Data quality committee - 2016 report

Revision	Final
Date of submission	31.12.2016
Author(s)	Valentine Charles (Europeana Foundation); Timothy Hill (Europeana Foundation); Antoine Isaac (Europeana Foundation)
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

Revision History

Revision No.	Date	Author	Organisation	Description
1	12/12/2016	Valentine Charles	EF	First draft
2	14/12/2016		EF	New version after feedback from Henning Scholz, Timothy Hill, Cécile Devarenne and Antoine Isaac, Europeana Foundation
3	31/12/2016	Antoine Isaac	EF	Wrapping up final comments

Table of contents

This document is an update on the activities of the Data Quality Committee (DQC) which kicked-off at the beginning of 2016. It follows a previous report submitted in June 2016 as part of the milestone 2: Updated Partner and Development[1]. The recommendations of the DQC for each activity are linked to the Europeana Product Development roadmap. The dependencies on that roadmap and other DSI2 activities are listed under each item and further detailed in annex.

The ongoing activities and efforts of the DQC are based on usage scenarios[2] created to reflect information-access user needs. Even though the DQC is now working on core metadata quality issues, all its recommendations are formulated in connection with these usage scenarios.

1. Mandatory metadata elements for ingestion of EDM data

The DQC continues its work on the definitions and use of the mandatory elements in the Europeana Data Model (EDM). The DQC has agreed on two new categorisations to characterise the current EDM elements: *mandatory* elements and *enabling* elements. This distinction has been proven necessary to convey in a clearer manner the idea of mandatoriness: while enabling elements support particular desirable but optional functionalities from a specific (set of) usage scenario(s), mandatory elements are required as a fundamental minimum for all metadata descriptions.

The DQC started by looking at the existing list of mandatory elements[3] and formulated new recommendations regarding them.The proposals went through a voting process within the DQC.

1.1. How to approach 'mandatoriness or mandatory groupings' in general

The EDM documentation defines several EDM elements as mandatory, either alone or as part of groupings, e.g., "Either dcterms:spatial or dc:type or dc:subject or dc:coverage must be provided" or "Either dc:title or dc:description must be provided". These groupings exist because data providers might not have metadata for all those elements and *artificially* filling them with values like “unknown” is not acceptable. However documenting these groupings in a clear manner remains a challenge. The DQC discussed different options but finally decided that retention of the existing groupings represented the best possible compromise to meet a wide range of requirements. There was general agreement, however, that effort should be made to clarify EDM documentation on this point. Most immediately, the wording of the groupings will be changed from, say "one of element dc:description or dc:title is

[1] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms-2-partner-and-data-plan.pdf

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

[3] http://pro.europeana.eu/share-your-data/edm-documentation

mandatory" to "If there is no dc:description for an object, there must be a dc:title. If both are available, provide both".
In addition, the DQC will further discuss the idea of having a “digest for beginners” to help data providers in their mapping choices (e.g., a summary table and/or use cases based on the user scenarios).

1.2. Recommendations about the mandatoriness of specific elements

EDM elements

dc:coverage and dcterms:temporal	
Current practice	"Either dcterms:spatial or dc:type or dc:subject or dc:coverage must be provided."
Recommendation	The group agreed to replace dc:coverage by dcterms temporal in the mandatory grouping: "Either dcterms:spatial or dc:type or dc:subject or dc:coverage must be provided." dc:coverage won’t be mandatory anymore.

dc:creator/contributor	
Current practice	dc:creator and dcterms:contributor are recommended in the current EDM mapping guidelines.
Recommendation	We agreed to keep both dc:creator and dcterms:contributor as recommended.

dc:title	
Current practice	The current definition is “The title of the CHO”.
Recommendation	We agreed to change the current definition with the standard definition from Dublin Core[4] “A name given to the resource”. dc:description remains an alternative to dc:title in case and only if this information is missing: "dc:title SHOULD be present; but if there is no dc:title available (for example for archaeological collections), it is acceptable to have dc:description instead."

