Europeana – Core Service Platform

DELIVERABLE

D1.3: DESCRIPTION OF PLANNED WORK FOR THE AGGREGATORS ON IMPROVING THE DATA

Revision	1.1
Date of submission	25 March 2016
Author(s)	Henning Scholz (EF), Julia Welter (DIF), Maria Teresa Natale (MICHAEL), Marzia Piccininno (MICHAEL), Corinne Szteinsznaider (MICHAEL), Walter Berendsohn (FUB), Gisela Baumann (FUB), Kate Fernie (2Culture), Dimitris Gavrilis (AthenaRC), Marco Rendina (eFashion), Johan Oomen (NISV), Gariella Ivacs (IALHI), Anastasia Gasia (EF), Valentine Charles (EF)
Dissemination Level	Public

REVISION HISTORY AND STATEMENT OF ORIGINALITY
Revision History

Revision No.	Date	Author (Organisation)	Description
0.1	14/07/2015	Henning Scholz (EF)	Template and initial draft
0.2	27/07/2015	Henning Scholz (EF), Julia Welter (DIF), Maria Teresa Natale (MICHAEL), Marzia Piccininno (MICHAEL), Corinne Szteinsznaider (MICHAEL), Walter Berendsohn (FUB), Gisela Baumann (FUB), Kate Fernie (2Culture), Dimitris Gavrilis (AthenaRC), Marco Rendina (eFashion), Johan Oomen (NISV), Gariella Ivacs (IALHI), Anastasia Gasia (EF), Valentine Charles (EF)	Final draft
0.3	30/07/2015	Gordon McKenna (CT), Marie-Claire Dangerfield (EF)	Review
1.0	03/08/2015	Henning Scholz (EF)	Reviewers comments integrated, layout edits
1.1	25/03/2016	Henning Scholz (EF), Julia Welter (DIF), Maria Teresa Natale (MICHAEL), Marzia Piccininno (MICHAEL), Corinne Szteinsznaider (MICHAEL), Walter Berendsohn (FUB), Gisela Baumann (FUB), Kate Fernie (2Culture), Dimitris Gavrilis (AthenaRC), Marco Rendina (eFashion), Johan Oomen (NISV), Gariella Ivacs (IALHI), Anastasia Gasia, Nuno Freire (EF)	Extraction of minimum tasks, enrichment activities and additional tasks per aggregator; overview tables for all activities including indicators to measure progress over time

Statement of originality:

Table of Contents

Introduction

The variety of content that is accessible via the Europeana Portal is huge. Books, paintings, photographs, newsreels, sound recordings from more than 3,500 museums, libraries, archives and audiovisual archives from across Europe have been made accessible over the last few years, resulting in a corpus of more than 52.3 million metadata records (as of 25 March 2016). But not only the content itself is very diverse, the quality of the data a user can find in Europeana varies a lot across the Europeana database. When we say 'data' we mean both the metadata describing the cultural heritage objects and the digital objects linked from within the metadata. For more details about metadata quality in Europeana it is referred to the Report and Recommendations from the Task Force on Metadata Quality[1].

Improving data quality is the first priority for Europeana under the Europeana Strategy 2015-2020. It is clear from the work of the Metadata Quality Task Force how important the contribution of aggregators towards better data quality is. We see that aggregators that invest in more metadata checks achieve a higher quality compared to aggregators investing less in metadata checks. A much closer collaboration between Europeana Office and aggregators but also between different aggregators creates a better understanding of the needs for better data and how to achieve this. The set-up of the Europeana DSI is supporting this closer collaboration. Eight domain and thematic aggregators are full partners of the Europeana DSI to work more closely together on various aspects of aggregation, including data quality improvements. The purpose of this deliverable is to make plans for data improvements transparent between all aggregating partners of the DSI to enable a better alignment of activities and to address specific challenges together, using the experiences of other aggregators. In setting up this plan it also helps to advice aggregators on priority targets and potential risks of choosing a certain approach.

The following documents are meant to be a guide for good targets for quality improvements:
- Europeana Publishing Framework[2]
- Europeana Publishing Guide[3]
- Partner and Data Development Plan[4]
- Report and Recommendations from the Metadata Quality Task Force[5]

A few examples are given to illustrate potential targets for quality improvements, that are elaborated in more detail in the above documents. If aggregators want people to find, view and use content, however users want to do that, content has to be freely re-usable (PDM, CC0, CC BY, CC BY-SA), directly linked from the metadata and the technical quality needs to be high (images need to be 1,200 pixels wide, which is equivalent to about 1 megapixel). Enrichment is an important element that improves data quality, e.g. through the use of open vocabularies such as Iconclass and Getty Art and Architecture Thesaurus. Thumbnails (or other previews) should be made available for all digital cultural heritage objects. Metadata should have relevant, human-understandable fields, e.g. titles should provide meaningful access for the users of Europeana. Depending on the quality of existing data in Europeana, the aggregators focus their efforts on specific elements to improve their collections, which is specified in this document.

---

[1] http://pro.europeana.eu/publication/metadata-quality-task-force-report
[2] http://pro.europeana.eu/publication/publishing-framework
[3] http://pro.europeana.eu/publication/publication-policy
[4] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Version3/Mileston es/Ev3%20MS2%20MS4%20MS5%20Partner%20and%20Data%20Development.pdf
[5] http://pro.europeana.eu/publication/metadata-quality-task-force-report

Although quality targets should be measurable as much as possible, this document is not to elaborate on the new data to be ingested. The focus of the quality improvements listed and described below are the data submitted by the aggregators in earlier years.

Planned work for the aggregators

Minimum tasks for all aggregators

All Europeana DSI aggregating partners agreed a minimum set of data quality related tasks that will result in a visible improvement of the user experience. We know from user interviews and our experience with work in projects like Europeana Creative that access to the underlying digital content is important to all users of Europeana. A user exploring collections in Europeana prefers to have thumbnails to review search results before proceeding with looking into collections. When accessing objects, it needs to be easy to do so; ideally a direct link will be available. The rights information needs not just to be there but it needs to be correct for the user to know what use case is possible with that content. This is addressed in the Europeana Publishing Framework[6], which focuses on the value for data partners providing high quality data with thumbnails and direct links to objects (where media formats allow for this). With this framework in hand, it makes sense to focus with quality improvements on the metadata fields behind the four tiers of participation specified in the Europeana Publishing Framework. This way we are not only improving the user experience but provide value for our partners as well.

The metadata fields that are going to be addressed by all aggregating partners of the DSI at least until the end of the first year of the Europeana DSI are edm:isShownBy (direct links to objects, fix broken links), edm:object (thumbnails, incl fix broken links[7]) and edm:rights (rights statements). The range of media formats provided by the DSI partners mean that it will not be possible to provide direct access to underlying objects in any case, nor is it possible to generate thumbnails for all media types. So the targets set for each of the aggregators in Table 1 will reflect the media formats being provided (there are limitations in linking to sound, 3D and video objects).

It is important to monitor progress and keep track on data quality improvements. Correcting rights statements might be ambiguous and difficult to measure, but we can easily keep track of the change of rights statement per aggregator, assuming that all changes aggregators are implementing follow the goal of correcting existing rights statements.

Measuring the technical quality of the underlying content and monitoring progress of improving it has become a lot easier since Europeana is extracting technical metadata and make them available for search via the API. This way a user can not only find content that meets specific technical quality targets, but also to generate statistics and reports about the quality of the content provided to Europeana. This works across all media types: It is possible to find images larger 4 megapixel, text that is searchable or HD videos longer than 20 minutes. In extracting the technical metadata Europeana also collects information about links being broken or URL not being responding and the reason for failing with extracting technical information. This is to some extent still work in progress in the sense that not all reports are customised and available for data

---

[6] http://pro.europeana.eu/publication/publishing-framework

[7] Europeana is also able to create thumbnails out of edm:isShownBy, which means that for image collections a working direct link to a media file in edm:isShownBy will cover two quality targets at once, the direct link to the object and the thumbnail. The same works for texts in pdf.

partners yet. Once this is completed, all data partners will regularly have complete statistics and reports about the technical quality of their data to monitor progress and prioritise future activities. With this in place we have the right tools to consistently measure improvements. We can calculate all kinds of indicators based on the statistics and agreed to report on the number of records that have been improved per aggregator and the number of improved records as a percentage of the total records per aggregator. If data quality improvements happen across a full datasets, we can showcase them for others to view the improved records. We will also follow the Europeana Publishing Framework and feature improved collections using the channels that are suitable for the datasets.

