Towards automatic quality assessment of component metadata

Thorsten Trippel, Daan Broeder, Matej Durco, Oddrun Ohren
University of Tübingen, Max Planck Institute of Psycholinguistics, Austrian Academy of Sciences, National Library of Norway
E-mail: Thorsten.Trippel@uni-tuebingen.de, Daan.Broeder@mpi.nl, Matej.Durco@assoc.oeaw.ac.at, Oddrun.Ohren@nb.no

Abstract

Measuring the quality of metadata is only possible by assessing the quality of the underlying schema and the metadata instance. We propose some factors that are measurable automatically for metadata according to the CMD framework, taking into account the variability of schemas that can be defined in this framework. The factors include among others the number of elements, the (re-)use of reusable components, the number of filled in elements. The resulting score can serve as an indicator of the overall quality of the CMD instance, used for feedback to metadata providers or to provide an overview of the overall quality of metadata within a repository. The score is independent of specific schemas and generalizable. An overall assessment of harvested metadata is provided in form of statistical summaries and the distribution, based on a corpus of harvested metadata. The score is implemented in XQuery and can be used in tools, editors and repositories.

Keywords: Component Metadata Description Infrastructure; CMDI; metadata quality assessment; metadata score; quantitative quality metrics.

1. Need for quality assessment of metadata

The quality of metadata is essential for fulfilling the purpose for metadata, namely resource discovery including all aspects as identifying resources, selecting resources from a set of resources, acquiring resources. Also basic data management tasks require high quality (technical) metadata. When running archives and repositories, metadata is essential and has been used for catalogues and indexes in museums, archives and libraries. Lately, additional classes of resources have been made subject to archiving, e.g. digital research data from the humanities and social sciences, among others these are lexical databases, text collections and corpora, interview recordings. Typically, metadata for research data is not created by archiving specialists and librarians (ASL) but by the subject matter experts (SME) creating the resources on the first hand or sometimes by enthusiastic individuals contributing to a project. This process can be seen also in other areas facilitating for example crowd sourcing. With the SME or lay person creating the metadata, there is a need for providing immediate feedback on the quality of the metadata to the SME in order to support provision of accurate and complete information in the metadata. It could also be required by applications working with metadata to set a threshold of quality to process and include specific data.

2. State of the art

Bruce and Hillmann (2004) define seven characteristics or indicators of metadata quality, e.g. *completeness* and *accuracy.* To assess the quality of a metadata record in terms of the seven characteristics they provide a checklist of 18 questions, but no metric or measurement process. However, quantitative measures for one or more of the quality characteristics are defined in several evaluation studies, notably for completeness, see Ochoa and Duval (2009) and Kapidakis (2011).

Hughes (2004) attempts to measure the quality of metadata for the OLAC metadata schema. He measures the existence and absence of metadata and weighting these to create a sore of 0 to 10. As the OLAC metadata set is too restrictive for the description of general language resources, his approach can only contribute to assessing metadata quality within the Component Metadata Description framework (CMD, see the ISO 24622 standards family under development and Broeder, et al., 2010; Broeder et al., 2012). Indeed, it seems possible to classify CMD components by importance, but it should not be ignored that the context of the resource type that is being described influences the judgement if a component is relevant or not. Hence it would be required to include this when using importance ratings of components in an evaluation process.

Kapidakis (2011) also tries to define criteria for measuring the quality of metadata, counting the number of metadata elements per record, the number of distinct elements used, length of descriptions in free text fields, use of elements with specific meaning, especially with regards to closed vocabularies and the selection of closed vocabularies. With these criteria he analyses whole collections of metadata, also taking into account different metadata sets. As he takes into account various metadata sets, this approach is more useful, though the calculation is still based on implementing a function for each metadata schema used, and overlapping semantics of metadata schemas are not included in the assessment.

3. Quality assessment for component metadata

Current metadata editors provide assistance especially for syntactic correctness, for example by syntactic parsing of XML documents against the assigned and used schema. We assume that the syntactic correctness is a solved issue, though especially for closed vocabularies, date formats, etc. it is still easy to create useful but invalid metadata as it is still not uncommon that XML metadata is modified by non-XML editors.

For quality of metadata there are a number of challenges, some posing a dichotomy to each other. On the one hand, the definition of quality for metadata is often vague and contradicting.

