Europeana – Core Service Platform

MILESTONE

MS9: TECHNICAL INFRASTRUCTURE DEVELOPMENT PLAN

Revision	1.0
Date of submission	30 June 2015
Author(s)	Henning Scholz (EF), Kerstin Herlt (ACE), Julia Welter (DIF), Maria Teresa Natale (MICHAEL), Marzia Piccininno (MICHAEL), Corinne Szteinsznaider (MICHAEL), Walter Berendsohn (FUB), Joerg Holetschek (FUB), Gisela Baumann (FUB), Kate Fernie (2Culture), Dimitris Gavrilis (AthenaRC), Marco Rendina (eFashion), Johan Oomen (NISV), Andrew Ormsby (BUFVC), Afelonne Doek (IALHI), Gariella Ivacs (IALHI)
Dissemination Level	Public

Funded by
the Connecting Europe Programme
of the European Union

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author (Organisation)	Description
0.1	27/05/2015	Henning Scholz (EF)	Template and initial draft
0.2	26/06/2015	Henning Scholz (EF), Kerstin Herlt (ACE), Julia Welter (DIF), Maria Teresa Natale (MICHAEL), Marzia Piccininno (MICHAEL), Corinne Szteinsznaider (MICHAEL), Walter Berendsohn (FUB), Joerg Holetschek (FUB), Gisela Baumann (FUB), Kate Fernie (2Culture), Dimitris Gavrilis (AthenaRC), Marco Rendina (eFashion), Johan Oomen (NISV), Andrew Ormsby (BUFVC), Afelonne Doek (IALHI), Gariella Ivacs (IALHI)	Final draft
1.0	30/06/2015	Henning Scholz (EF)	Final comments integrated, 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This milestone contains original unpublished work except where clearly indicated otherwise.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.

Table of Contents

Introduction

The aggregators participating in the Europeana DSI will need to develop and maintain their own technical infrastructures, allowing for a continuous supply chain of cultural heritage data from their partners to the Europeana DSI. The aggregators can then increase the number of participating data providers, processing an increasing amount of data while improving the quality of data submitted to the Europeana DSI. Procedures will be streamlined, workflows scaled up and toolkits and software packages reviewed to work more efficiently (including fixing bugs in workflows). Where possible technical development will be shared and should be open source. The goal is to secure efficient provision of metadata to the Europeana DSI. Thus, the purpose of this milestone document is to specify the plan for the technical infrastructure to support the operation of each aggregator. In specifying the elements of the infrastructure that are necessary to be developed, it is also an inventory of tools, software components or workflow orchestrators that are in use by each individual aggregator. This inventory will then feed in to the Aggregator Forum working group on technologies, tools and workflows.

The implementation of this technical infrastructure development plan will also inform the work on the requirements for the new ingestion management tool METIS. Defining the requirements as well es monitoring them as part of the development process and later on testing first components of METIS will happen in close collaboration with the aggregators. The work on this task during the first year of the Europeana DSI will be incorporated into the work and implementation plan to innovate the aggregation infrastructure (D1.1 of the Europeana DSI).

Technical infrastructure development per aggregator

European Film Gateway

EFG will further improve its Metadata Editor, which allows for adding and editing of data already ingested in EFG. The plan is to enhance functionalities in a way so that the tool allows adding new records to the EFG database manually. While not the preferred or sensible way to get in larger amounts of data into the EFG repository, it will prove valuable when it comes to integrating data from archives not yet connected to the EFG infrastructure that can only provide small collections (less than 100 items possibly). Costs for integration of smaller sets of data can be reduced severely this way. Especially smaller film archives often struggle to provide XML exports from their local databases without involving external database specialists. Efforts for writing and implementing mappings in some cases seem also too high, when archives can only give access to small collections. The possibility to manually add records helps to lower the thresholds for those archives to contribute to EFG and thus Europeana.

