Europeana – Core Service Platform

MILESTONE

MS2: UPDATED PARTNER AND DATA DEVELOPMENT PLAN (INCORPORATING OUTCOMES OF SUBTASK 1.1.2)

Revision	1.0
Date of submission	30 June 2016
Author(s)	Henning Scholz, Valentine Charles (EF)
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
1	01/06/2016	Henning Scholz	EF	ToC and outline
2	10/06/2016	Henning Scholz	EF	First draft of chapters excl DQC
3	17/06/2016	Valentine Charles	EF	Adding DQC chapter
4	24/06/2016	Joris Pekel, David Haskiya, Antoine Isaac, DSI aggregating partners and DQC members	EF	Review of the milestone document with a focus on the Data Quality Committee chapter
5	28/06/2016	Henning Scholz, Valentine Charles	EF	Incorporation of reviewer comments, final edits

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

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

Table of Contents

Executive summary

In the last years, the partner and data development of Europeana was focussed on a balanced representation of countries, domains and themes in Europeana. Although we still work with national, domain and thematic aggregators to improve their representation in Europeana, this plan has a much stronger emphasis on data quality. This is very much supported by the Europeana Publishing Framework, setting out the four scenarios for sharing collections with Europeana.

The Data Quality Committee (DQC) was set up with the ambition to tackle data quality issues at every level of the data exchange chain. One area of work of the DQC is on discovery/usage scenarios, to focus on data quality from the perspective of its intended use rather than as a theoretical effort. A metadata completeness measure is under development by members of the DQC to make data quality measurable and more visible. The DQC has started to gather a list of metadata problem patterns that are impacting search and interfering with any kind of ranking algorithms. The first potential recommendations have been identified, but still need to be refined. These potential recommendations address the multilinguality of the metadata and the mandatory elements for ingestion of EDM data. The data quality plan for Europeana DSI-2 is also summarised in here, which is now approved to be part of the Description of Action of the DSI-2 project.

A content strategy is in progress to be developed, to have a strategic plan for the development of the Europeana database, building on the Europeana Publishing Framework and the recommendations of the DQC. In addition to data quality, the content strategy will also address the type of content that Europeana should include, its geographic origin and to what extent and how user demands and a thematic focus should be considered.

With regards to partner development, Europeana is now in a state of transition. We are still building on the Aggregator Forum and the aggregator model is still key for the development of Europeana. However, concepts are under development on how to improve the value for the collection holders more directly. Developing the content base for the Europeana Thematic Collections is one approach, to actively connect with collection holders and offer them to be featured in the thematic collections. Europeana is also working on the technical infrastructure in order to make it as easy as possible to publish with Europeana. The outcomes of this work should turn the hierarchical aggregator model into a web of interconnected nodes of different sizes and capabilities that complement each other and work together to provide value to Europe’s memory institutions and citizens.

Introduction

This Partner and Data Development Plan is an update of the partner and data development plan published under Europeana v3.[1] It uses a similar structure and to some extent the same wording in some chapters and paragraphs that are still relevant today (e.g. content by Member State, content by domain and theme, 20th century content). Since we have published monthly content reports during DSI-1 and recently published a partner development report (DSI D1.2)[2], the current status of Europeana's collections, metadata and partners is not reported in here, as it was in the partner and data development plan under Europeana v3. Instead it defines the general aims and specific targets for future development, based on the Europeana Business Plan 2016[3] and the Europeana Strategy 2015-2020[4]. As some elements that are relevant for the partner and data development plan were addressed in much more detail in other Europeana DSI deliverables (D1.1, D1.2), they are kept very brief in here summarising the plans documented in these deliverables. Links to the deliverables are given in the relevant chapters for reference.

The first part of this report elaborates on data development, which includes the development of metadata and content.[5] The second part elaborates on partner development, although there is sometimes quite an overlap as data development can not always be planned without thinking about the most suitable partner for data. Finally we present some conclusions and make some recommendations for the further improvement of the Europeana database, for further discussion with the Europeana aggregators and data providers.

With more than 53 million objects in the database, we can say that we have assembled a 'critical mass' of cultural artefacts. However, having a critical mass by itself is not enough. Improving the quality of the database is of crucial importance and we aim to in 2016 add at least 2.5 million records to tiers 2, 3 and 4 as defined by the Europeana Publishing Framework[6]. End-users and re-users are increasingly demanding of 'quality' in the broadest sense of the word. Not only should what they are looking for be easy to find, but the information about the object has to be accurate and informative, and preferably in a number of languages too. They also desire that the digital object is of high resolution, as well as re-usable and sharable across the web. The improvement of the quality of the data, with a strong focus on accessibility, accuracy, consistency and re-usability of metadata and content, is key to fuel the creative industry and reach our end-users, which then gives as high return as possible to the memory institutions opening up their data.

[1] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Version3/Milestones/Ev3%20MS2%20MS4%20MS5%20Partner%20and%20Data%20Development.pdf

[2] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d1.2-amount-of-data-partners-and-outreach-to-major-institutions.pdf

[3] http://pro.europeana.eu/publication/creating-cultural-connections-business-plan-2016

[4] http://strategy2020.europeana.eu/

[5] The term 'content' here refers to "a physical or digital 'object' that is part of Europe's cultural and/or scientific heritage, typically held by the data provider or by an aggregator of the data provider" (description taken from the Europeana Data Exchange Agreement (DEA), article 1)

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

To achieve these aims, we continue to establish new partnerships, with a focus on organisations who can provide high-quality content in the areas that are defined as high priority in the annual business plans, i.e. music, art history, fashion, newspapers, First World War. We will nurture the relationships with our current partners by facilitating knowledge transfer and the sharing of experiences, and by intensifying the collaboration and participation of the contributing organisations during the data provision cycle. Moreover, we will help our partners make the most out of their participation in Europeana and maintain a dialogue as to how we can help and simplify the work done by data providers and aggregators.

