Data Quality Assessment in Europeana: Metrics for Multilinguality

Valentine Charles[1], Juliane Stiller[2], Péter Király[3], Werner Bailer[4], Nuno Freire[5]

[1] Europeana Foundation, The Hague, NL, valentine.charles@europeana.eu
[2] Humboldt-Universität zu Berlin, DE, juliane.stiller@ibi.hu-berlin.de
[3] Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, DE, peter.kiraly@gwdg.de
[4] Joanneum Research, Graz, AT, werner.bailer@joanneum.at
[5] INESC-ID, Lisbon, PT, nuno.freire@tecnico.ulisboa.pt

Abstract. Europeana.eu aggregates metadata describing more than 55 million cultural heritage objects from libraries, museums, archives and audiovisual archives across Europe. Quality (and particularly multilingual quality) of metadata is crucial for enabling search and data re-use in digital libraries. Capturing multilingual aspects of the data requires us to take into account the full lifecycle of data aggregation including data enhancement processes such as data enrichment. Multilinguality cannot be captured as one measure, but needs to be considered as an intersection of several measures, bringing challenges when interpreting and visualising the results. The paper presents an approach for capturing multilinguality as part of data quality dimensions, namely completeness, consistency and accessibility. We describe the measures defined and implemented, and provide initial interpretations of the results.

Keywords: metadata quality, multilinguality, digital cultural heritage, Europeana, data quality dimensions

1 Introduction

Europeana.eu[6] is Europe's digital platform for cultural heritage. It aggregates metadata describing more than 55 million cultural heritage objects from a wide variety of institutions (libraries, museums, archives and audiovisual archives) across Europe.

The need for high-quality metadata is particularly motivated by its impact on search, the overall Europeana user experience, and data re-use in other contexts such as the creative industries, education and research. One of the key goals of Europeana is to enable users to find the cultural heritage objects that are relevant to their information needs irrespective of their national or institutional origin and the material's metadata language.

[6] http://www.europeana.eu/portal/en

As highlighted in the White Paper on Best Practices for Multilingual Access to Digital Libraries [3], most digital cultural heritage objects do not have a specific language and can only be searched through their metadata, which is text in a particular language. The presence of multilingual metadata description is therefore essential to improving the retrieval of these objects across language spaces. Understanding Europeana's cross-lingual reach is essential to enhancing the quality of metadata in various languages.

In the Data Quality Committee (DQC), a body formed at Europeana's initiative, we specify functional requirements that define the purpose of the metadata and guide data-quality evaluation – following a core principle for metadata assessment (e.g. [7]). Europeana's international scope means that multilinguality is an inherent aspect of these requirements. To assess multilinguality in metadata, we look at commonly-agreed data quality dimensions: completeness, consistency and accessibility, and define and implement quality measures. We capture data quality along the full data-aggregation lifecycle, taking also into account the impact of data enhancement processes such as semantic enrichment. The model the data is represented in, namely the Europeana Data Model (EDM)[7], is also a key element of our work.

In the next section, we present data quality frameworks, dimensions and criteria that are commonly referred to in the context of data quality measurement. Section 3 describes how multilingual data is presented in Europeana's data model and the functional requirements that are the basis of our assessment. The data quality dimensions we use will also be presented. In section 4, we describe the implementation of the different measures. Section 5 describes first results and measures that were taken to improve data along the different quality dimensions. We conclude this paper with an outline of future work.

2 State of the art

Addressing data quality requires the identification of the data features that need to be improved and this is closely linked to the purpose the metadata is serving. Libraries have always highlighted that bibliographic metadata enables users to find material, to identify an item and to select and obtain an entity [1]. Based on this, Park [11] expands functional requirements of bibliographic data to discovery, use, provenance, currency, authentication and administration, and related quality dimensions. The approach to metadata assessment for cultural heritage repositories presented in [4] also starts from use for a specific purpose. While the work mentions the issue of multilinguality, it does not propose specific metrics to measure it.

Bruce and Hillmann [5] define the following measures for quality: completeness, accuracy, provenance, conformance, logical consistency and coherence, timeliness and accessibility, grouped in three tiers. Shreeves et al. [12] focus on three quality dimensions (completeness, consistency and ambiguity) for their explo-

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

rative study of four collections, which included metadata from several institutions. Not surprisingly, they found greater variances for collections aggregated from multiple sources as well as inconsistencies in the encoding scheme – a result also experienced by Europeana.

