On assessing metadata completeness in digital cultural heritage repositories

Matteo Lorenzini
Digital Humanities Fondazione Bruno Kessler and University of Trento, Italy

Marco Rospocher
Department of Foreign Languages and Literatures, University of Verona, Italy

Sara Tonelli
Digital Humanities Fondazione Bruno Kessler, Italy

Abstract

Metadata allows access to a wide variety of cultural heritage resources made available through repositories, digital libraries, and catalogues. Usually taking the form of a structured set of descriptive elements, metadata assist in the identification, location, processing, tracking, preserving, sharing, and retrieval of information, while facilitating content and access management. However, low metadata quality, such as the lack of mandatory information, incorrect information, or inconsistency, is still an open issue in many repositories. In this article, we present our ongoing work aiming at automatizing the metadata quality analysis, and the preliminary results on metadata completeness for the Italian digital library 'Cultura Italia'.

Correspondence:
Matteo Lorenzini, Digital Humanities Fondazione Bruno Kessler and University of Trento, Italy.
E-mail:
m.lorenzini@fbk.eu

1. Introduction

In the last years, the number of digital repositories has remarkably increased in the cultural heritage domain. As a consequence, metadata has become the backbone through which users can navigate information and improve their knowledge of specific topics, reusing also data coming from external sources (Tani et al., 2013). However, despite the massive use of metadata and their key role, the process of quality control still lacks a clear definition and workflow (Bellini and Nesi, 2013).

In the literature, metadata quality has been presented as a way to measure how much a cultural heritage object supports a given purpose[1] (Bruce and Hillmann, 2004). In that sense, the curation framework developed by Bruce and Hillmann (2004) is considered as a benchmark in the pursuit of quality assessment of digital repositories. This framework defines seven qualitative dimensions to measure metadata quality: Completeness, Accuracy, Conformance to Expectations, Logical Consistency and Coherence, Accessibility, Timeliness, and Provenance. Although no existing approach has tried to operationally

Digital Scholarship in the Humanities, Vol. 36, Supplement 2, 2021.
doi:10.1093/llc/fqab036

measure them on real repositories, they would be all helpful to systematically identify metadata problems, applying them for instance to the Europeana Digital Library[2] or Europeana content providers.

Few attempts have been proposed to automatically compute quality metrics (Margaritopoulos et al., 2009, 2012; Ochoa and Duval, 2009; Király, 2015; Ostojic et al., 2017). However, the existing approaches do not consider three important factors as follows:

- *Metadata creation process*: often carried out manually by human operators following the guidelines provided by an aggregator or its corresponding data providers. So, each metadata element from the used metadata schema could be interpreted in a different way depending on the operator's point of view.
- *Aggregation process*: the aggregation process of digital resources takes place when metadata made available by one or more data providers are harvested and merged. As metadata aggregation has increased critically over the last years in the cultural heritage domain, the curation process should be re-contextualized with the goal to check and fix metadata in a large-scale repository, which is not always homogeneous.
- *Context*: often low metadata quality depends on the fact that metadata curators and creators are not able to retrieve the information about a specific resource or, in a specific context, that metadata elements are simply not useful to cover a domain of interest, so they are not used. For example, an archaeologist and a philologist have a different perception of an epigraph: for the first, an epigraph documents an archaeological finding, while for the second the epigraph is perceived as a text. Also, the metadata definition provided by NISO[3] (NISO, 2004) points to this issue describing metadata as non-static entities: they have multiple different interpretations and they should be considered in relation to their context.

