Data Quality Committee
Mission Statement

General goals and principles

Quality is a key priority for the Europeana community. "Less is more" will be our mantra for the coming years, as we try to push less but higher quality data into Europeana that is fit for purpose. With the Europeana Publishing Framework[1] and the recent Task Forces on metadata quality and enrichment, we have general recommendations for quality improvements. What is missing is a good plan to agree on new standards based on these recommendations, and how providers and aggregators can implement these standards. The Data Quality Committee works to address these issues over time. Formally defined as a Europ[2]eana Network / EuropeanaTech[3] Working Group, the committee comes up with proposals on quality improvements.

We believe it is essential to tackle data quality issues and feedback at every level of the data exchange chain, especially before data reaches Europeana (i.e., data providers and aggregators). Therefore, many recommendations will be intended for implementation at the source of the data. In the Europeana context such recommendations will be submitted for ratification at each Europeana Aggregator Forum meeting, twice a year. We especially

---

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

[2] http://pro.europeana.eu/structure/our-network

[3] http://pro.europeana.eu/structure/europeana-tech

expect recommendations to feed into strategy documents such as the Europeana Content Strategy, or technical guidelines or rules such as the EDM data validation rules[4].

The Committee considers that data quality is always relative to intended use and cannot be analysed or defined in isolation from it, as a theoretical effort. Recommendations should be connected to usage scenarios. Quality models for specific domains (archaeology, art...) may be promoted, but we do not expect that the Committee will create or even extensively discuss these models. Our focus should be on (re-)use scenarios and not on doing work that rather belongs to the level of the domains themselves. However, we expect that many usage scenarios in the Europeana context are shared across domains and other data aggregation initiatives. While the Europeana case(s) will be the base for our work on identifying quality needs, we will seek to recognize these needs in other contexts and formulate recommendations at a level that can be beneficial to all parties involved. The committee may even provide guidelines on selecting data based on quality for re-users of (Europeana) data.

To properly take into account these principles into our work, the committee consists of representatives of data aggregators and providers, data re-users, technology providers, as well as from the Europeana office and international partners.

As a general rule, the Committee is not expected to create recommendations from scratch. We expect to mostly refer to work done elsewhere - sometimes by Committee members in other circles. It is of course perfectly fine if Committee members explore specific areas in a more detailed way outside of the Committee proper, e.g. in the context of initiatives in Europeana (e.g., EuropeanaTech task forces) or outside.

Scope of work

The following areas are expected to structure the Committee's work. Specific tasks will be defined and prioritized within these areas as the Committee sees fit.
For discussions on more detailed aspects within each of these working areas refer to the document at:
https://docs.google.com/document/d/1ZOSKrX7X9oU7Ob3wjinXVWenx8HCq5qnhUzljcpFKs8/.

1. Mandatory metadata elements for ingestion of EDM data

---

[4] http://pro.europeana.eu/edm-documentation

- The Committee needs to investigate if the current mandatory elements are relevant and sufficient. It also needs to propose how to make legacy data compliant with the agreed list of mandatory elements, or whether to disable it in some (data) exposition channels. Europeana still has several hundred thousand records that don't have all current mandatory elements.

- The work will include recommendations on measures of completeness for descriptive metadata (i.e., not content) based on presence/absence of fields, not their values (which is the topic of 'meaningful metadata values') below.

- Generally the Committee could have a critical look on the EDM definitions and guidelines, so that it makes EDM more precise and helps data providers to make consistent, less ambiguous choices. Guidance to make impactful changes (esp. on legacy data) a realistic prospect for data providers will be welcome.

2. Data checking and normalization

- The Committee should look into ways and rules to normalise metadata. This includes the use of vocabulary based values or normalised values.

- Recommendations for tools and services to validate or detect anomalies in EDM data (either at Europeana or provider/aggregator level) using technology such as XML validation or RDF Shapes.

3. Meaningful metadata values (in the context of use)

- Metadata values intended for human consumption (e.g. titles, descriptions in object metadata or labels attached to URIs) that are not unique or understandable are a concern, to just give one example. The Committee should look into ways of recommending meaningful metadata values (where 'meaningful' needs to be defined in the context of use) and indicators to measure improvements.

- This work includes measures for information value of statements (informativeness, degree of multilinguality...)

- Focus can be put on making recommendations about and assessing the impact on quality of contextual links, including these created via semantic enrichment. I.e. the info we get from data around the object. For instance, the number of labels in a DBPedia or Wikidata entity attached to an object, or obtained through aligning two concepts from different thesauri. This work could be done in parallel with other work on metadata enrichment and evaluation[5].

4. Quality of content (digital media)

---

[5] http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments

- The Committee may (probably at a later stage and with appropriate expertise on digital media formats) look at the quality levels for all types of content that are proposed by the Europeana Publishing Framework and work on improvements.

5. Coordination with other quality-related initiatives

- For example the DPLA-initiated discussions on quality[6]

First work items and milestones

NB: The Committee can present the results of its work (e.g. recommendations) in any form of output that it judges relevant: a technical report, a presentation...

M1: approve mission statement
M1-2: Operational principles for the group: communication channels, meeting, defining sub-groups if required.
M1-2: read and comment D1.3 from DSI Year 1
M1-2: agree on living document for discovery user scenarios employing metadata[7], stating the known problems we have to fix and identifying the gaps in current recommendations and guidelines.
M1-4: contribution to the Europeana Content Strategy (in fact this is 'data strategy'), due in May 2016

Organization

Co-chairs: Valentine Charles, Antoine Isaac
Secretary -Kirsten de Hoog
Communication channels:
- Basecamp: https://basecamp.com/1768384/projects/11189810
- Bi-weekly calls

---

[6] https://docs.google.com/document/d/15pmA276_fxShkCEagoIoJwCXH89PhrF3qWBgB8xSrag/

[7] https://docs.google.com/document/d/1ej0ouDg_uhOVnE1LE2–IEtI9xNhMpeqzNIIwSjLoAbI/