# Multilinguality of Metadata

## Measuring the Multilingual Degree of Europeana's Metadata

*Juliane Stiller*[1], *Péter Király*[2]

[1] Berlin School of Library and Information Science
Dorotheenstr. 26, 10117 Berlin, Germany
juliane.stiller@ibi.hu-berlin.de

[2] Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Am Faßberg 11, 37077 Göttingen, Germany
peter.kiraly@gwdg.de

**Abstract**

Digital cultural heritage portals provide universal access to cultural objects and associated metadata originating in divers countries and language milieus. Offering an equally heterogeneous audience access to this content is a challenging endeavour. To ensure accessibility for audiences with different linguistic backgrounds, it is crucial that the underlying metadata offers the same information in several languages. This paper presents the conceptualisation and implementation of a metric for measuring the multilinguality in the digital cultural heritage portal Europeana. For every field in each record across the entire collection, the level of multilinguality can be assessed. Quantifying the multilingual richness of data has significant benefits for increasing metadata quality, improving multilingual access to cultural collections and reaching multilingual audiences.

**Keywords:** multilinguality; measurement; big data; Europeana; digital libraries; metadata quality; evaluation

In: M. Gäde/V. Trkulja/V. Petras (Eds.): Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017), Berlin, 13th—15th March 2017. Glückstadt: Verlag Werner Hülsbusch, pp. 164—176.

# 1 Introduction

Metadata is at the heart of information systems intended to provide broad access to cultural heritage material which is often heterogeneous and published in diverse forms. One important information system offering access to digital cultural heritage is Europeana[1] – a platform that aggregates metadata from over 3,500 different memory institutions such as museums, libraries, archives and galleries. Originating in institutions from across Europe, the metadata is not only linguistically diverse but reflects the differing indexing practices of providing institutions. This heterogeneity is often a barrier to Europeana's goal of offering broad access to its collection across languages for use and reuse.[2] The linguistic diversity of the describing metadata affects browsing, retrieval and display of the material and can be considered to be one dimension of metadata quality. High quality metadata ensures frictionless functionality; it is accordingly crucial to understand the factors that work to contribute to metadata quality. Research on this topic indicated that several different metrics (completeness, accuracy, timeliness to name a few) had previously been suggested; the topic of metadata multilinguality however, has not received much attention.

Multilinguality for Europeana means that a metadata record contains the same information in different languages and that values in certain fields are annotated with their respective language. It is evident, that records with more translations satisfy a greater number of functional requirements, e.g. search and access for a broad range of users, along with support for existing and anticipated functionality. The question is how this language diversity can be quantified to drive strategic decisions that improve multilingual functionality, such as display of content in users' preferred languages, search across languages and semantic linking, in the long run.

This paper describes the conceptualisation and implementation of a quantitative measure of multilinguality for Europeana's metadata. In the next section, section 2, related work is presented focusing on multilinguality in metadata as a dimension of metadata quality. Section 3 presents the concept of the multilingual score, while section 4 describes the implementation of the score

1 http://www.europeana.eu/

2 Europeana Strategy 2015–2020: http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana Strategy 2020.pdf

into the Quality Assurance Framework (Király, 2015b). The paper ends with a discussion and prospectus for further developments.

# 2 Related work

## 2.1 Multilingual metadata in Europeana

Europeana aggregates over 54 million metadata objects[3] from European cultural institutions. These metadata objects describe either digitized physical cultural heritage objects or born-digital material. Each object page in Europeana provides metadata describing that object, a thumbnail preview and a link to the owning institution's page for the object in question. Objects can be textual representations, images, audio or audiovisual content; of these, some 55% of the content consists of images.

For collections such as Europeana's, multilinguality is a key factor of fundamental importance. Not only is the described cultural content multilingual in itself, but so is its metadata. This linguistically diverse data is searched and accessed by an equally diverse audience from a wide range of countries and speaking many different languages. Offering information access independent of language is a challenging endeavour requiring solutions for the multilingual enrichments of metadata, features for multilingual user interactions such as search and browse functionality and an adapted graphical user interface design. Europeana has published a White Paper incorporating community input from various sources detailing all aspects which contribute to truly multilingual access provision (Stiller, 2016). Solutions Europeana has so far introduced to overcome language barriers are the automatic enrichments of metadata with multilingual vocabularies such as GeoNames (Manguinhas, 2016), language-independent access options such as colour search and the introduction of a knowledge graph for entities (Petras et al., 2017).