We argue that existing works on metadata quality have also another limitation, in that they either focus on one dimension, or concern specific repositories or metadata schema/profiles. For example, to compute the completeness of a repository three main approaches have been presented in the literature:

- The presence or absence of metadata elements is computed with a binary assessment, assigning either 0 or 1 depending on the presence of specific metadata (Ochoa and Duval, 2009).
- For each metadata element a custom score is defined according to its importance with respect to the metadata profile (Király, 2015).
- The metadata completeness is evaluated at the field level, following two dimensions of analysis. The first dimension classifies a field as 'single' or 'multi-value' (e.g. dc: language). A multi-value field is considered complete if all the values indicated by the metadata profile specification are filled. The second dimension of analysis goes deeper into the hierarchical structure of the metadata schema, taking into consideration also the sub-elements of a given root element (e.g. the 'file' section from METS[4] metadata schema, which is composed of eight additional attributes). In this second dimension, a field is considered complete if all the sub-elements are filled. In both cases, completeness is computed as the weighted average of the filled elements with respect to the metadata schema (Margaritopoulos et al., 2009, 2012).

These approaches, even if they allow metadata curators to check the quality status of a single record or, more generally, of a dataset, do not try to embed in the computation also elements assessing whether low metadata quality is related to the lack of few metadata with high relevance or to the lack of many metadata elements with low relevance. We believe, however, that such metrics should enable users and curators to define in a flexible way what metadata they deem more relevant in the overall evaluation of completeness, to make this value comparable across different repositories, and to allow for a fine-grained analysis of the metadata elements. This is indeed the main contribution of this work, we propose a flexible way to compute completeness that takes into account mandatory and optional elements of a metadata scheme as well as the specific topic of a collection. We also provide an evaluation based on 'Cultura Italia' repository and show how, with the help of a graphical representation, our approach can support experts in assessing the quality of digital cultural heritage records.

2. Fine-Grained Completeness Assessment

Our long-term goal is to develop a framework that automatically checks metadata quality of a repository along different dimensions (Bruce and Hillmann, 2004). To develop such framework, two main activities are foreseen:

> Definition of metadata quality metrics, capturing the status of metadata both at object level (i.e., how good are the metadata of a single entry in the repository) and, aggregated, at repository level;
> Definition of algorithms to compute the aforementioned quality metrics, and (possibly) return suggestions on how to fix low-quality metadata.

In this article, we present the first results related to 'completeness'. In general terms, completeness is computed as the ratio of filled elements with respect to a metadata profile. In this computation, several variables should be taken into account, for instance, the elements that are mandatory and those that are optional, the context and the domain of a collection, as well as the preferences of curators when evaluating completeness.

Our approach to evaluate metadata completeness consists of the following key steps:

- Given a repository to be evaluated, metadata elements are divided into groups, representing their importance (e.g. compulsory/recommended/optional metadata);
- For each object o in the repository, a separate completeness score $c_G(o)$ is computed for each defined metadata group G as follows: the number of filled G metadata for that object is divided over the total number of G metadata elements. For instance, if an object o has 3 out of 10 of the compulsory metadata filled, $c_{compulsory}(o) = 0.3$. The resulting value is a real number between 0 and 1: the closer this value is to 1, the more complete the description of the object for that metadata group;
- We compute these completeness scores (one for each metadata group) for each element of the dataset. In order to provide also an overview of the completeness of a dataset, for each metadata group we draw a separate barplot, having on the x-axis 10 intervals representing completeness score ranges (e.g. 0–0.1, 0.1–0.2, . . ., 0.9–1.0) and on the y-axis the percentage of the objects in the whole dataset having that completeness score.

The datasets we used to test our methods consist only of non-aggregated and single fields. However, depending on the granularity of the quality check that curators want to apply, completeness can be computed with our approach both at the level of the root element and at the level of the aggregated elements (e.g. defining mandatory only some sub-elements, and optional the others), as well as for multi-value fields (e.g. considering a multi-value field complete if it consists of at least a value, or an expected number of values), coherently with Margaritopoulos et al. (2009) proposal. The extension of the assessment with multi-values and aggregated elements will be considered for future work.

Checking metadata quality according to completeness scores for the various metadata groups gives to metadata curators the possibility to have a complete view about the overall status of metadata quality. Curators can subsequently fix the objects with a low score, evaluating the different problems which contribute to the quality of the dataset.