On the other hand, for interpreting the metadata content, the semantic quality is essential. Thieberger (2012) asks for metadata "to make the data comprehensible". The semantic quality of metadata is rather hard to measure; techniques like clustering, topic detection, keyword extraction do not seem appropriate for highly structured metadata. Some of these measures could be applied to individual fields of a metadata schema, but not for a whole metadata file in which some fields may be filled with anything between restricted values to controlled vocabularies, patterns or may even be optional. Occhoa and Duval (2009) suggest the cosine difference between the resource vector and the metadata vector as a metric for measuring accuracy or correctness of the metadata. But this is probably more suitable for metadata of information objects than that of research data.

Kapidakis (2011) points out that the evaluation of the quality of metadata also depends on the underlying metadata schema even in cases where the shared metadata is not available in the original data format. Questions of descriptiveness and detail are often dependent on the underlying schema of the data. The metadata schema used may simply be a mismatch for the resource type, the available information and the particular purpose of the metadata. Often this becomes clear only from information beyond the direct context of the metadata and resources themselves.

In the case of the CMD framework the evaluation has to take the profile as such into account as different types of resources may be described by different profiles. In CMD terminology profiles correspond to XSchemas used for syntactic definition of the metadata instances, also containing concept links to the metadata categories. CMD allows for structured profiles, in which parts are bundled in (possibly nested) components, to be reused in other profiles. These components play their own role in the quality discussion, contributing differently to the overall metadata record quality when used in different profiles, but having the same expressive power.

From recent discussions around metadata quality for CMD based metadata schema the following considerations emerged.

The metadata quality is applying the following criteria:

1. the expressive power of the metadata schema
2. The use of concept links in the schema that allow semantic interoperability with other metadata schema.
3. the fullness to which the schema has been put to use, i.e. sparseness of the record.
4. the quality of the provided information
5. the truth of the description
6. the presence of recommended general data categories as used by search interfaces such as the VLO (www.clarin.eu/vlo) and often related to metadata schemas such as Dublin Core or OLAC.

In the conceptualization of Bruce and Hillmann (2004), Metrics 1-3 could be seen as a measure for completeness (on model and instance level), whereas Metrics 4-5 measure (syntactic and semantic) accuracy. To form a complete measure of the expressive power of a metadata instance, metrics 1-3 must be combined. For instance, a meager metadata model invariably result in meager metadata instances, even if all its elements are assigned values in the instances.

4. Defining measurable quality indicators

When defining measureable quality indicators for CMD-based metadata we follow the approach by Kapidakis (2011) and Hughes (2004) in defining measures based on the presence of metadata elements, length of content in free text fields and penalizing missing elements. We also take into account the presence of recommended data categories independent of their location in the hierarchical structure of the instance as the concrete name of the element is not fixed within the framework, also allowing features of multilinguality and community preference.

We distinguish as quality indicators for component:

- Number of defined elements
- Number of unique components used
- Total number of components used
- Number of elements defined from a core set of metadata categories including those explored by search tools such as the VLO
    - resource title or name,
    - modality,
    - resource class,
    - genre,
    - keywords or tags,
    - country,
    - contact person,
    - publication year,
    - etc.
- Public accessibility of the profile
- Number of references to distinct data categories defined externally
- Ratio of elements with data categories

Though it can be argued that none of these criteria in itself are sufficient indicators for the quality of a profile, it seems evident that good profiles will define a variety of elements, provide the concept links, (re-)use components and have some elements that can be mapped to Dublin Core or similar central schemas. A score that is based on these factors will rank high if the profile includes such general data categories and provides additional bonuses for resource type specific additions that are being used for the description.

For metadata instances, the following criteria are being used:

- XML validity according to the profile
- Number of elements
- Number of filled in elements
- Length of the description of description elements

As an important use of metadata records is to give access to the actual resources, one important indicator is the existence and validity of links (penalty for "dead links") – primarily the resource links ("Resource Ref"), but ideally all URLs in a record should be checked. This feature is yet to be implemented.

5. Implementing the score

Due to the variability of the CMD framework in defining profiles, a flexible approach will have to be used that is automatically adjusting to the profile. Using the metadata instance as the starting point this is achieved by extracting the information on the schema from the metadata instance and retrieving the component description from the component registry. For ease of use and to adjust for a higher level of abstraction independent of the CMD implementation, the CMDI[1] Component Specification Language (CCSL) for the profile is being used. The CCSL is an XML description of the elements and components with their respective reference to data category registries and the nesting of elements into components. The CCSL is used to generate the schema documents for evaluating the CMD instance.

