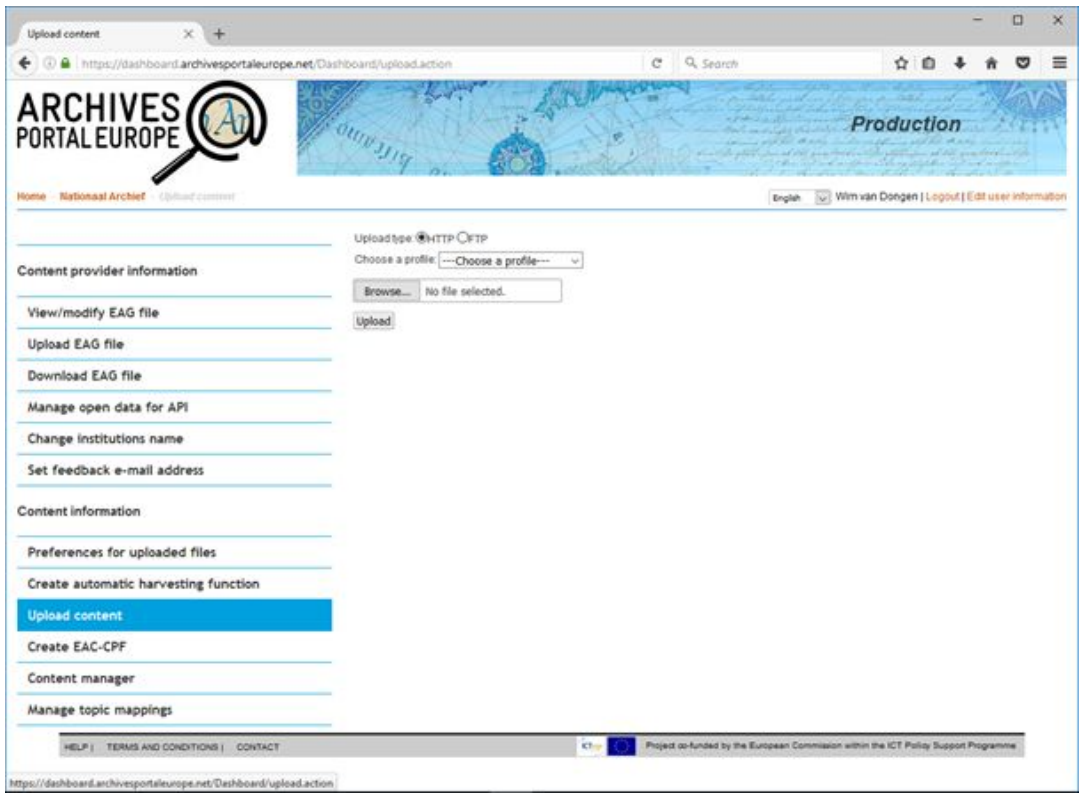

Figure 10. APE Dashboard: uploading content in the Archives Portal Europe via HTTP (or FTP).

[6] The source code of the Dashboard will in due time be available on APEF's Gitub account: https://github.com/archivesportaleuropefoundation; over there the source code of the local version of the Dashboard, the Data Preparation Tool (DPT), can already be downloaded. The DPT is available under a European Union Public License.
About the DPT itself, see: http://wiki.archivesportaleurope.net/index.php/Category:Tools_DPT_manual

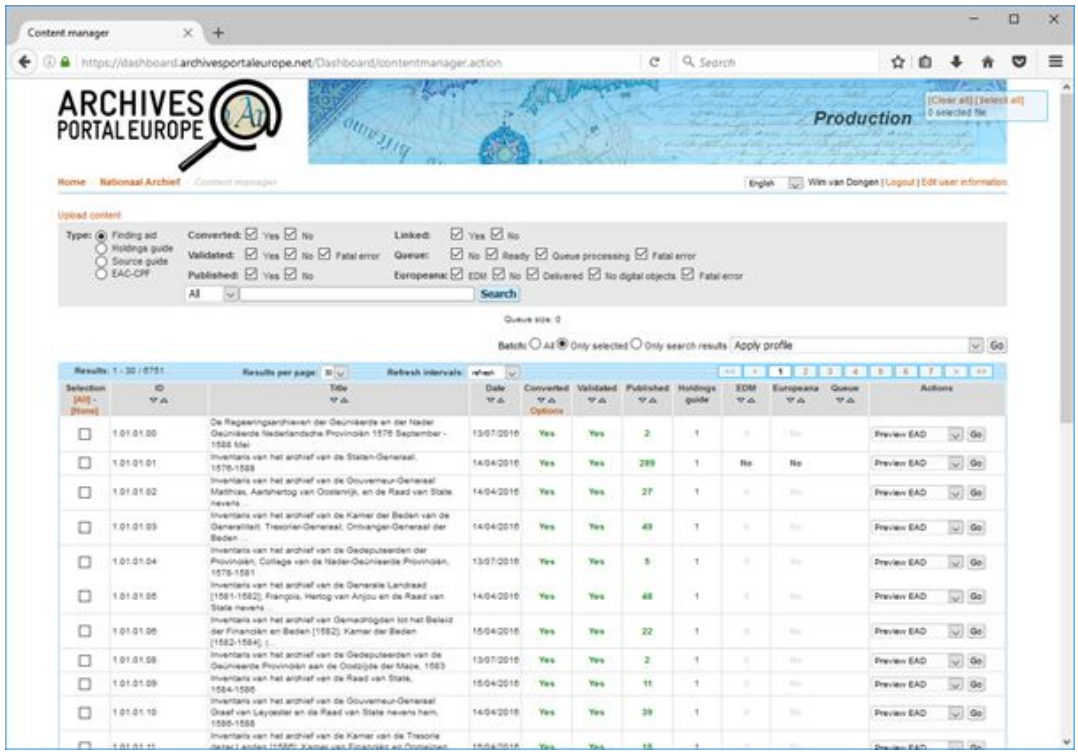

Figure 11. APE Dashboard: content manager screen showing all available XML files and their status.

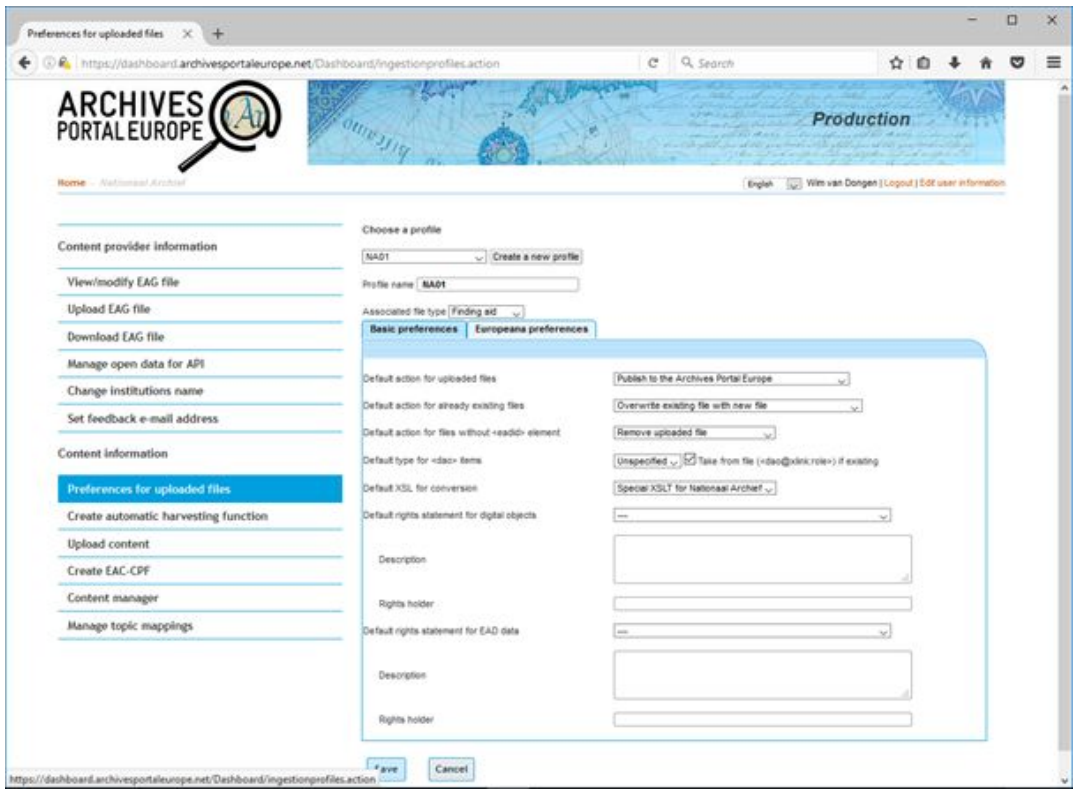

Figure 12. APE Dashboard: define a processing profile.