3. Use Case: Completeness in Cultura Italia

Cultura Italia[5] is an online aggregator of Italian cultural heritage records and gives access to a metadata repository, which gathers and organizes the information harvested from Cultura Italia's providers. It consists of around 4,500,000 records including images, audio visual content, and textual resources.[6] The repository is accessible via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) handler or via the SPARQL endpoint (Di Giorgio, 2015). The metadata is ingested into Cultura Italia using the PICO[7] metadata schema (Buonazia et al., 2007; Buonazia and Masci, 2007), a qualified Dublin Core[8] (DC) which consists of ninety-four elements. The ninety-four PICO elements are divided into compulsory (eight elements), recommended (ten elements),

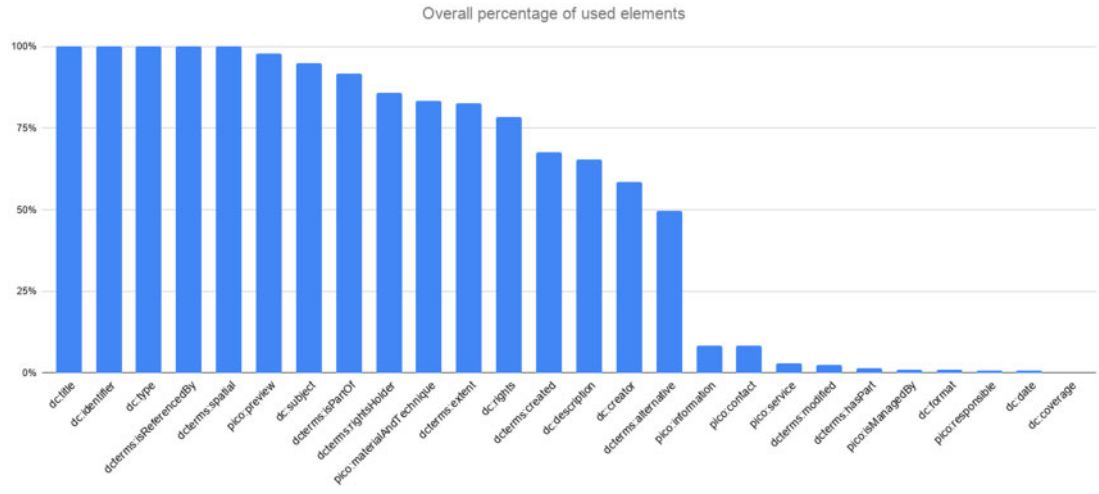

Fig. 1 Percentage of records in the MuseID-Italia dataset having a given metadata element

and optional (seventy-six elements). The records in the repository are indexed using the PICO SKOS thesaurus[9]: given the nature of the records aggregated in Cultura Italia, most of the terms used are related to the Art Object, Archaeology and Architecture, and Sound and Video domains.

Of all the datasets aggregated in Cultura Italia, in this work, we focus our completeness analysis on two specific datasets: MuseID-Italia (containing 76,828 records) and Regione Marche datasets (containing 90,602 records). Both datasets mainly deal with the Art Object domain.

To define the metadata groups needed to apply our approach, we start from the initial division of metadata elements provided by the PICO profiles. Then, among the optional PICO elements, we identify some metadata that is relevant (and should be filled) for objects of datasets in a specific domain (e.g. the Art Object domain). The selection of these metadata elements was conducted consulting the categorization provided by the PICO SKOS thesaurus and the PICO profile specification, and later validated by the technical unit of Cultura Italia. Therefore, for the considered datasets, we identify the following four metadata groups:

- Compulsory elements (eight elements): the compulsory PICO elements, that is, dc: title, dc: identifier, dc: subject, dc: type, pico: preview, dc: isReferencedBy, dcterms: license, pico: licenseMetadata.
- Recommended (ten elements): the recommended PICO elements, such as dc: description, pico: author, and dcterms: spatial.
- Domain-specific: the PICO optional elements that are relevant for a specific domain, and therefore should be preferably filled for objects of datasets in that domain. For the Art Object domain, we identified the following eleven elements: pico: commissioner, pico: matherialAndTechnique, dcterms: created, dcterms: isPartOf, dcterms: alternative, dcterms: modified, dc: contributor, dc: coverage, dcterms: bibliographicCitation, pico: printer, dcterms: replaces.
- Optional (seventy-six elements): the remaining PICO optional elements, such as dcterm: bibliographicCitation, pico: commissioner, pico: performer, etc.