Table 1. Targets for corrected rights statements, direct links to media files and thumbnails per aggregator. For a full description of how aggregators will approach this and when they plan to deliver results, see the task by task description further below in the document.

Aggregator	Rights Statements	Direct Links	Thumbnails
EFG	90% of items validated with the providers and delivered to Europeana with rights statements proposed by the provider	85% of all EFG items will have a direct link	100% of EFG collection will be checked for missing thumbnails, 85% of EFG items will have a thumbnail.
MUSEU	100% valid rights statements	50% minimum of the thumbnails and the direct links to verify broken links through the Europeana Media Discovery. Provision of edm:isShownBy for all data partners that are able to provide direct links	88,6% of the providers provided the edm:object field (over 6M records). 50% minimum of the thumbnails checked to verify broken links
OpenUp!	100% of OpenUp! collections with valid rights statement, 20% of provided objects freely reusable	100% of OpenUp! collections will be checked for broken links, 40% of records with direct links	100% of OpenUp! collections will be checked for missing thumbnails, 50% of missing items added
CARARE	100% valid rights statements	CARARE's collection of 2,239,699 items will be checked for broken links, 60% with direct links in the isShownBy property.	Metadata mappings are being corrected to address issues with the display of thumbnails, 100% of thumbnails will be checked for size; 60% of records will have thumbnails.
Europeana Fashion	By the end of DSI-1, 100% of aggregated items will have a valid rights statement. More than 15% will use a Creative Commons license	100% of the aggregated content (about 700,000 records) will have working direct links in the isShownBy property	100% of the aggregated content (about 700,000 records) will have working thumbnails

Aggregator	Rights Statements	Direct Links	Thumbnails
EUscreen	100% of items will have a controlled rights statement to clarify the terms of access and re-use relating to the digital object. Over 19.000 digital objects are labelled under open Creative Commons licenses.	100% of items will have links to digital objects, ca. 90% of which is video.	75% of items will have an illustrative preview (thumbnail image) for their content.
HOPE2	100 % of digital objects will have standardised rights statements	100% of the 759,323 records will be validated, 100 % will be corrected to have direct links in the isShownBy property.	100% of the digital objects will have working thumbnails
The European Library	10%-20% of the TEL collections checked, validated with the provider and delivered to Europeana with updated rights statements	30%-50% of the TEL collections checked and delivered to Europeana with better images (isShownBy)	Improved isShownBy will lead to the same extent also to improved thumbnails

Semantic enrichment activities

Enrichment activities are included in the full description of work per aggregator further below, but we agreed to present our semantic enrichment activities in a more consistent and harmonised way to have an aligned overview of work that is planned and outcomes that can be expected. This also helps to develop a common understanding of what we mean by enrichment and what vocabularies are used for what purpose by the DSI aggregating partners. For more context and a full overview of enrichment and evaluation, the Europeana Tech Task Force on Enrichment and Evaluation has concluded the work and published a report.[8]

Semantic enrichment activities always include the use of vocabularies. We differentiate two different types of semantic enrichment. (1) making vocabulary links in the metadata explicit by adding URI's to the vocabulary mentioned in the metadata already, (2) creating entirely new semantic enrichments for metadata where no vocabulary is mentioned explicitly in the metadata by recognising suitable vocabularies for the data. The risk of errors for type 1 semantic enrichment is very low and validation is not necessary. For the type 2 semantic enrichment, validation is required.

---

[8]

http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation//FinalReport_EnrichmentEvaluation_102015.pdf

Table 2. Overview of vocabularies used by the DSI aggregating partners to enrich their data.

Aggregator	Vocabulary	ML / AE[9]	Usage in data (fields effected)	Status quo (number of enriched records)	Target for May 2016 (number of enriched records)
eFashion	EF thesaurus (based on AAT, with fashion extensions)	AE	dc:type; dcterms:medium; edmfp:technique; gr:color	0[10]	at least 300,000 records enriched through semantic analysis; at least 430,000 records enriched through image analysis
TEL	TGN[11]	ML	dc:subject; dcterms:coverage; dcterms:spatial	0	1001 records
TEL	AAT[12]	ML	dc:subject	0	8188 records
TEL	Iconclass[13]	ML	dc:subject	0	36823 records
MUSEU	AAT	ML	dc:type, dcterms:medium	0	Min. 180,000 records, potentially 1 million.
MUSEU	TGN	ML	dcterms:spatial, edm:currentLocation		
MUSEU	TGN	ML	dcterms:spatial, edm:currentLocation		Min. 180,000 records, potentially 1 million.
MUSEU	ULAN[14]	ML	dc:creator, dc:contributor, dc:subject		Min. 180,000 records
EUscreen	IPTC[15]	ML	dc:subject	37,964 records	60,319 records
HOPE2	Social History Thesaurus	ML	dc:subject		
CARARE	AAT	ML	dc:subject	0	Min. 100,000 records, potentially more
CARARE	Perio.do[16]	ML	dc:temporal	0	Test on 30,000 records potentially more if it works with edm timespan
CARARE	-	AE		0	Test on 10,000 records potentially more

---

[9] Materializing existing metadata strings as links ('ML', type 1 semantic enrichment) or automatic enrichment with new references ('AE', type 2 semantic enrichment).

[10] In Europeana no record has been enriched. In europeanafashion.eu enrichments have been tested on about 50,000 records for semantic enrichment and 100,000 for image analysis enrichment. Results needs to be validated. Precision for enrichment: 90% for semantic enrichment; hopefully 70% for image analysis enrichment (still under study).

[11] http://vocab.getty.edu/tgn/

[12] http://vocab.getty.edu/aat

[13] http://www.iconclass.org/help/lod

[14] http://vocab.getty.edu/ulan/

[15] http://thesaurus.euscreen.eu/EUscreenXL/v1/

[16] http://perio.do/

To facilitate the future collaboration on enrichment activities, all aggregators will list the vocabularies they are using for semantic enrichment (see Table 2) and add informal notes on the quality of these vocabularies (in particular if very domain specific vocabularies are used that are not known outside a domain). We will inventorise the metadata of vocabularies (name, URL, access conditions) in order to prepare for a more serious quality assessment work of vocabularies in the future. Every aggregator will keep track of the semantic enrichment work and regularly report the number of records that have been enriched by what vocabulary. If type 2 semantic enrichment will be performed, every aggregator will pick a good sample of the enriched data and evaluates how many percent of data are correctly enriched. This validation is work intense as it requires appropriate methods and human raters to carefully check the enriched data, so the sample size has to be chosen wisely to make the task feasible.

Overview of additional tasks per aggregator

The next chapter is giving a full description of the work planned by the Europeana DSI aggregating partners within the first year of the DSI. Most aggregators do not only complete the minimum tasks and invest in semantic enrichment of their data, but do a lot more to improve the quality of the data. These additional tasks to improve the data are selected based on the needs of the domain and the quality of the underlying data provided previously to Europeana.

Most partners started their work with an evaluation task, to fully understand the gaps they need to fill with regard to data quality (EFG, MUSEU, CARARE, Europeana Fashion, HOPE2). Some data partners decided to invest in the quality of the titles and/or descriptions of their data to make them more meaningful and human readable, to ensure that that the digital cultural heritage objects submitted previously can be discovered more easily (EFG, OpenUp!, CARARE, Europeana Fashion, HOPE2). Deduplication and normalisation work has been proposed by a few aggregators (TEL, Europeana Fashion) to clean up the database. A very specific activity is proposed by EUscreen, to secure that all objects contributed to Europeana by EUscreen have keyframes.

Full description of work per aggregator

In preparing for this data quality plan, every aggregator provided a detailed description of all work that contributes to improving the quality of metadata and content. This description of work is structured in tasks according to the needs of each aggregator. In any case it includes at least the minimum tasks specified above and describes the enrichment activities in more detail. It also indicates dates of completion for each task and deliverables that aggregators foresee as an outcome of their work.