Automatic enrichment of metadata was evaluated (Olensky et al., 2012) with regard to its impact on retrieval. This process adds authority-type infor-

3 The size of Europeana is constantly increasing. On 24 January 2017, Europeana provided access to 54,217,972 million objects.

mation (person and place names, subject headings, date-time descriptors) from external vocabularies (such as VIAF, Wikipedia and GeoNames) that are often multilingual. On the one hand, the obtained results show that these processes need to be implemented with care to avoid negative impact on the user experience caused by incorrect enrichments, on the other hand, multilingual enrichments contribute to a higher visibility of documents in search results (Stiller et al., 2014), similar to the way query translation contributes to cross-lingual retrieval (Király, 2015a).

## 2.2 Multilinguality as a dimension of metadata quality

Attempts to improve multilingual information access cannot be undertaken without a holistic understanding of the multilinguality of metadata; however, there has heretofore been no methodology available for quantifying the multilinguality of a given record's metadata. Commonly accepted metadata quality dimensions and metrics do not include multilingual aspects – an astonishing omission given that access to data is one of the major motivators for improving and enhancing for data quality (Srivastava, 2011).

Eppler (2006: 71) identified up to 70 cited quality criteria for metadata; multilinguality, however, is not among them. Of course, multilinguality might be understood as a subcategory of other criteria such as completeness, accessibility and consistency. Such an approach, however, risks underestimating the multilingual problem, which remains a significant barrier for users who do not understand the language of the textual content– especially in digital collections (Chen, 2016: 17).

Taking an approach similar to Eppler's, Knight and Cowen (2005) reviewed 12 information quality frameworks in order to identify shared conceptualisations of information quality and derive a schema for assessing quality in the context of the World Wide Web. Again multilingual aspects are not mentioned. One can argue that the multilinguality of metadata is part of the user's perception of metadata belonging to the "subject criteria class" (Naumann & Rolker, 2000); here, a criterion such as "understandability" strongly depends on the language skills of a user. More objectively, the marking of language information in metadata can be determined and therefore measured. To the authors' knowledge, the only research which proposes a multilingual metric at all, measured the distribution of individual languages within a data collection based on language attribution (Vogias et al., 2013,

cited by Palavitsinis, 2014); the metric does not include whether a field's content is available in multiple languages.

To close this research gap, we have developed a model for measuring the multilinguality in metadata giving digital library administrators a means to assess their potential to reach multilingual audiences. This score and its implementation can further be used to support visualizations displaying patterns in the data which otherwise remain hidden. This development is part of the Quality Assurance Framework and Completeness measures of Europeana data (Király, 2015b). While the implementation described here focuses specifically on Europeana and the cultural heritage domain, the model itself is of course potentially applicable to other digital libraries.

## 3 A model for a multilingual score

To be able to calculate a multilinguality score for Europeana, one first needs to understand the organisation's information architecture and the potential multilingual dimensions which are reflected in it. The Europeana Data Model (EDM) is the metadata schema for the Europeana records (Isaac, 2013). It is based on RDF (Resource Description Framework), so a field value might have three types of values: literal (string or numeric), literal with language notation, and a resource identifier (URI) which points to another RDF statement. For example:

1. The value is a literal, e.g.
   `Subject: "Berlin Wall graffiti (writing)" .`
2. The value has language annotation (the language should be encoded as an ISO-639 language code, here *en* for English, and *de* for German languages), e.g.
   `Subject: "Brandenburger Tor"@de ,"Brandenburg Gate"@en .`
3. The value is a resource identifier, pointing to a multilingual vocabulary such as GeoNames, e.g.
   `Subject: <http://sws.geonames.org/2661886/> .`

These three potential formats of values populate the fields in an EDM record. They express a certain degree or level of multilinguality. According to the schema in table 1, we assigned scores to a field value ranging from 0 to 2.6 – the more multilingual information the higher the score.