This plan is also supposed to incorporate the outcomes of DSI subtask 1.1.2, which is about the investigation and development of the expert hub concept. This work is closely linked to the work on the innovation of the aggregation infrastructure. A summary is given in the last chapter under partner development, with a reference to the work and implementation plan to innovate the aggregation infrastructure (D1.1).[7]

Data development

Content report

Since a couple of years we are reporting the ingestion and publication progress in a monthly content report. With the start of the Europeana DSI, these content reports are also publicly available on Europeana Pro[8]. These reports give an overview of the development of some key aspects of the database and also allow to monitor our progress with key performance indicators. No content report is available for summer and early autumn 2015, as we did not published any data in that time period. We worked on our technical infrastructure instead to improve our tools and processes for ingestion and data publication.

Europeana Collections is offering a browse entry point to explore all the recent additions to the database[9]. In addition to the statistics available in content reports it is a way to see examples of new collections, see which cultural institution has contributed new items, the number of items they've added and when.

Content by Member States

Europeana aims to offer access to the cultural heritage of all European countries. However, not all EU Member States are equally well-represented in the collections in Europeana's database. Indicative targets for the minimum contribution of content to Europeana by each Member State, to be achieved by the end of 2015, were set by the European Commission in a recommendation in 2011.[10] Table 1 shows for each Member State the original targets, the results at the end of 2015

[7] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d1.1-work-and-implementation-plan-to-innovate-the-aggregation-infrastructure.pdf

[8] http://pro.europeana.eu/get-involved/projects/project-list/europeana-dsi

[9] http://www.europeana.eu/portal/browse/newcontent

[10] European Commission recommendation 2011/711/EU of 27 October 2011: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32011H0711

and the current situation (see also statistics per country at the Europeana Statistics Dashboard[11]).

Table 1. Progress per country towards the 2015 content targets.

Country	**Goal 2015**	**Achieved in % (27/03/2015)**	**Achieved in % (15/12/2015)**	**Status in % 01/06/2016**
Austria	600,000	299.2	329.4	348.8
Belgium	759,000	142.6	175.8	179.4
Bulgaria	267,000	35.7	37.4	43.6
Cyprus	45,000	45.6	77.5	86.2
Czech Republic	492,000	108.7	110.7	151.4
Denmark	453,000	214.6	285.1	368.8
Estonia	90,000	418.5	607.5	607.5
Finland	1,035,000	21.2	87.3	87.4
France	4,308,000	98.4	105.3	109.6
Germany	5,496,000	92.4	100.3	101.5
Greece	618,000	77.8	95.2	100.2
Hungary	417,000	162.4	185.6	220.5
Ireland	1,236,000	19.1	20.5	22.0
Italy	3,705,000	89.3	113.0	114.4
Latvia	90,000	147.3	158.6	158.3
Lithuania	129,000	112.7	124.4	131.2
Luxembourg	66,000	230.5	243.1	243.1
Malta	73,000	85.3	85.3	85.4
Poland	1,571,000	118.0	121.3	127.4
Portugal	1,575,000	45.7	15.5	14.9
Romania	528,000	13.9	23.6	34.3
Slovakia	789,000	43.1	14.2	14.3
Slovenia	243,000	98.9	217.9	227.0
Spain	3,180,000	150.3	136.0	141.5
Sweden	2,676,000	172.2	135.7	143.3
The Netherlands	1,936,000	346.7	325.5	324.2
United Kingdom	3,939,000	77.9	84.2	96.6

Some countries made substantial progress between March 2015 and the end of 2015 and passed their targets (France, Germany, Italy, Slovenia). Other countries also made progress, but

[11] http://statistics.europeana.eu/countries

have passed their targets either in early 2016 or are expected to pass them later this year (Cyprus, Finland, Greece, UK).

We are still working with countries that have not reached their targets to either support the aggregator in the country in finding new partners or supporting institutions in finding the most suitable aggregator to publish in Europeana. Specific activities were carried out under DSI-1 in particular with partners in Ireland, Romania and Slovakia. Tangible progress in these countries is still small in terms of numbers of new items in Europeana, but it looks promising to change over the next months.

Content by domain and theme

One of the unique aspects of Europeana is that it brings together, in one database, a wealth of material from four key domains: libraries, archives, audiovisual collections and museums. In addition, Europeana brings together collections representing a variety of themes, e.g. fashion, natural history, archaeology, photography, and labour history. As well as aiming for a geographically balanced distribution of content, we also strive for an even representation of these different domains and themes in our collection. One third of the Europeana database is currently (June 2016) populated by domain and thematic aggregators that are full partners of the Europeana DSI. The close partnership with these partners has lead e.g. to a substantial increase of audiovisual content, where now about two million objects can be found in Europeana. The partnership with domain and thematic aggregators will be continued under DSI-2, which will greatly help to balance the amount of records per domain. As explained below, we are working with these partners not only on getting more and better data but also on transforming aggregators into expert hubs (see chapter on innovating the aggregation infrastructure).