1-European Film Gateway (EFG)

Description of work
EFG will clean and enrich metadata coming from the film archive domain to EFG and Europeana. The focus of enrichment work will be on improving discoverability of digital objects by enriching keywords and content descriptions. To support the user experience on EFG and Europeana, EFG archives will be encouraged to provide direct links to digital objects and thumbnails, where this is not already the case, and to improve the rights labelling of individual objects so that their rights status becomes more transparent to users. Together with the EFG partners from the film archives, Deutsches Filminstitut (DIF) will work on these tasks. In addition, EFG seeks to connect new archives as well as new collections from existing EFG archives.

Subcontracting	
Task (Europeana DSI)	1.5.2 Improve the Europeana DSI data and metadata quality
	1.7.2 Develop, establish and maintain technical infrastructure
Amount (EUR)	48.000 EUR
Description	

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 (e.g. Metadata Editor Tool).

Task	T01.1 Evaluation of possible contributions and data improvements through EFG archives
Description	Association des Cinémathèques Européenes (ACE) and DIF carried out an EFG information meeting to inform all ACE member archives connected or not yet connected to EFG about the activities under the Europeana DSI. The meeting was used to bring attention to data quality issues raised by Europeana under the DSI. As a follow-up of this meeting held in Bologna on the 3rd of July 2015, ACE circulated a survey among all ACE member archives to find out about what they offer to work on under the Europeana DSI, where ACE has reserved a dedicated subcontracting budget for to pay ACE member archives and others for their activities. Deadline for this survey is 14 August 2015.

	The input provided by the archives will serve as the basis for further negotiations with the archives about concrete tasks to be carried out. At the end of the negotiations, there will be a concrete activities plan for all ACE archives that have indicated their interest to contribute to the Europeana DSI. All further tasks described for EFG in this deliverable depend on the outcome of the individual negotiations with the archives. Once the negotiations are finished, it is also possible to provide concrete numbers of objects to be enriched or newly provided.
Start Month	03[17]
End Month	11
Milestone(s)	M01.1 EFG Information meeting held (Month 03) M01.2 Survey circulated (Month 03) M01 3 Surveys evaluated and tasks identified (Month 05) M01.4 Data enrichment and contribution activities by archives to EFG completed (Month 11)
Deliverable(s)	D01.1 Activity plan for EFG archives (Month 07) D01.2 Delivery of improved metadata and new collections to EFG (Month 11)

Task	T01.2 Improvement of rights labeling
Description	Metadata in EFG is completely rights-labelled but not in all cases correctly. This is especially true for contributors who delivered larger amounts of content without rights labels within the original EFG project (2008-2011). In these cases the status was set to the default value "rights protected" for all their content. Seeing that the majority of film-related is still in copyright today, the number of items incorrectly labelled can be considered comparably small. Nevertheless, EFG will aim to specify rights labels more clearly, especially with regard to the public domain. This activity is dependent on the support of the archives contributing the content in question, as re-labelling of rights cannot be done without the individual content provider.
Start Month	04
End Month	11
Milestone(s)	M01.5 Identification of all objects that need relabeling and of archives that have to provide the respective information (Month 04) M01.6 Start of rights labelling (Month 06) M01.7 Identified incorrect rights labeling adjusted for correct display on Europeana (Month 13)
Deliverable(s)	D01.3 Rights labels corrected where possible and ingested into EFG (Month 12)

Task	T01.3 Direct links

---

[17] 00=Apr 2015, 01=May 2015, 02=Jun 2015, 03=Jul 2015, 04=Aug 2015, 05=Sep 2015, 06=Oct 2015, 07=Nov 2015, 08=Dec 2015, 09=Jan 2016, 10=Feb 2016, 11=Mar 2016, 12=Apr 2016, 13=May 2016

Description	EFG is already providing the majority of its content with a direct link. Under this task EFG will evaluate the exact numbers of IsSHownBy links provided so far (estimates: c80% of the content coming to Europeana from EFG has an IsShownBy link. c70% of videos aggregated by EFG have a direct link.) EFG will continue to communicate the need for direct links to the archives and engage in increasing the amount of items, especially video, with an IsShownBy link and identify archives that can provide edm:isShownBy in addition to edm:isShownAt. DIF will encourage all contributing archives to deliver edm:isShownBy where archives do not claim strategic reasons for not doing so, or where technical obstacles hinder the delivery of edm:isShownBy. In the latter case, DIF will support archives in finding solutions to overcome these obstacles if possible.
Start Month	04
End Month	11
Milestone(s)	M01.8 Identification of missing edm:isShownBy (Month 04) M01.9 Negotiation with respective archives about possibility to provide edm:isShownBy (Month 05) M01.10 Delivery of edm:isShownBy through involved archives to EFG completed (Month 11)
Deliverable(s)	D01.4 Delivery of edm:isShownBy completed for all EFG archives involved in the ACE subcontract that are able to provide direct links to EFG (Month 12)

Task	T01.4 Providing previews (thumbnails)
Description	ACE and DIF will put a focus on the provision of previews (thumbnails). There are still a number of especially video collections in EFG that do not provide them. In contrast to images and texts where thumbnails can be created automatically by the EFG and Europeana system, this does not work for video files. In the case of videos, the contributing archive has to create thumbnails and deliver them in the form of a link in the exported metadata. DIF will contact the partners in question and encourage them to provide thumbnails if they can.
Start Month	04
End Month	11
Milestone(s)	M01.11 Identification of collections missing thumbnails completed (Month 04) M01.12 Negotiations with archives completed that need to provide thumbnails for videos (Month 06)
Deliverable(s)	D01.5 Delivery of links to thumbnails to EFG completed by archives involved (Month 12)

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
European Film Gateway															
Task															
T01.1	*Evaluation of possible contributions and data improvements through EFG archives*				■	■	■	■	■	■	■	■	■		
Deliverables															
D01.1	*Activity plan for EFG archives*							■							
D01.2	*Delivery improved metadata and new collections to EFG*												■		
Milestones															
M01.1	*EFG information mtg. held*				■										
M01.2	*Survey circulated*				■										
M01.3	*Surveys evaluated and tasks identified*					■									
M01.4	*Data enrichment and contribution activities by archives to EFG completed*												■		

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
European Film Gateway															
Task															
T01.2	*Improvement of rights labelling*				■	■	■	■	■	■	■	■	■		
Deliverables															
D01.3	*Rights labels corrected where possible & ingested into EFG*												■		
Milestones															
M01.5	*Identification of objects that need relabeling and of archives that have to provide the respective information*					■									
M01.6	*Start rights labelling activities*							■							
M01.7	*100% of identified incorrect rights labeling adjusted for correct display on Europeana*														■

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
European Film Gateway															
Task															
T01.3	*Direct links*					■	■	■	■	■	■	■	■		
Deliverables															
D01.4	*Delivery of edm:isShownBy completed for all EFG archives involved in the ACE subcontract that are able to provide direct links to EFG*												■		
Milestones															
M01.8	*Identification of missing edm:isShownBy*					■									
M01.9	*Negotiation with respective archives about possibility to provide edm:isShownBy*						■								
M01.10	*Delivery of edm:isShownBy through involved archives to EFG completed*												■		

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
European Film Gateway															
Task															
T01.4	*Providing previews (thumbnails)*					■	■	■	■	■	■	■	■		
Deliverables															
D01.5	*Delivery of links to thumbnails to EFG completed by archives involved*												■		
Milestones															
M01.11	*Identification of collections missing thumbnails completed*					■									
M01.12	*Negotiations with archives completed that need to provide thumbnails for videos*							■							

2-MUSEU

Description of work

MUSEU represents three main sources of records for Europeana - the former projects ATHENA, Linked Heritage, and AthenaPlus. MUSEU will guarantee the sustainability of the records of these projects (and their presence in Europeana) over Europeana DSI. Moreover, MUSEU will involve new datasets and data providers within Europeana DSI.

The metadata improvement process will be applied to the new records. Former ATHENA, Linked Heritage, and AthenaPlus partners will be involved on a voluntary basis to enrich and refine the data they have already supplied with.