For the implementation, two additional data sources need to be used together with the CMD instance and the matching CCSL, namely a mapping of data categories to core categories and some information on expected values for the individual score factors in order to normalize the score. These two resources will be described in the next to sections.

To allow for the greatest flexibility and the use within different implementations, it was decided to use XQuery as the implementation language as it is standardized and can be used with various processors and repository frameworks independent of their runtime environment. The XQuery can be retrieved via http://hdl.handle.net/11022/0000-0000-2067-8 .With an appropriate processor such as saxon or BaseX this query takes a CMD file as input and produces a table with the values described above.

5.1 Mapping of ISOcat data categories onto central categories

In CMDI the name of elements and their position in the document instance is fixed only for each profile. As various profiles can differ in naming elements, a semantic grounding is achieved by concept links referring to a data category in the definition of an component or element within the profile. The granularity of data categories can be very different here, but central elements reappear. For example a date reference can vary between a specification of a year or specific dates. For a quality assessment we therefore map data categories defined in ISOcat and in Dublin Core onto 14 core categories. This mapping is not formally defined, but meant to be an indication of possible mappings. It is assumed that the results indicate the quality of the metadata, not that this provides a fully generalized mapping from ISOcat to Dublin Core. For example the distinction of Project Name (ISOcat DC-2536) and Project Title (ISOcat DC-2537) is very specific and can for quality reasons be ignored in general. Similarly the resource name (ISOcat DC-2544) and resource title (ISOcat DC-2545) can be seen as specifications of the Dublin Core title category. The full list of mappings is provided in Table 1 as it is used in the current evaluation implementation. The mapping may be seen as useful for other mappings of detailed CMD data sets to general, not as expressive metadata schemas.

Core category	**Data category identifiers in ISOcat (DC-) or Dublin Core**
Project name	DC-2536 DC-2537 DC-5414
Resource name	DC-5428 DC-5127 DC-4160 DC-4114 DC-2544 DC-2545 DC-6119 Dublin Core: title
Date indication	DC-2509 DC-2510 DC-2538 DC-6176 Dublin Core: created, date, issued
Continent	DC-2531 DC-3791
Country	DC-2532 DC-3792 DC-2092
Language	DC-2482 DC-2484 DC-5361 DC-5358
Organisation	DC-2459 DC-2979 DC-6134 DublinCore: publisher
Genre	DC-2470 DC-3899
Modality	DC-2490
Subject	DC-2591 DC-6147 DC-5316 Dublin Core: subject
Description	DC-2520 DC-6124 Dublin Core: description
Resource class	DC-5424 DC-3806 Dublin Core: type
Format	DC-2571
Keywords	DC-5436

Table 1: Mapping of data categories to core categories; origin of data categories is either Dublin Core or ISOcat, the latter indicated by DC and their identifier

5.2 Normalization

To calculate the quality of metadata resources, the scale is normalized by the average values or gold standard of the individual factors. Based on 247278 CMDI files harvested via OAI-PMH in 2013 from various CLARIN centres[2]. The average was calculated with XQuery

[1] CMD Infrastructure

[2] For the data see http://hdl.handle.net/11022/0000-0000-206D-2 Note that this data was harvested and as such may be outdated.

using BaseX 7.8.1[3].

	Gold standard	Average
Profile part		
Number of data categories[4] mappable to core data categories	14	3.76
Number of unique data categories linked to by elements defined in profile	NA	31.62
Number of data categories linked to by elements defined in profile	NA	35.47
Number of components used in profile	NA	5.87
Number of unique Components used in profile	NA	5.05
Total number of elements defined in profile	NA	47.89
Number of unique elements defined in profile	NA	33.30
Instance part		
Length of first description (number of characters)	NA	203.00
Number of elements in instance	NA	27.83
Number of non empty elements	100%	15.90

Table 2: Average values calculated based on 247278 CMD records from various data providers, harvested via OAI-PMH; NA indicates values that do not have a perfect score by definition

6. Calculating the score

The score takes into account the score for the profile and the score for the instance to form a common score.
The profile score S_P is defined as the sum of the normalized indicators I_{norm}. The normalization function is the ratio of the indicator and the average as listed above. The instance score S_I is defined as the sum of the normalized (character) length of the first description element, the normalized number of elements in the instance, and the normalized number of filled elements. Empty elements are penalized with regards to the filled in fields. The filled in fields contribute to the score in normalized form. This leads to the following formula:

$$S_P = \frac{\#DatCatMap}{14} + \frac{\#\ Unique\ data\ cats}{31.62} + \frac{\#\ data\ cats}{\#\ Unique\ data\ cats} + \frac{\#\ Unique\ components}{5.05} + \frac{\#\ Elements}{47.89} + \frac{\#elements}{\#Unique\ elements}$$

$$S_I = \frac{\#\ instance\ elements}{27.83} - \frac{(\#\ instance\ elements - \#\ filled\ elements)}{\#\ filled\ elements} + \frac{\#filled\ elements}{15.9}$$

The total rating is then calculated as the product of S_P and S_I. Currently the length of the free text examples is not included in the calculation, but should be included in the future.

