Europeana DSI 2– Access to Digital Resources of European Heritage

DELIVERABLE

D6.5: REPORT ON QUALITY METRICS AND IMPROVEMENT OF MULTILINGUALITY IN EUROPEANA

Revision	V1.0
Date of submission	31.08.2017
Author(s)	Juliane Stiller (Humboldt-Universität zu Berlin), Hugo Manguinhas (Europeana Foundation) Nuno Freire (INESC-ID), Timothy Hill (Europeana Foundation), Antoine Isaac (Europeana Foundation)
Dissemination Level	Public

Co-financed by the European Union
Connecting Europe Facility

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision No.	Date	Author	Organisation	Description
1	01.08.2017	Juliane Stiller	Humboldt-Universität zu Berlin	First draft
2	07.08.2017	Juliane Stiler	Humboldt-Universität zu Berlin	Incorporated comments by Valentine Charles, David Haskiya
3	07.08.2017	Nuno Freire	INESC-ID	Appendix 3
4	11.08.2015	Hugo Manguinhas	Europeana Foundation	section 7.2
5	14.08.2017	Timothy Hill	Europeana Foundation	section 2.5
6	15.08.2017	Juliane Stiller	Humboldt-Universität zu Berlin	Incorporating comments by Antoine Isaac and Valentine Charles
7	30.08.2017	Juliane Stiller	Humboldt-Universität zu Berlin	Incorporating comments by Pablo Uceda

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

‘Europeana DSI is co-financed by the European Union's Connecting Europe Facility’

Table of Contents

Introduction 5

2. Multilingual Dimensions of Europeana Components 6

2.1 Multilingual aspects of metadata 6

2.1.1 Language tags indicating the language of metadata values 6

2.1.2 edm:language for language of metadata record 7

2.1.3 dc:language for language identification of objects 8

2.2 Impact of multilingual information in metadata on system components 8

2.3 Users' understanding of multilingual information 9

2.4 Usage scenarios 10

2.5 Entity Collection (EC) 11

3. A Model of Multilingual Saturation in Europeana 12

4. Metrics for Multilinguality of Data 13

4.1 Multilinguality as a facet of data quality dimensions 14

4.2 Indicators and Measures 15

5. Metrics for Multilinguality of System Functionalities 17

6. Measuring Multilinguality of Metadata in Europeana 18

6.1 Preliminary considerations and terminology 18

6.2 Calculating the score 19

6.3 Determining the multilinguality of a record throughout its Europeana life cycle 19

6.4 Preliminary results 21

7. Improvement of Multilinguality in Europeana 23

7.1 Normalizing values in the dc:language fields 23

7.2 Multilingual Entities 24

8. Conclusion & Future Work 27

References 28

Appendix A: EDM fields where use of language tag and/or multilingual resources is encouraged 30

Appendix B: requirements for multilingual logging in Europeana 31

Appendix C: Language Normalization Evaluation Report 33

Introduction 33

Functionality 33

The language normalization algorithm 34

Language vocabularies 34

Matching algorithms 34

Evaluation procedure 35

Results 36
Conclusion 37

1. Introduction

The subtask *6.8.2. Measure and enhance multilingual performance* is set out to develop metrics to determining the “multilingual degree” of Europeana - including its services, its data and the user interface. Goal is to enable Europeana to track progress with regard to the development of multilingual features (e.g. the multilingual enrichment of data) and determine the impact of processes which contribute to a richer multilingual experience for users, namely enrichment, metadata translation and language detection. In this deliverable, the work on determining the multilingual degree of Europeana is presented.

Europeana aggregates content from various European institutions with differing indexing practices and languages used for describing the digital objects. Not only system functionalities, such as language change buttons, need to be in place that allow multilingual data to be accessed but also means need to be implemented that allow users to understand relevancy of given objects in relation to an information need. So far, Europeana has little knowledge about its multilingual reach and how much of the metadata has multilingual information. A thorough analysis and interpretation of the data also needs to take into account the quality of the inspected data. Malformed data and data that is not standardized (e.g. language codes) can lead to incorrect results. We quantified multilingual data and identified means to improve the quality of information available in different languages.

Additionally to measuring multilinguality in Europeana, this deliverable is going to report on the efforts to further increase and improve multilinguality in Europeana. Firstly, the biggest change in this regard is the introduction of the Entity Collection - a curated database of individuals, places, and concepts. Matching entities with multilingual labels in several languages to metadata content vastly increased the amount of multilingual metadata - making it easier to access content across languages. Secondly, we report on the improvement of existing data that is relevant for multilingual access through normalization of values denoting languages in the *dc:language* field.

The deliverable is structured as follows: Section 2 describes the different components of Europeana that influence the multilingual experience of users - special emphasis is given to the formal expressions of language in the data and the features of Europeana. A model of the multilingual saturation of Europeana aggregating the perspectives of the different components is presented in section 3. Section 4 develops specific metrics for the system component data, whereas section 5 looks at metrics for the system functionalities. In section 6, the implementation of scores for multilingual data in the Data Quality Framework is outlined and first results are presented. Section 7 summarizes efforts to improve multilingual data based on the aforementioned developed metrics. The deliverable ends with section 8 discussing the results and referring to future work.