20th century content

Adding more content from the 20th century has always been a concern for Europeana. Copyright related issues make it very difficult for cultural heritage institutions to open up and submit their data to Europeana. This result in what is often referred to as the '20th century black hole'. This 'black hole' was analysed in more detail end of last year[12]. This study concluded that "in order to promote the online availability of cultural heritage from the 20th century, it is necessary to reduce the burden of rights clearance for these institutions"[13].

The Europeana Publishing Framework and Guide

The Europeana Publishing Framework (EPF) is setting out four scenarios for sharing collections with Europeana[14]. The four scenarios or tiers are based on what data partners want to achieve when publishing with Europeana and what they are able to provide. The more data partners are providing to Europeana in terms of quality, the more Europeana can offer to the data partners in terms of services to make collections more visible and usable. Once a data partner has decided

[12] http://pro.europeana.eu/files/Europeana_Professional/Advocacy/Twentieth%20Century%20Black%20Hole/copy-of-europeana-policy-illustrating-the-20th-century-black-hole-in-the-europeana-dataset.pdf

[13] quoted: http://pro.europeana.eu/blogpost/the-missing-decades-the-20th-century-black-hole-in-europeana

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

on how to publish with Europeana, the Europeana Publishing Guide should be consulted for the practical information to make data Europeana-ready.

With the EPF in mind, we are aiming to improve the quality of what gets published in Europeana. For 2016, we are aiming to add at least 2.5 million records to tiers 2, 3 and 4 as defined by the EPF. This means we have more digital objects that are directly linked from their metadata, more digital objects with higher quality thumbnails and more digital objects labelled with licenses that allow re-use and which are directly accessible via links. This will help re-users to find more suitable collections for research, education and tourism applications. It will also help with the development of the Europeana Thematic Collections, which are an incentive for data partners that would like to make their collections more visible within a thematic context. See also below the chapter on Europeana Thematic Collections.

After the EPF was published end of 2015, we are now working on the full adoption and implementation of the framework. Many aggregators have already invested in communication activities towards their partners, several aggregators even translated the EPF and/or the Europeana Publishing Guide and/or created supporting documents for the EPF in the language of the country they are operating in. These activities will be continued in 2016 and 2017, followed by a survey to get more feedback about the framework. In parallel, we need to become more clear on how the EPF is applied across the Europeana database and remove ambiguities as much as possible to allow data partners having a clear understanding of which tier their data are compliant with and what will happen with the data they submit to Europeana.

Data Quality Committee

The Data Quality Committee (DQC), formalised as a Member Council Working Group, was kicked-off at the beginning of 2016. This standing committee has been formed to tackle data quality issues at every level of the data exhange chain - from its creation to its publication. It therefore gathers experts from various backgrounds: metadata experts, software developers, search and retrieval experts. The main directions of work have been formalised as part of a Mission Statement[15] covering items such as mandatory elements for ingestion of EDM data, data checking and normalisation, and data completeness. The DQC had its first physical meeting during the Aggregator Forum in April 2016.

The status of the discussions and the current efforts are described below. Given the breadth of the work undertaken in the past months, these ideas will need further maturation before reaching recommendation status. Future recommendations proposed by te DQC will be implemented as part of the KPIs defined for Europeana DSI-2 (see section below).

[15] http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_WG/DataQualityCommittee//DataQualityCommittee_MissionStatement_032016.pdf

Current efforts of the Data Quality Committee

Discovery/User scenarios

The need for high quality metadata is particularly motivated by its impact on search and the overall Europeana user experience. The higher the quality of metadata in Europeana is, the more accurate the search on the Europeana dataset will be. The DQC has therefore defined its main requirements in terms of discovery and information-retrieval requirements. A series of usage scenarios[16] have been created reflecting information-access user needs (based on the Europeana user personas[17]), listing current metadata issues for a given scenario and then proposing future actions. These scenarios focus specifically on metadata and are not tackling any challenges regarding the user interface or the user experience in Europeana.

Table 2. Example of discovery-usage scenario.

Scenario title	**Browse by Concepts** Concepts in the broad sense which would include also e.g. genres of art and music and resource types.
Scenario	As a user I want to be able to browse an index (or visualised browse entry points) of concepts represented in the Europeana corpus.
Motivation	This satisfies the general need expressed in the personas for ‘findability’; and see in particular Paul’s need to ‘search for content under a clear structure’.
Metadata analysis	This scenario requires consistent use of term-based subject and resource type classifications. In Europeana’s case the terms also must have multilingual labels. Developing this would be possible if: All Europeana partners used the same SKOS-compliant terms for subjects and resource types and supplied URIs for it. The terms would need to have labels in all official languages of the EU and ideally also some regional European languages. BUT this is not the case and so we need to take specific actions to be able to support browse scenarios within an acceptable time frame (=within 2016).
Proposed actions	In order to begin supporting this scenario we will develop the Europeana Entity Collection[18] of aligned terms drawn from Dbpedia, Wikidata and specialised vocabularies like Getty AAT and use them for semantic enrichment of dc:subject and dc:type.