In this formula the filled in description has a comparatively large effect (i.e. descriptions of more than 200 characters will contribute a lot to the rating). The penalty of elements not filled in is comparatively high because it is assumed that this also indicates that the metadata was either not properly filled in or an inappropriate profile was selected.

7. Results from applying the score

Applying the scoring algorithm to CMD data harvested via OAI-PMH from various data providers shows that the score leads to expected results. Table 3 shows basic statistical information on the score.

Minimum score	-449.09
Maximum score	2142.41
Average score	17.47
Standard Deviation	34.55

Table 3: Statistical overview of the distribution of scores

The distribution of the scores is illustrated in Figure 1.

8. Usage/Application scenarios

We can distinguish between three types of applications of the score:

1.*Measuring the quality at editing time:* this provides a person creating the metadata with immediate feedback on the amount of additional data necessary to receive a better score. However, though in principle a score could be computed in a way that the scale is closed, we propose an open scale, especially as a full score would seem problematic as leaving out or filling in optional elements would lead to distortions of the score.

2. *Evaluation of a specific repository:* to provide a feedback to those running a repository and evaluating the overall quality of the metadata included in a given repository, also showing shortcomings, it is anticipated that the repository as such should also be evaluated by the average score of the resources provided and the number of profiles and resources.

3. *To give an overall assessment of the quality of the metadata provided by CMD*-providers and a more central assessment of profiles, the overall comparison of available CMD-documents also seems to be appropriate.

[3] http://basex.org/

[4] Concepts defined in some external concept registry, e.g. ISOcat, Dublin Core reference