2. Multilingual Dimensions of Europeana Components

As already expressed in the White Paper (Stiller (ed), 2016), there are several dimensions of multilinguality in Europeana: the data, the interface and the interactions which can be multilingual or have a multilingual perspective. If we want to develop metrics to measure the multilingual degree of Europeana services, we need to distinguish several levels: firstly, there is the metadata and its multilingual information; secondly, the impact of the multilingual information in the metadata on system interactions; and thirdly, the users' understanding of multilingual content. To understand the interplay of these levels, we present several usage scenarios with a multilingual dimension. The newly introduced Entity Collection will bring many multilingual labels that will impact the usage scenarios.

2.1 Multilingual aspects of metadata

In general, multilinguality in metadata can be measured on record level, field level or system level. There are several options where multilingual information can occur in metadata: (a) the language of metadata values is indicated by a language tag[1], and field values can be translated if their language is indicated by language tags and (b) the *edm:language* to indicate the assumed language of the metadata record as a whole, (c) the language of the object as indicated by values of fields such as *dc:language*.

2.1.1 Language tags indicating the language of metadata values

Metadata in Europeana are textual information describing the digital cultural heritage object. With language tags (or language attributes), providers can indicate the language of a particular value. Appendix A shows the different fields for which language tags are encouraged.
For instance, metadata from an institution can look like this:

```
<#example> a ore:Proxy ; # data from provider
  dc:subject "Ballet",   # literal (with no indication of language)
  dc:subject "Opera"@en  # literal with language tag
```

No language information is given in the first dc:subject statement, the second has language information that tells us that the literal is in English.
Under ideal circumstances, for each literal in a given field the language is known indicated by a tag. In a case where several different language tags in a field exist, more multilingual information is available. Different language tags can indicate translations:

```
<#example> a ore:Proxy ; # data from provider
    dc:subject  "Opera"@en,  "Oper"@de   #  literals  with  language  tags
representing translations
```

But this does not always need to be true:

```
<#example> a ore:Proxy ; # data from provider
  dc:subject   "Ballet@en",   "Opera"@it    #   literals   with   different
language tags but not translations
```

[1] expressed in an XML attribute with an ISO code (ISO639)

Automatically distinguishing these different cases is very difficult.

As extensively described in the Europeana Semantic Enrichment Framework[2]. Europeana enriches values in particular fields such as *dc:subject* automatically with controlled and multilingual vocabularies, such as GeoNames or DBpedia. The following example shows an automatically added link in the *dc:subject* field. The dereferencing of the link will retrieve all multilingual data attached to the particular concept defined in a linked data service. The language variants for this particular keyword will be added to Europeana's search index enabling cross-lingual search and display of keywords in a user's preferred language.

```
<#example> a ore:Proxy ; edm:europeanaProxy true ;
  # enrichment by Europeana 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
```

Compared to the source data, the record has now more multilingual information. Using multilingual controlled vocabularies as enrichment means has another advantage: added keywords in different language versions are very likely to be translation variants. The enrichment of metadata with multilingual vocabularies is an important process for Europeana to increase their multilingual reach.

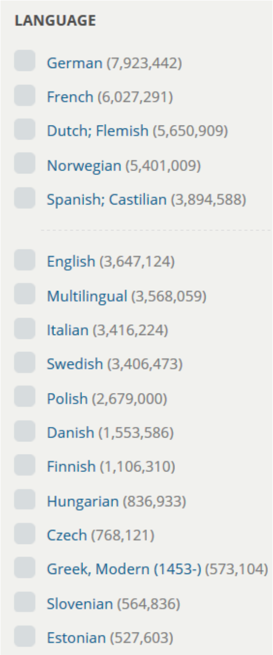
Also the development of Europeana's entity collection (Petras et al., 2017, D6.3 Search Improvement report, section on the entity collection) relies on the use of multilingual labels for language dependent display and for identifying the right language variants in the entity autocomplete implementation (see MS30 – SEARCH IMPROVEMENT PLAN, p. 4-5[3]).

2.1.2 edm:language for language of metadata record

The *edm:language* field denotes the assumed language of the metadata record populating with a constant value for all the records in the dataset. It is used to populate the language facet in the Europeana portal (Fig. 1). As the field's value often matches the language of the provider's institution, it can only denote the assumed language of the metadata record as a whole.

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

[3] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI/Milestones/europeana-dsi-ms30-search-improvement-plan.pdf



LANGUAGE

□ German (7,923,442)
□ French (6,027,291)
□ Dutch; Flemish (5,650,909)
□ Norwegian (5,401,009)
□ Spanish; Castilian (3,894,588)

□ English (3,647,124)
□ Multilingual (3,568,059)
□ Italian (3,416,224)
□ Swedish (3,406,473)
□ Polish (2,679,000)
□ Danish (1,553,586)
□ Finnish (1,106,310)
□ Hungarian (836,933)
□ Czech (768,121)
□ Greek, Modern (1453-) (573,104)
□ Slovenian (564,836)
□ Estonian (527,603)

Figure 1: Language facet in Europeana portal populated by edm:language field.

2.1.3 dc:language for language identification of objects

Digital cultural objects as they can be retrieved in Europeana can also have a language. If the object is of textual or audiovisual nature or in any other way a linguistic artefact, data providers are urged to indicate the language of the object. The Europeana Data Model (EDM) intends the *dc:language* field to be used in the following way: `<dc:language>de</dc:language>`. In this example the language field indicates that the language of the described object is in German.