[16] http://pro.europeana.eu/get-involved/europeana-tech/data-quality-committee

[17] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d6.2-requirementsforeuropeana.pdf

[18] https://docs.google.com/document/d/16Lcuddgw7fNV0EQ7gnIJW5-C76z4HUIvRA8aunY13U4/

	Gradually make dc:subject and/or dc:type mandatory elements. Without values in these fields Europeana has nothing to base its semantic enrichment on.[19] Encourage provision of terms based on select vocabularies that are aligned with the Europeana Entity Collection.

This approach allows the DQC to focus on data quality from the perspective of its intended use rather than as a theoretical efforts. All the items and further recommendations from the DQC are therefore formulated in connection with these usage scenarios. They will also encompass domain-specific considerations due to the cross-domain nature of Europeana.

The relation between data quality and the representation of Events has been for instance the subject of more detailed discussions. Metadata descriptions provided to Europeana are at the moment focused on the description of a Cultural Heritage Object and this approach is supported by the Europeana Data Model (EDM). EDM also provides the way to describe CH objects as an Event which has not been implemented by Europeana. For some data partners represented in the DQC, this absence can contribute to a decrease of quality for some CH objects. Archeological artefacts for instance would see their metadata quality enhanced with an Event based modelling. The group has started to list requirements for browsing and searching by Events in the context of this discussion.

Mandatory metadata elements for ingestion of EDM data

EDM specifies a series of mandatory elements that have been initially created to ensure a minimal level of data quality in Europeana. However these mandatory elements are the subject of lot of discussions as they can result in quality issues if not understood or used properly. The main issues discussed by the DQC were the following:

- the definitions and use of the mandatory elements are not always well understood by data partners which often lead to duplications of values;
- the required mandatory elements are not always in the source metadata and needs to be created while transforming the source metadata to EDM;
- the current EDM documentation could be improved;
- some data partners argue for more manadatory elements, other for less.

Metadata completeness measure

One of the first effort initiated by Europeana as an attempt to measure the quality of the Europeana dataset was to apply a formula for calculating and assigning a completeness rating to every object that is provided. This rating could then be used as an indicator so that more visibility was given in the portal to objects that show a higher level of completeness.

[19] Look at e.g. this music manuscript object. As it has no dc:type or dc:subject value classifying that it is indeed a music manuscript Europeana semantic enrichment can’t connect it to a vocabulary term.

As the need for more granular quality measures increased, the DQC decided to focus on updating this completeness measure. This work was very much motivated by the individual effort done by Péter Király, a member of the DQC as part of his Metadata Quality Assurance Framework[20]. The DQC has so far focused on:

1.) Identifying criteria for measuring the completeness of a EDM record in Europeana such as the presence of all the mandatory fields, the presence of language tags, etc. Those criteria are based on the requirements outlined by the usage scenarios and further work is being done to associate them with some quality dimensions that would 'qualify' the quality.

> Following the approach of the Europeana Publishing Framework, the DQC has started to investigate ways to better measure, interpret and communicate the completeness result score to data partners. The approach taken is to group the EDM fields per usage dimensions. The currently proposed completeness measure, under discussion, includes the following dimensions:
>
> - *Descriptiveness:* a complete EDM record should contain rich descriptive information;
> - *Searchability and Findability:* a complete EDM record should have all the properties that correspond to Europeana's users search pattern;
> - *Contextualisation:* a complete EDM record should have all the properties that provide contextual information or can trigger the creation of contextual information (e.g. enrichment framework);
> - *Identification:* a complete EDM record should have all the properties that will allow a user to distinguish a CHO from another one within Europeana;
> - *Browsing:* a complete EDM record should have all the properties that will allow a user to navigate within a graph of CHO through a series of relationships;
> - *Viewing:* a complete EDM record has all the properties that allows a user to view, play, listen a given CHO;
> - *Re-usability:* a complete EDM record has all the properties that allows a user to know how to re-use a CHO;
> - *Multilinguality:* a complete EDM record should have language information.
>
> The DQC also suggested the addition of a *Isness* and *Aboutness* dimensions. These dimensions will be further discussed by the DQC in the coming months and will be aligned with the Discovery/Usage scenarios mentioned earlier.

2.) Implementing these criteria as part of completeness measures so that further reports and statistics can be built on top of them. One challenge in the implementation of these criteria is to decide on the weight attributed to each EDM fields; which will define what makes a record complete or not.

[20] http://www.slideshare.net/pkiraly/metadata-quality-assurance-framework-at-qqml2016-full

At the time of writing, several metrics are investigated and tested on the Europeana dataset[21] based on the distribution of (all available) EDM fields. The following parameters have been included in the calculations and are available per datasets and data partners:

- Distribution of mandatory elements (including the conditions specified by EDM such as at least dc:title or dc:description must be present)
- Fields frequency: how many time a field occurs in a dataset.
- Fields cardinality: cardinalities are based on the EDM specifications.
- Degree of multilinguality per datasets based on the presence of language information.

A visual representation (graph, bar chart) is available for each parameter and allows an easy analysis of the results. Now that these parameters are included in the completeness measure algorithm, the DQC is refining the requirements for each parameter to make sure their interpretations reflect metadata practices. Discussions are for instance focusing on the weighting given to individual EDM fields and how it would affect any functionalities based on the completeness measure such as results ranking.

The work of the DQC on completeness will support the KPI defined for DSI-2: "Have a completeness measure in the API output" (see below).