Moreover, it is desirable to investigate measures that help to simplify the sharing of metadata from data providers' cataloguing systems and online platforms with the EFG infrastructure. This evaluation needs to be done in close cooperation with our technical partner that runs the underlying D-Net platform. DIF will assess its current aggregation mechanism and compare it with those of other aggregators within the workflow working group of the Europeana Aggregator Forum. Based on the results, DIF will investigate possible solutions for easier data contribution by existing and new providers.

DIF will further enhance its tools that are in use to facilitate editing of existing data in EFG. DIF will investigate possibilities to modify the tools in a way that they allow not only the interaction of archivists with their data in EFG, but also to provide them with a possibility to release the modified data to the live portal in an easy way, not requiring the intervention of a technical partner. The tools that are currently in use in connection with the EFG infrastructure are:

Content Checker. The Content Checker is a validation tool that allows low-level searching and browsing of the EFG pre-production Information Space in order to check if metadata records have been correctly harvested and mapped before they are release to the live EFG portal and Europeana. This tool is used by all archives connected to EFG.

Vocabulary Checker. The Vocabulary Checker is used to check whether vocabulary terms in local databases have been properly matched to the EFG vocabularies and are displayed in a harmonised way on the EFG portal. The Vocabulary Checker displays the number, the types and the positions of possible matching errors in the Information Space. Depending on the type of error, archivists can decide if an error can be solved directly in the Information Space via the Metadata Editor Tool or locally in the original source archive.

Metadata Editor Tool. The Metadata Editor Tool (MET) is a cataloguing tool for the enrichment of the Information Space. It allows data curators to edit and delete metadata records in the Information Space, as well as to establish relationships between existing (authority) records, even if coming from different sources. The MET is aware of controlled vocabularies, hence supports archivists while editing controlled elements by proposing a drop down list with the terms defined as controlled EFG vocabulary.

Vocabulary Editor Tool. This curation tool allows the creation of vocabularies in the EFG Information Space. Terms can be increasingly added to a vocabulary, modified or deleted. Synonyms are managed as well. This tool is handled by the EFG's technical partner in charge of data ingestion and DIF.

Cleaning Rules Editor: This curation tool allows the definition of rules to be associated to an XPath and applied by the system to implement the cleaning phase of the aggregation workflow. It it used by EFG's technical partner in charge of data ingestion.

Tasks to be subcontracted:

- Hosting and maintenance of EFG database
- Participation in discussion and planning of issues concerning T1.1 "Data and aggregation infrastructure innovation".
- Ingestion work for new data providers and enriched data sets from existing providers
- Improvement of tools as described above

MUSEU

MICHAEL will review existing technical infrastructure. It will maintain MINT for 2015, with an extension until April 2016, if needed for the whole duration of the project, while working on the new workflow of METIS, under the Europeana DSI; and it will, where appropriate, make use of the technologies created for the Europeana Connection Kit.

Subcontract for aggregation platform integration & development: Rather than develop a separate aggregator workflow the museums aggregator MUSEU will become part of the METIS workflow in the framework of the Europeana DSI. Requirements for the data provision workflow for museums as part of METIS, will be developed in the course of 2015, under MICHAEL responsibility, with help from Collections Trust (under task 1.3 of this project). MCA will contribute to task 1.3 with the 'Requirements for Visualization and Indexing of Museum Content' that are being prepared as final contribution of the AthenaPlus project. They will also be the basis for the technical specifications to be included in the subcontracting.

To maintain the current infrastructure, using MINT, a subcontract to a suitable provider will be put in place to carry out the following work:

- Support the ingestion of new data and the data delivery to Europeana;
- The provision of a user-friendly LIDO profile to be used by cultural heritage institutions when mapping their data structure to the Europeana Data Model;
- After a review of workflow and feasibility a subcontract might also be needed for the configuration of the Europeana Connection Kit (ECK) to integrate with the METIS workflow.

OpenUp!