A language facet for the language of the object (similar to the one shown in Figure 1) would be useful but would require a central normalization process (see section 7.1). For objects that do not obviously have a language, the Europeana's Data Quality Committee[4] recommends[5] the identification of non-linguistic content with *ZXX* as indicated by the ISO 639-2 standard.[6]

2.2 Impact of multilingual information in metadata on system components

The quality and richness of multilingual information as described in section 2.1 will have an impact on the system and interactions users with different language backgrounds have with Europeana. Metadata is necessary to fulfill functional requirements regarding search and browsing capabilities. The impact of multilingual information on these functionalities could be measured and tracked over time.

Depending on the user query 'mode', the multilingual variety of metadata presented in the search results changes, e.g. at the moment textual queries often lead to search results less varied in languages whereas queries for entities result in metadata more linguistically varied. Also: the

[4] http://pro.europeana.eu/get-involved/europeana-tech/data-quality-committee
[5] https://docs.google.com/document/d/1ej0ouDg_uhOVnE1LE2-IEtI9xNhMpeqzNIIwSjLoAbI/edit# (p.11)
[6] https://www.loc.gov/standards/iso639-2/php/code_list.php

more multilingual the metadata, the more languages could be reflected in the search results. Perhaps even more fundamental to Europeana's core mission of providing access to the diversity of Europeana culture: the more multilingual the data is, the more objects can show up among the results for a given query (as more query-object matches can be done in Europeana's index). A reflection of the variety of multilingual interactions could be the potential for language crossings within a search or browsing task. How likely is a user query to retrieve results different than the query language? A hypothesis is that searches for named entities, such as location and person names, may result in more variety (in languages) than topical searches. Users might click on results different from their native language. To make an informed statement about the language crossings of users on the portal, sophisticated logging needs to be put into place (Appendix B lists several requirements for multilingual logging in Europeana).

Additionally, suggestions how the search system could influence measures on multilinguality, e.g. boosting documents with links to vocabularies in SERP can be listed. This might result in greater language variety and higher visibility of small languages. Also the provided user interface languages play a role here.

2.3 Users’ understanding of multilingual information

When speaking about multilingual information, it is essential to look at the user perspective. If a system offers cross-lingual search and browsing functionalities it needs to make sure that users also understand the information. The questions to ask here is: how much translation and understanding does a user need to perform a given task? In Europeana, this understanding has several levels:

A. Understanding why an object ranks for a query. Users might be irritated by the occurrence of objects in language different from the query language. It is beneficial to make transparent why a result in a language different from the user language ranks. For example, this could be indicated by highlighting the term (which is a translation of the user query) in the search results.

B. Understanding the metadata to determine relevance of the object. When a user enters a query and gets results in languages different from her query language, functionalities need to be in place to allow her to determine the relevance of a given object to her query.

C. Understanding the object to satisfy a given information need.

One part of multilingual aspect is how much the user can understand or take away from a record across languages. Can the record satisfy a given information need?

Adapting the hierarchy of text handling task (Taylor & White, 1998) to Europeana, the following tasks could be defined starting with the most difficult task to the easiest task.

Task	Description	Basic requirement
Gisting	Produce a summary of the object in a language	dc:title / dc:description / thumbnail / in native language + picture/thumbnail
Extraction	Named entity recognition: identify names of people,	Completeness of dc:creator, dc:publisher and all other field

	places, concepts and time	with entities in native language
Categorize	Sort documents to a given topic	Some keywords in native language + Picture
Detection	Find documents of interest	Few keywords in native language + Picture
Filtering	Discard irrelevant documents	Picture

Table 1: Text handling task and information required for it adapted from Taylor and White, 1998.

2.4 Usage scenarios

In the work of the aforementioned Data Quality Committee, several usage scenarios were developed that require multilingual information to let users find and explore content across languages. For supporting these usage scenarios, enabling elements were determined. Table 2 lists these scenarios and the requirements needed in the data to enable them.

Scenario	**Enabling elements**
cross-language recall	● all the EDM elements supporting literals SHOULD be provided with language tags. ● using EDM elements in combination with (i.e. which link to) a contextual entity with multilingual features is RECOMMENDED.
improved facet based on language of metadata	● all the EDM elements supporting literals SHOULD be provided with language tags. ● using EDM elements in combination with (i.e. which link to) a contextual entity (with link to a multilingual features) is RECOMMENDED.
improved facet based on language of content	● For enabling this scenario the dc:language element, describing the language of a CHO, MUST be provided when a resource is of edm:type TEXT and SHOULD be provided for these other types (AUDIO, IMAGE, VIDEO, 3D). ● the Europeana Data Quality Committee (DQC)[7] RECOMMENDS[8] the use of the ISO 639-2 code[9] for no linguistic content (ZXX).

[7] http://pro.europeana.eu/page/data-quality-committee

[8] If the dc:language will populate a facet in future, the necessity for the use of the ISO-standard needs to be a MUST and not only a RECOMMEND.

[9] https://www.loc.gov/standards/iso639-2/php/code_list.php

Table 2: Usage scenarios and enabling elements developed by the DQC in relation to multilinguality.