Designing an information quality framework, Stvilia et al. [14] have proposed a taxonomy of 22 measures, grouped into three categories: intrinsic, relational/contextual and reputational information quality. In this taxonomy, completeness/precision exists both in the intrinsic (e.g. the number of elements, the non-empty elements) and in the relational/contextual categories (completeness wrt. a recommended set of elements). Although multilinguality is not explicitly mentioned in this paper, it fits in this model as part of contextual completeness.

Zaveri et al. [16] propose dimensions for quality assessment for linked data, grouped into the accessibility, intrinsic, contextual and representational categories. Completeness is listed as an intrinsic criterion, while multilinguality is covered by versatility, which is considered a representational criterion. The ISO/IEC 25012 standard [2] defines a data quality model with 15 characteristics, discriminating between inherent and system-dependent ones, but putting many of the criteria in the overlap between the two classes. Completeness is defined as an inherent criterion, while accessibility and compliance are in the overlapping area. Multilinguality could be seen as being compliant to providing a certain number of elements in a certain number of languages, and as enabling access to users who are able to search and understand results in certain languages.

Mader et al. [9] include issues related to multilinguality, such as incompleteness of language coverage and missing language tags, in their analysis of SKOS vocabularies. Multilinguality is covered by Dröge [6] as part of the eight classes of criteria defined for assessing the quality of vocabularies. However, no metric is developed in detail.

It becomes apparent that while multilinguality is considered in some works, it is usually not treated as a separate quality dimension, but rather as part of other criteria or dimensions existing at quite different levels in different quality models. To the best of the authors' knowledge, the only resource which measured multilingual features in metadata is [15], cited by [10]; here language attributions across a collection were counted. A first iteration of a multilingual saturation score that counted language tags across metadata fields in the Europeana collections as well as the existence of links to multilingual vocabulary was introduced by Stiller and Király [13]. The further development of this metric as well as the integration of measures for all aspects of multilingual data is described in this paper.

3 Approach

Firstly, to determine the multilingual degree of metadata across several quality dimensions we have to understand the different ways multilingual information is expressed in Europeana's data model. Secondly, multilingual information is at the core of the functional requirements that constitute the services Europeana

is providing for an international audience. Thirdly, these requirements and the structure of multilingual data inform the criteria and metrics that enable us to measure multilinguality across several metadata quality dimensions.

3.1 Multilingual information in Europeana's metadata

Multilinguality in Europeana's metadata has two perspectives: concerning the language of the object itself, and the language of the metadata that describes this object. First, the described cultural object, insofar as it is textual, audiovisual or in any other way a linguistic artefact, has a language. The data providers are urged to indicate the language of the object in the *dc:language* field in the Europeana Data Model (EDM) in this way: `<dc:language>de</dc:language>`. This information is then used to populate a language facet allowing users to filter result-sets by language of objects. The language information is essential for users who want to use objects in their preferred language. Second, the language of metadata is essential for retrieving items and determining their relevance. Metadata descriptions are textual and therefore have a language. Each value in the metadata fields can be provided with a language tag (or language attribute). Ideally, the language is known and indicated by this tag for every literal in each field. If several language tags in different languages exist, the multilingual value can be considered to be higher. For instance, consider as an example this data provided by an institution:

```
<#example> a ore:Proxy ; # data from provider
  dc:subject \Ballet",   # literal
  dc:subject \Opera"@en  # literal with language tag
```

The first *dc:subject* statement is without language information, whereas the second tells us that the literal is in English. Europeana assesses data in particular fields to enrich it automatically with controlled and multilingual vocabularies as defined in the Europeana Semantic Enrichment Framework.[8] As shown in the following example, the deferencing of the link (i.e., retrieving all the multilingual data attached to concepts defined in a linked data service) allows Europeana to add the language variants for this particular keyword to its search index.

```
<#example> a ore:Proxy ; edm:europeanaProxy true ;
  # enrichment by Euroepana with multilingual vocabulary
  dc:subject <http://data.europeana.eu/concept/base/264>

<http://data.europeana.eu/concept/base/264> a skos:Concept .
  # language variants are added to index
  skos:prefLabel "Ballett"@no, "Ballett"@de, "Balé"@pt,
                 "Baletas"@lt, "Balet"@hr, "Balets"@lv
```

The record now has more multilingual information than at the time of ingestion into Europeana. The different language versions from multilingual vocabularies are likely to be translation variants.

[8] https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBj1pEx0Y/edit

3.2 Functional requirements for multilingual services

In order to better understand the impact of multilinguality on the overall quality of the data, we have identified the multilingual aspects of several functional requirements relevant to Europeana. Note that these requirements are also crucial to better communicating the results of our measures to our data providers, and hence to motivating improvement of the data.