Gathering and detecting problem patterns with metadata values

When starting working on completeness measures, the DQC identified errors in the values provided as part of the EDM metadata fields in the current Europeana dataset that have a strong impact on search. While completeness focuses on the presence of a given field in a metadata record, it doesn’t provide any quality indicators on the value a field hold.

The Committee has gathered a list of problem patterns that are impacting search and interfering with any kind of ranking algorithms. These problem patterns can reflect data normalisation issues. The following table lists some of the problem patterns identified:

Title contents same as description contents
Systematic use of the same title
Bad string: "empty" (and variants)
Shelfmarks and other identifiers in fields
Creator not an agent name
Absurd geographical location
Subject field used as description field
Unicode U+FFFD (�)
Very short description field

The DQC will continue collecting problem patterns with metadata values and will investigate ways to automatically detect those errors in the metadata and express them as validation or normalisation rules. Technolgies such as SHACL[22] and RDFUnit[23] will be tested in this context.

[21] http://pkiraly.github.io/2016/04/09/third-report/

Coordination with other quality-related initiatives

The DQC is constantly trying to link its work to other projects from the Europeana Network and beyond. The work of projects and initiatives such as EEXCESS[24], Europeana Fashion[25] or Europeana Food and Drink have been used as references. The involvement of representatives from the Digital Library of America[26] (DPLA), Digital NZ[27] or the Digital Library of Australia[28] also connects the DQC to projects overseas.

Potential recommendations of the Data Quality Committee

The discussions of the past months allowed the identification of first recommendations. Those recommendations are still being refined and will require more work before being submitted to the Aggregator Forum for validation.

Improvement of multilinguality in the metadata

Most of the discovery-usage scenarios created by the DQC require the metadata to be multilingual. The recommendations of the DQC are aligned on the latest recommendations published as part of the White Paper “Best Practices for multilingual access to digital libraries”[29].

- Language tags and language identification - Language tags[30] in the metadata identify the language of the metadata values. If the metadata is available in several languages, a language tag must be provided for each language. Identification of metadata language is crucial for (automatic) data enrichment, and processes such as named entity recognition.
- Multilingual vocabularies - metadata can be made multilingual by using multilingual vocabularies. Mapping monolingual vocabularies to multilingual ones or translating them in other languages is recommended to facilitate the access to specialised datasets.
- Multilingual Semantic Metadata Enrichment - Multilingual semantic enrichments add new links to equivalent or semantically related (e.g. broader or narrower) resources to the metadata, their descriptions being available in different languages. The newly created links will provide further related keywords and other translations.
- Translating Multilingual Metadata and Multilingual Objects - translated metadata (e.g. titles, descriptions or abstracts, keywords) should be provided when available. The DQC suggests that better recommendations should be provided to distinguish the language of the metadata and the language of the object. Works can have several languages (e.g. in the case of subtitles in different languages) or none (most visual works). The title (dc:title) element, can be really ambiguous as many works can have multiple titles in many languages.

[22] https://www.w3.org/TR/shacl/
[23] http://aksw.org/Projects/RDFUnit.html
[24] http://eexcess.eu/
[25] http://www.europeanafashion.eu/portal/home.html
[26] https://dp.la/
[27] http://www.digitalnz.org/
[28] http://trove.nla.gov.au/
[29] http://pro.europeana.eu/publication/best-practices-for-multilingual-access
[30] Current Europeana recommendations http://pro.europeana.eu/share-your-data/data-guidelines/edm-case-studies/data-multilinguality

For use withing these recommendations, the DQC advocates controlled vocabularies for refering to languages themselves like ISO 639-2[31] or the Name Authority List (NAL) of the Publications Office of the European Union[32]. Further investigations will support the realisation of the DSI-2 KPI: Normalise values in dc:language (see below).

Mandatory elements for ingestion of EDM data

The topic of mandatory elements was the subject of many discussions in the DQC over the past months. One of the key issues was related to the definition of the concept of 'mandatory' element which is not always understood by data partners or well-communicated by Europeana.

The DQC tried to find different ways to express the idea of 'mandatory-ness' and proposes to distinguish mandatory elements (properties or classes) from enabling elements.

Mandatory elements:

- Elements required for key Europeana functions;
- These elements support key functional requirements of the Europeana metadata aggregation infrastructure such as identification, link to digital representations, provenance;
- The communication around these elements should motivate data providers in providing good data for those elements (similar to the Europeana Publishing Framework), not just any 'filler' value.
- For instance, Europeana as a digital library requires the presence of a link to a digital object, therefore edm:isShownBy or edm:isShownAt must be present.

Enabling elements:

- Elements whose mandatory-ness is 'qualified' by a specific (set of) usage scenario(s);
- Elements that enable functions and improve services in Europeana or from third parties;
- These elements are highly desirable as they increase user satisfaction;
- Europeana functions requiring enabling elements won't reach their full potential as well as expected if the enabling elements are missing from CHO metadata descriptions. For example, cultural objects lacking these elements may not be shown or rank as well in the results on the Europeana portal.
- Missing enabling elements do not make a dataset invalid;
- Enabling elements can be content specific. Note that if a content object lacks a certain element (e.g., creator or language) and isn't included in the results this is still acceptable.
- It is possible to show what these elements 'do' in specific situations to motivate providers to provide them.

The next step for the DQC will be to distribute the current EDM fields across these two categories while considering the requirements raised by the Discovery-Usage scenarios.