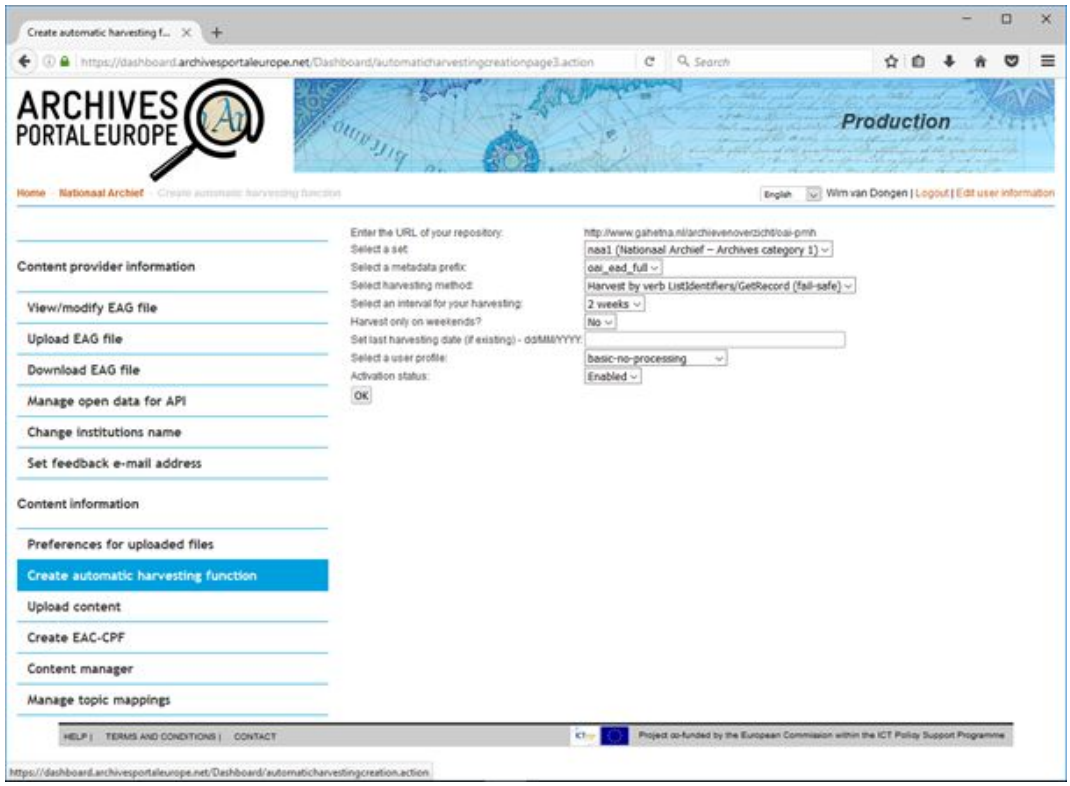

Figure 13. APE Dashboard: configure automatic connection with an OAI-PMH repository.

The most efficient approach is to define a profile for instructing the Dashboard framework to perform all basic processing actions (converting, validating, indexing and publishing) automatically after the XML files have entered the Dashboard (see Fig. 12) and to connect that to an OAI-PMH repository, which also can be done automatically (see Fig. 13).

After defining a processing profile and connecting that to an OAI-PMH repository, the whole harvesting and processing process will be performed fully automatic at regular intervals (minimum interval period: two weeks), only fetching and processing changed or new XML files (and deleting obsolete ones if necessary). The Dashboard will automatically inform the data partner about the results of each of these harvesting and processing rounds, so he/she can relax and only has to take action in case the email notification mentions problems, which rarely happens. The Dashboard keeps track of the harvesting/processing schedule (see Fig. 14), so the data partner can also always check the status of the harvesting/processing rounds in the Dashboard.

The (meta)data that are published via the Archives Portal Europe are mainly finding aids, which are overviews of the descriptions of archival material that is kept in archival institutions throughout Europe. Since all archival institutions still only have managed to digitise a small percentage of all their archival material, not all individual descriptive units of these finding aids have links to digital objects (scans of the archival material itself). So only a part of the content that each archival institution has published via the Archives Portal Europe can be forwarded to Europeana.

The Dashboard contains information on which finding aids have links to digital objects, so on which information from the finding aids can be converted into EDM records to be forwarded to Europeana. This information is available in the columns “EDM” and “Europeana” in the Content manager screen (see Fig. 11 above). In case the information in these columns is respectively “0” and “No” in grey, that means that there are no links to digital objects available in that particular finding aids. However if the information in these columns says respectively “No” and “No” in black (like in Fig. 11 for finding aid number 1.01.01.01), then that means that there are links to digital

objects in that specific finding aid, but “No” that information is not converted into EDM yet and “No” there are no EDM records forwarded to Europeana for this finding aid yet.

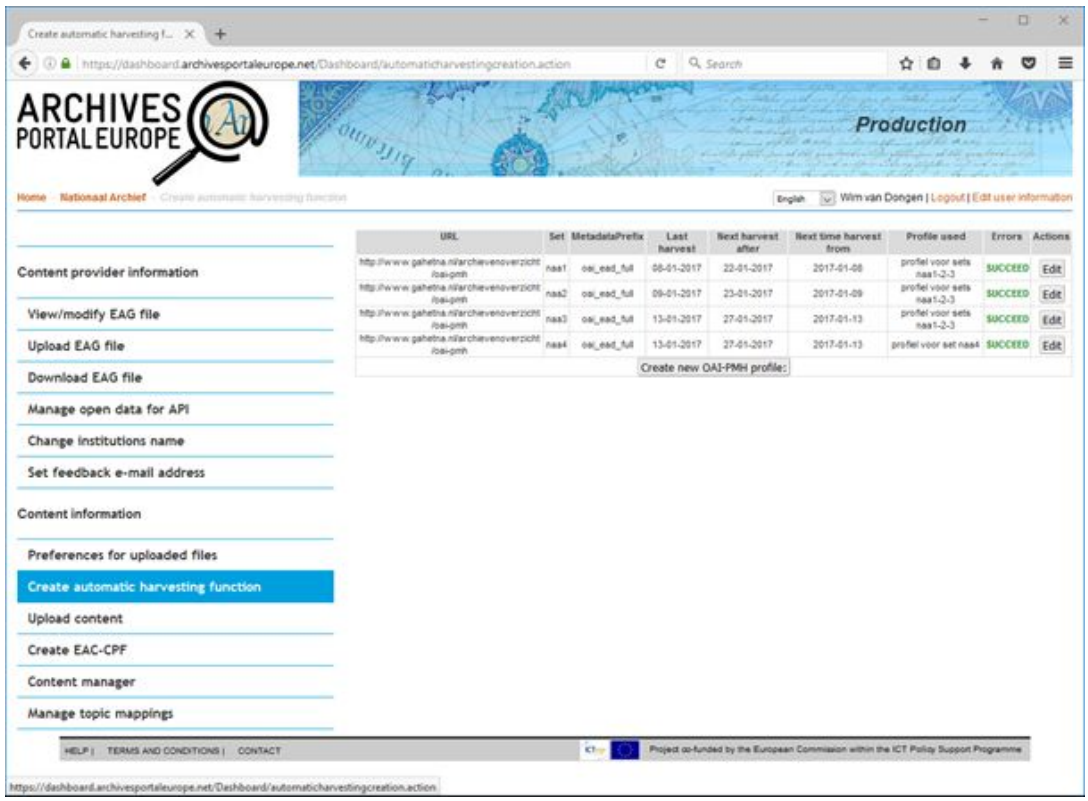

Figure 14. APE Dashboard: overview of harvesting/processing schedule.

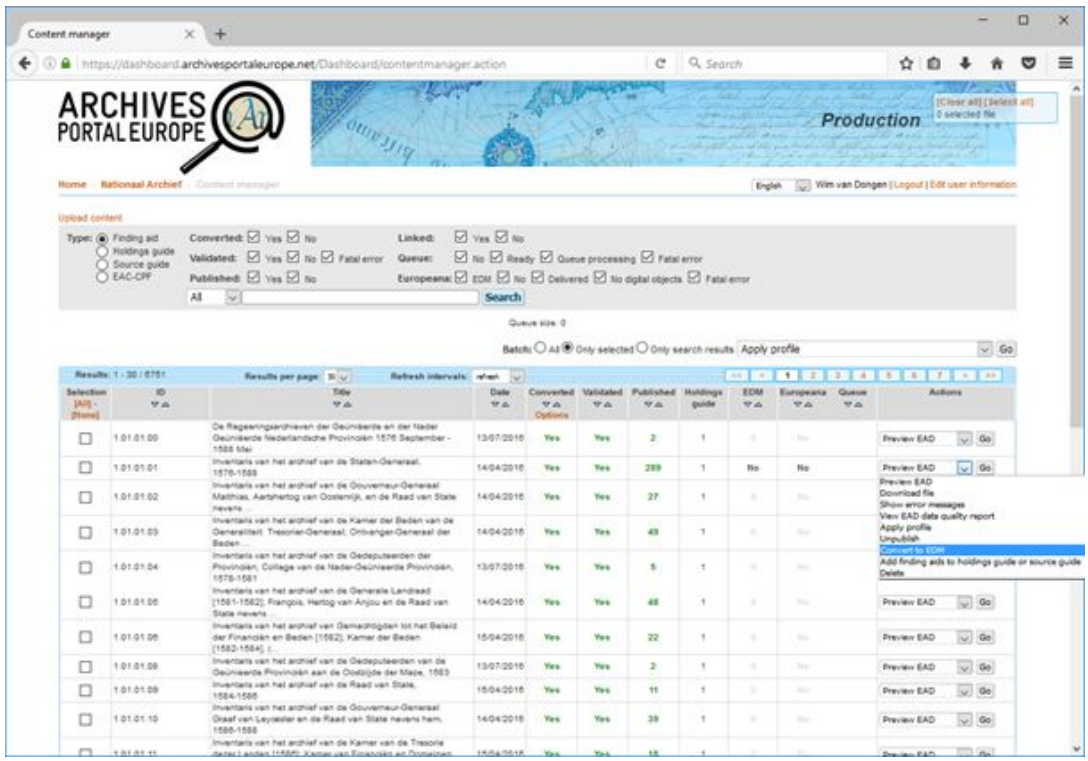

Figure 15. APE Dashboard – start EDM conversion for an individual finding aid.

This action starts an automatic apeEAD to EDM conversion script of which the outcome can be influenced a bit via choices in a dialogue screen (see Fig. 16).