Freie Universität Berlin will improve the data flow that constitutes the OpenUp! data provision process in order to sustain the supply of metadata to Europeana DSI in the highest quality possible (see Fig. 1). In detail: (1) the data quality will be improved by adding additional data sources to the metadata enrichment process; (2) the BioCASE provider software will be adapted to ease data provision to the Europeana DSI; (3) the data harvesting and enrichment procedure will be adopted to the new BioCASE pipeline using HIT+ (advanced Harvesting and Indexing Toolkit), opening it up to further data standards and protocols; (4) in collaboration with subcontractor (A) the OpenUp! Natural History Aggregator will be further automated to allow batch processing of multiple sources with HIT+ and increase the efficiency of metadata enhancements in collaboration with subcontractor (B); and (5) FUB will investigate how LOD / Semantic Web and new database technologies can be integrated with the process.

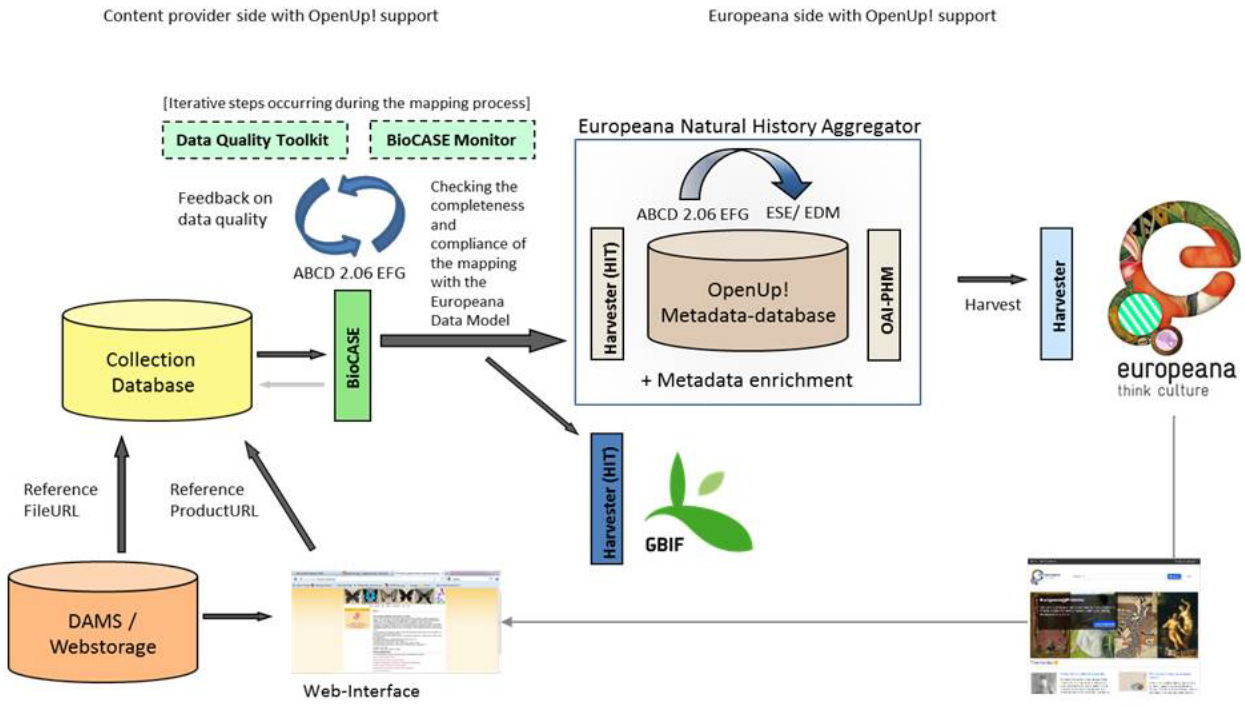

Fig. 1. Aggregation workflow of OpenUp!. The Biological Collection Access Service for Europe (BioCASE) is a transnational network of biological collections (http://www.biocase.org/). GBIF is the Global Biodiversity Information Facility (http://www.gbif.org/). The Access to Biological Collections Data (ABCD) Schema is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data).[1]

[1] http://www.bgbm.org/tdwg/CODATA/Schema/default.htm

The OpenUp! project created a software suite and workflow for the harvesting of community standard data, their conversion, enrichment and provision as EDM to the Europeana DSI. This “OpenUp! Natural History Aggregator” is contained in a Virtual Machine and can be operated at different places. For a sustainable operation of this process, outsourcing has proven to be the most cost-effective option and a subcontract will be put in place. The aggregation platform under Europeana DSI will continue to use the OpenUp! sustainability model, i.e. the operation is financed by the data providers under an SLA with the operator of the platform.