A helpdesk and training service will be available for the whole length of the project that will support the content providers in the:
  - mapping process to extent the use of the LIDO elements and also of the normalization functionalities of MINT in order to improve metadata consistency;
  - enrichment of the metadata with the Getty vocabularies AAT, TGN, and ULAN in order to increase the linking of Europeana content to the most relevant LOD sources for the museum sector;
  - application of licenses suitable for the reuse of digital content.

Museum sector requirements (Task 02.1)

MUSEU will address the improvement of the museum data quality, in particular in terms of better visualization and indexing in Europeana. This will be done by promoting the '***Requirements for Visualization and Indexing of Museum Content***' elaborated within the AthenaPlus project. Using these requirements for improving the quality of museum data on the provider's side as well as their representation in Europeana will enhance searches for museum items, and hence their accessibility in Europeana (the **Museum Collections Pilot** with enhanced and quality metadata is available at http://54.247.69.120/build/movio_musem/).

The requirements were created to meet the needs of the museums in the Europeana environment, following the Europeana metadata quality recommendations. They will also empower the user's visual experience through the provision of multiple and better quality images, and also by dramatically improving the quality of metadata with the use of well established vocabularies that will make the retrieval process more accurate. The following activities take place:
  - definition of the museum sector requirements in terms of expressivity and quality of metadata using LIDO as common framework;
  - implementation of an EDM-LIDO profile in the MINT platform and mapping of the LIDO records to it;
  - integration of the museum requirements in the MINT aggregation platform in order to give content providers the appropriate tool to deliver high quality data;
  - configuration of the Europeana Connection Kit to allow a wider range of museums to submit content to Europeana enriched in their data back into their core systems.

MUSEU experts work with the Europeana R&D team for the inclusion in EDM the event type.

Metadata improvement (Task 02.2)

The following activities address the minimum tasks for the aggregators as requested by Europeana:

➢ **Direct links to digital objects**

MUSEU asks to its own providers to include direct links because this allows the final users to have a better experience when using Europeana and the objects to be re-used in creative ways.

As to the legacy records, we estimate that 52.4% has the <edm:IsShownBy> value. The percentage is higher (66.2%) for the most recent datasets provided within the AthenaPlus project because the other records are mostly in the ESE format and were provided where less strict rules in terms of quality were in force.

The Europeana statistics, although they do not cover yet the whole sum of the MUSEU records, shows that some are not working or are not direct. The use of these statistics (the MUSEU platform can't detect the broken links) are fundamental to clarify the entity of the work to be addressed.

As to the provision of new datasets, it was agreed that the <edm:IsShownBy> field can't be set as mandatory for the content providers because this is not always technologically or legally possible because:

- sometimes content providers are not in the position to give such a value because they do not want to make the files downloadable (like for the recordings and video files);
- some data management systems can't generate the URL corresponding to the <edm:IsShownBy>;
- the digital objects are part of collections and the provider wants that the users get even more useful information seeing the context of the object in the collection as a whole.

Nevertheless MUSEU will continue to inform the content providers about the necessity to provide, whenever possible, direct links to their digital objects applying the <edm:IsShownBy> label in addition to <edm:IsShownAt>.

➢ **Thumbnails**

MUSEU estimates that the over the 88% of the providers used the <edm:object> to provide the URLs for the thumbnails. Through the Europeana Media Discovery Tool MUSEU will check the 50% minimum of the thumbnails to verify broken links.

➢ **Rights labeling**

All records that the MUSEU providers published in Europeana within the ATHENA, Linked Heritage and AthenaPlus projects (2008-2015) have a correct rights label.

For the legacy records (1,982,215 from ATHENA and 1,332,927 from Linked Heritage) the consortium has already put in place a strategy of rights correction and integration when the value was missing following the campaign that Europeana launched in February 2014 on the occasion of the change of some rights statements. Some of the legacy records in fact did not display the <edm:rights> value because it was not mandatory when they were published. All the values that these legacy datasets now display in Europeana was added or modified accordingly to the current rights statements.

All last datasets published within AthenaPlus (3,584,788 items) have the <edm:rights> value.

Sometimes the content providers accepted the modification Europeana proposed to the applied value, when this was considered as appropriate, some other times the providers themselves asked to change the <edm:rights> in the records already published into a new value.

In DSI we plan to update the rights label of only one dataset (63 records) the Lithuanian Art Museum provided in 2010 under the umbrella of the ATHENA project. The value will switch from Rights Reserved-Free access to CC-BY 4.0.

All the new records that will be published in Europeana during the DSI project will have suitable rights statements. The MINT platform for the data aggregation embeds all the 13 current values so that the content providers can easily chose and apply the correct one.

The MUSEU helpdesk provides constantly information about the 13 rights statements, also during the training workshops.

Semantic enrichment (Task 02.3)
Taking into account considerations from both Europeana's side, as expressed by their task forces on a Multilingual and Semantic Enrichment Strategy and on Metadata Quality, and from the data providers' side, we identified the following important factors for a successful, high-quality and sustainable enrichment of metadata with links to external sources:
1. The enrichment should be adopted as early as possible in the process of metadata production.
2. The enrichment should be based not only on string matching mechanisms, but rather exploit further contextual information.
3. The enrichment workflow should allow the data providers to validate and use the enrichment results according to their own specific criteria.

The core strategy we therefore are following is to support the enrichment of underlying local (i.e. on the providers' side) terminologies with links to the external sources, with the ultimate goal to feed these links into the metadata provided, both updating existing metadata and enriching any newly produced metadata.

To this end, we support the enrichment of local authority data with two approaches. On one hand, we developed a web service for linking to two Getty vocabularies, the Union List for Artist Names for agents, and the Getty Thesaurus for Geographic names for places.
The evaluation of the web services in the Athena Plus project suggests that the linking process to ULAN is of high validity, though only using the life dates of an agent as additional context information to the preferred and alternative names. Decreasing confidence values as well as the number – one or multiple – of matches returned by the service clearly correlate with imprecise or incomplete data in either the source or the target record(s). For the linking process to TGN, however, further refinements are suggested as the context information as currently exploited – continent, nation, and other places the actual place in question is part of – is not sufficiently ensuring unique and valid results. In any case, the decision for which confidence level the results of the automatic matching are fed back into the local source database is up to the data providers, based on their own quality criteria and quality checks.

However, the use of these web services requires specific technical steps in the source database which are not easy to perform for the data provider. We therefore support on alternative approach in the MINT platform: When performing a mapping to LIDO in MINT, the data provider can use the value mapping functionality of MINT for certain LIDO elements, which are connected in MINT to subsets of AAT or TGN. The provider can reuse these mappings for updating datasets, and as well for providing new datasets when they have the same data structure than previously provided datasets.

In general, the strategy of enriching underlying terminologies, particularly in the local metadata production environment, with links to external sources proves to be very efficient for increasing the number of links and thereby the quality of the metadata.
Europeana benefits in the most efficient and sustainable way from the linking results, i.e. by

harvesting records that have been annotated by the semi-automatic way here described that guarantees good results.

MUSEU plans to do an enrichment work with two content providers, Bildarchiv FotoMarburg (Germany) and Nationalmuseum (Sweden) using the TGN, ULAN and AAT vocabularies. Their experience will be a driving factor for other content providers in DSI-2.