Cross-language recall: When a user wants to search for a document or documents relevant to her needs, and to do so irrespective of the language of the documents and/or their metadata, multilingual metadata is a necessity. It is therefore required that all searchable freetext metadata elements[9] be consistently tagged with their language.[10]

Language based facets: In a language-based facet requirement a clear distinction needs to be made regarding whether the language refers to the language of the metadata or the digital representation of the work. The method of measuring the multilingual saturation is different in both cases. While the language of the metadata will be measured against the presence of language tags, the language of the content will be captured in the *dc:language* field in EDM. The existence of multilingual metadata records in Europeana demands that language-tagging occur at the level of the metadata element rather than the record as a whole.

Entity-based facets (people, places, concepts, periods): while users will be primarily searching for entities in their own language (the city of "Paris"), they will need to access results in different languages (Parijs, Parigi, ...). To support this scenario we will rely on the translated labels provided by data providers as part of their data or on multilingual vocabularies to which the data can be linked. In both cases measuring multilinguality is crucial to identify language gaps.

It is also important to determine the impact of workflows contributing to the increase of multilingual labels (such as machine learning and natural language processing techniques for language detection, automatic tagging, or semantic enrichment) in the data in order to define the measures but also to interpret the results.

3.3 Multilinguality as a facet of quality dimensions

For measuring multilinguality, we look at three quality dimensions: completeness, consistency and accessibility. Each of these dimensions assesses multilinguality from a different perspective. The results reflect how well the functional requirements work.

[9] `https://docs.google.com/spreadsheets/d/1f6vbeo4mt0stl0yAbVCyp_-5bzBBCq4Dy6p7xhMcjSI/`

[10] for instance, in accordance with the ISO 639-1 or 632-2 (`https://www.loc.gov/standards/iso639-2/php/code_list.php`) list, the IANA language tag registry (`http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry`), or the Languages Name Authority List (NAL) (`https://open-data.europa.eu/en/data/dataset/language`)

Completeness. Completeness is a basic quality measure, expressing the number (fraction) of fields present in a dataset, and identifying non-empty values in a record or (sub-)collection. For a fixed set of fields completeness is thus straightforward to measure, and can be expressed as the absolute number or fraction of the fields present and not empty. However, the measure becomes non-trivial when data is represented using a data model with optional fields (that may e.g., only be applicable for certain types of objects), or with certain fields for which the cardinality is unlimited (e.g., allowing zero to many subjects or keywords). These characteristics apply to EDM. In such cases the measure becomes unbounded, and a few fields with high cardinality may outweigh or swamp other fields.

In the context of measuring multilingual completeness, the metric is two-fold. First, the concept of completeness can be applied to measuring the presence of fields with language tags or, second, the presence of the *dc:language* field. Any measure of multilinguality must be seen in relation to the results of measuring completeness. Only fields both present and non-empty can be said to have or lack language tags and translations. A record which is 80% complete can still reach 100% multilingual completeness if all present and non-empty fields have a language tag. The completeness affects all three functional requirements as they all work better the more multilingual information exists in a given set of fields. **Consistency.** Consistency describes the formal conformity to set standards and field rules as well as the logical coherence of the metadata. With regard to multilinguality, the dimension assesses the variety of language values in the *dc:language* field. The goal is to define a standard for language notations and normalize the field in this regard (see section 5.2). The consistency measure is mainly relevant for the language based facet. The better normalization works, the better the user can filter objects by their language.
Accessibility. Accessibility describes the degree to which multilingual information is present in the data, and allows us to understand how easy or hard it is for users with different language backgrounds to access information. So far, Europeana has little knowledge about the distribution of linguistic information in its metadata – especially within single records. To quantify the multilingual degree of data and measure cross-lingual accessibility, the language tag is crucial. Resulting metrics can be scaled to the field, record and collection levels. In practical terms, the accessibility measure serve to gauge cross-language recall and entity-based facet performance.

To summarize: with regard to multilinguality, we identified the following dimensions and quality criteria (Table 1 3.3).

4 Operationalizing the metrics for multilinguality

Implementation requires a good understanding of EDM in order to not exclude relevant data. The source data go through several levels of data aggregation before being aggregated into Europeana. Data processes can take place at each of theses levels and affect the final data output. EDM allows us to distinguish

Table 1. Dimensions, criteria and measures for assessing multilinguality in metadata.