For each field, the scoring in table 1 is used. If a field has a simple string value the scoring is 0, if the string value is marked with a language tag it gets a 1. If there are 2–3 different language tags the score 2 is applied, for 4–9 different language tags the score 2.3 and for more than 10 different language tags the score 2.6. Resource identifiers can be contributed by the Europeana data providers (those cultural heritage institutions, which share their records with Europeana), or they can be automatically added as the result of Europeana's internal semantic enrichment process (Manguinhas, 2016). If a resource identifier is dereferenceable, the labels associated with the dereferenced entity are counted as though they belonged natively to the record; that is to say, the labels are considered to be in a sense 'folded in' to the record. On the other hand, if the identifier cannot be successfully dereferenced, then its contribution to the score is 0.

*Table 1: Scores for field values with regard to multilinguality*

| Levels of multilinguality per field | Expressed in numbers |
|---|---|
| Missing field | NA |
| Text string without language tag (language not known) | 0 |
| Text string with language tag (language known) | 1 |
| Text string with 2–3 different language tags (language known with potential translations) | 2 |
| Text string with 4–9 different language tags (language known with potential translations) | 2.3 |
| Text string with more than 10 different language tags (language known with potential translations) | 2.6 |

Obviously, in an ideal case the different language tags per field indicate translations of certain string values, but we are well aware that this is not always going to be the case. For some fields where one would expect a unique value (such as dc:title), we can assume that several labels with different tags indicate translations. For other fields where we often have several values (such as dc:subject), however, we cannot infer that the different instances are translations of each other. We are accordingly here simply counting distinct language tags, rather than translations per se.

Each field in a record is scored without a weighting. That means that a value in e.g. the dc:title field is not rated as more important than one in the dc:type field, and all fields are considered equal. This is the practice even in situations where one might expect some biasing – for example, with fields

such as dc:title, which will typically not contain links to a controlled vocabulary and thus tend to count lower than other fields.

## 3.1 Normalization

Normalization (scaling the scores to the range of 0 to 1) is considered beneficial for comparing, displaying and visualizing the data. One of the challenges is to determine how to normalize the score accurately. For now, a scaling of scores through

```
normalizedScore = 1 - 1/(score + 1)
```

is implemented. To avoid information loss during normalization, both scores – original and post-normalization – are stored and displayed.

## 3.2 Aggregating scores by instance, field, record or collection

The various approaches taken to score aggregation can be best illustrated by means of an example:

| Subject field in record 1 with 2 instances of text strings with 2–3 different language tags each | Subject field in record 2 with 3 instances of text strings with 2–3 different language tags each |
|---|---|
| Instance 1: "table"@en, "tafel"@nl, "tisch"@de .<br><br>Instance 2: "book"@en, "boek"@nl . | Instance 1: "flowers"@en, "bloemen"@nl, "blumen"@de .<br><br>Instance 2: "cup"@en, "tasse"@de .<br><br>Instance 3: "woman"@en, "frau"@de . |
| Sum: 4, Average: 2 | Sum: 6, Average: 2 |

Here, we have several instances of the same field. Each one of the instances yields different scores. We calculate both the sum of the individual scores, and the average. Since there are more instances in record 2, it gets a higher sum, but the average will be the same.

For deeper investigation, the tool supports the retrieval of the aggregated scores and the list of values for the individual instances (cf. table 2).

The “instances” section contains the type and score of individual instances of fields; the “score” is for the final scores based on all instances. At the top and collection level, calculations are based on the scores of all in-

stances; only when inspecting individual records is information on particular fields preserved, displayed, and the score aggregation made explicit.