3.1 Results

Before analysing in detail, the completeness of the considered datasets according to the proposed metadata group, we investigate the frequency of usage of PICO metadata elements in the records of the collections. Figures 1 and 2 graphically represent the percentage of records in the two datasets (MuseID-Italia e Regione Marche) having the given metadata elements. We can note that many (but not all) of the compulsory metadata are filled for all records in the datasets. For the metadata elements in the other

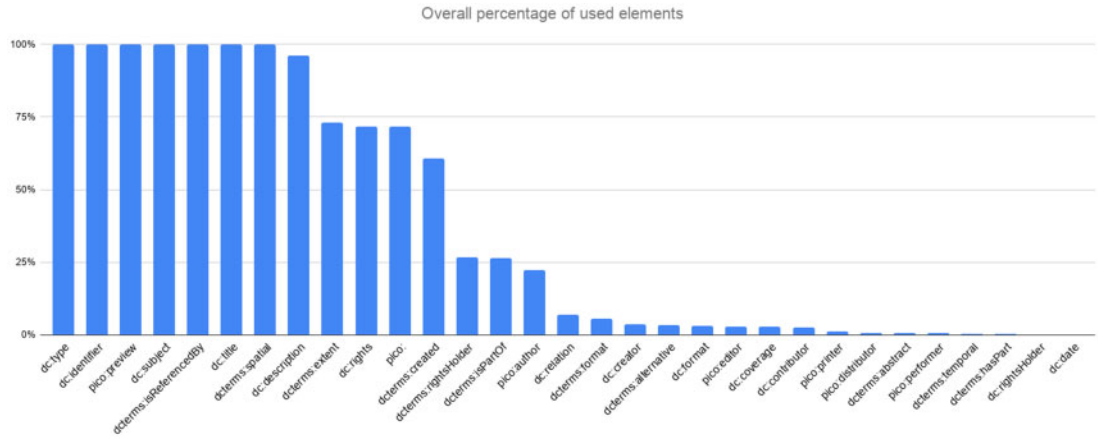

Fig. 2 Percentage of records in the Regione Marche dataset having a given metadata element