Task	T02.1 Museum sector requirements
Description	MUSEU will revise the museum sector requirements as defined within the AthenaPlus project in order to enhance the accessibility of the museum content in Europeana. The museum sector requirements are structured on the LIDO schema as common framework that allows to show the metadata in their full expressivity. The LIDO to EDM mapping will complete the work and assure the integration in the Europeana workflow.
Start Month	02
End Month	07
Milestone(s)	M02.1 Release of the museum requirements (Month 10) M02.2 Workshop about the requirements and LIDO and their application in MINT (Month 10)
Deliverable(s)	D02.1 MUSEU data quality museum requirements and LIDO-EDM mapping (Month 10)

Task	T02.2 Metadata improvement
Description	Old and new data providers will supply Europeana with metadata records displaying richer information. This improvement will happen taking into account the common agreed tasks. **a)** Direct links to digital objects (edm:IsShownBy) and **b)** Thumbnails (edm:object): although the 88.6% of the MUSEU content partners provided the edm:object fields, MUSEU will identify the data providers that did not give the <edm:IsShownBy> value and promote its application if there are no technical or political constraints. As to task **c)** Rights labeling (edm:rights), MUSEU will constantly monitor the application of the correct value and provide support when needed. The quality of the metadata will be also enhanced through: - a more extensive use of the LIDO elements and also of the normalization functionalities of MINT in order to improve metadata consistency and compliance to the Europeana guidelines on metadata quality - the use of the standard and multilingual vocabularies, in particular the Getty vocabularies such AAT and TGN that empower the retrieval process These activities will be done using a new and dedicated MINT instance shaped on the basis of the museum sector requirements as described in MS9. MICHAEL will provide through MUSEU-Hub training and helpdesk support for this task as

	presented in MS8.
	MICHAEL will ensure that the Europeana Connection Kit (ECK), together with its 'dark aggregator' continues to be supported and potentially be integrated into Europeana's workflow. Quality is ensured for those using the ECK because of its control of metadata (based of SPECTRUM) and in particular of IPR licensing labelling.
Start Month	03
End Month	13
Milestone(s)	M02.3 Aggregation starts (Month 08) M02.4 50% minimum of the thumbnails and the direct links to verify broken links through the Europeana Media Discovery Tool (Month 13) M02.5 Provision of edm:isShownBy for all data partners that are able to provide direct links (Month 13)
Deliverable(s)	D02.2 Report on MUSEU aggregation and metadata improvement (Month 13)

Task	T02.3 Semantic enrichment
Description	Enrichment work through external web services on the providers' side, Bildarchiv Foto Marburg (Germany) and Nationalmuseum (Sweden), using the TGN, ULAN and AAT vocabularies. The minimum number of enriched items is ca. 180,000.
Start Month	10
End Month	13
Milestone(s)	M02.6 180,000 enriched records (Month 13)
Deliverable(s)	D02.3 Report on semantic enrichment (Month 13)

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
MUSEU															
Tasks															
T02.1	*Museum sector requirements*			■	■	■	■	■	■						
T02.2	*Metadata improvement*			■	■	■	■	■	■	■	■	■	■	■	■
T02.3	*Semantic enrichment*											■	■	■	■
Deliverables															
D02.1	*MUSEU data quality museum requirements*											■			
D02.2	*Report on MUSEU aggregation and metadata improvement*														■
D02.3	*Report on semantic enrichment*														■
Milestones															
M02.1	*Release of the museum requirements*											■			
M02.2	*Workshop about the requirements and LIDO and their application in MINT*											■			
M02.3	*Aggregation starts*									■					
M02.4	*50% minimum of the thumbnails and the direct links to verify broken links through the Europeana Media Discovery Tool*														■
M02.5	*Provision of edm:isShownBy for all data partners that are able to provide direct links*														■
M02.6	*180,000 enriched records*														■

3-OpenUp!

Description of work
Freie Universität Berlin (FUB) will facilitate the standardisation of the metadata provided. Data provision in OpenUp! has quite strictly adhered to the indications given by the Europeana DSI operators, so most if not all OpenUp! records fulfil the criteria set by the guidelines and policies in place. OpenUp! does provide native EDM; OpenUp!'s Natural History Aggregator checks the URL's of providers with every harvest, and there is a continuously working link-checker in place that flags subsequent errors. All multimedia objects have policy-conformant rights-statements, and the records that contain coordinate data are enhanced with place names in the aggregation process. The application of the existing enrichment measures (addition of common names of organisms, synonyms of scientific names, place names etc.) as well as the control of licensing and the technical quality control are included in the current sustainability setup of OpenUp!. Enrichment can be further improved by strengthening the data basis of the web services provided by OpenUp! partners. All work on data quality improvements will be done in close collaboration with the data partners that provided the data. This includes lots of communication on data quality between all parties including the subcontractor and the data partners.

Subcontracting	
Amount (EUR)	20,000

Access to digital objects in the Europeana DSI is substantially facilitated by enriching the metadata provided with synonyms and multilingual translations. For natural history objects, the names of organisms provide an entry point for the general public and scientists alike. Common names are of particular importance for the general public. During OpenUp!, a web service was created providing common names of organisms when fed with their scientific names. Currently, 470,000 common names in 250 languages from 28 original sources, enhancing the data for 150,000 scientific organism names are included and available for the Europeana DSI. This greatly enhanced access, because the collections themselves do provide only scientific names with their objects. The subcontractor will further strengthen the data basis of the Web service that is used to enhance the metadata in the OpenUp! Aggregator. FUB expects the subcontractor to include at least 20 additional data sources. The actual number of new common names is difficult to estimate; currently the data sources provide an average of more than 16,000 names each. The new sources will probably somewhat smaller, because a special accent will be placed on common names in languages currently under-represented in the Europeana DSI (currently, the bulk of the data are in French, Greek, Italian, Polish, and Portuguese).
The subcontractor must be willing to sustain the service in the long term and needs to be sufficiently ingrained in the natural history community to obtain and judge the quality of new data sources.

Task	T03.1 Integration of additional web sources
Description	Integration of additional web sources into the *CommonNames* web service to facilitate access to natural history content in more languages, e.g. in Dutch, Icelandic, and Chinese
Start Month	04

End Month	12
Milestone(s)	M03.1 Ten additional web sources integrated (Month 08) M03.2 Twenty additional web sources integrated (Month 12)
Deliverable(s)	D03.1 Documentation of additional integrated sources (Month 13)

Task	T03.2 Integration of often queried generic terms
Description	Unlike scientists, interested users of Europeana DSI may query natural history objects not by scientifically correct names but by generic terms like 'mouse', and 'chicken'. To satisfy this demand, often queried generic terms will be integrated into the *CommonNames* web service and linked to relevant scientific names. Important terms include for example domestic animals and well-known vermin.
Start Month	04
End Month	12
Milestone(s)	M03.3 Additional generic terms integrated and linked (Month 12)
Deliverable(s)	D03.2 Documentation of additional linked terms (Month 13)

Task	T03.3 Implementation *All-in-one-Service*
Description	The export of common names to enrich OpenUp! metadata will be altered in cooperation with the OpenUp! aggregator. A service will be implemented that allows requesting common names and their sources via a terminal list of scientific names and downloading the results all-in-one shortly after the request. Thus, metadata enrichment is expected to operate more independently.
Start Month	04
End Month	12
Milestone(s)	M03.4 First delivery of common names via *All-in-one-Service* (Month 05) M03.5 Second delivery of common names via *All-in-one-Service* (Month 09)
Deliverable(s)	D03.3 *All-in-one-Service* running stable and with high performance (Month 13)

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
OpenUp!															
Tasks															
T03.1	*Integration of additional web sources*					█	█	█	█	█	█	█	█	█	
T03.2	*Integration of often queried generic terms*					█	█	█	█	█	█	█	█	█	
T03.3	*Implementation All-in-one-Service*					█	█	█	█	█	█	█	█	█	
Deliverables															
D03.1	*Documentation of additional sources*														█
D03.2	*Documentation of additional terms*														█
D03.3	*All-in-one-Service running*						█	█	█	█	█	█	█	█	█
Milestones															
M03.1	*10 additional sources*									█					
M03.2	*20 additional sources*													█	
M03.3	*Additional terms integrated*													█	
M03.4 & M03.5	*Delivery of common names to the OpenUp! Aggregator via All-in-one-Service*						█				█				

23

4-CARARE

Description of work
2Culture will work with the CARARE network to plan a roadmap to improvements in data quality, such as the removing barriers to our partners for use of controlled vocabularies and historic place name gazetteers made available as linked open data, and implementing recommendations such as licensing and direct links to digital objects, to make collections compliant to the Europeana Publishing Framework. 2Culture, with Athena RC, will evaluate the enrichment of metadata offered at CARARE aggregator level to communicate the benefits to our partners. Athena RC will improve existing CARARE data by enriching them in the following ways: enabling SKOS, adding concepts from standard thesauri, adding various types of relations, normalising content. Furthermore, the quality of the provided metadata will be measured, and provide link validation and identity management services. Athena RC will establish and operate a helpdesk service for the duration of the project, providing metadata support to the CARARE data providers covering all aspects from metadata mapping (e.g. using the MINT tool) to the use of metadata quality / enrichment services in the MORe aggregator infrastructure.