The subcontract will (i) support the switching from harvesting the data provider’s BioCASE web service record-by-record to harvesting the aggregated standard dump now provided by the BioCASE software as an option; and (ii) further automate the OpenUp! Natural History Aggregator with respect to metadata ingestion, transformation and enrichment workflow; (iii) investigate the consequences of scaling up the workflow to a steadily growing numbers of data providers. The subcontractor must also be an operators of the OpenUp! Natural History Aggregator platform.

CARARE

Athena RC will provide input to the understanding of user requirements for knowledge transfer and information sharing, based on its work on digital curation and experience in CARARE, LoCloud, Europeana Cloud, and also in DARIAH-EU Virtual Competency Centre.

Athena RC will build on the cloud version of MORe (Metadata & Object Repository) and extend it to support multiple intermediate schemas, multiple projects and providers per project, and to provide extensive statistics and reporting. The harvesting process will be made more user-friendly. Validation checks will be fully automated and pro-active, informing providers of possible mistakes or misconceptions. The automation of the ingestion process will have a great impact in sustainability as it will allow ingestion from a large number of providers with minimum resources. Pro-active checking and validation routines will save time and resources. Issues such as tracking down broken links and vocabulary enrichment are already provided by MORe, this process will be improved to include multiple schemas and more accurate enrichment services. New enrichment micro-services, that are of interest to archaeology and architecture will be incorporated into MORe. Finally, services for measuring and reporting metadata quality will be included.

Athena RC will contribute its expertise, and real world experience based on the actual implementation and operation of the MORe aggregator and service architecture, on ingestion, metadata enrichment and services, towards the specification and trialling of a distributed Europeana infrastructure.

A subcontract will be put in place for the maintenance of MINT for use by CARARE.

MORe Aggregator specifications

The MORe aggregator (as it has evolved through LoCloud) specifications include a scalable, elastic and de-centralized architecture where a micro-service oriented architecture is used to provide flexibility for the various aggregation tasks such as: validation, enrichment and publications. A flexible storage provides scalability and high availability for the services that also run on a scalable architecture.

MORe supports multiple metadata schema suchs as: Dublin Core, LIDO, EAD, EDM, ESE and others. It also provides flexible validation through a micro-service oriented architecture where the different validation micro-services are used on demand and per case. Example of validation micro-services include: schema validation, structural validation, link checking, Schematron rule validation.

Regarding enrichment, again a micro-service oriented architecture is employed where the various enrichment services can perform complex enrichment tasks such as:

- language identification
- geo-normalization
- geo-coding
- reverse geo-coding
- coordinate transformation
- automatic and semi-automatic vocabulary matching

These enrichment micro-services are orchestrated through enrichment plans. Enrichment plans provide a way for streamlining the execution of the services for each case separately.

Publication is not limited to OAI-PMH and Archive download but allows for direct publishing (through the respective APIs) to RDF stores, SolR index servers etc.

The aggregation workflow is shown in Figure 2.

