Measuring completeness as metadata quality metric in Europeana

Péter Király
pkiraly@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
Germany

Introduction

The functionalities of an aggregated metadata collection are dependent on the quality of metadata records. Some examples from Europeana, the European digital library, to display the importance of metadata: (a) Several thousands records have the title „Photo" without further descriptions; how can a user find these objects?, (b) Several data providers listed in the „Institution" facet under multiple different names, should we expect that the user will select all name forms of an organization?, (c) Without formalized date value, we are not able to use the functionality of interactive date range selectors. The question is how can we determine which records should be improved, and which are good enough? The manual evaluation of each record is not affordable. This paper proposes a methodology and a software package, which can be used in Europeana and elsewhere in the domain of cultural heritage.

Background and foundations

Europeana collects and presents cultural heritage metadata records. The database contains more than 53 million records from more than 3200 institutions (figures extracted from the Europeana Search API) in the Europeana Data Model (EDM) schema. The organizations send their data in EDM or in another metadata standard. Due to the variety of original data formats, cataloging rules, languages and vocabularies, there are big differences in the quality of the individual records, which heavily affects the functionalities of Europeana's services.

In 2015 a Europeana task force investigating the problem of metadata quality published a report (Dangerfield et al., 2015), however – as stated – „there was not enough scope … to investigate … metrics for metadata quality …." In 2016 a wider Data Quality Committee was founded. The current research is conducted in collaboration with it, having the purpose of finding methods, metrics and building an open source tool (see also, the project's Github page) to measure metadata quality.

State of the art

The computational methods of metadata quality assessment emerged in the last decade in the domain (Bruce and Hillmann, 2004, Stvilia et al., 2007, Ochoa and Duval, 2009, Harper, 2016). Papers defined quality metrics and suggested computational implementations. They however mostly analyzed smaller volumes of records, metadata schemas which are less complex than EDM, and usually applied methods to more homogeneous data sets. The novelty of this research is that it increases the volume of records, introduces data visualizations, and provides open source implementation to use in other collections.

Methodology

For every record, features were extracted or deducted which somehow related to the quality of the records. The main feature groups are:

- **simple completeness** – ratio of filled fields,
- **completeness of sub-dimensions** – fields groups support particular functions, such as searching, or accessibility,
- **existence and cardinality of fields** – which fields are filled and how intensively.

The measurements happen on three levels: on individual records, on subsets (e.g. records of a data provider), and on the whole dataset. On second and third level we calculate aggregated metrics; the completeness of structural entities (such as the main descriptive part and the contextual entities – agent, concept, place, timespan – connecting the description to linked open data vocabularies).

The final completeness score is the combination of two approaches. In the first one the weighting reflects sub-dimensions. In the second one, the main factor is the normalized version of cardinality to prevent biasing effect of extreme values.

The tool – built on big data analytics software Apache Spark, the R statistical software and has a web front-end – is modular. There is a schema-independent core library and schema specific extensions. It is designed to be used in continuous integration for metadata quality assessment.

Results

Comparison of the scores of the field importance and field cardinality approaches shows that they give different results (however they correlate by the Pearson's coefficient of 0.52.). Because of the nature of calculation the compound score is quite close to the first approach: the functionality based scores lie in the range of 0.186 and 0.76 and cardinality scores are in the range of 0.031 and 0.335, and it has smaller effect on the final score.

There are data providers, where all (in some cases more than ten thousand) records get the same scores: they have uniform structure. The field-level analysis shows (what one simple score is not able to testify) that in these collections all the records has the very same (Dublin Core based) field set. On the other end there are collections where both scores diverge a lot. For example in the identifying sub-dimension a data provider has five distinct values (from 0.4 to 0.8) almost evenly distributed while one of the best collection (of the category) is almost homogeneous: 99,7% or the records have the same value: 0.9 (even the rest 0.3% has 0.8). It means that in the records of the first dataset the corresponding fields (dc:title, dcterms:alternative, dc:description, dc:type, dc:identifier, dc:date, dcterms:created and dcterms:issued in the ore:Proxy part and edm:provider and edm:dataProvider in the ore:Aggregation) are frequently not available, while they are almost always there in the second. The tool provides different graphs and tables to visualize the distribution of the scores.