Task	T04.1 Evaluation
Description	This task will evaluate the CARARE data that is already accessible via Europeana (more than 2 million records) and new data to be submitted with the aim of identifying data quality issues, and making a plan of action for improvements. CARARE organised a network meeting which brought data quality issues and Europeana's new publication policy to the attention of network members on 18-19th June 2015. As a follow-up to this meeting: ● 2Culture will circulate a survey to members to find out what content they aim to work on under the Europeana DSI including data and metadata quality issues. ● Athena RC will carry out a review of existing CARARE data with the aim of identifying data quality issues ● Athena RC will review potential enrichment services and innovative technologies that could be used to improve data quality for the domain. ● The results of the survey and reviews will be analysed and an activities plan will be produced.
Start Month	03
End Month	06
Milestone(s)	M04.1 CARARE network meeting (Month 06) M04.2 Survey carried out and results evaluated (Month 06)
Deliverable(s)	D04.1 Activity plan (Month 06)

Task	T04.2 Improve data quality
Description	This task will build on the results of T04.1 and will be responsible for improving metadata and data quality of the aggregated content. Under this task:

	● Mappings of existing datasets will be updated by members of the network using MINT, to take developments in both the CARARE schema and the implementation of EDM by Europeana.
	● Updated and new datasets and mappings will be ingested to MORe where providers can make use of the services and technologies available to analyse the completeness and otherwise quality assure their metadata.
	● Particular attention will be paid to:
	   ○ Titles and descriptions;
	   ○ The provision of links to digital objects
	   ○ The availability of digital files to generate thumbnails;
	   ○ The availability of place names and geographic coordinates;
	   ○ subject and period indexing;
	   ○ The coherence of rights statement.
	● The potential to enrich the metadata using the services available through MORe will be evaluated.

Start Month	05
End Month	08
Milestone(s)	M04.3 Improved data quality plan (Month 11)
Deliverable(s)	D04.2 Report on improved content (Month 12)

Task	T04.3 Automatic metadata enrichments
Description	This task will be utilise the microservices available within the MORe repository to validate and enrich the metadata, and will be responsible for executing enrichment plans and delivering enriched content. This task will focus on: ● Normalization incl: language identification, place names, dates & periods; ● Utilization of vocabularies to thematically enrich records; ● Identification of relations between records and contextual resources in Wikipedia; ● Link checking; ● Quality measurement of thumbnails and object links.
Start Month	08
End Month	13
Milestone(s)	M04.4 Metadata enrichment services (Month 13)
Deliverable(s)	D04.3 Report on metadata enrichment (Month 13)

Task	T04.4 Maintenance and update of the CARARE metadata schema
Description	Under this task a working group will be established to maintain and develop the CARARE metadata schema. The working group will: ● Review requests for updating of the schema and make recommendations; ● Maintain the schema documentation and mapping guidelines; ● Maintain the schema XSD; ● Develop an RTF version of the schema; ● Disseminate and promote the CARARE metadata schema.

Start Month	03
End Month	13
Milestone(s)	M04.5 CARARE schema (Month 12)
Deliverable(s)	D04.4 CARARE schema documentation (Month 12)

Task	T04.5 Vocabularies and the domain
Description	This task will build on the results of LoCloud and encourage members of the network to collaborate in the development of domain vocabularies. It will: ● Identify and make available domain related vocabularies; ● Enable collaborative development of domain related vocabularies; ● Enable multilingual developments; ● Provide vocabulary services; ● Enable thematic enrichment of records.
Start Month	08
End Month	13
Deliverable(s)	D04.5 Domain vocabularies and vocabulary services (Month 12)

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
CARARE															
Tasks															
T04.1	*Evaluation*				X	X	X	X							
T04.2	*Improve data quality*						X	X	X	X	X	X	X		
T04.3	*Automatic metadata enrichments*									X	X	X	X	X	X
T04.4	*Maintenance and update of the CARARE metadata schema*				X	X	X	X	X	X	X	X	X	X	X
T04.5	*Vocabularies and the domain*								X	X	X	X	X	X	X
Deliverables															
D04.1	*Activity plan*							X							
D04.2	*Report on improved content*													X	
D04.3	*Report on metadata enrichment*														X
D04.4	*CARARE schema documentation*													X	
D04.5	*Domain vocabularies and vocabulary services*													X	
Milestones															
M04.1	*CARARE network meeting*							X							
M04.2	*Survey carried out and results evaluated*							X							
M04.3	*Improved data quality plan*												X		
M04.4	*Metadata enrichment services*														X
M04.5	*CARARE schema*													X	

5-Europeana Fashion

Description of work
Europeana Fashion will work to improve the aggregated fashion metadata (700,000+ records) according to the Metadata Quality Task Force Recommendations, in particular: ● It will extend the use of the Europeana Fashion multilingual thesaurus for the classification of all the ingested records (maintaining and extending the fashion thesaurus);   ● It will experiment and apply automatic enrichment techniques (data semantification and image analysis) to add properties such as "colour" to the ingested records (in collaboration with Task 4.5.4);   ● It will apply refinement techniques (using the MINT group annotation tool) to further harmonize ingested metadata, especially for names of designers and brands and dates;   ● It will add more specific licensing information, encouraging data providers to use Creative Commons licenses and in particular Free Culture licenses.   ● It will support data providers in implementing direct links to their provided digital objects (a practice already largely in use in the Europeana Fashion network of data providers)

Task	T05.1 Maintenance and update of the Europeana Fashion Thesaurus
Description	In this task, the existing Europeana Fashion thesaurus will be refined, extended and enriched to meet the needs of the Fashion community at large. This work will be carried on in collaboration with the thesaurus working group, established during the former Europeana Fashion project. The thesaurus will be published online in a machine-readable format (SKOS), to be re-used openly by all the fashion community.
Start Month	02
End Month	13
Milestone(s)	M05.1 Publishing of the updated version of the Europeana Fashion thesaurus (Month 13)
Deliverable(s)	D05.1 Updated version of the Europeana Fashion Thesaurus (Month 13)

Task	T05.2 Harmonisation, refinement and cleaning of the aggregated fashion data sets
Description	The aggregated fashion datasets (more than 700 thousand records) will be cleaned and harmonised using the group annotation tool developed in the MINT platform. Best practice and recommendations from the Metadata Quality Task Force will be applied during this process. In particular, the refinement work will focus on normalisation of brand and designer names, dates and places using when possible open vocabularies (e.g. DBpedia, Geonames). An increased metadata check will be performed on EDM-mapped metadata prior their publication. Metadata enrichment processes will be applied and validated on all the aggregated datasets (see also T05.3). Also licensing information will be harmonised and validated for the 100% of the aggregated content, trying to encourage the use of Creative Commons licenses. Special attention will be given also to the provision of direct links to digital objects in the metadata records, aiming at reaching 100% of direct links in our metadata records and consequently also the generation of thumbnails for 100% of the aggregated content. This task will work in collaboration with Task 1.6.2 "Raising competency levels

	of data providing partners", in order to disseminate and implement best practices about metadata quality and publication through also the organisation of a workshop for data providers. A Basecamp group on content enrichment will be also set-up and used to communicate and offer support to data providers.
Start Month	02
End Month	13
Deliverable(s)	D05.2 Report on fashion data sets cleaning and refinement (Month 13)

Task	T05.3 Automatic metadata enrichment of the aggregated data sets
Description	In this task we will continue the work, started in the former Europeana Fashion project, about automatic enrichment of metadata. In particular we will fine tune and apply image analysis techniques to extract dominant colors from about 430.000 fashion catwalk images (using deep learning techniques), and we will fine tune and apply data semantification techniques to enrich about 300.000 metadata records extracting relevant terms from the textual descriptions to be added to elements like "Materials", "Techniques" and "Colors". This task will also collaborate with Task 4.5.4 on data enrichment.
Start Month	07
End Month	13
Deliverable(s)	D05.3 Report on semi-automatic metadata enrichment (Month 13)

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
Europeana Fashion															
Tasks															
T05.1	*Maintenance and update of the Europeana Fashion Thesaurus*			■	■	■	■	■	■	■	■	■	■	■	■
T05.2	*Harmonisation, refinement and cleaning of the aggregated fashion datasets*			■	■	■	■	■	■	■	■	■	■	■	■
T05.3	*Automatic metadata enrichment of the aggregated datasets*								■	■	■	■	■	■	■
Deliverables															
D05.1	*Updated version of the Europeana Fashion Thesaurus*														■
D05.2	*Report on fashion data sets cleaning and refinement*														■
D05.3	*Report on semi-automatic metadata enrichment*														■
Milestones															
M05.1	*Publishing of the updated version of the Europeana Fashion thesaurus*														■

6-EUscreen

Description of work
The quality of data will be improved through three areas of work. They benefit the overall Europeana infrastructure. In effect, activities in this task will also support the work to reach out to new content providers.   ● Europeana Publishing Framework: The aggregators will be informed about the Europeana Publishing Framework and its applicability in their respective working contexts.   ● Enrichment: This area of work focuses on enriching the metadata using manual and automatic methods and tools.   ● Keyframe extraction.