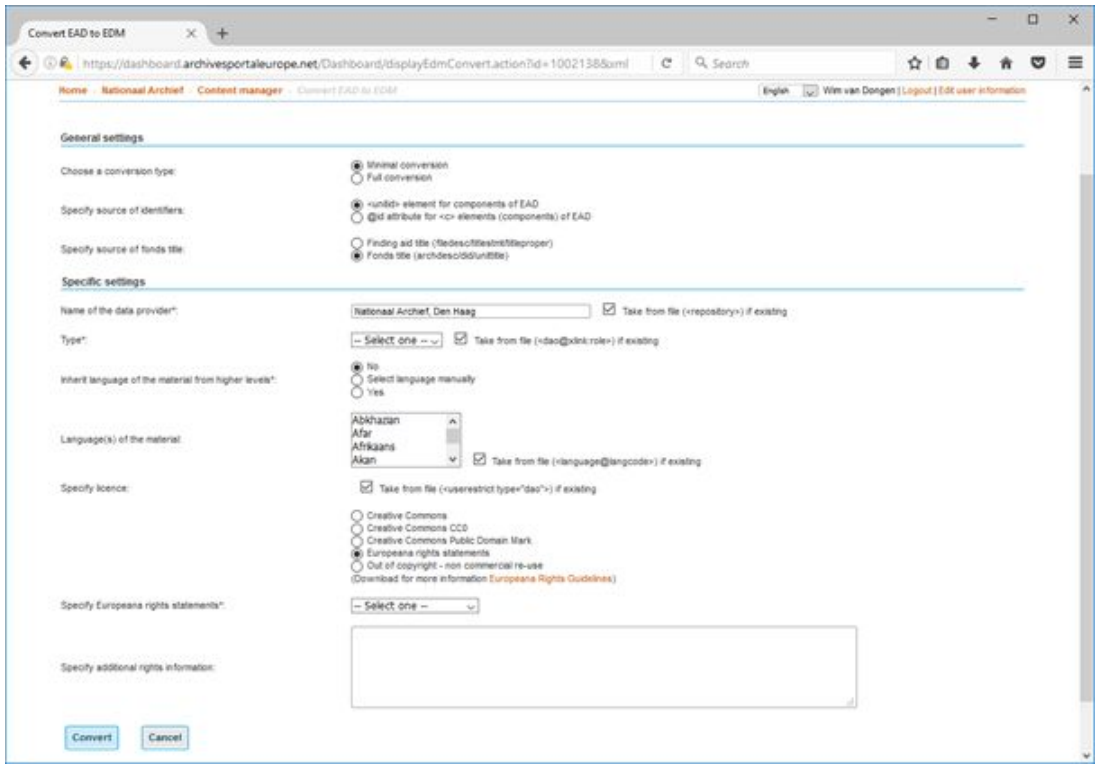

Figure 16. APE Dashboard – fine-tuning the apeEAD to EDM conversion.

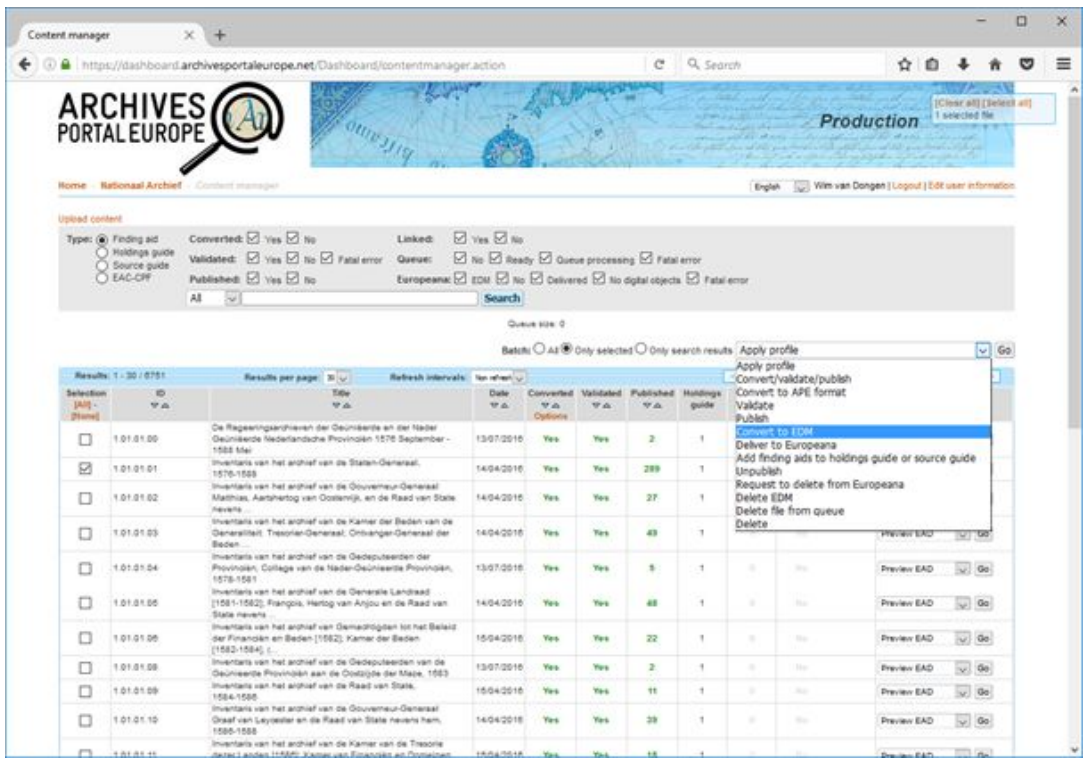

Figure 17. APE Dashboard – start EDM conversion for more than one finding aid.

Of course it is possible to produce EDM records for more than one finding aid. For that purpose the finding aids of which information can be forwarded to Europeana have to be selected in the first column of the Content manager screen in the Dashboard, after which the batch option “Convert to EDM” can be chosen. Then the same dialogue screen as shown in Figure 16 will appear and all choices made in this screen will then be applicable for the EDM conversion of all selected finding aids. This makes it necessary to use this option only for finding aids that give access to the same kind of digital objects.

After the conversion of the finding aid information on digital objects to EDM, the EDM records can be forwarded to Europeana by choosing “Deliver to Europeana”, either in the dropdown list in the column actions for an individual finding aid or in the dropdown list for batch actions in case this has to be done for more than one finding aid at once. This action takes care of putting the newly created EDM records in the Archives Portal Europe’s OAI-PMH repository for Europeana to come and fetch them. When this is finished, the status of the finding aid(s) are updated in the Content manager screen, indicating the amount of EDM records and digital objects (web resources) produced. In the case of finding aid number 1.01.01.01 that is 4 EDM records and 15 digital objects, out of the 289 descriptive units the whole finding aid contains (see Fig. 18).

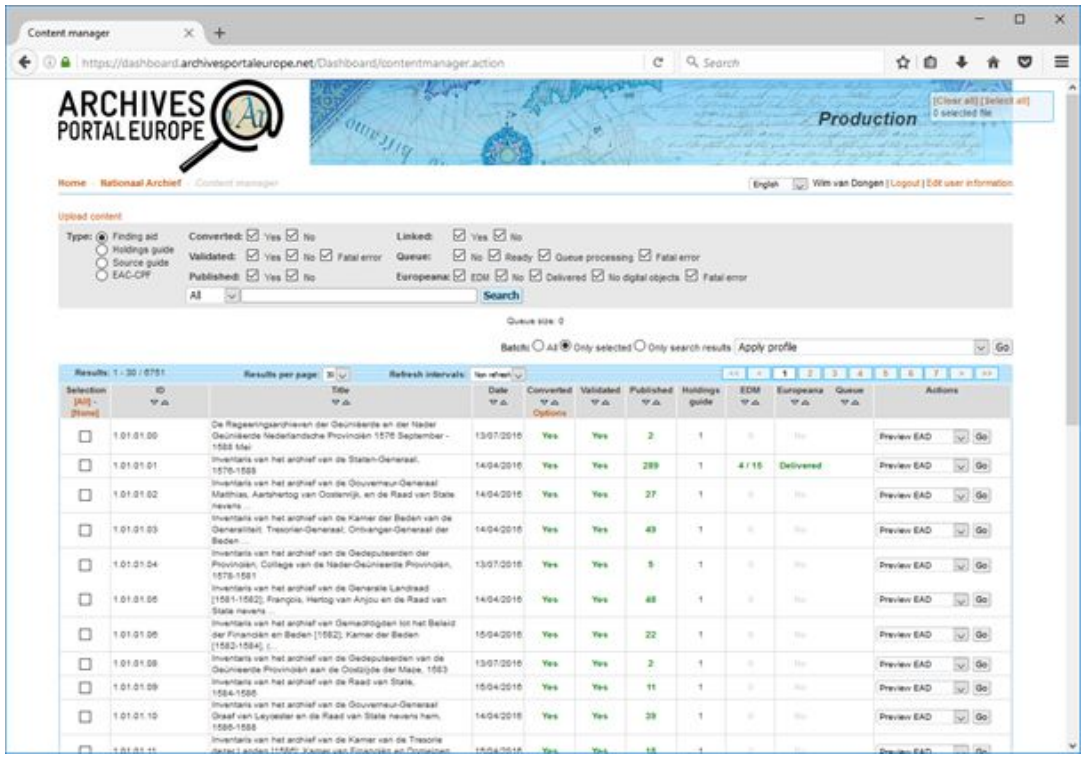

Figure 18. APE Dashboard – result of the EDM conversion of finding aid number 1.01.01.01.

This process of forwarding content to Europeana can also be fully automated. All that takes is to expand the processing profile as shown in Figure 12. In case in the tab “Basic preferences” of that profile for the option “Default action for uploaded files” the default value “Publish to Archives Portal Europe” is replaced with “Publish to Archives Portal Europe and Europeana”, then the second tab “Europeana preferences” will become available (see Fig. 19).