[31] https://www.loc.gov/standards/iso639-2/php/code_list.php
[32] https://open-data.europa.eu/en/data/dataset/language

Possible recommendations for specific metadata elements were also discussed and will need to be refined and agreed in the coming months, helping to realize DSI-2 KPIs:

- clearer definition of the dc:title element. A title is key to identify an object in a list of results. The type of title expected by Europeana is however not always clear. What qualifies as a good title in Europeana?
- clearer distinction between dcterms:temporal and dc:date and its sub properties dcterms:created and dcterms:issued.
- normalisation of dates. Further investigations will be carried out to provide more recommendations on this point and support the DSI-2 KPI: Normalise date information for specific datasets or data partners (see below).
- clarification (for data providers) of the difference between describing a contextual resource such as an agent or a concept as a subject or as a property of a CHO (for instance an Agent can be the subject of a painting or the creator of that painting). The current discovery-usage scenarios are already providing hints to make this distinction clearer.
- removing dc:type from the mandatory element group “at least one of dc:subject, dc:type, dc:coverage or dcterms:spatial is mandatory” which is so far very much about record “aboutness”. Refinements of the definition of dc:type will inform the DSI-2 KPI: Have a vocabulary agreed and available to normalise dc:type. The DQC also discussed the possibility to have dc:type and dc:subject in the same group of mandatory elements (see below).
- having dc:creator present when the creator of a given CHO is known. this will allow a user to browse the Europeana dataset by creators.
- removing dc:coverage from the mandatory group “at least one of dc:subject, dc:type, dc:coverage or dcterms:spatial is mandatory”.
- dc:subject contains key information for resource discovery. The DQC asked whether it should be recommended to create subject information when it is not available in the legacy metadata. In this case it would be recommended to use controlled vocabularies.
- removing edm:type for the ProvidedCHO to add it in the Aggregation and/or the WebResource, as many data providers and aggregators duplicate the information required by edm:type in dc:type.
- providing more recommendations on using EDM elements to describe contextual resources.

Data quality plan for DSI-2

For Europeana DSI-2 we have agreed to adopt and implement recommendations and standards for metadata and content quality, including those proposed by the Data Quality Committee. In this context the Europeana Publishing Framework and Guide will be reviewed and updated and the EDM Mapping Guidelines will be aligned with the new standards. In collaboration with the DSI aggregating partners quality issues of legacy data will be addressed and new standards will either be gradually adopted or datasets will be depublished that are not possible to improve accordingly. As part of this work under DSI-2, more vocabularies are going to be used for recurring terms.

The following targets have been agreed within the Europeana DSI consortium:

- Normalise and deduplicate organisations providing data to Europeana (edm:provider, edm:dataProvider), incl. make sure organisations to have a unique identifier;
- Normalise values in dc:language;
- Normalise date information for specific datasets or data partners;
- Enriching metadata with synonyms and multilingual translations of scientific names for species (OpenUp!);
- Foster semantic enrichment of records using AAT, ULAN and TGN (MUSEU);
- Develop customised enrichment plans per data provider based on validation and quality reports (CARARE);
- Adding SKOS concepts for subject indexing for key CARARE data partners;
- Add subtitles for pre-teletext video sources to enhance access to audiovisual heritage (EUscreen);
- Improve accuracy, precision and specificity of titles and descriptions (Europeana Photography);
- Have a completeness measure in the API output, based on the Metadata Quality Assurance Framework developed by Péter Király;
- Have a vocabulary agreed and available to normalise dc:type.

Beyond the above data quality improvements, we will also improve (international) interoperability to switch to international rights statements (rightsstatements.org), improve implementation of EDM and offer support to distribute and display images according to the IIIF protocol.

Europeana Content Strategy

The Europeana services for use and re-use can only be as good as the content they have access to. While this data development plan gives an idea about the direction the Europeana database is going to develop, a strategic planning of data development is still missing. This content strategy must build on existing frameworks like the Europeana Publishing Framework and the recommendations of the Data Quality Committee, in order to improve data quality over the next years. A content strategy must, however, go beyond these aspects to identify the most effective and useful content for Europeana and make informed decisions about new content areas. It will help to identify how to "get the right content to the right users at the right time"[33]. It will also help to create an approach to quantify and show the value of content within our organisation and those of our partners. Further, a content strategy will help to create a process for efficient and effective content publishing, from creation through to publication.

Although data quality is an important aspect of the content strategy, several other concerns need to be addressed as well, in order to cover the many facets that are important to consider for the development of the Europeana database. It is mentioned in several places in this document that a thematic focus is getting more important for Europeana but a strategy on how we populate the Europeana Thematic Collections with the best possible data is not yet available. The same is true for how we systematically ensure that user demands are considered for the development of the

[33] From the Content Strategy Alliance definition of content strategy, http://contentstrategyalliance.com/the-beginnings/csa-charter/

Europeana database. It is also not clear whether Europeana should focus on publishing the few and well known masterpieces (like the Mona Lisa) or the lesser known cultural heritage collections. Non-digital objects are another priority, to e.g. understand the relevance of bibliographic data for Europeana. A clear approach on how to deal with non-European sources does also not exist to date. For all these aspects and some others, we also need to be clear how we want to incorporate and consider new developments (e.g. new metadata standards) and how the content strategy is implemented across the Europeana database (new and legacy data). As for the implementation, it is important to be clear on priorities for publication (of new data) and criteria for depublication (of legacy data).