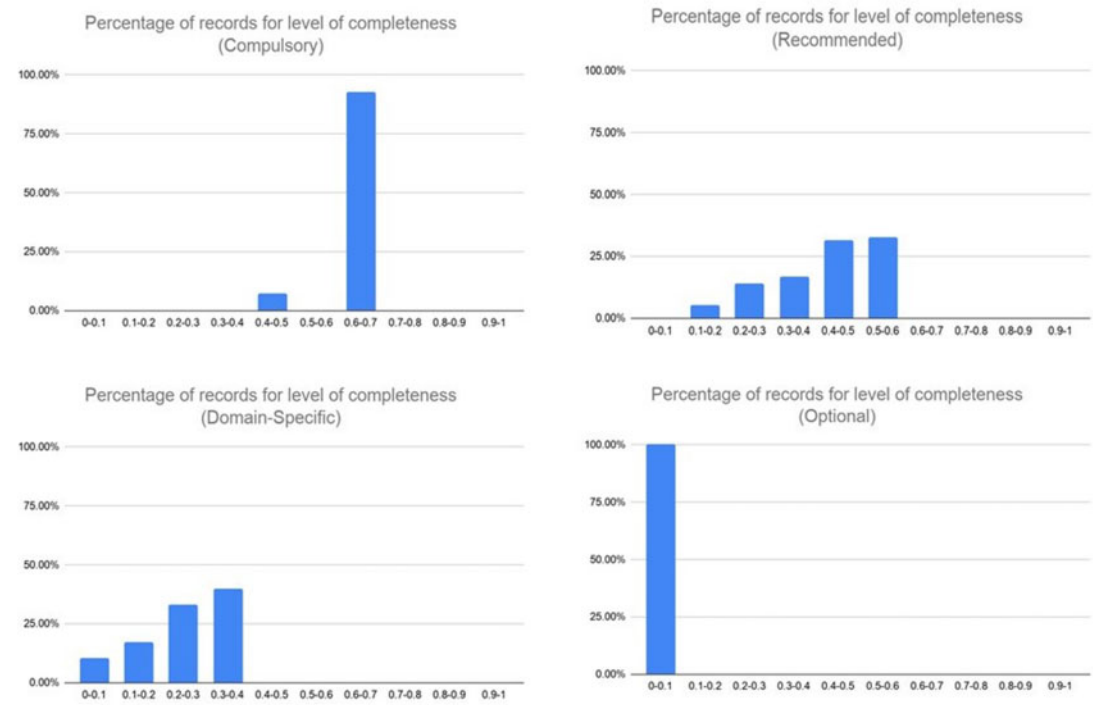

Fig. 3 Completeness plots for MuseID-Italia dataset

groups, the percentage of records having those elements filled is substantially lower. For example, in the Regione Marche dataset, only 21% of the resources are filled by using the pico: author element. This means that, in most cases (79%), the end-users will not be able to filter the resources by 'Author' or perform a free search by typing the name of the artist.

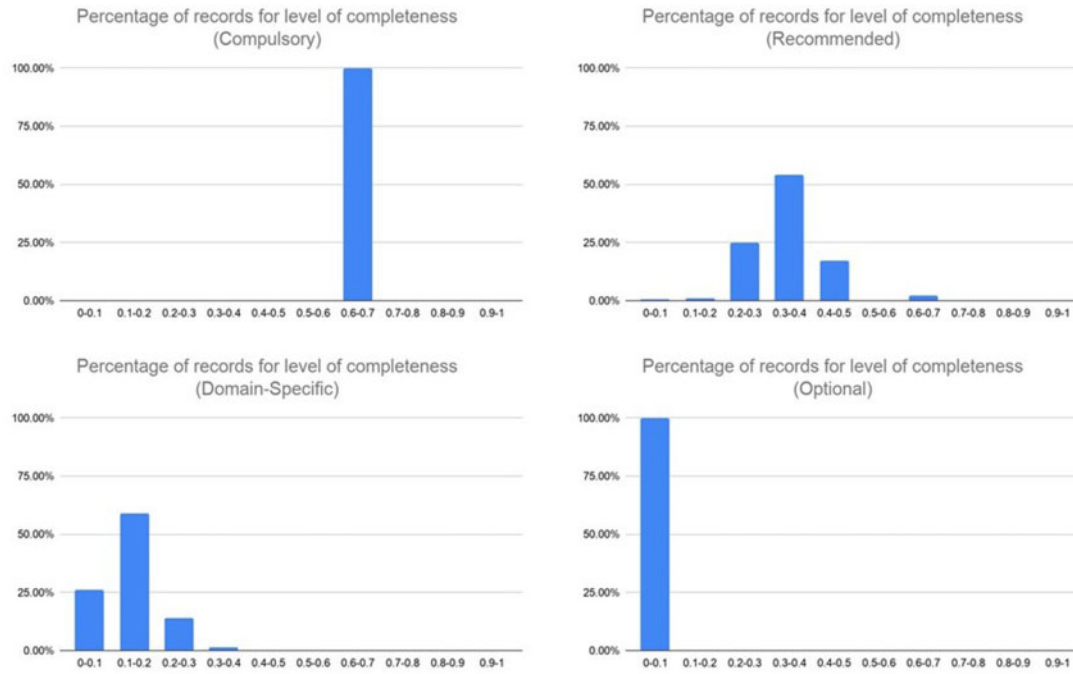

Fig. 4 Completeness plots for Regione Marche dataset

Then, given the four metadata groups proposed in Section 3, we compute for each record in the datasets four completeness scores (one for each metadata group), and separately for each metadata group, we analyse the distribution of the resulting completeness scores over the datasets, by plotting the aggregated bar plots as described in Section 2. These plots are reported in Figs 3 and 4.

