Automatic metadata curation of the cultural heritage resources

Matteo Lorenzini[1−2]

[1] Universitá degli Studi di Trento matteo.lorenzini@unitn.it
[2] Fondazione Bruno Kessler
m.lorenzini@fbk.eu

Abstract. This paper presents my thesis proposal on automatic metadata curation of cultural heritage resources. Metadata curation represents one of the aspects of the "metadata management life-cycle" together with ingestion, maintenance and dissemination. The objective of metadata curatorship in general is to ensure the user can access objects of interest to him/her from a repository, digital library, catalogue, etc. using well-assigned metadata values aligned with an appropriately chosen schema. Ideally, in a repository or digital archive, all the objects should be described using the same accuracy and data structure. In this way the user can retrive all the informations, objects of interest and the related items as result of a single search.
However, this is very rare. Objects are described using different levels of details and different data model design might be used by the cataloguers. In such cases, after the metadata quality control process, the curator must correct or normalize all the objects with errors or incosistencies in the metadada schema.
New approaches based on the use of the semantic enrichment of the digital resources can help the resources aggregator during the metadata curation process having an impact also in the improvement of resources discovery.

Keywords: CIDOC-CRM · Metadata Curation · Dublin Core · Digital Library

1 Introduction

The traditional concept of library has undergone a profound change: from a collection of physical information resources (mostly books) to a collection of digital resources. In addition, the notion of digital resource includes not only texts in digital form, but also, in general, any kind of multimedia resources[3]. Hence, making accessible cultural heritage resources requires metadata schemas rich in semantics and a structure able to cover the material heterogeneity and the variety of memory institutions (libraries, archives, museums). However, often

[3] Such collections may be composed of text, written on different materials, paintings, photographs, 3D objects, sound recordings, maps or even digital object.

we are facing with problems related to the low quality of the metadata used for the description of the digital resources: wrong definition, inconsistence of the resources or resources described according only to the minimal mandatory metadata entities[4]. There may be many reasons for that, all completely valid, e.g.; in many cases these institutions have few human resources to work on improved metadata, they are often not themselves the sources of the metadata.

The goal of the present work is to improve metadata quality integrating semantic web principles into the metadata curation process.

2 State of the art

Digital curation, broadly interpreted, is about maintaining and adding value to a trusted body of digital information for both current and future use: in other words, it is the active management and appraisal of digital information over its entire lifecycle.

The necessity of a lifecycle approach, to ensure the continuity of digital material, is discussed by Pennock [5]. A lifecycle approach ensures that all the required stages are identified and planned, and necessary actions implemented, in the correct sequence. This can ensure the maintenance of authenticity, reliability, integrity and usability of digital material, which in turn ensures maximisation of the investment in their creation.

The curation framework developed by Bruce and Hillmann [3] is considered as a benchmark in the pursuit of quality assessment. This framework defines seven parameters to measure the quality of the metadata: *Completeness, Accuracy, Conformance to Expectations, Logical Consistency and Coherence, Accessibility, Timeliness, Provenance.* In the digital libraries domain, these parameters are fundamental for the evaluation of metadata quality before the curation processes. The evaluation helps various curators to systematically identify metadata problems. This could be straightforwardly applied to Europeana Digital Library[5] [2] or the ARIADNE poroject[6] [1].

In the context of linguistic resources, the CLARIN[7] consortium supports metadata curation developing a metadata curation module to facilitate the metadata ingestion and curation process of the Virtual Language Observatory (VLO). In the context of Semantic Web the metrics were often not explicity defined and did not consist of precise statistical metrics. Moreover, only few approaches were actually accompained by an implemented tool and none of them covered all the data quality dimensions [7].

From the literature analysis it can be inferred that the existing approaches are either too abstract or extremely focused on one dimension e.g.; completeness [1,2].

[4] The minimal metadata set able to guarantee the harvesting process from the content provider to the resource aggregator.

[5] `https://www.europeana.eu/portal/it`

[6] `http://www.ariadne-infrastructure.eu`

[7] `https://www.clarin.eu`

3 Problem statement

Wrong and incomplete mappings affect the discoverability and acessibility of the resources for the users: metadata curation plays an essential role to the improvement of the metadata quality. In a standard cycle, the resurces are checked by the curators which, applying the metrics described in the previous paragraph, analyze the metadata quality. Depending on the types of issues curators intervene as follows :

- Send back to the content provider the datasets or records in order to fix the incosistence/lacks/errors.
- Fix by hand the incosistence/lacks/errors.
- Normalize the resources using a controlled vocabulary.

This approach, even if it can be considered as a standard procedure, leads to problems related to effective impact of the curation process: In most of the cases curation is done by hand by the curators, quality analysis involves just the mandatories metadata elements[8] and metadata quality analysis does not consider all the parameters suggested by Bruce and Hillmann.

For a better contextualization of the problems we can briefly refers to Cultura Italia[9], the italian digital library. Resources are integrated in Cultura Italia in the form of metadata using PICO[10] profile. It is based on the international standard language Dublin Core[11] that can describe, in a single scheme, every type of cultural resource, both physical and digital. Dublin Core consist of 15 main elements. PICO profile adds to these others 37 elements conceived for the application of the Dublin Core in Cultura Italia. We can identify two types of errors:

- Low metadata completeness: Objects are described using the 6 mandatory metadata elements from PICO for metadata harvesting. Elements like dc:description or dc:author which are considered as optional are under represented E.g in the dataset from "Regione Piemonte" which consists of 71.710 records the element dc:description is never used.
- Low accuracy: Metadata are filled using non exaustive informations like dc:title "photo".