The content strategy is being developed by a team of Europeana Foundation staff supported by four experts representing the key domains, i.e. libraries, museums, archives, audiovisual archives. The goal is to present the content strategy at the Europeana Annual General Meeting in November 2016, with 2017 being the first year the content strategy is formally in effect.

Partner development

This chapter looks at how Europeana will continue to develop its network of data partners. At the moment, over 3,700 cultural institutions from across Europeana have made (parts of) their collection available via Europeana. This great achievement could not have been done without a dedicated network of aggregators and projects. While aggregators are still key data partners for Europeana, we are exploring how we can engage more directly with cultural heritage institutions in Europe.

Aggregator Forum

Continuing our practice of the last years, we will facilitate interaction and the sharing of knowledge and experiences through several meetings for and with our partners, in particular the meetings of the Aggregator Forum. Over the years the Aggregator Forum meeting has grown from a relatively small get-together, to a full three day event with over 50 participants. It has become a crucial event where the aggregators can share their experiences, learn from each other and decide on next steps to be taken by them and Europeana. From having one meeting per year organised and facilitated by Europeana Foundation we are now running two Aggregator Forum meetings per year, with an aggregator being responsible for hosting and running the autumn edition.

Besides having meetings of the entire Aggregator Forum, we also organise meetings or workshops with and for individual aggregators and participate at events aggregators organise for their partners. This helps to increase the participation from a particular country or domain.

Fostering the relationship with aggregators in a spirit of cooperation and mutuality, emphasising the benefits of the partnership for each individual member and the ecosystem as a whole is a constant challenge. Some of the challenges have to do with the way the aggregator model works (see also chapter below on innovating the aggregation infrastructure). However, even with the current set up, we need to aim for a real partnership on equal terms between all players in the aggregation landscape and clear benefits for data partners. Establishing a partner satisfaction indicator should help to make this more visible and measurable.

Europeana Thematic Collections

Thus far, content acquisition for Europeana has been largely supply-driven where Europeana served as a 'broker' between the institution and the aggregator. We have started to also take a more active approach towards institutions that are not yet working with Europeana. While we plan to remain inclusive and open to everybody and every type of cultural heritage collection, we will also focus our attention more on institutions that hold material that fits in the themes that have been prioritised. Together with the aggregators we will try to define potential partners and actively approach them to see if they have an interest in sharing their collection via Europeana. This way we will improve the offering to our end user in the Europeana thematic collections (see D1.2 [36]).

Europeana thematic collections show a filtered view, based on a broad topic, of the Europeana database for specific audiences. Each one also contains editorial and curated content (e.g. online exhibitions, blogs) intended to meet the needs and interests of a user community with a specific interest in the theme of the collection. The first thematic collection created was Europeana Music[34], curated by Europeana Sounds, which was launched in a beta version in January 2016 and will be continually improved during 2016. On 30 April 2016 Europeana launched another thematic collection, Europeana Art History Collections[35], the purpose of which is to promote discovery of Europe's art collections for public enjoyment, education and research.

The Collections and Data Partner Services teams are currently evaluating, in collaboration with the respective aggregator, several collections that are considered to be of major importance for art history, but are not in Europeana yet or not in the quality we would need for the thematic collection. We are reaching out to the collection holders either directly or through aggregators, depending on the aggregation landscape in the country the institutions are based in. The success of these outreach efforts will be reviewed in DSI-2. In that sense the work on the Art History Collection is a pilot to set up and test a more structured approach to reaching out to major cultural heritage institutions in Europe and bring them into Europeana. This is described in more detail in the report on the amount of data partners and the outreach to major institutions (DSI D1.2).[36]

Innovating the aggregation infrastructure

The Europeana DSI is using an aggregation model that was created more than five years ago. While a few key elements changed along the way (data model, metadata licensing, ingestion tools), the fundamental principles of the aggregation model are still the same. This includes the way the current model supports a quantitative growth of the database instead of improvements of data quality. However, technology and demands have changed and so must the way data is collected and shared by cultural heritage organizations. This does not only affect the technical infrastructure but also the way we collaborate on organizational and individual levels.

[34] http://www.europeana.eu/portal/collections/music

[35] http://www.europeana.eu/portal/collections/art-history

[36] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d1.2-amount-of-data-partners-and-outreach-to-major-institutions.pdf

We need to transform the way we make Europe’s cultural heritage available by turning the aggregator model upside down: from a hierarchically organized top-down approach, we need to change and start collaborating as interconnected nodes that support each other and work together to provide value to Europe’s memory institutions and citizens: a web, not a pyramid (Fig. 1). The expert hub concept is key to this change: domain and thematic aggregators will become expert hubs of the Europeana DSI, recognizing the expertise they already provide and allowing for an increased emphasis on expertise-based services. Over time, we hope to replace the concept of aggregators with the concept of hubs (e.g. expert hubs for domain and thematic aggregators), with the consequence that ‘aggregators’ as defined in the Europeana Glossary of Terms will no longer exist.[37]