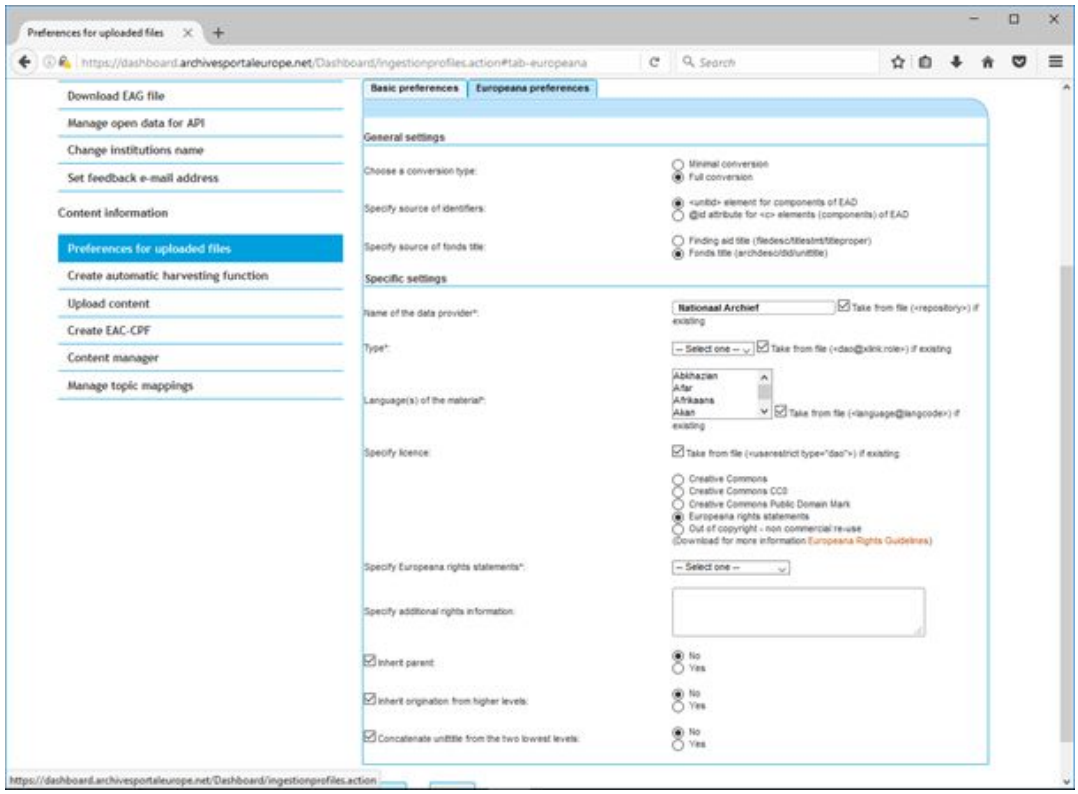

Figure 19. APE Dashboard – adapt the processing profile for automatic delivery to Europeana.

This tab basically contains the dialogue screen of Figure 16, so the possibility to fine-tune the standard apeEAD to EDM conversion script. Once this tab is filled in and the profile is saved and activated in combination with an OAI-PMH repository configuration as described above, the automatic harvesting and processing process not only takes care of automatically publishing content in the Archives Portal Europe, but also of automatically forwarding the digitised part of it towards Europeana.

It is possible to create more than one OAI-PMH harvesting profile as well as more than one harvesting and processing profiles, so it is possible to use different EDM conversion fine-tuning for different batches of finding aids.

Finally: this whole set-up assumes Europeana harvests EDM records from the Archives Portal Europe's OAI-PMH repository on a regular basis, but unfortunately this is not the case yet.

Technical infrastructure maintenance

The fact that Europeana is still not harvesting EDM records from the Archives Portal Europe's OAI-PMH repository on a regular basis is due to the fact that the current Europeana ingestion system is not robust enough for the large amount of huge EDM records which the Archives Portal Europe provides. This has been recognised as an issue by the technical teams of both parties and a strategy to overcome this has been developed. Basically the problem will be dealt with in three ways.

The first one: at the moment the Europeana Data Partner Services team is evaluating a set of sample files provided by the Archives Portal Europe's technical team, which contains all possible varieties of the current apeEAD to EDM conversion that is being offered to Archives Portal Europe data partners in the Dashboard, so actually the results of all possibilities of the above

described apeEAD to EDM conversion fine-tuning dialogue screen for one particular test finding aid. The outcome of this might be that the apeEAD to EDM conversion functionality has to be adapted, which the Archives Portal Europe's technical team will take care of.

A second way of resolving this situation might come out of the fact that the Europeana and Archives Portal Europe teams have come to an agreement to focus more on data quality instead of data quantity. A fully prepared proposal based on the principles as discussed by both teams still has to be drafted and accepted (it's on the agenda for the next APEF face-to-face Country Manager Network meeting on Malta in April 2017), but one of the principles is to aim for less hierarchy in the Archives Portal Europe EDM records and by consequence not too many digital objects attached to individual EDM records, which will probably lead to lighter and handier EDM record sets. These two principles (less hierarchy and less digital objects connected to descriptive units) will also be fostered by the implementation of EAD3 for so called 'additional finding aids' which the Archives Portal Europe team has recently started. At the same time this will eventually also lead to more meaningful metadata records for Europeana, so to better metadata quality.[7]

The third and probably most promising way of overcoming this 'deadlock' situation is to test and pilot with Europeana's future ingestion and workflow systems Metis and Operation Direct. A pilot with a first version of the Operation Direct system has already been successful: during the Europeana AGM in Riga on the 9th of November 2016, Europeana's Operation Direct team was able to demonstrate 'live' a 'real time' connection with Archives Portal Europe data by making use of the new Archives Portal Europe API services.[8] The possibilities of this 'real time' connection will be explored any further in 2017 and in addition a pilot with Europeana's Metis system will start early February 2017.

In the meantime the Archives Portal Europe technical team will adapt the apeEAD to EDM conversion fine-tuning dialogue screen to implement the new set of rights statements Europeana introduced in November 2016 and will also look into possibilities to enrich the Archives Portal Europe's EDM records with information on topics/themes, in order to make it easier for Europeana to select Archives Portal Europe content for re-use in its thematic channels.

Finally the Archives Portal Europe's technical team intends to adapt its system in such a way that in the ingestion phase it will automatically detect whether a finding aid contains links to digital objects at one or more of its hierarchical levels or not and store that information in the database next to basic (high level) information of finding aid. This will make it easier to tailor the Archives Portal Europe's API services to the needs of Europeana, i.e. Operation Direct.

[7] For more information about the Archives Portal Europe's EAD3 implementation see: http://wiki.archivesportaleurope.net/index.php/EAD3_implementation_guideline
[8] For more information about the Archives Portal Europe's API services see: http://wiki.archivesportaleurope.net/index.php/APE_API_documentation

Photoconsortium

Ingestion workflow overview

Photoconsortium got the legacy from Europeana Photography project, in which the aggregation was performed via the MINT software, customized in a special version that is called “EP MINT”.

The workflow of MINT consists of 4 main steps:

- Import: the data partner accesses the MINT online environment via user authentication, and then uploads his/her dataset in the system
- Mapping: the content provider associates specific fields of the metadata schema to his/her metadata
- Transformation: with just a click, the content provider launches the transformation process, that converts the dataset to EDM and stores it on NTUA’s OAI-PMH server

Publication: once informed, the Europeana Data Partner Services team harvests the dataset from NTUA’s server and then checks and processes it for publication to Europeana portal (Fig. 20).

Figure 20. Ingestion workflow of Photoconsortium.

Photoconsortium offers to the members a service of support for the aggregation process, which is provided via:

- Helpdesk for general inquiries on the use of the system, and on the procedure and requirements for publication to Europeana (about the DEA, the Europeana Publishing Framework, IPR support, etc.);
- Direct one-to-one assistance on metadata, helping the content providers understand if their source database is suitable for mapping in MINT or if it is necessary to amend/enrich/modify the metadata fields, and how;
- Technical support for the use of MINT, problem-solving, and follow up after Europeana’s feedback on the provided metadata;
- Periodic checks on the published metadata.

The workflow of MINT is described in detail in the following section, while here it is relevant to illustrate what are the features of MINT that were designed for Europeana Photography project.

The EP MINT is a version of MINT software specifically customized according to the needs and peculiarities of photographic archives databases. During the Europeana Photography project, 16 photographic archives and collections joined forces to design user-based requirements for a tool that would help with the harmonization and enrichment of partners’ datasets, with the final aim to provide to Europeana rich, descriptive metadata, searchable in a variety of languages.

What resulted from this iterative process, which was translated in a customization of MINT:

- setting mandatory metadata elements for all content partners, larger than the Europeana mandatory fields in EDM, in order to provide richer descriptions and information;
- the creation of bookmarks, named after the Europeana Photography mandatory fields, that link to the relevant LIDO elements, helping providers in that way to easily map their metadata to the intermediate schema;

- 16 language vocabulary to be included in the original datasets, or to be added during the mapping process. The integration of the content specific vocabularies in EP MINT facilitates the mapping of “source” vocabularies to the Europeana Photography vocabulary. The vocabulary is published in SKOS format for easy re-use, and it is easily extendible and enlargeable with additional languages and features.

Indeed, the Multilingual Vocabulary is a major feature: it allows not only for translation and semantic enrichment of the metadata at the time of publication on Europeana, but helps guide all partners in the metadata creation; it is a hierarchically structured vocabulary of techniques, keywords and photographic types, dedicated to early photography, and currently contains 561 concepts.
The first part of this thesaurus lists, in a hierarchical way, the photographic techniques used to create the object. At the top of the hierarchy, it mainly distinguishes between positive and negative types of photography, but what it really gives value to the vocabulary is to specify photographic processes, carriers and formats. The techniques are linked to the development process of photography in the period covered by the project (1839-1939), and therefore limited to the techniques available between the first photographs and the Second World War.