For a full description of the scenarios, please refer to the document “Discovery- User scenarios and their metadata requirements - v.3” aggregated by the DQC.[10]

2.5 Entity Collection (EC)

Analysis of Europeana's query logs has shown that users query the site for entities - that is to say, for particular individuals, places, and defined topics - much more often than they enter keywords corresponding to, say, work titles or broad thematic concerns.[11] As a result of this finding, Europeana has begun to implement an Entity Collection for its services including the Europeana Collections portal - that is to say, a curated database of individuals, places, and concepts, through which users can query the document collection with precision.[12]

While Entity Collections are generally useful for improving information retrieval, they are particularly valuable in a multilingual context. The nature of Europeana's metadata makes it a poor candidate for machine translation across languages: while the collection as a whole is too large (50 milllion+ items) for comprehensive automated translation, every individual record within it contains insufficient context to guarantee high linguistic quality on its own. By contrast, the relatively sparse and stable data associated with entities in the EC means that translation can be highly effective, while still being of a sufficiently small scale to guarantee easy enrichment and evaluation. In addition, the fact that the entities of the EC are harvested from Linked Open Data sources means that they are often already associated with labels in several dozen languages, which simply need to be ingested into our datastores to be made available to our users. Accordingly, the Entity Collection forms the chief strategy currently being pursued by Europeana to improve multilingual access to its services.

[10] The contents of the document may be subject to change in future as the usage scenarios are still work in progress: https://docs.google.com/document/d/1ej0ouDg_uhOVnE1LE2-IEtI9xNhMpeqzNIIwSjLoAbI/

[11] https://europeanadev.assembla.com/spaces/europeana-r-d/documents/dxtIrqGpSr55TMdmr6CpXy/download/dxtIrqGpSr55TMdmr6CpXy

[12] The alpha release of the Entity API is available here: http://labs.europeana.eu/api/entities-collection

3. A Model of Multilingual Saturation in Europeana

To offer a holistic view on the different aspects that influence multilinguality in an information system, components shaping the multilingual experience need to be analyzed. Fig. 2 shows the mutually reinforcing components metadata and system functionalities that impact the level of multilinguality the users is experiencing. The multilingual degree of metadata influences interactions across languages, whereas the system functionalities can influence how much of the (content associated with) multilingual data is visible to the user. Also, the multilingual degree of metadata impacts the level of tasks a user can fulfill (table 1 & 2). The system functionalities that supports multilinguality enable the user to cross languages during their task, e.g. when the user finds objects in languages different from the query language.

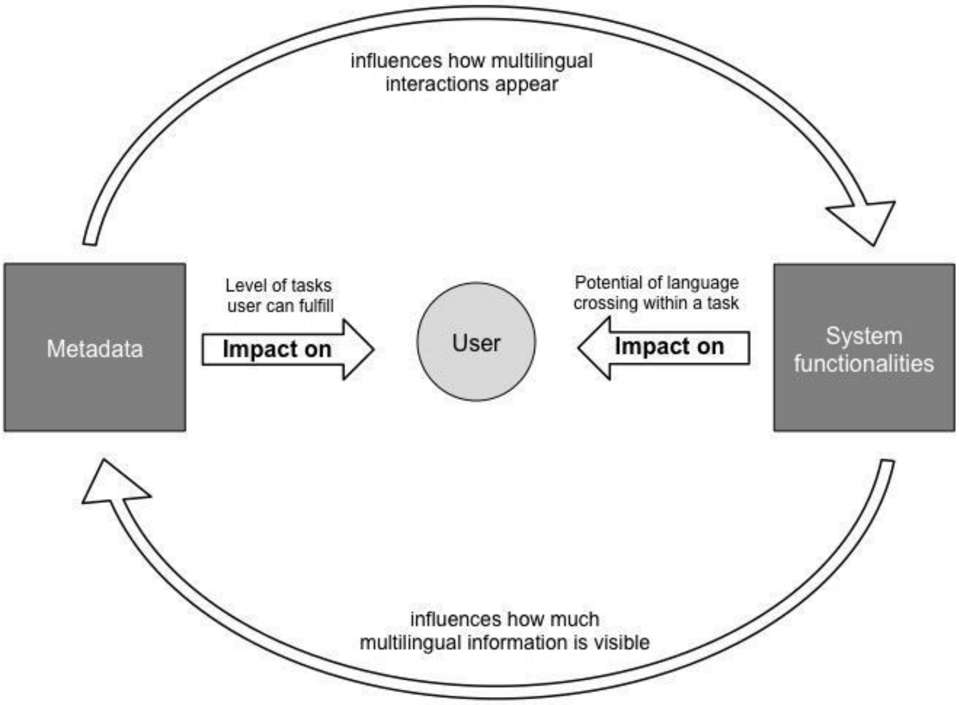

influences how multilingual interactions appear
Metadata
Level of tasks user can fulfill
Impact on
User
Potential of language crossing within a task
Impact on
System functionalities
influences how much multilingual information is visible

Figure 2: Schema presenting influence of multilingual system functionalities and metadata on user.

Level of multilinguality in	Impact on users
Object and its describing metadata	Level of tasks user can fulfill
System functionalities	potential of language crossings within a task

Table 3: Model for multilingual quality measures for Europeana.

4. Metrics for Multilinguality of Data[13]

In the literature of metadata assessment or quality, multilinguality is sometimes mentioned but it is often not considered a data characteristic that should be measured. We understand multilinguality as a facet of different quality dimensions. In metadata assessment it is accepted practice that quality dimensions are defined in accordance with functional requirements of a given system. For Libraries, the functional requirements of bibliographic metadata can be defined as users to find material, to identify an item and to select and obtain an entity (IFLA, 1998). Park (2009) builds upon this idea and determines discovery, use, provenance, currency, authentication and administration as the main functionalities good quality metadata should support.