Dimension	Criteria	Measure
Completeness	Presence or absence of values in fields relating to the language of the object or the metadata	Share of multilingual fields to overall fields
		Presence or absence of *dc:language* field
Consistency	Variance in language notation	Distinct language notations
Accessibility	Multilingual Saturation	Numbers of distinct languages
		Number of language tagged literals
		Tagged literals per language

between values provided by the data provider(s) from the information added by Europeana (for instance by semantic enrichment), doing so by leveraging the proxy mechanism from the Object Re-use and Exchange (ORE) model. It enables the representation of resources in the context of aggregations offering different views on the same resource [8]. Any implementation of quality measures, and in particular of multilingual saturation, needs to take into account this distinction depending on the results it is intended to yield. For instance, the contribution to the score might be higher if we only consider the Europeana proxy where a value was enriched with a multilingual vocabulary (e.g. DBpedia).

4.1 Implementation of the measures

The different metrics for the assessment of multilinguality in metadata are implemented in the Data Quality Assurance framework.[11] The measurement has four phases. (1) data collection and preparation: the EDM records were collected via Europeana's OAI-PMH service[12], transformed to JSON where each record was stored in a separate line, and stored in Hadoop Distributed File System[13] (2) record-level measurement: the Java applications[14] measuring different features of the records run as Apache Spark jobs, allowing them to scale readily. The process generates CSV files which record the results of the measurements such as the number of field instances, or complex multilingual metrics. (3) statistical analysis: the CSV files are analyzed using statistical methods implemented in R and Scala. The purpose of this phase is to calculate statistical tendencies on the dataset level and create graphical representations (histograms, boxplots). The results are stored in JSON and PNG files. (4) user interface: interactive HTML and SVG representations of the results such as tables, heath maps, and spider charts. We use PHP, jQuery, d3.js and highchart.js to generate them.

[11] `http://144.76.218.178/europeana-qa/multilinguality.php?id=all`

[12] `http://labs.europeana.eu/api/oai-pmh-introduction`. Our client library: `https://github.com/pkiraly/europeana-oai-pmh-client/`.

[13] Snapshot created at the end of 2015, consisting of more than 46 million records (392 GB disk space): `http://hdl.handle.net/21.11101/0000-0001-781F-7`.

[14] Source code and binaries: `http://pkiraly.github.io/about/#source-codes`.

5 Interpretating the output of the measures

The status of these metrics with regard to interpreting their results and thus deriving strategies for metadata improvement differs considerably. Whereas the measures for the quality dimensions of completeness and accessibility need more refinement, as well as more discussion of appropriate display of the results, the consistency metric has already served to drive a successful project for normalizing languages. These various developments and their outcomes are reported in this section.

5.1 Completeness

The completeness measure indicated that 904 (out of 3548) collections have no value in the *dc:language* field, which shows the field is missing. On a record level, 58,03% of the records have a *dc:language* field.[15] Another pattern one can detect in the data relates to the misuse of fields. For example, the metric "cardinality" allowed us to identify collections that have metadata fields with more than 3 instances of *dc:language*. In one example, as many as 153 language values were found associated with this field, owing to duplication of the language tag. One recommendation to avoid this in future would be to check the field for duplicated entries and reduce them to a single language tag.

5.2 Consistency

Measuring consistency in the *dc:language* field shows that values are currently quite heterogeneous in the Europeana dataset. *dc:language* values are predominantly normalized in ISO-639-1 or ISO-639-3, but, in contrast, values nevertheless sometimes occur in natural language sentences that cannot be processed automatically. We also find language ISO codes without their reference to the ISO standard in use, or references to languages by their name. Additionally, over 400 different language variants are present in the field. This table presents some general statistics about the presence of ISO-639 codes in the values of *dc:language* in the Europeana dataset:

Total values in the Europeana dataset	33,070,941
Total values already normalized (ISO-639-1, 2 letter codes)	23,634,661
Total values already normalized (ISO-639-3, three letter codes)	4,831,534

The metric helps us to design further language normalization rules which in turn can be used to further improve the results of the quality measures. Any visualisation implementation of the quality measures would also benefit from such normalisation. We have developed a language-normalization operation

[15] http://144.76.218.178/europeana-qa/frequency.php

on the Europeana dataset which itself comprises a range of operations - i.e., cleaning, normalization and enrichment of data. The following exemplify some cases of the different types of operations:

Input value	Normalization output (ISO 639-1)
"English"	"en"
"eng"	"en"
"English and Latin"	"en", "la"
Greek; Latin	"el", "la"

The output of the normalization algorithm is a value from any of several authoritative language vocabulary. The algorithms work based on a core vocabulary, the Languages Name Authority List (NAL) published in the European Union Open Data Portal[16]. All normalization operations are internally performed using the core vocabulary and the alignements to other vocabularies it contains [17]. The final output can be given in accordance with any of the aligned vocabularies and can be configured to use URIs from the NAL vocabulary or any of the aligned ISO code-sets.