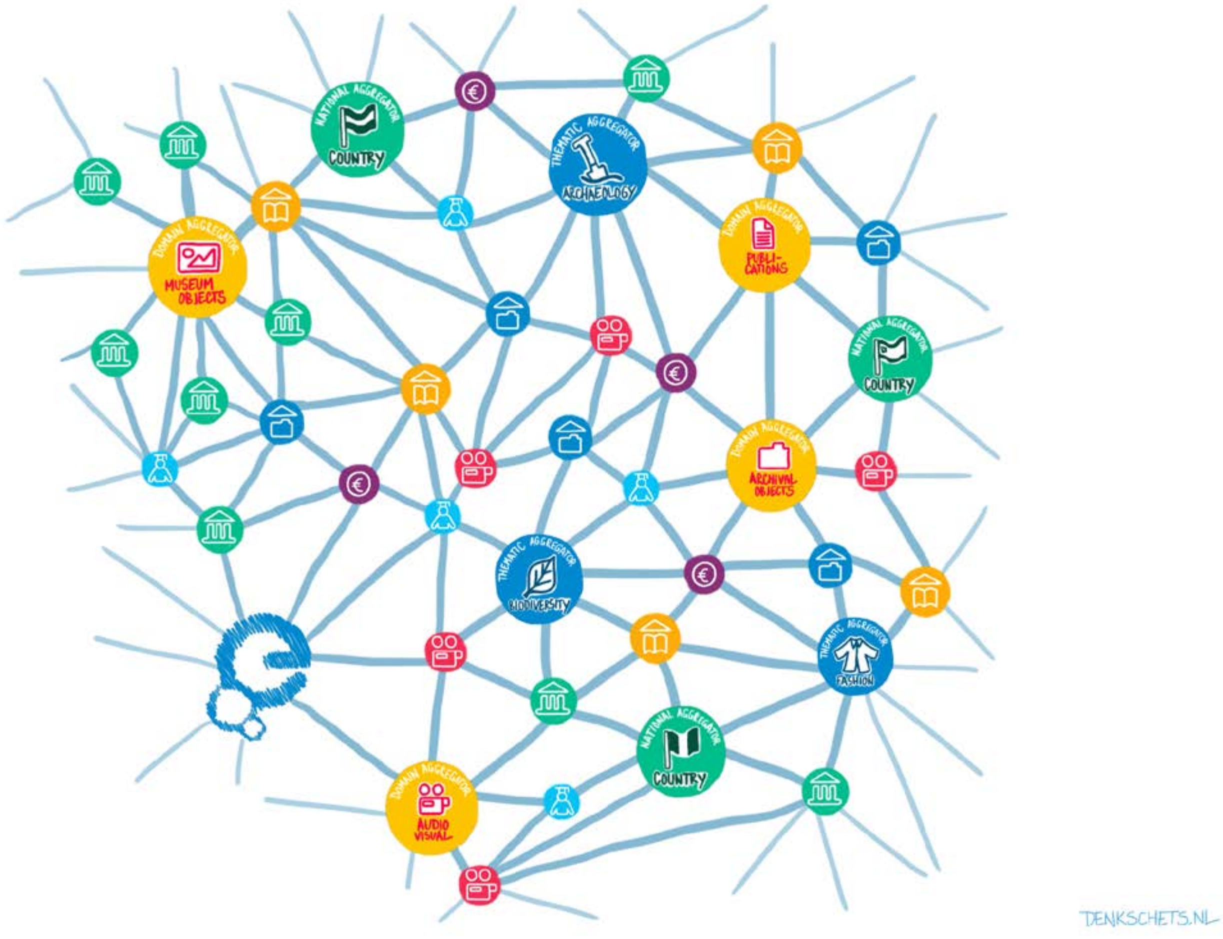

Figure 1. The hierarchical aggregator model should turn into a web of interconnected nodes of different sizes and capabilities that complement each other and work together to provide value to Europe’s memory institutions and citizens.

[37] See the definitions further above or http://pro.europeana.eu/page/glossary.

In 2015, we started to investigate how domain and thematic aggregators could shift their focus from pure data collection towards becoming expert hubs for their partners. The expert hub concept is going to be further developed in 2016 and the first expert hubs will launch in 2017. In 2017, we will also create a sustainability plan with domain aggregators who are becoming expert hubs. We will also continue to work with national aggregators on their role in the aggregation landscape. We will develop a joint plan on how best to manage the exchange of data between national aggregators, expert hubs and the Europeana Foundation.

The technical infrastructure we are developing aims to facilitate this change from aggregators to hubs and allow us to bring the highest possible quality of Europe's rich cultural heritage online. Together with the data partners, we aim to create a shared, dynamic, efficient and cost-effective metadata aggregation for the Europeana DSI. As part of the work on a new metadata aggregation infrastructure we are developing Metis, a toolset for harvesting, analysing, transforming, enriching and publishing data on the Europeana platforms (e.g. Europeana Collections). This happens in close consultation with the Europeana DSI aggregators, and we are holding individual meetings and workshops with them to work on the requirements for Metis. We are also prototyping more innovative lightweight and user-friendly solutions to acquire data from our partners. These solutions would allow data partners to directly interact with the Europeana data platform and empower them to improve the quality of their data through immediate feedback. For this work we are also consulting existing and potential data partners across Europe to understand their needs and consider their requirements for the first prototypes.

For a much more detailed overview of the concepts and roadmap behind this infrastructure innovation, see the work and implementation plan to innovate the aggregation infrastructure (Europeana DSI-1 D1.1).[38]

[38] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Deliverables/europeana-dsi-d1.1-work-and-implementation-plan-to-innovate-the-aggregation-infrastructure.pdf