Then the vocabulary helps to identify and capture the reason for taking a specific photograph. The photographic practice describes in keywords, of which one or more can be selected, the way a photo was conceived and for which goal was intended. This goal often holds information that is not immediately visible on the actual picture itself, but is necessary or interesting for research. Finally, the vocabulary includes a keywords section containing a large number of subjects, identified and regrouped to describe the subject of the photographs. This section is mainly based on the IPTC model, and the Europeana Photography partners’ vocabularies. Regarding the themes, it is an issue that goes beyond the field of photography and, in fact, the existence of free thesauruses such as AAT (Art and Architecture Thesaurus) from the Getty Institute solves this issue. The differentiating factor is found in technical concepts. Europeana Photography Multilingual Vocabulary is a meritorious achievement in a strategic aspect, for today but mainly for the future online re-use of digitised photographic collections.

Technical infrastructure maintenance

The maintenance of MINT and the developments towards the MINT4ALL are managed by NTUA and described in a following section. As with other MINT-powered aggregators, the EP-MINT would need to be updated for the Europeana new rights statements

From the side of Photoconsortium, the association takes care of the maintenance of the Photoconsortium portal (www.photoconsortium.net consisting of a website with various information about the members and the activities of the association, and including a blog), which doesn’t have a database of content. The metadata are in fact stored in NTUA’s OAI-PMH server.

Furthermore, an analysis will be done in the future for the possibility to integrate the IIIF standard and guarantee the ingestion of possible repositories with these contents.

Maintenance and development of MINT - towards MINT4ALL

Scope of MINT

The MINT aggregation platform facilitates the ingestion of semi-structured data and offers the ability to establish crosswalks to the reference schemas (i.e. EDM fashion profile, EDM, Carare, Lido, EUscreen schema, EBUcore, EDM sounds profile, DC, EAD) 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 DSI-2 project is to enable users to:

- Provide metadata records in a range of “source” formats
- Convert metadata to selected target schema
- Monitor the progress of content provision

At the moment MINT is used by seven thematic aggregators (six in DSI-2), two national aggregators and additionally European broadcasting Union and Europeana.

Thematic or National Aggregator	Metadata schema	Publishes to	Additional services
EUscreen (DSI-2 partner)	EUscreen schema Based on EBUCore	EUscreen Portal Europeana Collections	Manual metadata enrichment
Europeana Fashion (DSI-2 partner)	EDM Fashion Profile EDM	eFashion Portal Europeana Collections	Semi-automatic metadata enrichment; Extraction of visual information from images (colours, shapes) and semantic enrichment (Named Entity Recognition and extraction for Materials, Techniques and colours)
Europeana Photography / Photoconsortium (DSI-2 partner)	LIDO schema	Europeana Collections	Automatic metadata enrichment with 16-languages vocabulary on photography
CARARE (DSI-2 partner)	Carare schema LIDO schema EDM schema	MORe	
Europeana Sounds (DSI-2 subcontractor)	EDM Sounds profile	Europeana Collections	Manual metadata enrichment
Museu (DSI-2 partner)	Lido Schema	Europeana portal	Automatic metadata enrichment

Hope	various	Internal use	
German Digital Library	various	Internal use	
European Broadcasting Union	EBUcore schema	Internal use	
National Library of Israel	EAD	Internal use	
Europeana	EDM	Internal use (part of UIM)	

NTUA manages the MINT instances of the thematic aggregators that are part of DSI-2. The rest MINT instances and publications modules are managed by the organisations and NTUA is only providing the technical support.

Workflow

The metadata ingestion workflow handled by MINT, as illustrated in Figure 21, 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. Following is the Schema Mapping procedure, during which the imported metadata are mapped to the selected target schema. A graphical user interface assists content 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 also supporting the share and reuse of metadata crosswalks and the establishment of template transformations. The third step is the Transformation procedure, during which providers' metadata is 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 the target schema 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 Operations Officers 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 target schema functions as an anchor, to which various data partners 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.

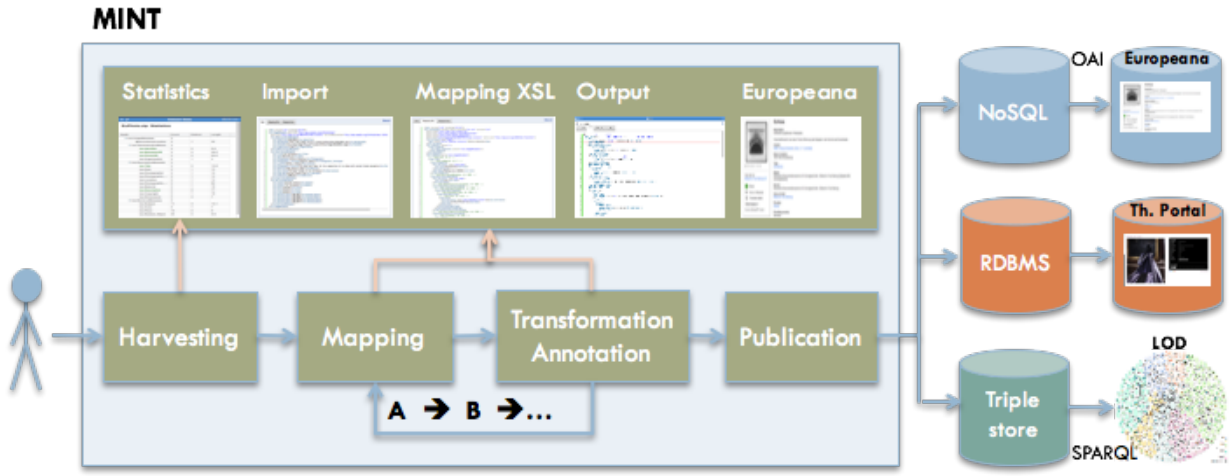

Figure 21. MINT Ingestion workflow.

Mapping process

Mapping of metadata into the desired schema is the basic functionality of the MINT ingestion platform. A visual mapping editor for the XSL language is used and mappings are performed following a drag-and-drop procedure, selecting items from input schema area and dropping them in 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 22. By setting preview options users have the ability to preview XML code of import and transformed items.

Figure 22. The various mapping options in MINT

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.

Current maintenance work

Even with no development happening, the MINT instances need a minimum of maintenance work. Even if not a single user logs in, the servers need regular disc backups and instances have to be moved from failing machines to newer ones. But people do use the instances and make the following works necessary:

- *Bugfixing.* MINT is fairly stable and mature, so major problems don't occur. This is therefore a rare event and usually not much effort has to be spent.
- *Ingestion and mapping support.* Although all our users have gone through courses and are quite familiar with the process, each new upload, each slightly changed input format has the potential of a creating a problem that normal users cannot easily solve (or more likely have forgotten how to do it). This is a more frequent event, still usually easily solved.
- *Schema evolution.* All schemas evolve to adapt to new data needs. Sometimes this means, that older datasets have to be adjusted by updating their mapping and re-transforming them into new XML. Depending on the type of change and the amount of retrofitting needed this process varies from being easy to being fairly involved.
- *Thesaurus maintenance.* As with schemas, useful thesauri evolve and acquire new words and corrections, which may need to be backported to old datasets or may just be available for new data.

MUSEU and Photoconsortium Aggregator

Although MUSEU and Phtoconsortium are using different MINT instances, they are using the same schema and workflow to publish metadata to Europeana.
The MINT instance of the MUSEU aggregator can be found in
http://mint-projects.image.ntua.gr/museu
The MINT instance of the photoconsortium aggregator can be found in
http://mint-projects.image.ntua.gr/photography

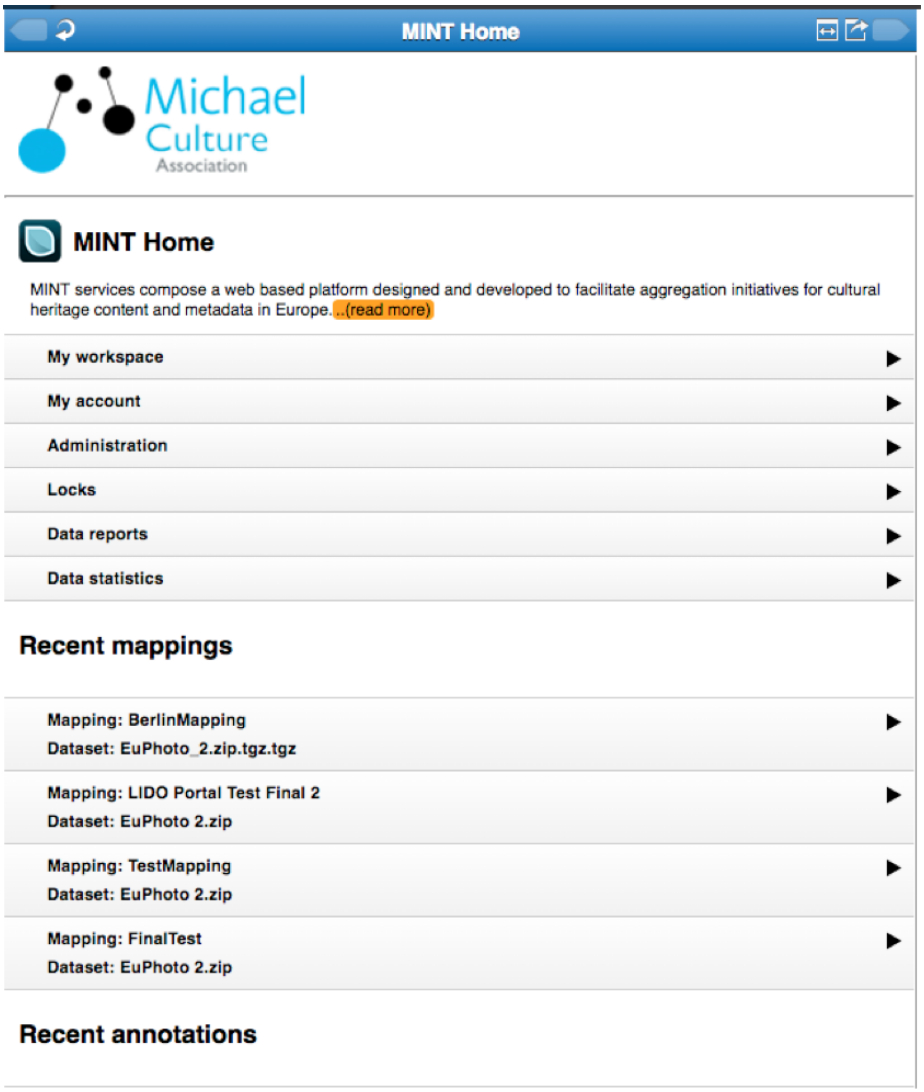

Figure 23. MUSEU MINT instance.