5.3 Accessibility

As noted earlier, our approach to measuring multilingual saturation in metadata allows us not only to measure the data's quality as it is provided by contributing institutions, but also provides us with insight into the effectiveness of Europeana's data enhancement processes, such as semantic enrichment. The numbers provided as part of the implementation of the quality measures are either not enough or sometimes too complex to interpret the results and draw conclusions. At the time of writing, we are in the process of refining the saturation score analyzing first results of the calculations. Visualisation is therefore a necessary tool to supplement the measure and guide the data providers in their data analysis.

6 Conclusion and future work

In this paper, we presented our approach to assessing the multilingual quality of data in the context of Europeana. This approach takes into account the functional requirements identified for digital libraries and translates them into actionable measures. These measures look at the data dimensions of completeness, consistency, and accessibility. Finally, we also present the initial outcome of our quality measures, such as the definition and implementation of normalisation rules for languages. The numbers provided as part of the quality measures

[16] `https://open-data.europa.eu/en/data/dataset/language`

[17] NAL is aligned with ISO 639-1 (currently in use at Europeana), ISO 639-2b, ISO 639-2t and ISO 639-3, providing human-readable labels in all European languages

still need to be refined. Future work on refining visualisation reports will help us both to interpret the numbers and to adjust our measurements. For instance, in order to get a comprehensive view of the quality of a data, the different metrics will need to be presented together (e.g. multilinguality on top of completeness) so that the interrelation between the different metrics is made visible.

References

1. Functional requirements for Bibliographic records: final report / IFLA Study Group on the Functional Requirements for Bibliographic Records. No. vol. 19 in UBCIM publications ; new series, K.G. Saur, München (1998)
2. ISO/IEC 25012. , Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model (2000)
3. White Paper on Best Practices for Multilingual Access to Digital Libraries. Tech. rep., Europeana (2016)
4. Bellini, E., Nesi, P.: Metadata Quality Assessment Tool for Open Access Cultural Heritage Institutional Repositories. In: Information Technologies for Performing Arts, Media Access, and Entertainment. pp. 90–103 (2013)
5. Bruce, T.R., Hillmann, D.I.: The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: Hillmann, D., Westbrooks, E. (eds.) Metadata in Practice, pp. 238–256. ALA Editions, Chicago (2004)
6. Dröge, E.: Criteria for vocabulary evaluation and comparison. Tech. rep., Humboldt-Universität zu Berlin (2012)
7. Guy, M., Powell, A., Day, M.: Improving the Quality of Metadata in Eprint Archives. Ariadne (38) (2004)
8. Isaac, A.: Europeana data model primer. Tech. rep. (2013)
9. Mader, C., Haslhofer, B., Isaac, A.: Finding quality issues in SKOS vocabularies. In: Proc. Intl. Conf. onTheory and Practice of Digital Libraries. pp. 222–233. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
10. Palavitsinis, N.: Metadata Quality Issues in Learning Repositories. Ph.D. thesis, Alcala de Henares, Spain (Feb 2014)
11. Park, J.r.: Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. Cataloging & Classification Quarterly 47(3-4), 213–228 (2009)
12. Shreeves, S.L., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B., Cole, T.W.: Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practices for Federated Collections. In: Proc. ACRL Twelfth National Conference. pp. 223–237. Minneapolis, Minnesota (2005)
13. Stiller, J., Király, P.: Multilinguality of Metadata. Measuring the Multilingual Degree of Europeana's Metadata. In: Proc. 15th International Symposium of Information Science (ISI). pp. 164–176 (2017)
14. Stvilia, B., Gasser, L., Twidale, M.: A framework for information quality assessment. J. American Society for Information Science & Technology 58(12) (2007)
15. Vogias, K., Hatzakis, I., Manouselis, N., Szegedi, P.: Extraction and Visualization of Metadata Analytics for Multimedia Learning Object Repositories: The case of TERENA TF-media network (2013), `https://www.terena.org/mail-archives/tf-media/pdf547CE3lKFt.pdf`
16. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for Linked Data: A Survey. Semantic Web 7(1), 63–93 (2015)