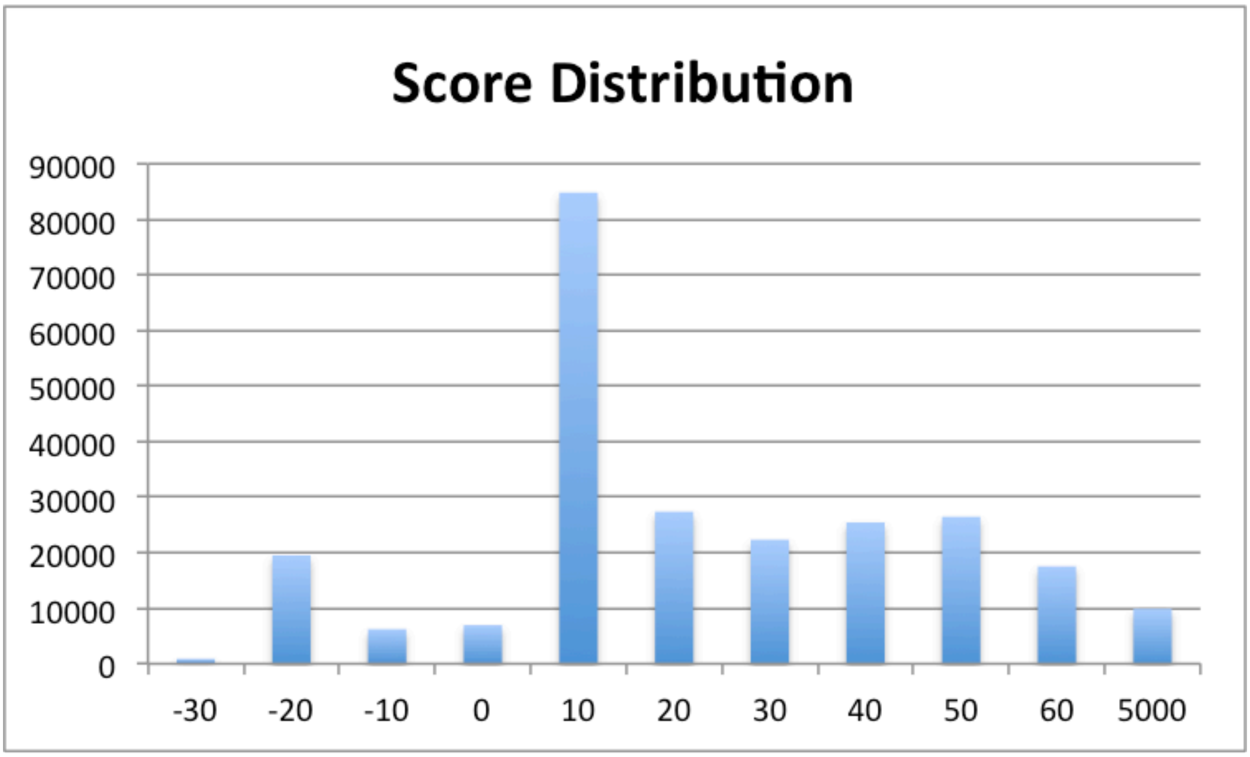

Figure 1: The distribution of the score, the x-axis shows the lower bound of the interval

Additionally to the actual evaluation at instance level, the pre-processing results of the profiles and components evaluation are valuable information for the metadata modeler:

1. Applying the schema level assessment on all defined profiles and components would yield a ranking that would help the modeler to choose the right one
2. On the fly quality assessment of the profiles being constructed, would give the modeler an indication of its quality (relative to defined criteria).

8.1 The formula as basis for further adaptation

The core set of data categories listed in Table 1 contains very general information elements, assumed to be relevant for language resources of any kind. Hence this score is suited for broad assessment of metadata quality of heterogeneous collections or aggregations. To configure a score targeted towards specific types of resources (for instance audio corpora), the core set may be extended with data categories considered important for the type in question.

8.2 The score in retrieval applications

Retrieval applications such as the VLO sometimes suffer from data sparseness of the underlying metadata when providing a structured search for language resources. Using the score, it is possible to set a threshold from which on a resource would be included in such an application. Additionally, it would be possible to rank the resource in search applications according to the quality of the metadata. These are possible options that are not implemented yet.

9. Future work

The score presented here is only a first approach to the problem of assessing the quality of highly variable metadata schemes and instances within the CMD framework. At first glance, the distribution and the quality score are as one would expect, especially looking at the outliers which show a low score for metadata records that are regarded as poor by the authors and high scoring instances that are seen as very well done. However, a formal evaluation and comparison to a manually imposed score is not yet available. A test will be conducted in the future with a small sample of CMD records originating from different providers and using different profiles and evaluated by human metadata experts to assign a score. These scores will then be compared to the automatically generated ones to validate and possibly further calibrate the assessment formula itself.

10. Acknowledgement

This paper presents resources and technology that are made available to the community via CLARIN, the Common Language Resources and Technology Infrastructure, which is a European Research Infrastructure.

11. References

Broeder, D; Kemps-Snijders, M.; Van Uytvanck, D.; Windhouwer, M.; Withers, P. & Wittenburg, P. (2010). A Data Category Registry- and Component-based Metadata Framework. In: Calzolari et al. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta: 43-47.

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., & Trippel, T. (2012). *CMDI: a Component Metadata Infrastructur*. Talk presented at LREC

2012: 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey: 1387-1390.

Bruce, T. and Hillmann, D. (2004). The Continuum of metadata quality: defining, expressing, exploiting. In: Metadata in practice. ALA Editions., pp 238-56.

Hughes, B. (2004). Metadata Quality Evaluation: Experience from the Open Lnaguage Archives Community. In Z. Chen et al. (eds). ICADL 2004, LNCS 3334, pp 320-329. Berlin and Heidleberg. Springer.

ISO/DIS 24622-1 (2013). Language resource management - Component Metadata infrastructure - Part 1: The Component Metadata Specification Model (CMDI-1)

Kapidakis, S. (2011). Measuring Metadata Quality for Europeana Local. Corfu, Greece, Ionian University, Department of Archive and Library Sciences, Laboratory on Digital Libraries and Electronic Publishing.

Ochoa, X and Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. International Journal of Digital Libraries. 10(2-3): 67-91.

Ohren, O. and Verling, K. (2010). Measuring Metadata Quality. Report, National Library of Norway, Draft

Thieberger, N. (2012), Counting Collections. In: Endangered Languages and Cultures, URL: http://www.paradiscec.org.aug/blog/2012/11/counting-collections/ , 2012-11-29T07:37