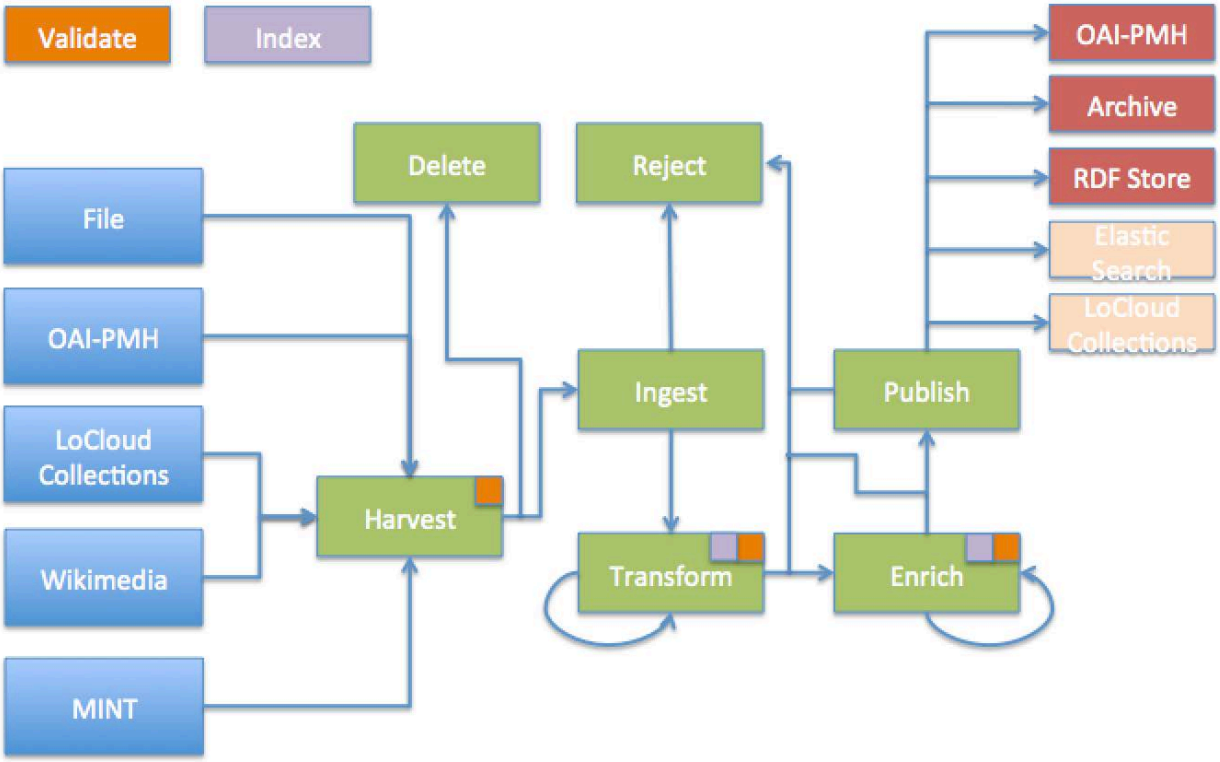

Fig. 2. MORe ingest workflow.

Europeana Fashion

In this task Europeana Fashion focuses on the improvement of the Europeana Fashion ingestion platform based on the open source platform MINT. In particular we will contribute to the improvement of:

- the usability of the metadata mapping tool, improving its UI. In this way metadata mapping from the native schemas to EDM/EDM-fp will be faster and more accurate;
- the usability and the functionalities of the group annotation tool used to “refine” set of records in a semiautomatic way. Allowing data providers to harmonize their metadata in a faster and more accurate way.

Europeana Fashion will also focus on improving the actual Europeana Fashion publication infrastructure (integrated with the MINT ingestion platform) in order to:

- automatically check and report for broken direct links pointing to the provided digital objects
- enrich ingested records through automatic linking to authority files (e.g. for fashion designers)

The cloud service provider will maintain and fine-tune the Europeana Fashion cloud infrastructure already in place (portal and digital objects repository). Estimation of the costs has been made looking at the first six months of service in 2014.

The subcontractor will maintain and support the MINT instance used for the Europeana Fashion aggregator and it will enhance it for the specific needs of fashion data, adding and improving the functionalities listed above. It will also implement and experiment automatic enrichment techniques (like colour extraction) on the fashion data.