Common quality dimensions in the GLAM (galleries, libraries, archives and museum) domain are defined by Bruce and Hillmann (2004): completeness, accuracy, provenance, conformance, logical consistency and coherence, timeliness and accessibility, grouped in three tiers. For accessibility, they distinguish between technical and intellectual accessibility. The intellectual accessibility assess the fitness of metadata for multiple audiences[14]. Multilinguality is not particularly mentioned in this notion but can be understood as being a part of it. Similarly to Bruce and Hillmann, Shreeves et al. (2005) use three quality dimensions, namely completeness, consistency and ambiguity for their explorative study. Their focus was the quality of collections aggregated from different institutions.

Designing an information quality framework, Stvilia et al. (2007) developed measures grouped into three categories: whereas intrinsic measures are somewhat objective, relational measure always depend on the context of usage, the third group holds measures for reputational information quality. In this framework, precision is part of the intrinsic measures (e.g. the number of elements, the non-empty elements) and completeness is categorized as a measure of the and in the relational group (completeness wrt. a recommended set of elements). Multilinguality is not explicitly mentioned in this paper, but considering it to be an special application profile based on usage scenarios it can be part of contextual completeness.

Mader et al. (2012) analyzed the quality of SKOS vocabularies and identified problems related to multilinguality, such as incompleteness of language coverage and missing language tags. Similarly, Dröge (2012) defined criteria for evaluating the quality of vocabularies. Multilinguality is mentioned, however, no metric is developed in detail. In a similar vein, Albertoni et al. (2015) propose a measure for determining the quality of linksets in Linked Open Data. They assess the potential of these datasets of RDF links in complementing their linked datasets.

Multilinguality can be understand as a perspective or facet of different common metadata quality dimensions. Following the dimensions completeness, consistency and accessibility will be examined to developed metrics for multilinguality.

[13] Parts of this subsection are adapted from Charles et al., 2017.

[14] In other quality analysis frameworks like ISO/IEC 25012 - Data Quality model (http://iso25000.com/index.php/en/iso-25000-standards/iso-25012) this notion of 'intellectual accessibility' may correspond to other dimensions like 'understandability'.

4.1 Multilinguality as a facet of data quality dimensions

Completeness is a commonly referred quality dimension. A completeness measure can look at the fields present in a record or collection, measuring the share of non-empty fields. Multilingual completeness can be measured from two perspectives. Firstly, one can determine the share of fields that have values with language tags. In contrast to the “normal” completeness measure, the multilingual completeness will use the non-empty fields as a baseline. That means that only fields both present and non-empty can be said to have or lack language tags and translations. This is different to a completeness measure that uses all possible fields as the baseline to determine the share of non-empty fields. A record that has values in 80% of the possible fields can still reach 100% multilingual completeness if all present and non-empty fields have a language tag.
Secondly, the presence of the *dc:language* field as a binary measure can be determined. Does the metadata record identify the language of the objects it is describing?

Consistency refers to the coherence of the data across all fields and records. With regard to multilinguality, the dimension assesses the coherence of language values either in the *dc:language* field or the language tags specifying the language of values in certain fields. Are the same values used to denote the same language. Consistent values in the *dc:language* fields allow to correctly populate the language facet in Europeana, so user can better filter objects by their language.

Conformity is often described as the conformity to set standards and field rules. For example, a conformity test on the language codes discussed previously will check if these codes comply with a given (set of) standard(s), such as ISO-639-3 (NB: in this specific case, conformance can be tested as an extra quality requirement over consistency: not only values need to be homogeneous, but they need to follow the same standard).

Accessibility is a somewhat more fuzzy metrics which describes how well the data can be understood and how well it can be retrieved. Of course, in the realm of multilinguality, accessibility is about retrieval and sense-making across languages. Can users with different language backgrounds access information in Europeana? How well (or: how evenly) is linguistic information in Europeana’s metadata distributed allowing access across several languages? This notion follows an understanding of Accessibility that goes beyond the technical aspects of allowing particular user groups access to content. In our understanding of Accessibility, we follow Christine Harlow’s approach that subsumes different access points involving different languages as Accessibility (Harlow, 2017). Quantifying this is not a trivial task and the language tag is crucial here. The interpretation of such a measure should also be scrutinized as this particular multilingual score can be very biased: not all the countries provide the same amount of content so some small languages are less covered and less likely to show up search results. Even though Europeana tries to select multilingual vocabularies, languages are not always covered equally. Also, vocabularies tend to have more translations for broad concepts or well-known entities than for specific ones or small and/or unpopular entities. These observations need to be taken into account when interpreting any scores related to access to information.

Summarizing, the following measure can be applied in different dimensions with regard to multilinguality.

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
Conformity	Compliance to a given standard	Binary or share of values that comply or not comply
Accessibility	Multilingual accessibility	Numbers of distinct languages Number of languages tagged Tagged literals per language

Table 4: Dimensions, criteria and measures for assessing multilinguality in metadata.[15]

4.2 Indicators and Measures

Considering the smallest entity in Europeana - a record - we can determine the multilingual completeness of its metadata.

Given the value in certain fields, one can determine how multilingually complete a certain record is. The table 4 lists some possible indicators and their related measures on a record level.

Factor	**Indicates**	**Measure**
dc:language has a value	Language of the object	binary
dc:language has a value which is normalized (Match to 639-2 code)	Groups of objects with the same language	binary
For specific field, there is a language tag	language of field value	binary
For specific field, there are several language tags	field values in different languages	number of different language tags
For specific field, there is a link to controlled vocabulary	translations of field values	binary or after dereferencing of link: number of different language tags
Number of fields have multilingual information	multilingual completeness on record level	fraction of fields that have multilingual information

Table 5: Factors and measure to indicate multilinguality on record level.