Task	T06.1 Europeana Publishing Framework
Description	The aggregators will be informed about the Europeana Publishing Framework and its applicability in their respective working contexts.
Start Month	01
End Month	06

Task	T06.2 Enrichment
Description	This area of work focuses on enriching the metadata using manual and automatic methods and tools. In order to improve the quality of the data using enrichment tools, a group-edit functionality will be used that facilitates content 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.
Start Month	06
End Month	12

Task	T06.3 Keyframe extraction.
Description	Keyframe extraction are added to increase the quality of the Europeana service. An important element of the aggregated data is the visual signifier (i.e. thumbnails) that is sent to the Europeana portal. An automatic screenshot extraction service is developed, securing that all objects contributed to Europeana by EUscreen have keyframes, including correct rights labels.
Start Month	07
End Month	13

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
EUscreen															
Tasks															
T06.1	*Europeana Publishing Framework*	■	■	■	■	■	■	■							
T06.2	*Enrichment*							■	■	■	■	■	■	■	
T06.2	*Keyframe extraction*								■	■	■	■	■	■	■
Milestones															
M06.3	*EUscreen data improvement technical delivery*												■		

7-Social History Aggregator (HOPE2)

Description of work
The Social History Aggregator (HOPE2) aims to improve its practices on digital appraisal, materials selected by the new data providers will be reviewed by the core team to define their relevance for the Social History Aggregator and Europeana. The concept of the collection, its completeness and its context is also important to showcase. In case of the existing data providers, some collection descriptions need to be refined, while new partners will go through a stricter quality check prior to considering them for publication. A new policy defines one of the crucial components to provide hierarchical display to digital objects, achieved through elements such as dcterms:isPartOf, edm:isRelatedTo, dc:relation, edm:isNextInSequence.
Semantic enrichment tasks will include applying more multilingual elements, for example, for collection descriptions (English and the language of the content), in case of titles if applicable. Translations in 5 European languages for the domain specific themes/subject tags will be added. These tasks will also comprise adding more keywords, subjects, tags connected with skos:Concept class. All HOPE/HOPE2 metadata values using Linked Open Data will be enriched. Metadata should be enriched where possible using Linked Open Data vocabularies to ensure the greater online exposure.
The Social History Aggregator (HOPE2) will highlight the global geographical references to the collections and the extended geographical coverage. The use of the edm:Place class will allow these institutions to submit more specific geospatial information. In addition, ideally, the latitude and longitude of the object will be submitted in the edm:Place class, or edm:TimeSpan class can be also used.
Re-using social history collections is a key aspect of the digital object for both the general and specialist user. Rights statements need to be provided by selecting one of the 13 Europeana rights statements, or providing more clarity on the different licenses and convincing the partners to assign more open licenses. Restrictive statements will be revisited and discussed with existing data providers. The discussion around copyright is especially important when we encourage content providers to add direct links to their digital objects.
Four new data providers will be added (see also Europeana DSI task 1.6.1) by May 15. Their collections need to be ingested and enriched with themes, language, geospatial elements in addition to multilingual collection descriptions and metadata.
The end date for the proposed improvements is May 15, 2015.

Task	T07.1 Assessing existing metadata in the Social History Aggregator and evaluating possible improvements
Description	It looked at the dataset available on the Social History Portal and Europeana. Currently, the portal has 183 collections, while Europeana includes 98 HOPE collections, 5,597 mid-level records and 753,628 item-level records. Further analysis was carried out on rights values, PIDs, collection appraisal, direct linking, tagging tool functionality, the multilingual metadata values and geospatial information. The result of this analytical work enabled to renegotiate policies with the current partners and convince them to get engaged in the data improvement tasks.
Start Month	03
End Month	13
Milestone(s)	M07.1 International Association of Labour History Institutions (IALHI) Annual

	Meeting to be held in Edinburgh to share the first results (Month 05) M07.2 Produce an action plan (Month 05) M07 3 Execute action plan and tasks identified (Month 10) M07.4 Report back on quality improvement and contribution by current partners (Month 13)
Deliverable(s)	D07.1 Final action plan for social history institutions (Month 06) D07.2 Data delivery of improved metadata completed (Month 13)

Task	T07.2 Defining requirements for multilingual access and implementation
Description	Among the semantic enrichment tasks we considered multilingual access one of the most ambitious goals of the project. The current HOPE policy does not prescribe the provision of mandatory bilingual metadata (English and metadata of the content), however for the sake of consistent discovery experience we will highlighted the importance for the following elements: collections title and descriptions, item titles, subject terms.After the preliminary assessment we have decided to reduce the compulsory work to bilingual tags and collection descriptions, in specific cases: KEE-OSA, ASKI as new content provider we agreed on providing bilingual item-level descriptions. The task also includes all HOPE/HOPE2 metadata values using Linked Open Data. Metadata should be enriched where possible using Linked Open Data vocabularies to ensure the greater online exposure. Incorporating another.
Start Month	05
End Month	13
Milestone(s)	M07.5 Requirements for multilingual access (Month 05) M07.6 Interim report on site visits (Month 09) M07.7 Final report on implementation (Month 13)
Deliverable(s)	D07.3 Data delivery of multilingual metadata completed (Month 12) D07.4 Applying Linked Open Data vocabularies (Month 13)

Task	T07.3 Defining requirements for geospatial enrichment and implementation
Description	The current social history collections often lack country references while IALHI is a global organisation, we would like to make sure that our international network is well represented from the point of view of our collections as well. The use of the edm:Place class allow institutions to submit more specific geospatial information. The use of country coverage will enhance discovery in addition to provenance information. Ideally, the latitude and longitude of the object should be submitted in the edm:Place class, or edm:TimeSpan class can be also used. Wherever it is possible we will use automatic geotagging.
Start Month	05
End Month	13
Milestone(s)	M07.8 Requirements for geospatial enrichment and implementation (Month 05) M07.9 Interim report on site visits (Month 09) M07.10 Final report on the implementation of geotagging (Month 13)
Deliverable(s)	D07.5 Applying geospatial references (Month 12) D07.6 Data delivery of geospatial metadata completed (Month 13)

Task	T07.4 Defining requirements for resolving additional access issues: broken links, direct links, formats, technical constraints, licensing