In the light of the problems underlined above our research question will be: "Can we improve metadata quality with automatic curation techniques and semantic web technologies?". This topic will be treated in the domain of the cultural heritage.

[8] Crucial information like "description" lacks quality evaluation
[9] `http://www.culturaitalia.it`
[10] `http://www.culturaitalia.it/opencms/documentazione_tecnica_it.jsp?language=it&tematica=static`
[11] `http://dublincore.org`

4 Proposed solution

The solution we are going to propose is a framework which, using the metrics from Bruce and Hillmann, aims to check automatically[12] the issues derived from wrong cataloging processes in order to optimize the metadata curation workflow. Moreover it aims to improve the metadata quality by suggesting the missing metadata elements or errors to the curators. Considering the scenario from Cultura Italia described in the previous paragraph the curation process will be characterized by the following main tasks:

- Automatization of the quality metrics: This PhD project will be mainly focused on the quantitative definition of the *Completeness, Accuracy and Logical consistency and coherence* metrics. For each of these metrics we will be define a customized algorithm[3] able to measure the quality of the metadata compared to the metadata profile from the dataset object. *Completeness* parameter, statistical approach: Each metadata standard, for example Dublin Core or PICO, defines a number of possible fields (15 for Simple Dublin Core, 37 for PICO). Completeness obtained by computing will be to count the number of fields in each metadata instance that contain a no-null value. In the case of multi-valued fields, the field is considered complete if at least one instance exists. *Accuracy*, natural language processing approach: a vector space model will be defined. Here will be computed the distance between metadata instances and the domain of the resource or dataset. A shorter distance correspond to a higher accuracy of the metadata instance. *Logical consistency and coherence*, semantic web technologies approach: The consistency will be calculated compared to the degree to which the resource description matches with the metadata standard schema and definition. The coherence will be measured at the instance level. Will be computed the degree to which all the fields describe the same object in a similar way analyzing the correlation between text and metadata elements. Here, a challenging issue, will be the evaluation of the logical consistency and coherence of the textual entities with respect to the domain of the digital object.
- Suggestion of potentially correct metadata values : Usually, because of the licences regarding the re-use of metadata, the aggregator can not modify the metadata given by the data provider. So, the errors and the potentially correct metadata values, will be reported by a log to the metadata creator. Then, the metadata creator can decide whether to accept the suggestions.
- Evaluation methodology: validation will concern two aspects:
  - Metadata schema: compliance of the new elements with the standards structure of the metadata profile.
  - Consistence of the new elements with respect to the context of the digital object: evaluation made by the logical consistency and coherence parameter.

[12] Manual curation processes are of course limited in their coverage: the amount of objects that can be curated in datasets like the ones of Cultura Italia or other aggregators is too high for the human resources available

In order to achieve the best results two different methods will be tested:

- Definition of a "gold standard" dataset: will be used in order to train the framework about how the correct metadata schema should be in terms of structure and content.
- Insert (intentionally) some errors in the well defined resources in order to check if the proposed system can recognize the errors.

5 Conclusion

Considering the amount of digital archieves, problems related to metadata curation becomes evident. Reasons may be different: There is no curation task force, the metadata curation activity is delegated to the content providers or the metadata curation activity is made by hand. The development of an automatic process will enable the curators to not only obtain snapshots of the quality of a repository, but also to constantly monitor its evolution and how different events affect it without the need to run costly human effort. This could lead to the creation of innovative applications based on metadata quality that would improve the final user experience.

References

1. Bruce,T.R., Hillmann, D.: The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: Metadata in Practice, 2004. ALA Editions. (2004).
2. Kiraly, P.: A Metadata Quality Assurance Framework.(2015)
3. Ochoa, X., Duval, E.: Automatic evaluation of metadata quality in digital repositories. In:Internatinoal Journal on Digital Libraries, 10, pp. 67–91. (2009)
4. Ostojic, D., Sugimoto, G., Durco, M.: The Curation Module and Statistical Analysis on VLO Metadata Quality. In: CLARIN annual conference 2016, pp.90–100. CLARIN consortium. (2016).
5. Pennok, M.: Digital curation:A life-cycle approach to managing and preserving usable digital information.In: Library and Archives Journal, vol.1 (2008)
6. Sompel, H. V. D., Nelson, M., Lagoze, C. & Warner, S.: Resource Harvesting within the OAI-PMH Framework.In: D-Lib Magazine, 10. (2004).
7. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S.: Quality assessment for Linked Data: A Survey.In Semantic Web, 7, pp. 63–93 (2016)
8. Radulovic, F., Mihindukulasooriya, N., Garca-Castro, R. & Gmez-Prez, A. A comprehensive quality model for Linked Data.. Semantic Web, 9, 3-24. (2018).
9. Melo,A., Vlker,J.,Paulheim,H.:Type Prediction in Noisy RDF Knowledge Bases Using Hierarchical Multilabel Classification with Graph and Latent Features. International Journal on Artificial Intelligence Tools 26(2): 1-32(2017)
10. Paulheim,H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 8(3): 489-508 (2017)