EUscreen

The primary aim of this subtask is to maintain the technical infrastructure of the EUscreen aggregator and its connection to the central Europeana service (see Fig. 3). Next to this, two areas of development are identified as being essential to meeting the requirements outlined in the Europeana Publishing Guide[2]:

- The MINT metadata ingestion, mapping and aggregation tool. Checks and back-up possibilities will be 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 will expand its possibilities to cater for subtitle files as an additional form of (searchable and linkable) metadata. The information will be exploited through entity recognition and added to the linked data repository for access through the SPARQL end point, also for re-use in third-party platforms. This work is specifically relevant for Task 3.3. Establish content distribution partnerships. Further integration with EBUcore will ensure harmonization with the audiovisual industry. This is related to Subtask 4.5.2. Develop EDM. Publication of separate audiovisual segments.
- Media Fragment Aggregation Service for the publication of separate audiovisual segments. A fragment-based service is developed to make it possible to publish separate entities even if they are physically stored as one file. The best example is a news broadcast (for instance the evening news) that is currently stored with Europeana as one object, but basically contains several subjects. A great improvement in terms of relevance can be achieved by offering these subjects are different entities and linking from Europeana to the specific fragment inside the stream when users find media fragments on Europeana.

EUscreen will subcontract for three components related to the indexing part of the EUscreen aggregator. This specific work requires very specific technical expertise:

1. Support the addition of new data providers. Provide technical support for new aggregator partners to link their multimedia delivery infrastructure (i.e. streaming video) to the EUscreen aggregator
2. Keyframe extraction to increase quality of the Europeana service. An important element of the aggregated data, is the visual signifier (i.e. thumbnails extracted from video's) that are sent to the Europeana portal. At present, only a part of the objects send to the Europeana

[2] http://pro.europeana.eu/publication/publication-policy

DSI has keyframes. To address this issue, an automatic screenshot extraction service for videos pushed to Europeana. This service will work on all aggregated data provided by EUscreen. The quality of the thumbnail is further improved by applying an object recognition service that searches for a relevant image inside the video stream based on the title or other metadata of the clip.

3. Media Fragment Aggregation Service for the publication of separate audiovisual segments. Implementing the W3 Media Fragment recommendation into an aggregation service to make it possible to publish separate entities to Europeana even if they are physically stored at the partner as one file. The best example is a news broadcast (for instance the evening news) that is currently stored with the Europeana DSI as one object, but basically contains several subjects. A great improvement in terms of relevance can be achieved by offering these subjects are different entities and linking from the Europeana DSI to the specific fragment inside the stream if the user find this media fragment on Europeana. This streaming server will also support the Digital Rights Management framework that will be defined.