Description	The appearing problems with our current collections could be explained partly by technical constraints edm:isShownBy or edm:isShownAt, lack of persistent ID, technical measures applied to hinder downloads, obsolete formats, non-standard thumbnails. On the other hand the lack of clear license information, missing rights statements and supporting policies on the local collection pages including the Social History Portal represent a conglomerate of problems. Here the enrichment process is more complex, it entails training, consultation, technical work which should be all combined. At the same time the reluctance to rely on more open access principles may conflict with traditional institutional policies, economic exploitation etc. The result of these activities should be monitored through not only having more rights statements with more open provisions as metadata in the light of the Europeana Publishing Framework but communicating more transparent access policies, introducing notice and takedown approaches etc. The current statistics prove that major flaws in linking objects could be found at the following repositories: Confederazione Generale Italiana del Lavoro (1,234, isShownAt links), Bibliothèque de Documentation Internationale Contemporaine (BDIC) (2,888 isShownAt links), Maison des Sciences de l'Homme de Dijon (MSH Dijon) (1,778 isShownAt links), in total: 6,157. These links will be 100% converted to isShownBy. This is mandatory requirement for the four new content providers. In case of applying more standard licenses, among HOPE CPs we have one major content provider using 'http://www.europeana.eu/rights/unknown/' as their edm:rights: the Archiv der sozialen Demokratie (AdsD) has 43,095 records with unknown status. The shift to more open licenses represents a major challenge in case of archival institutions, collections often contain records with unknown copyright status. It is always the local law which provides the framework for the copyright and offers exceptions and copyright mitigation solutions. We are in the process of negotiating with the German institution to revise the edm:rights and try to apply Europeana standard license: http://www.europeana.eu/rights/rr-r/ Persistent Identifiers (PIDS) will be added to all metadata records and digital objects, the HOPE PID Service ensures the application of PIDs consistently for new content providers as well. .
Start Month	05
End Month	13
Milestone(s)	M07.11 Revisiting the HOPE IPR Best Practices Recommendations by acknowledging the Europeana Publishing Framework (Month 05) M07.12 Interim report on local IPR situations (Month 09) M07.13 Interim report on local technical problems and possible solutions (Month 09) M07.14 Final report on improved access (Month 13)
Deliverable(s)	D07.7 Updated HOPE WIKI on licensing and IPR including links to local sites (Month 13) D07.8 Data delivery of rights statements (Month 13)

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
Social History Aggregator (HOPE2)															
Tasks															
T07.01	*Assessing existing metadata in the Social History Aggregator and evaluating possible improvements*				■	■	■	■	■	■	■	■	■	■	■
T07.02	*Defining requirements for multilingual access and implementation*						■	■	■	■	■	■	■	■	■
T07.03	*Defining requirements for geospatial enrichment and implementation*						■	■	■	■	■	■	■	■	■
T07.04	*Defining requirements for resolving additional access issues: broken links, direct links, formats, technical constraints, licensing*						■	■	■	■	■	■	■	■	■
Deliverables															
D07.01	*Final action plan for aocial history institutions*						■								
D07.02	*Data delivery of improved metadata completed*														■
D07.03	*Data delivery of multilingual metadata completed*													■	
D07.04	*Applying Linked Open Data vocabularies*														■
D07.05	*Applying geospatial references*													■	
D07.06	*Data delivery of geospatial metadata completed*														■
D07.07	*Updated HOPE WIKI on licensing & IPR including links to local sites*														■
D07.08	*Data delivery of rights statements*														■

Milestones		1	2	3	4	5	6	7	8	9	10	11	12	13	14
M07.01	*International Association of Labour History Institutions (IALHI) Annual Meeting to be held in Edinburgh to share the first results*						X								
M07.02	*Produce an action plan*						X								
M07.03	*Execute action plan and tasks identified*											X			
M07.04	*Report back on quality improvement and contribution by current partners*														X
M07.05	*Requirements for multilingual access*						X								
M07.06	*Interim report on site visits*										X				
M07.07	*Final report on implementation*														X
M07.08	*Requirements for geospatial enrichment and implementation*						X								
M07.09	*Interim report on site visits*										X				
M07.10	*Final report on the implementation of geotagging*														X
M07.11	*Revisiting the HOPE IPR Best Practices Recommendations by acknowledging the Europeana Publishing Framework*						X								
M07.12	*Interim report on local IPR situations*										X				
M07.13	*Interim report on local technical problems and possible solutions*										X				
M07.14	*Final report on improved access*														X

8-The European Library

Description of work
The European Library (TEL) will finalise the migration of its datasets from the ESE format to EDM. The migration to EDM will improve the quality of the data on several important points, such as making all objects provided to Europeana to have persistent identifiers; all objects are labelled with a rights statement from the Europeana rights statements list; all text objects have a language; all metadata records pointing to a digital file or to a html page representing a digital object. TEL will work on the consistency of rights labels by making sure the rights statements chosen by data providers are consistent with the objects they are applied to. The focus will be made on making sure the objects in the Public Domain are labelled with the Public Domain Mark. TEL will develop validation rules for checking automatically not only by date of publication but also by the digitisation laws that exist per country. This will help to identify objects that belong to the Public Domain but they are not licensed accordingly. TEL will analyse its collections in Europeana to identify duplicate records and perform deduplication of datasets when needed (e.g. providers with new collections send also duplicate records, that is to say, records included in older sets and published already in TEL and Europeana). TEL will focus on the normalization of language data. It will investigate how un-normalized language data and language tags exist in its collections, and work in the provision of normalization tools for the most common cases. The tools will be made available as services for METIS so that they can be applied in the future by TEL, Europeana, other aggregators and data providers. TEL will also work on the enrichment of data with controlled vocabularies by increasing the amount of linked data identifiers from AAT, Iconclass, TGN used in the metadata. TEL ran a data mining report to identify the amount of identifiers used anywhere in the TEL dataset. TEL will also work with the data providers to improve the quality of the images (size, resolution, etc) published in Europeana but also to increase the number of direct links (edm:isShownBy) to the digital objects and to make all digital objects have a thumbnail. TEL identified the number of records with missing thumbnails in Europeana.

Task	T08.1 Consistency of rights labels
Description	Make sure the rights statements chosen by data providers are consistent with the objects they are applied to. A specific focus will be put on making sure the objects in the Public Domain are marked Public domain. Validation rules based on the publication date and digitisation laws per country will be put in place.
Start Month	00
End Month	13
Milestone(s)	M08.2 10%-20% of the TEL collections checked, validated with the provider and delivered to Europeana with updated rights statements (Month 13)

Task	T08.2 Deduplication of records
Description	TEL will perform deduplication of datasets when needed (e.g. providers with new collections send also duplicate records, that is to say, records included in older sets and published already in TEL and Europeana).
Start Month	00

End Month	13
Milestone(s)	M08.3 10%-20% of the TEL collections checked for duplicates, deduplicated and delivered to Europeana without duplicates (Month 13)

Task	T08.3 Normalisation
Description	Identification of collections which can provide test cases for the normalization of language data. Development of open source tools to increase the amount of language attributes in the metadata, and normalisation of language to language ISO codes. The tools will be developed for integration with the Metis ingestion system, currently under development, for Europeana and The European Library.
Start Month	00
End Month	13
Milestone(s)	M08.4 10%-20% of the TEL collections checked for existing cases for improvements of language tags and data. Open source tools available in the Europeana GitHub (Month 13)

Task	T08.4 Vocabulary enrichment
Description	Increase the amount of linked data identifiers from AAT, Iconclass, TGN in the data. Identification of number of identifiers used in the TEL dataset done in month 3.
Start Month	00
End Month	13
Milestone(s)	M08.5 100% of the TEL collections checked for existing data that can be linked to Europeana supported vocabularies. Delivery to Europeana of the identified collections, enriched with linked data identifiers, with a total expected of 40 thousand enriched metadata records (Month 13)

Task	T08.5 Quality images & thumbnails
Description	TEL will work with its data providers to improve the quality of the images aggregated to Europeana, as defined by the Europeana Publishing Framework. TEL will work with its data providers to increase the number of direct links (isShownBy) and the amount of thumbnails for all objects provided to Europeana. Identification of TEL records with missing thumbnails in Europeana done in month 3
Start Month	00
End Month	13
Milestone(s)	M08.6 30%-50% of the TEL collections checked and delivered to Europeana with better images, direct links (edm:isShownBy) (Month 13)

GANTT chart

Task/Deliverable/Milestone		00	01	02	03	04	05	06	07	08	09	10	11	12	13
The European Library															
Tasks															
T08.1	*Consistency of rights labels*	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T08.2	*Deduplication of records*	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T08.3	*Normalisation*	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T08.4	*Vocabulary enrichment*	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T08.5	*Quality images & thumbnails*	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Milestones															
M08.1	*10-20% of the TEL collections checked, validated with the provider and delivered to Europeana with updated rights statements.*														■
M08.2	*10-20% of the TEL collections checked for duplicates, deduplicated and delivered to Europeana without duplicates*														■
M08.3	*10-20% of the TEL collections checked and delivered to Europeana with increased number of language attributes & normalised ISO codes.*														■
M08.4	*100% of the TEL collections checked and 40.000 records delivered to Europeana with enriched vocabulary by increasing the number of linked data identifiers.*														■
M08.5	*30-50% of the TEL collections checked and delivered to Europeana with better images*														■