[15] Table copied and adapted from Charles et al., 2017

Similar indicators can be used to determine the multilingual degree of a whole collection.

Factor	**Indicates**	**Measure**
Languages in the metadata	multilinguality coverage of the metadata	number of different language tags
Quantity of information in a specific language across collection	how plenty or scarce language information really is	number of values per language tag

Table 6: Factors and measure to indicate multilinguality on collection level.

If trying to define score for multilinguality one needs to determine the fields which are relevant for multilinguality. This relevance is determined (1) by the field’s value for supporting multilingual usage scenarios and understanding for the user (e.g. description, subjects) (see table 2), (2) by its relevance for search (appendix A lists all fields with multilingual information).

5. Metrics for Multilinguality of System Functionalities

If we now shift the perspective from the metadata to the system functionalities and see how they can be assessed in terms of their ability to exploit multilingual information, the following measures could be taken into account.

Factor	**Indicates**	**Measure**
Variety of languages of metadata in SERP (e.g first result page)	Potential language crossing	number of different languages objects are in
Amount of documents per language in SERP	Potential language crossing	number of documents per language
Visibility of languages in SERP retrieved by various query sets	Which languages actually make it into the SERP	How often does a language result occur in search
Number of language crossing	Do user actually cross languages?, meaning: query in one language clicked result (metadata) is in a different language, object is yet in another language	Number of queries for which that happens

Table 7: Measure for assessing impact of multilingual metadata and system functionalities in users.

Given these theoretical framework for measuring the multilingual saturation of the Europeana portal, we now delve into the use case of quantifying the multilinguality within the metadata of Europeana.

6. Measuring Multilinguality of Metadata in Europeana

During the course of the project Europeana DSI-2, not only theoretical assumption about multilingual data quality was made, but also concrete measures were implemented. This was done in the realm of the Europeana Data Quality Assurance Framework that quantitatively assess the quality of Europeana's metadata (Király, 2015). This section describes the considerations and specificities of Europeana data that were taken into account to implement the measure for the different dimensions.

6.1 Preliminary considerations and terminology

Quantitatively determining the level of multilinguality solely looks at the values of the metadata and determines their multilingual variety or richness. The intention is not to incorporate a measure of completeness as it is already implemented in the Europeana Data Quality Assurance Framework[16] but measure the multilinguality of existing records and collections. The multilingual measure is bound to the completeness measure (Király, 2015) in the way that all existing statements are taken into account. So a missing property or statement should not harm the multilingual score, but rather the completeness score.

In terms of terminology, we used the terminology for the RDF data model[17]. A ***statement*** is a triple consisting of *subject, property, object*, for example *proxy1, dc:subject, "sculpture"@en*. A ***property*** is a resource that can be used as the predicate (2nd position) of a statement, e.g. *dc:subject*. A set of values for the *dc:subject* property for one Europeana object is referred to as "the objects of *dc:subject* statements for the proxy" (when the CH object is represented by a proxy).

A language tag is the 'en' in '"sculpture"@en' above. Throughout this deliverable, 'language tag' and not 'language' is used when talking about the technical definition of a measure. This is important as languages are not exactly in correspondence with language tags (a language can be represented by several language tags e.g. sculpture@en, sculpture@eng).

The idea is to determine a score for multilinguality which can be applied on statement, property or record level. A simplified schema was defined as the basis for the measurement assuming that each statement in a property can have one of the following values: a literal, a literal with a language tag, a URI (ideally to a controlled vocabulary):

Levels	Description	Contribution to the score
0	String value with no language indication	We have no multilingual information here, therefore a simple string contributes to the score with the value of 0
1	String value with language tag	The number of language tags and the number of distinct language tags is counted

[16] http://144.76.218.178/europeana-qa/

[17] https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/#section-data-model

2	URI[18]	If the URI was dereferenced the values will be counted according to level 1, if the URI was not dereferenced the value will be counted as level 0

Table 8: Levels of multilinguality within field values.

Each statement in a property can be assessed based on this schema. Weighting of different properties is not applied as it would have introduced another dimension profile. Same goes for restriction of properties, e.g. some properties are more likely to have URI than others, which were not taken into account. This first iteration of a metric to measure multilinguality in metadata was thoroughly described in Stiller & Király (2017) and partly implemented in the Data Quality Assurance Framework.

6.2 Calculating the score

For each property, the scoring in the table 6 is used. If a statement is a simple string value the scoring is 0. If the string value is marked with a language tag, this language tag contributes towards the multilingual score. The language tags overall are counted as well as the distinct ones regarding a certain entity (statement, property or proxy). If a statement has a URI to a controlled vocabulary, the dereferenced data that materialized in the Europeana data will be counted towards the property the URI occurred in (please refer to section 6.3 for a deeper discussion on how to measure multilinguality in the different stages of a record). This allows us to calculate the score based on the structure of RDF graphs, e.g. we look at the ProvidedCHO (or proxy) node and look at the multilingual degree for every statement that has that node as subject.

In an ideal case, the different language tags per statement indicate translations of certain string values, but this does not need to be the case. For some properties where one would expect a rather unique value (such as *dc:title*), one could assume that several literals with different tags indicate translations. For properties, where several values (such as *dc:subject*) exist, one cannot assume that these are translations of each other. The only certainty of a translation would be if the value is a resource (say, a Concept or Agent) that operates a "de facto grouping" of labels (these labels are related through the skos:Concept or edm:Agent classes that acts as a "hub"). And even then one can't be really sure. There can be alternative labels for a concept, which are not really translations. Therefore, we are speaking about different language tags rather than translations.