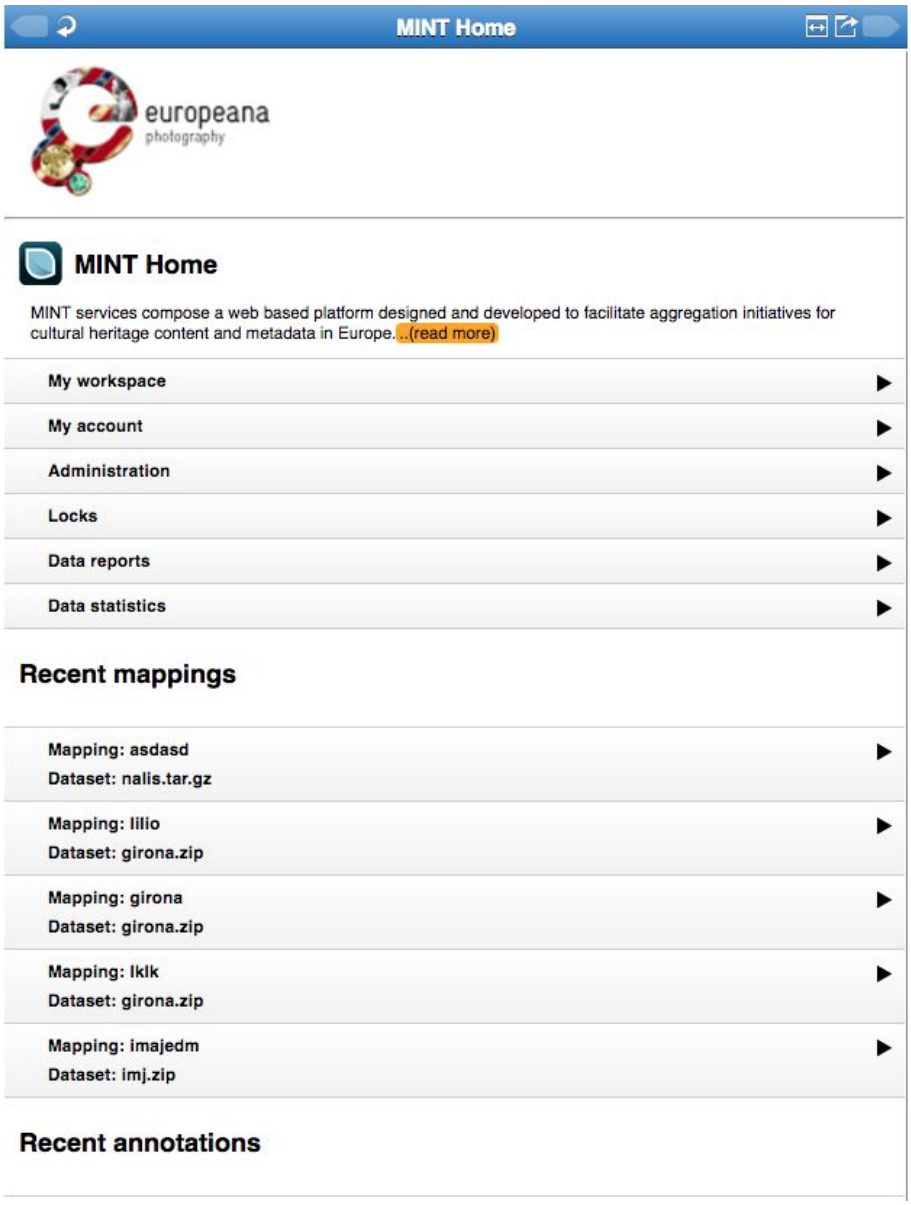

Figure 24. Photoconsortium MINT instance.

MINT customizations for the Museu and Photoconsortium Aggregator

LIDO specific customizations have been performed for reflecting the recent metadata quality requirements of Europeana in the production of both LIDO and EDM metadata. In detail the task group “visualization and indexing of museum content” of the AthenaPlus and Europeana Photography projects has authored the requirements for producing high quality LIDO metadata and the following bookmarks have been created on MINT (Fig. 25).

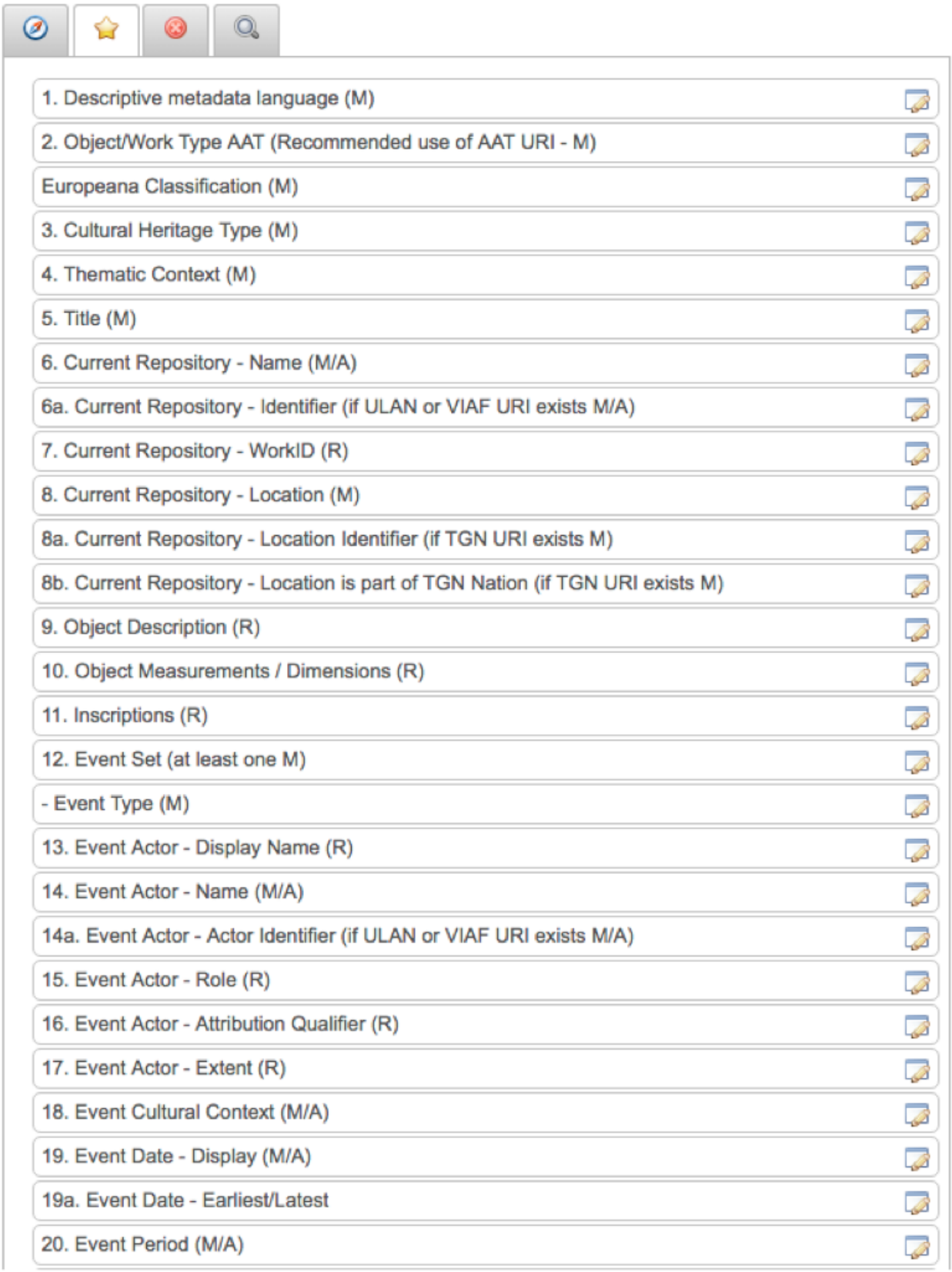

Figure 25. Bookmarks in MINT.

The MUSEU and Photoconsortium aggregators automatically perform metadata quality check controls ensuring a minimum standard of quality in the metadata produced by it. The quality check controls are on the cardinality restrictions specified in. The MINT instance for the MUSEU and Photoconsortium aggregators supports:

- EDM and LIDO schema validation through a user friendly interface with inline error detection that helps the content provider to easily detect and correct invalid files.
- Schematron rules that enable a more strict validation based on the museum domain specific criteria that will be specified by the experts.
- Preview interfaces that involve the content provider into the quality evaluation process by letting him view the produced metadata files and also the way that these would look like when published on the Europeana portal.

Maintenance, technical and functional support

Technical and functional support has been offered through a mailing list that is dedicated to this web tool. In detail a help desk support system has been set up for assisting the Content Providers during all the phases of the aggregation workflow. More specifically the Content Providers were able to get instructions on how to map their in-house metadata to LIDO for exploiting its full expressiveness, as well as technical information about MINT functionalities that permit them to fit their metadata perfectly according to the metadata quality requirements. The helpdesk service is also for the Content Providers of projects AthenaPlus, Europeana Photography, Athena and Linked Heritage that want to ameliorate the already published records using the new MINT instance.