*Table 2: Field level scores in REST API and in the web interface*

| REST API | Web interface |
|---|---|
| "Place/skos:altLabel": {<br>"instances": [<br>{"TRANSLATION": 2.0},<br>{"TRANSLATION": 2.0},<br>{"TRANSLATION": 2.0},<br>{"TRANSLATION": 2.0},<br>{"TRANSLATION": 2.0},<br>...<br>{"TRANSLATION": 2.40},<br>{"STRING": 0.0},<br>],<br>"score": {<br>"sum": 20.40,<br>"average": 1.85454545,<br>"normalized": 0.649681<br>}<br>} | *instances:*<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2)<br>▪ translation (2.40)<br>▪ string (0)<br>*score:*<br>▪ sum: 20.40<br>▪ average: 1.85<br>▪ normalized average: 0.65 |

## 4 Implementation

The multilingual saturation score is implemented within the completeness measures of the open source Metadata Quality Assurance Framework: http://144.76.218.178/europeana-qa/. The framework is written in a modular way: the record level feature extraction and calculation of the score is written in Java using the Apache Spark framework, the statistical analyses were written in R and Scala[4]. The data was ingested from Europeana's OAI-PMH server[5] and stored in Apache Hadoop's distributed file system as JSON files

4 Source codes are available from http://pkiraly.github.io/about/#source-codes, the workflow's details are described at http://pkiraly.github.io/cheatsheet/.

5 See http://labs.europeana.eu/api/oai-pmh-introduction for details of Europeana's OAI-PMH server. In order to make this research reproducible we published this snapshot

– one record per line. This way, the process could be easily parallelized and distributed over multiple processors and machines. The output consists of JSON files and PNG images. The web interface – written using PHP and d3.js – renders this output and provides rich navigation through, and interactive data visualizations of, the data.

Select dimension: Multilingual saturation (cumulative) groupped by Data providers

| | # records | Dataset | Minimum | Maximum | Range | Median | Mean | Standard deviation |
|---|---|---|---|---|---|---|---|---|
| 238 | 35134 | Parisienne de Photographie | 35.4 | 333.4 | 298 | 113.05 | 116.1796 | 34.2511 |
| 647 | 5630 | Deutsche Kinemathek | 30.2 | 562.6 | 532.4 | 110.6 | 116.539 | 37.8709 |
| 335 | 20651 | KU Leuven | 52.2 | 321.4 | 269.2 | 102.4 | 110.0678 | 27.2582 |
| 969 | 1739 | Bildarchiv Foto Marburg / Institut Mathildenhöhe Darmstadt | 42.9 | 158.6 | 115.7 | 101 | 102.7493 | 17.7996 |
| 809 | 3190 | Archiwum Muzyki Wiejskiej | 67.2 | 124.5 | 57.3 | 98.5 | 92.2154 | 9.7659 |
| 411 | 14350 | Universitätsbibliothek Leipzig / Digitaler Portraitindex | 15 | 154 | 139 | 92.6 | 89.1179 | 20.9254 |
| 292 | 31 | Germanisches Nationalmuseum Nürnberg / Digitaler Portraitindex | 55 | 122.3 | 67.3 | 91.6 | 83.5129 | 18.5291 |
| 690 | 4695 | MAK - Österreichisches Museum für angewandte Kunst / Gegenwartskunst | 13 | 195.9 | 182.9 | 91.1 | 89.5043 | 18.2391 |
| 1731 | 184 | Cinémathèque Royale de Belgique | 27.8 | 309.1 | 281.3 | 84.4 | 91.4761 | 42.0275 |
| 1031 | 1388 | www.esbirky.cz | 40.5 | 129 | 88.5 | 82.5 | 80.9622 | 15.9246 |
| 360 | 3 | LWL-Museum für Kunst und Kultur (Westfälisches Landesmuseum) / Digitaler Portraitindex | 74.9 | 98.6 | 23.7 | 81.9 | 85.1333 | 12.1763 |

*Fig. 1* Cumulative score of multilingual saturation per data providers, ordered by the median values

On the main page one can see the aggregated statistics for each dataset or data provider (fig. 1). By means of the drop-down menu, the multilingual saturation can be displayed for each field, plus the cumulated one showing the sum, the average or the normalized average of the multilingual saturation score. The table shows basic statistics, such as record count, minimum, maximum, mean, median values, range and standard deviation. The table is sortable, allowing ready exploration of the data; such an approach is particularly useful in identifying outliers that arise from data problems.[6]

The same information is available on a heatmap visualization (cf. fig. 2). Each dataset/provider in Europeana is represented by a square in the heatmap. When clicking on an individual square, i.e. data set/provider, statistics for this data set and its constituent fields are shown. All fields starting with "multilingual saturation" are of interest here.[7]

(created in the end of 2015, containing 46 million records, 1755 files, 420 GB in total) under this persistent identifier: http://hdl.handle.net/21.11101/0000-0001-781F-7.