The plots show that the results obtained for completeness on the two datasets are generally low. For example, in MuseID-Italia dataset (Fig. 3, top, left), most of the records (∼92%) obtain a completeness score for the compulsory group in the range between 0.6 and 0.7, while for Regione Marche (Fig. 4, top, left) all records achieve for the same metadata group a score between 0.5 and 0.6. The same can be observed for the domain-specific schema (between 0.1 and 0.4). All records in the datasets rarely use elements of the optional metadata group, while the usage of recommended and domain-specific metadata elements vary.

With our approach, we show that, thanks to the computation based on four different metadata groups, the system returns to the metadata curator and aggregator a precise and comprehensive picture of the overall dataset completeness, also allowing for comparisons across datasets.

4. Conclusion and Further Steps

In this work, we introduced a novel, fine-grained way to compute metadata completeness, performed organizing metadata elements in different groups accounting for their relevance for the considered dataset. This way, metadata curators can efficiently and effectively detect issues in digital repositories, optimizing the curation process. We concretely applied our approach in a use case comprising MuseiD-Italia and Regione Marche datasets.

In the future, we plan to extend this work by considering other dimensions contributing to metadata quality besides completeness. For example, we are experimenting with automatic quality assessment of textual descriptions associated with cultural heritage

records, as well as with approaches that support curators to check semantic consistency of the different fields. Our final goal is to offer a complete set of metrics to evaluate metadata quality paired with a suite of tools that compute them automatically both on single records and on aggregated data.

References

Bellini, E. and Nesi, P. (2013). Metadata quality assessment tool for open access cultural heritage institutional repositories. In *International Conference on Information Technologies for Performing Arts, Media Access, and Entertainment*. Springer, Berlin, Heidelberg, pp. 90–103.

Bruce, T. R. and Hillmann, D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. *ALA Editions*.

Buonazia, I. and Masci, M. E. (2007). Il pico application profile. un dublin core application profile per il portale della cultura italiana.

Buonazia, I., Masci, M. E., and Merlitti, D. (2007). The Project of the Italian Culture Portal and its Development. A Case Study: Designing a Dublin Core Application Profile for Interoperability and Open Distribution of Cultural Contents. *ELPUB*, pp. 393–404.

Di Giorgio, S., Felicetti, A., Martini, P., and Masci, E. (2015). Dati. CulturaItalia: A Use Case of Publishing Linked Open Data Based on CIDOC-CRM. *EMF-CRM@ TPDL*, pp. 44–54.

Király, P. (2015). *A Metadata Quality Assurance Framework*. Göttingen: Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen.

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., and Manitsaris, A. (2009). A fine-grained metric system for the completeness of metadata. In *Research Conference on Metadata and Semantic Research*. Springer, Berlin, Heidelberg, pp. 83–94.

Margaritopoulos, M., Margaritopoulos, T., Mavridis, I., and Manitsaris, A. (2012). Quantifying and measuring metadata completeness. *Journal of the American Society for Information Science and Technology*, **63**(4): 724–37.

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

Ostojic, D., Sugimoto, G., and Durčo, M. (2017). The curation module and statistical analysis on VLO metadata quality.

Pennock, M. (2007). Digital curation: a life-cycle approach to managing and preserving usable digital information. *Library & Archives*, **1**: 34–45.

Tani, A., Candela, L., and Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, **49**(6): 1194–205.

Notes

1 In the cultural heritage domain is it possible to find three main purposes: Preservation, Registering, and Discovery (Pennock, 2007).
2 https://www.europeana.eu/portal/en.
3 US National Information Standards Organization.
4 http://www.loc.gov/standards/mets/.
5 http://www.culturaitalia.it/.
6 CulturaItalia is also one of the national aggregators of the European digital library Europeana.
7 http://purl.org/pico/1.1/picotype.xsd.
8 https://www.dublincore.org/.
9 http://www.culturaitalia.it/pico/thesaurus/4.3/thesaurus_4.3.0.skos.xml.