Europeana Fashion

MINT customizations for the Europeana Fashion

The MINT instance (http://mint-projects.image.ntua.gr/fashion) used for the Europeana Fashion project has been appropriately extended for fulfilling the increased metadata quality requirements of Europeana. A lot of effort has been given to the quality of metadata through the corresponding Europeana taskforce, hence the MINT mapping tool has been further customized based on the EDM-FP schema. A set of bookmarks has been made to the EDM-FP elements that although are not mandatory are highly recommended and should be filled by the content providers. In that way content providers will produce more expressive EDM-FP metadata. This change has also affected the EDM-FP to EDM crosswalk that is used for the Europeana publication, as well as the EDM-FP previews on MINT that will have to be appropriately updated.

Another important development that has been performed is the update of the OAI-PMH server for hosting the EDM-FP metadata. Currently the OAI-PMH is used for the Europeana publication and hosts the EDM and the OAI_DC flavours of a record. The first is needed for the Europeana publication while the second one is necessary for implementing the OAI-PMH protocol. However, the EDM record is a downgraded version of EDM-FP record since all the additional information that EDM-FP records hold is expressed under more abstract semantics. For example the edmfp:photographer of the EDM-FP records becomes a dc:creator etc. The extension of the OAI-PMH server for hosting the EDM-FP metadata will not only support the efforts of the providers by sustaining and hosting the metadata they originally produced (EDM-FP) but it will also enable the Fashion channel on Europeana's side.

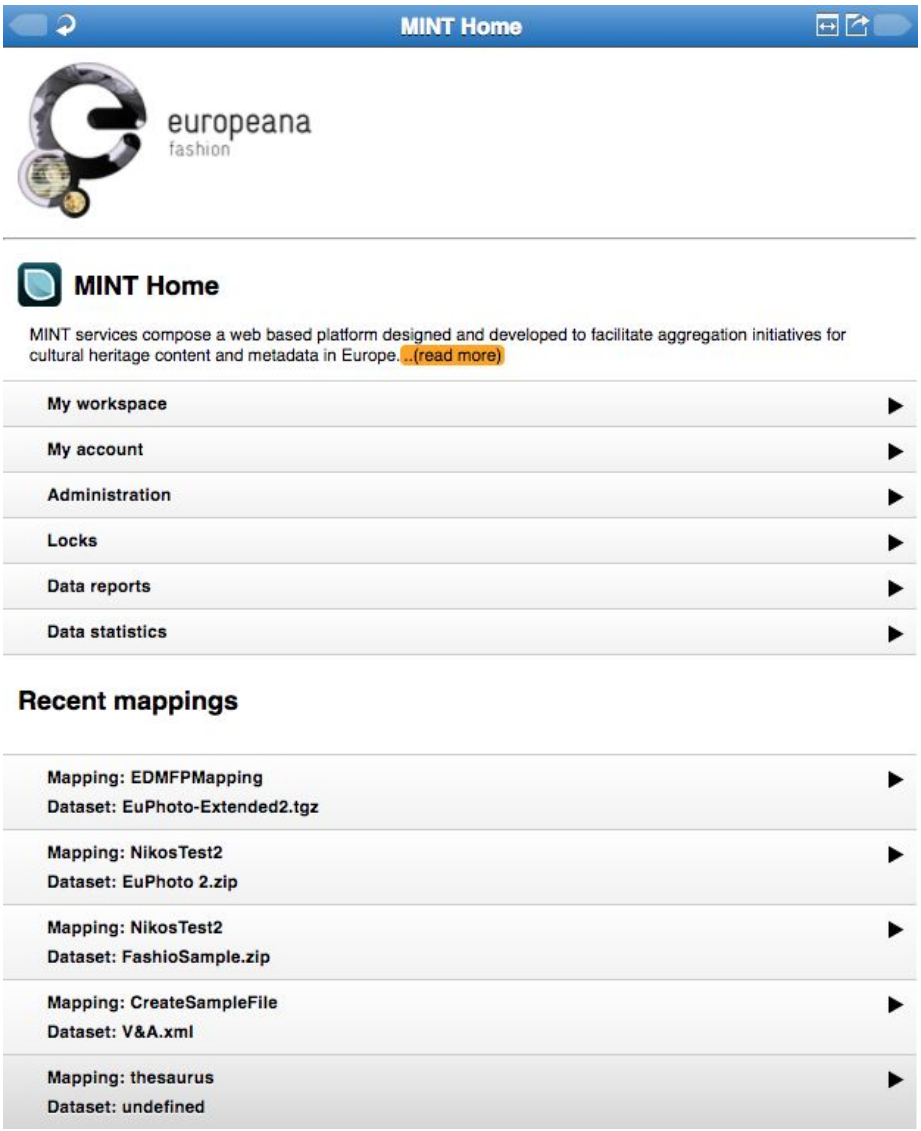

Figure 26. eFashion MINT instance.

Optimization of the publication process

The Europeana Fashion platform has a quite complicated publication process because its metadata need to be transferred to three different backend repositories that are used for different purposes. The complexity is further increased by the fact that these publications need to be synchronous – in other words they are all dependent to each other. First the RDBMS publication has to be performed for creating the URIs that are used as the isShownAt values at the Europeana publication and finally the enrichments need to be performed on the triplestore but then brought back to the RDBMS for displaying on the portal. This process is not only complex but also time consuming and it clearly needs to be optimized. The publication process can be more efficient by appropriately extending the RDBMS of the Europeana Fashion portal for holding the necessary information for all the publications and thus making it the central point into the publication process.

Semantic enrichment and normalization

The contents of the EDM-FP metadata fields contributed by the several content providers carry information that can be subjected to enrichment analysis. In particular, certain EDM-FP fields such as dcterms:medium, edmfp:technique, contain specialized information about the respective items, either in textual form or as URIs of the Fashion Thesaurus that has been developed for the Fashion Project in order to allow semantic characterization of the data. In the case these fields contain textual entries, the semantic linking may be recovered by analysing the entries in order to detect the corresponding Fashion Thesaurus entries. Other fields like dc:title and dc:description contain always textual information, that in most cases are rich enough and can be analysed in order to identify additional occurrences of Fashion Thesaurus entities that may refer to the type, colour, materials and techniques used in the respective item; such information can be used to expand the information contained in other special purpose fields. Finally, fields such as dcterms:spatial, mrel:dsr, mrel:pht and edmfp:model contain textual names of places, designers, photographer and models, respectively, that are associated with the respective item. All of these names can be the subject of analysis that will link them to a Linked Data resource (e.g. DBpedia), if possible.

Following the pilot experiment with content enrichment in the framework of the Fashion Project, we have systematically processed the contents of the abovementioned EDM-FP metadata fields in all the published datasets, in order to produce a set of semantic characterizations for each item, which may be either Fashion Thesaurus entities or Linked Data resources. The enhanced EDM-FP files will have been published according to the best practices for publishing Linked Data.

Image Analysis

The pilot phase of the Fashion project led to a processing pipeline of two stages, namely a segmentation phase and a colour extraction one, which enabled automatic colour annotation of clothing given a colour palette. By taking into consideration the pros and cons of the method we will implement an improved version of the tool based on recent methodologies (deep learning tools) that will produce improved segmentation masks in terms of higher precision clothing segmentation and - as a consequence - a more accurate colour representation. Furthermore, we will experiment with different colour pallets to qualitatively evaluate the results. Annotation has been applied to a set of approx. 435,000 catwalk images from the Europeana Fashion collection that lacks manual colour annotation.

Maintenance, technical and functional support

NTUA is maintaining the aforementioned web-tool and modify it for fulfilling the requirements and requests issued of the contracting authority. The requests will be negotiated with the ICCS and they must include modifications or customizations of existing features.

Technical and functional support is offered through a mailing list that will be set up dedicated to this web tool. In detail a help desk support system will be set up for assisting the Content Providers during all the phases of the aggregation workflow.

EUscreen

MINT customizations for the EUscreen aggregator

One of the main customizations made in the EUscreenXL project was the setup of different target schemas, according to the project requirements and the needs of the content providers. In particular three target schemas have been used by the content providers of the EUscreenXL project:

- EUscreenXL Item/Clip, used by the content providers for transforming metadata about clips and publishing them on the EUscreenXL portal;
- EUscreenXL Series, used by the content providers for transforming metadata about series and publishing them on the EUscreenXL portal;
- EUscreenXL - EDM, used by the content providers for transforming metadata about videos on their own websites into a Europeana suitable schema.

The EUscreenXL project has a quite sophisticated and complicated publication workflow. MINT acts as the metadata processing unit that prepares metadata that are then sent to Europeana, to the EUscreenXL portal and to a triplestore.

As described above, EUscreenXL content providers can use three different schemas for transforming their metadata while another schema has been employed for the migration of the EUscreen metadata to the portal.

The metadata mapped to EUscreenXL - EDM schema relate to multimedia records that are not published on the EUscreen portal. These go directly from the MINT mapping tool to the OAI server hosted by NTUA

On the other hand metadata mapped to the other schemas are published to the EUscreen portal. This makes the publication process for them a bit more complicated. In the previous case the mapping schema used was EDM and no further processing of the metadata was required. The content provider made the mapping to EDM using only information from the metadata imported. However, for the ITEM/CLIP, SERIES and OLD EUscreen schemas various transparent-to-the-user manipulation processes are performed. First, EUscreen OLD is mainly based on EUscreen ITEM/CLIP schema and it has been implemented for the publication of the EUscreen project metadata that is the predecessor of EUscreenXL to the portal. Since the metadata requirements have been raised from the time of the first EUscreen project, the metadata produced for that project were not as expressive as the metadata produced for EUscreenXL. Therefore, EUscreen OLD - implemented for acting as an intermediate bridging this expressivity gap between the old and the new schema - allows the EUscreen metadata to be validated and published according to the EUscreenXL metadata guidelines.