[4] http://dublincore.org/documents/dcmi-terms/#terms-title

dc:type and dc:subject	
Current practice	"Either dcterms:spatial or dc:type or dc:subject or dc:coverage must be provided."
Recommendation	These elements can't be made mandatory on their own as it will be difficult for data providers to provide both. The DQC agrees on the current level of mandatoriness of these two elements but would like to emphasize the need for them in the documentation. The DQC will work on guidelines as part of the discussion on the enabling elements.

edm:currentLocation	
Current practice	There are currently no specific recommendations for this element.
Recommendation	The DQC is discussing best practices for this element.

edm:type	
Current practice	This element is mandatory in the edm:ProvidedCHO class.
Recommendation	The DQC agreed to move away edm:type from edm:ProvidedCHO to edm:WebResource as the type AUDIO, VIDEO, TEXT and IMAGE describe more the WebResource than a cultural heritage object (CHO). This change will affect some functionalities of the Europeana Collections portal and will therefore need to be discussed with the Europeana Product Development team (see D5). edm:type will be made recommended for the WebResource used in edm:isShownBy and edm:hasView.

EDM contextual classes

Current practice

There are currently no mandatory elements for the contextual classes. The recommendations of the DQC apply only when a contextual entity is provided.

Recommendations

edm:Agent

The DQC makes skos:prefLabel mandatory, recommends to have several skos:prefLabel with different languages tags (for translations), and recommend the use of skos:altLabel, rdaGr2:dateOfBirth and rdaGr2:dateOfDeath.

edm:Place

The DQC recommends to make mandatory skos:prefLabel or the combination of wgs84_pos:long and wgs84_pos:lat and the use several skos:prefLabel with different languages tags.

edm:TimeSpan
It was agreed to make skos:prefLabel or the combination of edm:begin and edm:end mandatory and to recommend several skos:prefLabel with different languages tags. The DQC notes the difference between timespans as chronological values (different PrefLabel can be used e.g. "3200BC" is written "3200AC" is some languages, "20th century") and historical timespans (e.g. “Bronze Age”, “Middle Age”) . edm:TimeSpan shouldn’t be recommended for the latter.

skos:Concept
The DQC recommends to make mandatory skos:prefLabel, make several skos:prefLabel with different languages tags recommended and skos:altLabel recommended unless several prefLabel are already given with different language tags (especially for the case of translations: altLabel are indeed not suitable for translations of prefLabel.)

1.3. Proposals on other topics (not directly related to mandatory elements)

Adding new elements to existing classes

dc:type	
Current practice	This element is currently only available at the level of edm:ProvidedCHO.
Recommendations	Following the discussion and suggestions on the mandatoriness of edm:type, the DQC recommended to enable the use of dc:type for edm:WebResource.

edm:currentLocation and edm:Place	
Current practice	There are currently no specific recommendations for edm:currentLocation in relation to edm:Place.
Recommendations	Future guidelines for edm:currentLocation (see above) might require the addition of an address element to edm:Place.

Recommendations about values

The DQC agrees on the following guidelines for specific elements:

- CHOs should have distinct dc:titles

- dc:type should not be (strictly) identical to edm:type
- high-level dc:subject values like 'archaeology' are allowed, especially when there's no other subject that can be easily filled in.

Updates in the EDM documentation

The DQC agreed on a series of updates for given EDM elements described as part of the EDM mapping guidelines[5].

dcterms:isPartOf and dcterms:hasPart
Addition of the note: "It is possible to use either dcterms:isPartOf or dcterms:hasPart to express relation between objects in a hierarchy. However in many cases (especially when a parent object has many children) it is preferable to use dcterms:isPartOf".

dc:date, dcterms:temporal, dcterms:created and dcterms:issued
Addition of the note: "NB: other EDM elements are relevant for expressing dates of different events in the life of the CHO: dc:date, dcterms:temporal, dcterms:created and dcterms:issued. Be careful and choose the most appropriate one!"

edm:currentLocation and dcterms:spatial
The difference between edm:currentLocation and dcterms:spatial will be improved with the note: “edm:currentLocation is used only to record the place where the CHO is currently held (e.g. a museum or gallery); dcterms:spatial is used to record the place depicted in the CHO and other locations associated with it. Be careful to choose the most appropriate one”.