6 An example of the sum of the multilingual saturation of the dc.title field can be found here: http://144.76.218.178/europeana-qa/?feature=saturation_sum_proxy_dc_title&type=data-providers .

7 One example is the dataset of the Rijksmuseum, the relevant score can be found from row 147 downwards. http://144.76.218.178/europeana-qa/dataset.php?id=51&name=Rijksmuseum&type=d

On the level of record investigation, you can see in the table “analysed metadata fields”, how the values were scored and what the score is for each field.

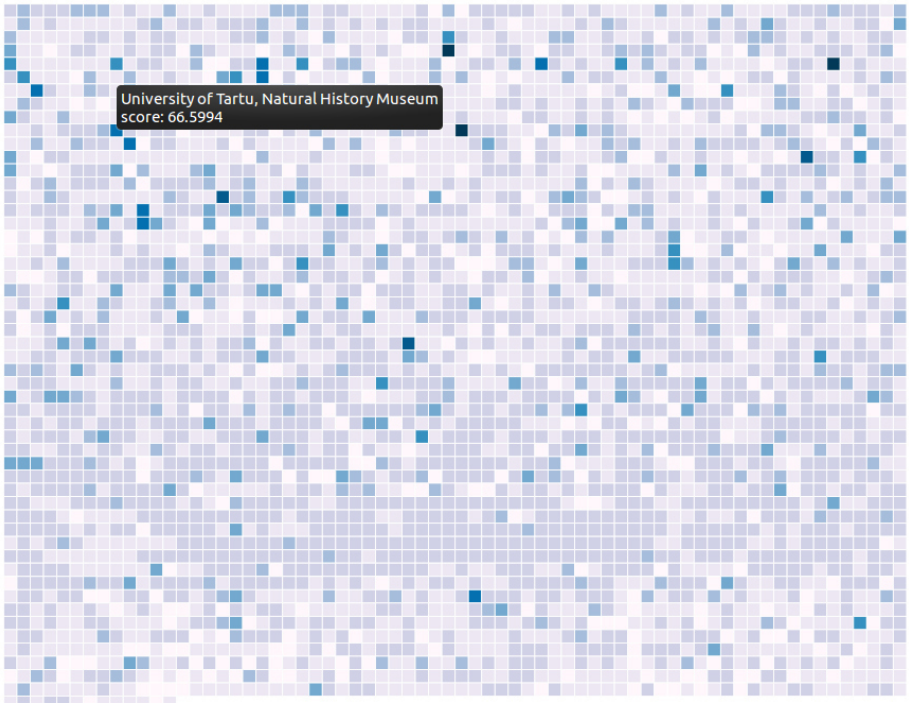


*Fig. 2* Cumulative score of multilingual saturation per data provider as interactive heatmap: the darker the colour, the higher the score. The squares are linked to in-depth analyses of the collection.

This score is part of the completeness measures and should not be understood as an independent score. All fields are taken into account but the multilingual measure is bound to the completeness metric. So a missing field should not harm the multilingual score, but be reflected in the completeness score.

## 5 Discussion and further development

In this paper, we introduced a score for measuring the multilinguality of metadata. This metric assesses the values of different fields and describes the fields' multilingual variety or richness. This is a first attempt in quantifying multilingual information in metadata in digital collections. It is intended that this will in turn assist in improving the quality of metadata in this regard and raise awareness of the importance of multilinguality for information access. The different aggregation levels of the score across a field, such as dc:title, over the whole collection help to reveal data quality problems – the statistical analysis of the information can be leveraged with the several visualizations offered.

Improvements to the score will be made in future by reviewing and comparing collections, their resulting scores, the representation of the scores and the visualizations. Additionally, the occurrences of multilinguality in metadata will be linked to the user experience and to their impact on search, browse and other functionalities in the portal. By doing this, the score can be harnessed to its full potential.