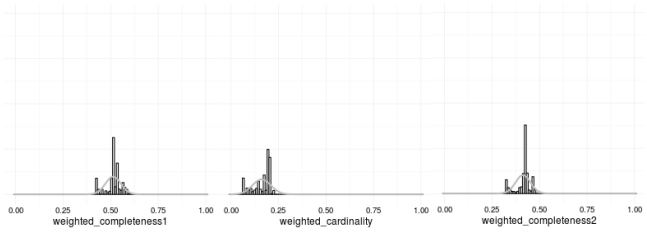

Figure 1. Distribution of completeness scores in a dataset. We can see the differences between the functionality based (left),the cardinality based (center) and the combined method (right).

From the distribution of the fields the first conclusion is that lots of records miss contextual entities, and only a couple of data provider has 100% coverage (6% of the records has *agent*, 28% has *place*, 32% has *timespan* and 40% has *concept* entities). Only the mandatory technical elements appear in every records. There are fields, which are defined in the schema, but not filled in the records and there are overused fields – e. g. *dc:description* is frequently used instead of more specific fields (such as table of contents, subject related fields or alternative title).

Users can check all the features on top, collection, and records level on the web interface. Data providers get a clear view of their data, and based on this analysis they can design a data cleaning or data improvement plan.

Europeana is working on its new ingestion system which integrates the tool. When a new record-set will arrive, the measuring will run automatically, and the Ingestion Officer can check the quality report.

Further work

We will examine other metrics (e.g. multilinguality, accuracy, information content, timeliness), and check known metadata anti-patterns. We plan to compare the scores with experts' evaluation and with usage data and to implement related W3C standards: Shapes Constraint Language (Knublauch and Kontokostas, 2016), and Data Quality Vocabulary (Albertoni and Isaac, 2016).

Conclusion

In the research we re-thought the relationship between functionality and the metadata schema, implemented a framework which proved to be successful in measuring structural features which correlate with metadata issues, and we were able to select low and high quality records. We remarkably extended the volume of the analyzed records by introducing big data tools, which were not mentioned previously in the literature.

I showed my research in case of a particular dataset and data schema but the method I follow based on generalized algorithms, so it is applicable to other data schema. Several DH researches based on schema defined cultural databases, and in those cases the research process could be improving by finding the weak points of the sources.

Acknowledgements

I would like to thank all of the members of the Europeana Data Quality Committee.

Bibliography

Dangerfield, M-C. et al. (2015). "Report and Recommendations from the Task Force on Metadata Quality." (http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf)

Bruce, T. R. and Hillmann, D. I. (2004). "The Continuum of Metadata Quality: Defining, Expressing, Exploiting." In

Hillman, D. and Westbrooks E. (eds), *Metadata in Practice*, Chicago, ALA Editions, 2004.

Stvilia, B., Gasser, L., Twidale, M. B. and Smith, L. C. (2007). "A framework for information quality assessment." *Journal of the American Society for Information Science and Technology*, 58(12): 1720-1733.

Ochoa, X. and Duval, E. (2009). "Automatic evaluation of metadata quality in digital repositories." *International Journal of Digital Libraries*, 10: 67-91.

Harper, C. (2016). "Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA)." *The Code4Lib Journal*, 33 (http://journal.code4lib.org/articles/11752)

Knublauch, H. and Kontokostas, D. (eds.) (2016). "Shapes Constraint Language (SHACL). W3C Working Draft 14 August 2016." (https://www.w3.org/TR/shacl/)

Albertoni, R. and Isaac, A. (eds.) (2016). "Data on the Web Best Practices: Data Quality Vocabulary. W3C Working Group Note 30 August 2016" (https://www.w3.org/TR/vocab-dqv/)