Identifiers
The guideline note on identifiers[6] should indicate that it is mandatory to provide a unique identifier for each class; that the identifier for the aggregation is required to be distinct from the identifier for the CHO; or that the identifier for the web resource is required to be a URI/pointer to the object/its landing page. The note on identifiers should be clearly referenced in the EDM mapping guidelines.

Structure of EDM mapping guidelines
The DQC agreed that this document should be improved, especially wrt. its rationale and structure, so as to highlight the semantic “affinities” of the EDM elements, e.g., by grouping the elements together when they enable similar discovery scenarios.

Dependencies with the Europeana Product roadmap and other activities (see Annex)

[5] http://pro.europeana.eu/EDM_Documentation
[6] http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/FAQs/URIs%20in%20EDM_pro.pdf

D1-Normalise date information for specific datasets or data partners
D2-Normalise values in dc:language
D3-Have a vocabulary agreed and available to normalise dc:type.
D5-Implementation of additional mandatory elements in EDM
D6-Update and improvement of the EDM documentation

The next step for the DQC will be to identify the elements that are characterized as "enabling elements" according to the new definition.

2. Metadata completeness measure and multilingual saturation

The work on metadata completeness driven by Péter Király[7] has progressed. Completeness calculations and visualisations are now used to inform the DQC's ongoing discussion of such issues as different completeness profiles, the contribution of mandatory fields to completeness, and the relative importance of language tagging. Once complete, this work is expected to yield a metric for metadata completeness that can be used to inform data re-users and Europeana officers of record quality and to improve functionality like search and ranking.[8]

At the time of writing, the DQC is focused on measuring the degree of multilinguality or 'multilingual saturation' of the Europeana datasets. It consists mainly in establishing measures for determining the multilingual richness of the metadata, based on the elements identified in the White Paper "Best Practices for multilingual access to digital libraries"[9]. Broadly speaking, the multilingual saturation score will be based on the number of appropriately-language-tagged translations provided for each field within a record. Specific concerns such as normalisation and scaling, or the contribution made by links to multilingual controlled vocabularies, are still under discussion. We expect to reach a conclusion by the end of Q1 2017.

Dependencies with other DSI2 activities (for more detail see Annex)
D4-Have *one* completeness measure in the API output, based on the Metadata Quality Assurance Framework developed by Péter Király
D7-Definition of a quality plan and targets for data partners

3. Gathering and detecting problem patterns with metadata values

[7] http://www.slideshare.net/pkiraly/metadata-quality-assurance-framework-at-qqml2016-full
[8] See below, Annex D4,
[9] http://pro.europeana.eu/publication/best-practices-for-multilingual-access

The Committee has gathered a list of problem patterns for metadata values (e.g., data normalisation issues) that affect search and interfere with ranking algorithms. Considerable effort has been spent identifying and organising the problems. The list contains now around 40 problems categorised as follows:

- **Description:** type of problem, e.g. identical title and description fields;
- **Evidence:** a metadata example illustrating the problem, e.g. http://www.europeana.eu/portal/record/2023702/35D943DF60D779EC9EF31F5DFF4E337385AC7C37;
- **Negative impact:** description of the negative effect of the pattern on the data, e.g. Distorts search weightings;
- **Notes:** general observations on impact, not covered under other headings;
- **Discovery Scenario affected:** the scenario affected, as outlined in the Metadata and Discovery Scenarios document[10].;
- **Action:** a remedy that could be be taken against the problem such as reporting, correcting or normalising it;
- **Severity:** how serious the problem is;
- **Solutions for checking:** possible solution(s) to detect and/or solve the problem. Includes notes on the anticipated method, technology and difficulty of tackling the problem.

The next steps will consist in (1) prioritising the problems and start testing some technologies to recognize them in the data (e.g. SHACL[11]) and maybe some of them; (2) start disseminating this list of problem patterns to the Europeana data providers to raise awareness.

Dependencies with other DSI2 activities (for more detail see Annex)
D7-Definition of a quality plan and targets for data partners
D8-Data prototyping and investigations of new technologies

4. Coordination with other quality-related initiatives

The DQC is constantly working to link its efforts to other projects from the Europeana Network and beyond. More recently the work of the DQC was shared with the DLF Metadata Assessment Working Group[12] in the USA with which we hope to collaborate. Our work was also presented during the last Semantic Web for Libraries conference (SWIB)[13].
The public page of the DQC provides an overview of all our activities[14].