Metadata mapped to ITEM/CLIP, SERIES and OLD are all published on the EUscreen portal. Once these records are published they get a unique EUscreen portal identifier and an EUscreen landing page. This information is vital for the metadata sent to Europeana - both the record identifier and the landing page are mandatory in EDM - but it cannot be calculated until the metadata are actually published on the portal. Furthermore, the EUscreen portal is based on the EUscreen schemas, while a transformation to EDM was required for publishing these records on Europeana. Therefore a publication process has been implemented that first publishes metadata to the EUscreen portal, then retrieves the portal identifiers and creates the EDM records. These records are sent at the following OAI server that Europeana uses for harvesting and publishing.

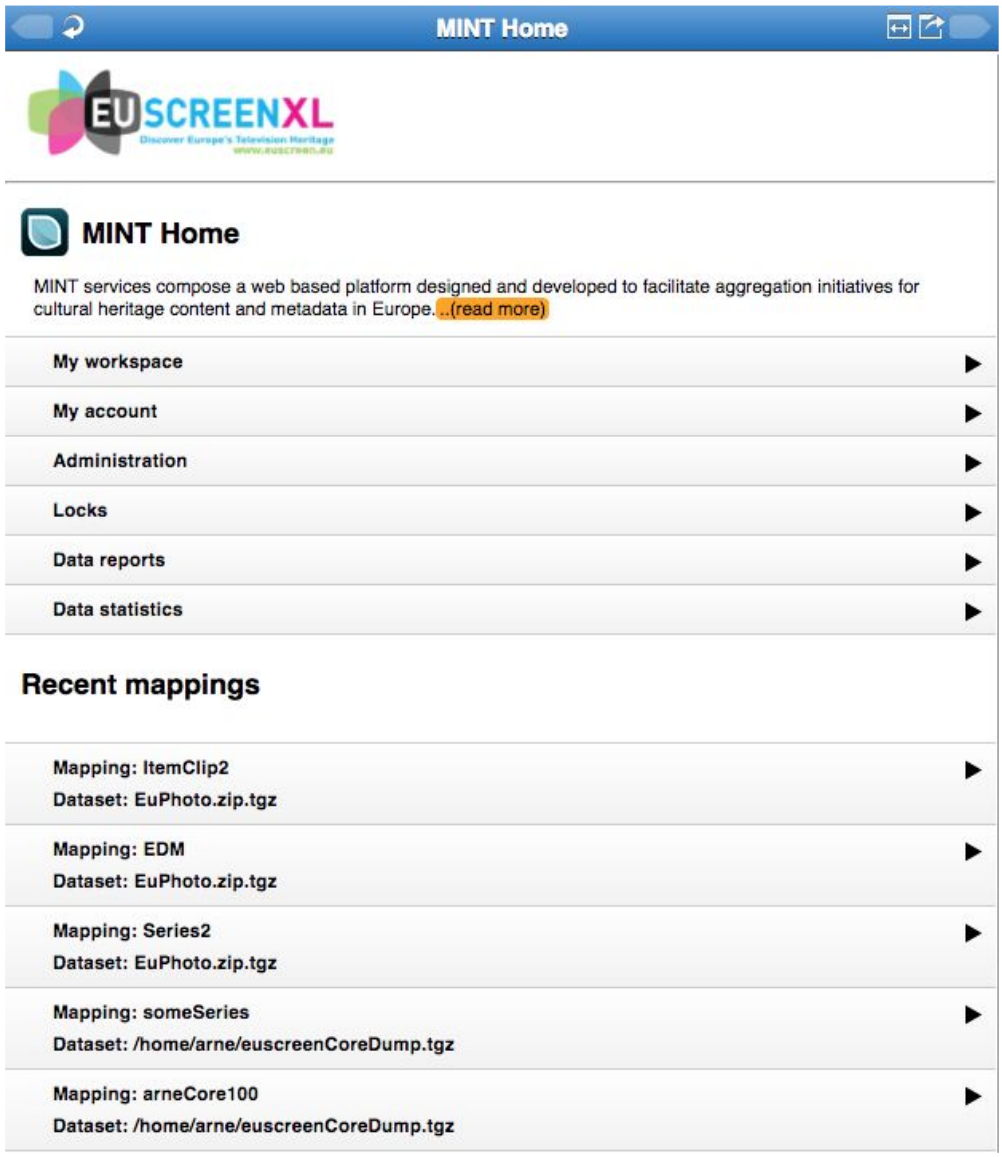

Figure 27. EUscreen MINT instance.

Metadata Enrichment

This area of work focuses on enriching the metadata using manual and automatic methods and tools. During ingestion providers are manually enriching metadata fields using the EUscreen subject terms thesaurus (based on IPTC thesaurus), ISO thesaurus for languages and the Geonames thesaurus for countries, cities and geographical places. Both EUscreen and Geonames thesauri are multilingual. After the manual enrichment, the metadata are automatically enriched from relevant sources in the Web. Access to the RDF data is provided through a SPARQL end-point. In order to improve the quality of the data using enrichment tools, a group-edit functionality has been used that facilitates data partners and data moderators to clean and normalise the data in a user friendly way. It gives the ability to group items using special search filters and normalise/edit the values using established standards as well as provide suggestions for enrichment from Web resources (e.g. DBpedia) that the users can accept or reject.

Maintenance, technical and functional support

Ongoing aggregation works and new partners needed support with the MINT platform, as there have been fewer opportunities to learn about the process for them. The publication process from the EUscreen portal dataset to Europeana is a manual process that involves development of supporting scripts. During the lifetime of the various EUscreen projects, different versions of the thesaurus have been created and have to be synced with the final version that is in use at Europeana. On provider side, URLs cannot always be kept stable, so another effort is made to adapt data at Europeana to new URLs. Further, efforts are made to improve the presentation of the existing metadata, which involves manual and semi-automated cleaning procedures.

Checks and back-up possibilities have been put in place at partner institutions to ensure that the assembled metadata and ingestion possibilities remain stable at all times. Specifically for audiovisual media, the platform has expanded its possibilities to cater for subtitle files as an additional form of (searchable and linkable) metadata. The information has been exploited through entity recognition and adding them to the linked data repository for accessing them through the SPARQL end point also for re-use in third-party platforms.

EUscreen has also contributed to the activities initiated by Europeana, specifically the inclusion of time-based media in the Europeana ecosystem and use of automatic enrichment technologies. We will assess developments in the area of W3C Media Fragments, and HTML5 and how these relate to EDM.

MINT4All

Mint was widely used by many aggregators and Europeana before the start of the Digital Service Infrastructure and has split into many slightly different instances. It is a fairly old software and developed over many years, so many components are quite outdated. Old software versions have many problems, from unfixed bugs to security vulnerabilities.

Over the years the requirements have evolved, but the general architecture hasn't evolved equally with them. Functionalities have been added to MINT, sometimes only specific to one aggregator´s workflow, sometimes in the code that is shared by other aggregators, sometimes in separate components.

To remedy the situation, MINT is undergoing modernization inside the DSI-2 framework.

1. MINT and all the libraries necessary to operate it, are upgraded to Java 8. The build system is adjusted to make incorporating new libraries and updating existing ones easier. Major parts of this upgrade are
    a. Move from Java 6 to java 8
    b. Move to Ivy dependency management
    c. Upgrade Hibernate from version 3.x to 4.x (5.x if possible)
    d. Move from Tomcat 6 to Tomcat 8
2. The divergence of MINT into different systems has following implications
    a. Organizations that take part in different aggregation strands have multiple unconnected logins and workspaces. They are duplicated and may have duplicated data.
    b. Although code is shared, not all instances are updated to the latest code and bugs can be fixed in one version, but not in others.
    c. Not all options of code sharing and reuse are fully exploited.

    d. No functionality can be offered to detect data duplication.
    e. Maintenance effort is multiplied.
3. The MINT data model is upgraded to allow all aggregating workflows to be hosted in the same MINT instance (Spring 2017).
4. The MINT user interface is upgraded to handle multiple aggregation workflows. (Summer 2017)
5. During the migration to the unified MINT instance organisation workspaces will be integrated (hence MINT4All). Modules for data deduplication can be employed, inside one workflow or across multiple domains.

The general MINT architecture still follows best practices from when it was first designed, which means that the backend server is mostly responsible for view generation (HTML). This cannot be considered practical by modern standards. All adjustments to the platform are now aimed at generating JSON data based web services, which leave the correct rendering to the user´s browser and appropriate JavaScript libraries. This will automatically improve the scriptability of the MINT server and its usability as a service to other applications.

Integration with Metis and Operation Direct

Although MINT is aimed at providing aggregators with the easiest possible route to get their data into the Europeana infrastructure, integration with Europeana services was minimal. This will change in the near future:

1. The original MINT provided content owners with a preview, how their data would be rendered in the Europeana portal, something all content owners care about. Meanwhile, the rendering process in Europeana has been improved a few times, and these changes couldn't be reflected in MINT. The Europeana preview service, offered from their in-house data processing system Metis, will change this. MINT will use it to allow CPs to get a realistic preview of their data and a chance to optimize their presence on the portal. This will improve data quality on the portal.
2. Similarly, Metis will offer a validation service on EDM data, which cannot be replicated on the MINT system. When it becomes operational, MINT will use it to tell providers exactly which records will be accepted by Europeana and which won't. (MINT will as well relay any further information about rejection reasons that the validation service will offer).

With "Operation Direct" Europeana aims to give data partners a direct route to their content at Europeana. It will facilitate a quicker publishing and withdrawal cycle, and for providers with access to the right tools, eliminate the need for aggregation or mapping altogether. Until this becomes fully operational, NTUA will closely cooperate with its development and support it as additional publication target from MINT. Preliminary tests have been performed successfully.

Conclusion

Many different tools are used and many different ingestion workflows are implemented by all the DSI aggregating partners. Some commonalities exist between all the partners using MINT, but even here, customisations were necessary due to special requirements of the partners. Only on a high level it is possible to identify that in principle the same steps are addressed in each ingestion workflow, but on implementation level the differences are significant.

The workflow descriptions but also the plans for technical work on all the tools will be important information to shape the future development of ingestion workflows and tools, used to operate an aggregator.