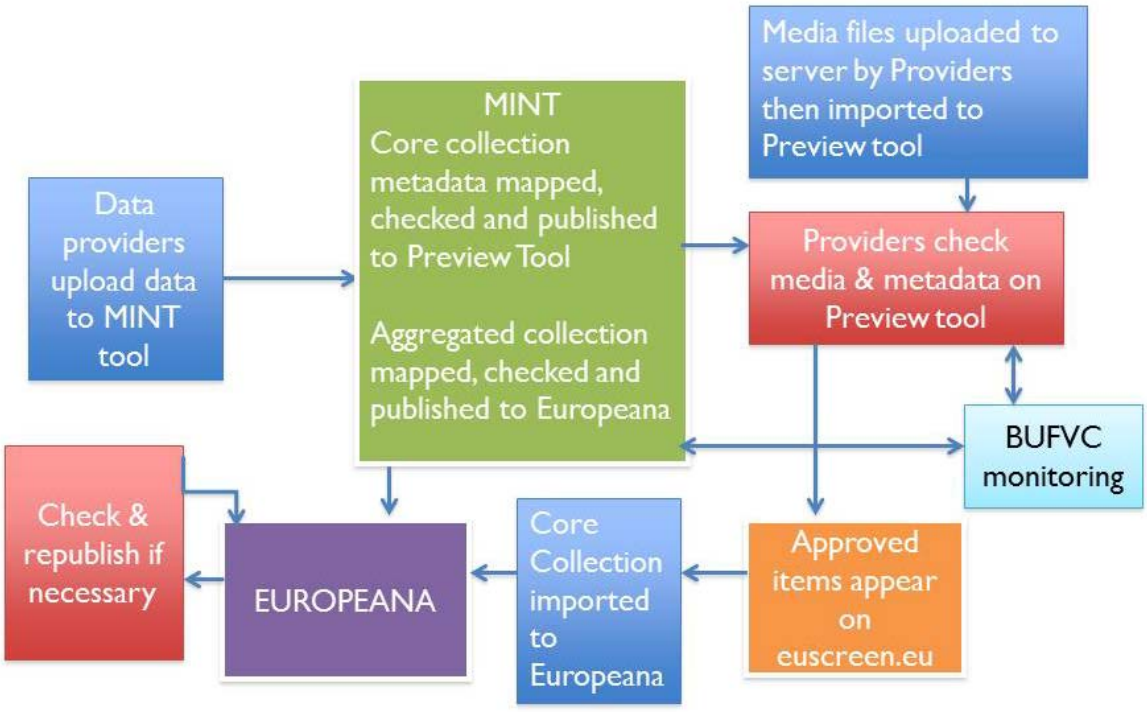

Fig. 3. Aggregation workflow of EUscreenXL.

Similar to the work in indexing (above) this requires very specific expertise, not available at Sound and Vision. The subcontract consists of four components related to the back-end of the EUscreen aggregator:

1. 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 will be used that facilitates data providers and data moderators to clean and normalise the data in a user friendly way. It will give 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. This work will take the outcomes of Subtask 4.5.2. Develop EDM into account.

2. Specific maintenance of the MINT metadata ingestion, mapping and aggregation tool. Checks and back-up possibilities will be 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 will expand its possibilities to cater for subtitle files as an additional form of (searchable and linkable) metadata. The information will be 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.
3. Infrastructure Innovation. Contribute to the activities initiated by Europeana in this area, 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.

HOPE2.0

IALHI will improve the current Social History Aggregators' infrastructure and tools to enable a better and more user friendly ingest process. IALHI will start with a short assesment phase for the current infrastructure and tools of SHAPE and the tools that are available in the Open Source Community or within the Europeana DSI Network. Based on the outcome of the assesment IALHI will decide on the tools and applications that will be used for upgrading the infrastructure and make it more sustainable. The current mapping procedures are complicated, time consuming and it requires quite an amount of human interaction and effort to map the data providers collections to EDM. Furthermore IALHI will focus on improving the Online Content Checker Tool and the Online Tagging Tool. Both tools are providing feedback to the data providers on their ingested collections and ways to enrich the data with themes applicable to collections in the field of Social History.
The Persistent Identifier Service that was developed during the HOPE project and is used by almost all of the current data providers will also be available to new data providers. If necessary the current service will be extended.

Conclusion

All DSI partners collaborating under the DSI use different sets of tools and operate different ingestion workflows that meets their needs. Based on the above inventory, some commonalities are obvious. The MINT platform is used or going to be used by a number of aggregators (MUSEU, CARARE, Europeana Fashion, EUscreen). It is essential to start coordinating MINT development efforts for each aggregator to avoid duplicating efforts. Although the development of domain specific services connected to MINT is core of the ongoing development, a close alignment of all partners working with MINT is necessary. Therefore, a coordination group is going to be set up, that not only makes sure that development activities are aligned, but also that training needs for data providing partners are coordinated (see MS8 - Plan for training and workshops of aggregators for data providing partners).

There are more components that some aggregators seem to have in common. Both European Film Gateway and Hope2.0 work with a content checker, for example. As a content checker is on the list tools that Europeana need to offer to all aggregators, it needs to be investigated how partners with experience in content checker tools can help to develope and implement such a tool for the wider network.

More commonalities might exist that are not that obvious from the above descriptions. An Aggregator Forum working group was set up that will investigate aggregator workflows in more detail and to also standardise workflow description for easier alignment between aggregators. This will also help to identify requirements and priorities for the development of METIS. METIS will become the shared infrastructure of the Europeana DSI for publishing with Europeana. With METIS in place, aggregator workflows will be standardised and aggregators don’t need to maintain separate infrastructures in order to allow for a continuous supply chain of cultural heritage data from their partners to the Europeana DSI.