Of special concern is the distinction between the multilingual potential of the metadata that is submitted by the providers and the multilinguality Europeana is able to add to the data automatically. Currently, the score reflects the multilinguality of a metadata record in a modified version of what the data providers submit – that is to say, after Europeana has optionally added information from external semantic vocabularies in the ingestion process. Since these are multilingual data sources, this process improves the overall score, obscuring the multilingual character of the original record. In future, we want also to measure the original records and determine the multilinguality of the record during different stages of the ingestion process. This will help identify strategies to exploit the multilingual potential of data more fully.

**Acknowledgements**

This research was conducted under the aegis of the Europeana Network's Data Quality Committee (DQC)[8] formed early in 2016. We would like to thank all members of the DQC for their contributions to this research. We would also like to thank Timothy Hill from the Europeana Foundation for his feedback and the

8 http://pro.europeana.eu/page/data-quality-committee

review of this paper.

## References

Chen, J. (2016): *Multilingual Access and Services for Digital Collections*. Santa Barbara, CA: Libraries Unlimited.

Eppler, M. J. (2006): *Managing information quality. Increasing the value of information in knowledge-intensive products and processes*. Berlin, New York: Springer.

Király, P. (2015a): Query Translation in Europeana. In: *The Code4Lib Journal*, 27. http://journal.code4lib.org/articles/10285

Király, P. (2015b): A Metadata Quality Assurance Framework. https://pkiraly.github.io/metadata-quality-project-plan.pdf

Knight, S., and J. M. Burn (2005): Developing a framework for assessing information quality on the World Wide Web. In: *Informing Science: International Journal of an Emerging Transdiscipline*, 8 (5), 159–172.

Manguinhas, H. (2016): Europeana Semantic Enrichment Framework. Documentation, Europeana. http://bit.ly/2ijmKYM

Naumann, F., and C. Rolker (2000): Assessment Methods for Information Quality Criteria. In: *Proceedings of the 5th International Conference on Information Quality* (pp. 145–161).

Olensky, M., J. Stiller, and E. Dröge (2012): Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy. In: *6th Research Conference, MTSR 2012, Cádiz, Spain, November 28–30, 2012*. Proceedings, Berlin: Springer (pp. 252–263).

Palavitsinis, N. (2014): *Metadata Quality Issues in Learning Repositories*. PhD thesis. Universidad de Alcalá. http://dspace.uah.es/dspace/handle/10017/20664

Petras, V., T. Hill, J. Stiller, and M. Gäde (2017): Europeana – a Search Engine for Digitised Cultural Heritage Material. In: *Datenbank Spektrum*. doi:10.1007/s13222-016-0238-1

Srivastava, M. D. (2011): *Metadata creation in digital libraries*. Delhi: Pacific Publication.

Stiller, J. (Ed.) (2016): White Paper on Best Practices for Multilingual Access to Digital Libraries. Europeana White Paper. http://pro.europeana.eu/files/Europeana_Professional/Publications/BestPracticesForMultilingualAccess_whitepaper.pdf

Stiller, J., M. Olensky, and V. Petras (2014): A Framework for the Evaluation of Automatic Metadata Enrichments. In: S. Closs et al. (Eds.): *Metadata and Semantics Research: 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27–29, 2014*. Proceedings (pp. 238–249). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-13674-5_23

Vassilakaki, E., and E. Garoufallou (2013): Multilingual Digital Libraries: A review of issues in system-centered and user-centered studies, information retrieval and user behavior. In: *International Information and Library Review*, 45 (1–2), 3–19. https://doi.org/10.1016/j.iilr.2013.07.002

Vogias, K., I. Hatzakis, N. Manouselis, and P. Szegedi (2013): Extraction and Visualization of Metadata Analytics for Multimedia Learning Object Repositories: The case of TERENA TF-media network. In: *Workshop on Learning Object Analytics for Collections, Repositories and Federations, 9 April, 2013*. https://www.terena.org/mail-archives/tf-media/pdf547CE3lKFt.pdf