6.3 Determining the multilinguality of a record throughout its Europeana life cycle

An issue is the multilinguality of an object or its metadata in its different stages from the creation of the metadata at the provider institution to the ingestion into Europeana where provided URIs might get dereferenced and data is automatically enriched with the Europeana semantic enrichment process. One thing to consider is the structure of a Europeana record. A "***record***" can

[18] For a definition and use of URIs in the context of EDM, please see: http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/FAQs/URIs%20in%20EDM_pro.pdf

be seen as an RDF graph, i.e. a set of statements about several resources. In Europeana we can consider a record to be always a graph 'centered' around a main resource: the Europeana (cultural heritage) object. But there will be two main kinds of EDM records with a different base structure:

1. the ones sent to Europeana by providers, where the "Europeana object" is represented by a ProvidedCHO and all descriptive metadata for the object is attached to it.
2. the ones Europeana creates internally and shares via the API, where the original descriptive metadata and new data created by Europeana for the "Europeana object" are attached to "proxies" not the ProvidedCHO directly.

So far, the implementation of the multilingual score is working with API records that are represented in 'EDM internal' schema. In practice, providers have to submit data according to the 'EDM external' schema. Figure 3 shows the different possibilities for measuring the multilinguality of statements based on the data used for the assessment.

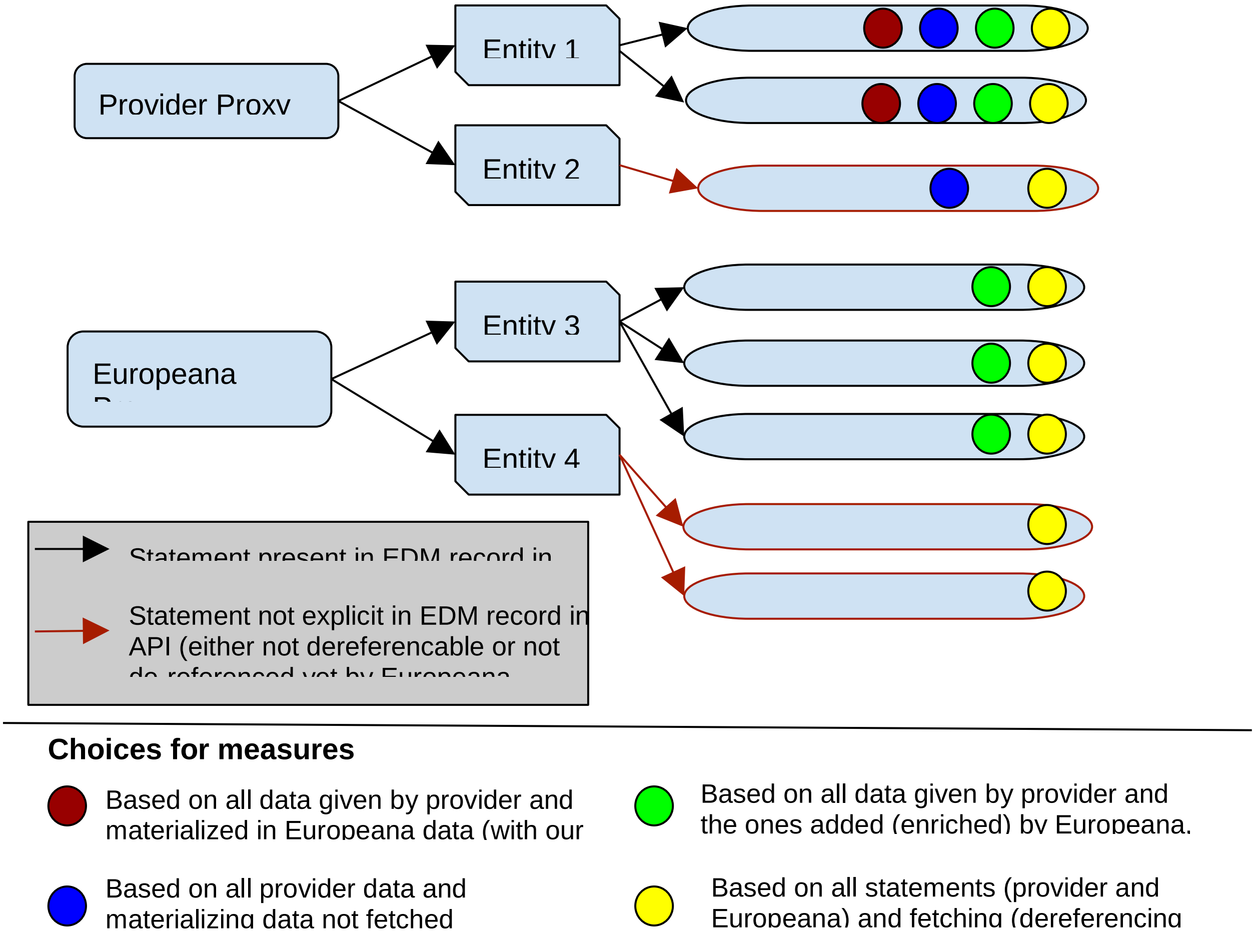

Provider Proxy
Entity 1
Entity 2
Europeana Proxy
Entity 3
Entity 4
Statement present in EDM record in
Statement not explicit in EDM record in API (either not dereferencable or not de-referenced yet by Europeana
Choices for measures
Based on all data given by provider and materialized in Europeana data (with our
Based on all data given by provider and the ones added (enriched) by Europeana.
Based on all provider data and materializing data not fetched
Based on all statements (provider and Europeana) and fetching (dereferencing

Figure 3: Different levels for inclusion of data as basis for multilingual measures.