[10] https://docs.google.com/document/d/1ej0ouDg_uhOVnE1LE2-IEtI9xNhMpeqzNIIwSjLoAbI/edit#
[11] https://www.w3.org/TR/shacl/
[12] http://dlfmetadataassessment.github.io/
[13] http://swib.org/swib16/
[14] http://pro.europeana.eu/get-involved/europeana-tech/data-quality-committee

Annex: Dependencies with the Europeana Product Development roadmap and other DSI2 activities

The work of the DQC has dependencies between several activities and products defined as part of the DSI2 programme, either in the data quality plan reported in MS2 or the Europeana Product Development roadmap (defined in the context of WP6). This annex reports on these dependencies.

D1-Normalise date information for specific datasets or data partners
Normalisation of date values will be handled by a new cleaning and normalisation service developed as part of the Metis data services. This work happens outside of the DQC, however the DQC will provide (and has already begun discussing) recommendations on the normalisation rules adopted for the EDM elements supplying dates information.

D2-Normalise values in dc:language
Language normalisation is also handled as part of the Metis data services. The DQC can provide input on the vocabularies to choose to perform the normalisation of dc:language.

D3-Have a vocabulary agreed and available to normalise dc:type.
The DQC will be consulted on the vocabulary after initial discussions have taken place within the Europeana Foundation. The choice of a vocabulary for dc:type could be guided by an analysis of the logs from the Europeana Collections, which would provide a good overview of the terms used by users when searching for a type of objects.

D4-Have *one* completeness measure in the API output, based on the Metadata Quality Assurance Framework developed by Péter Király
One measure for completeness will be included in the Metis statistics data service delivered as part of DSI2 and included in the Europeana datasets after re-index. The DQC is directly involved in this task by:

- Discussing how the metadata should be measured.
- Investigating ways to analyse, interpret and visualise the results.

The final product will result in a completeness score which will be included in the metadata, and set of statistics (such as the multilingual saturation) which could be included in a quality statistic dashboard as part of the Metis development.

D5-Implementation of additional mandatory elements in EDM
The DQC identified new mandatory elements that will need to be implemented in the official EDM. We will need to discuss with the Europeana Product Development team whether the recommendations can be implemented and when the implementation would take place.

DQC discussions and recommendations on improving the data will also support the work on the Europeana Search improvement[15] and the plan for populating and curating the Europeana Entity Collection[16].

[15] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Milestones/ms6.6-search-improvement-plan.pdf

The work of the DQC has also dependencies with other activities that do not directly belong to product development.

D6-Update and improvement of the EDM documentation
The EDM documentation will have to be updated to include the recommendations from the DQC on mandatory and recommended elements and on data quality. In the longer term this update could result in the definition of a new conceptual framework, which will articulate data quality recommendations together with the user scenarios defined by the DQC.

D7- Definition of a quality plan and targets for data partners
The list of problems patterns identified by the DQC will be used to define new quality targets for data partners. A new quality plan will allow for the incremental adoption of the recommendations and proposals of the DQC.

D8-Data prototyping and investigation of new technologies
R&D efforts will be pursued as part of the DQC. We will investigate extensions of the data model - notably, the support of Events. New technologies for validating and reporting on EDM data problems will be tested.

Beyond the dependencies mentioned above, we will also follow the work done by the Content Strategy team[17].

NB: the following targets from the DSI2 data quality plan are out of scope for DQC:

- Normalise and deduplicate organisations providing data to Europeana (edm:provider, edm:dataProvider), incl. make sure organisations to have a unique identifier;
- 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)

[16] https://docs.google.com/document/d/1A5Rb3Oe9edin5gdRpqFILlR0YPUodVOel3SdcBP00dA/
[17] See the presentation on work in progress on the Content Strategy at the October 2016 Aggregator Forum: https://docs.google.com/presentation/d/1MZ88iUa02viDvG3xByOn44DFiy3oGQbwskDMWhvpW7A/