To make this clearer, table 8 gives examples of different options for measuring the multilinguality. In theory this different levels exist in the Europeana data. Practically, the main distinction should be made between taking all proxies into account or not (1 vs. 3 below). The numbers in the table relate to the following categories:

① data provider's proxy and enrichments with resources from Europeana's dereferencing list or for which providers already submitted descriptions in their EDM records

② data provider's proxy and enrichments considering all data that could be fetched for them

③ all proxies and standard enrichments from Europeana

④ all proxies and enrichments considering all data that could be fetched for them

Source	**Entity + Label**	**Link to vocabulary**	**Example value**	①	②	③	④
ex:providerProxy	dc:subject "aSubject"@en	-	aSubject"@en	①	②	③	④
	dc:creator <http://vocab.getty.edu/aPersonNumber>	on Europeana's dereferencing list	skos:prefLabel "thePersonName"@en	①	②	③	④
	dc:type <http://udcdata.info/rdf/065280>	Not on Europeana's dereferencing list (only for English terms)	(in theory the record could include skos:prefLabel "Painting"@en)		②		④
ex:europeanaProxy	dc:subject <http://dbpedia.org/aSubjectID>	Enriched by Europeana	skos:prefLabel "aSubject"@en ; "unSujet"@fr .			③	④

Table 9: Options for measuring multilinguality in metadata according to source.

All resources that are not connected via statements of the ProvidedCHO or proxies are skipped in the measuring process.

6.4 Preliminary results

For multilingual **completeness**, 904 (out of 3548) datasets have no value in the *dc:language* field or the field is non-existent. Looking at the record level, for 58,03% of the records a *dc:language* field exists (figure 4).[19] Delving deeper into the numbers offered by the Data Quality Assurance Framework, one can detect misuse of fields. For example, the metric "cardinality" supports the identification of 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.

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

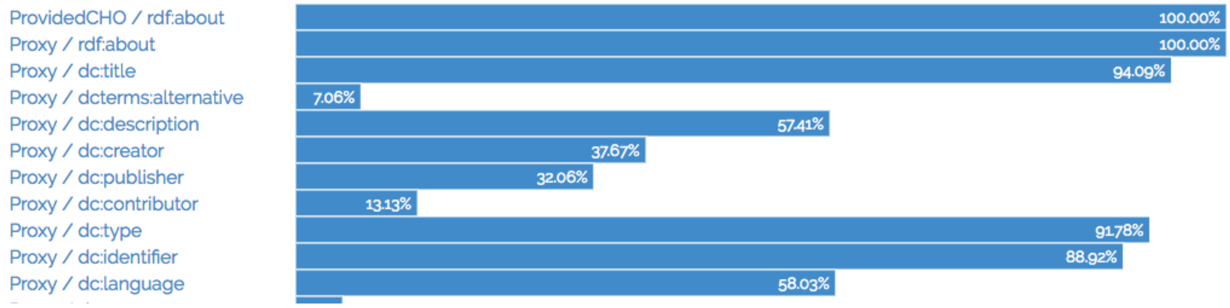
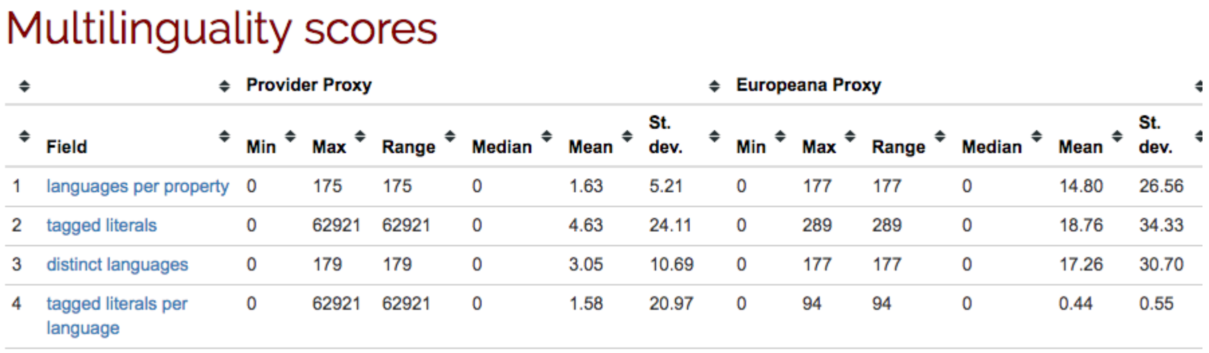


Figure 4: In 58% of the records the dc:language exists.

For Accessibility, preliminary results were obtained which still need to be tested and adapted if necessary. Figure 5 shows the resulting table in the Data Quality Assurance Framework where scores for the languages per property, tagged literals, distinct languages and tagged literals per language are given. To enable thorough analysis of the data the tool offers various graphics and diagrams to identify patterns and outliers.



Multilinguality scores

		Provider Proxy						Europeana Proxy					
	Field	Min	Max	Range	Median	Mean	St. dev.	Min	Max	Range	Median	Mean	St. dev.
1	languages per property	0	175	175	0	1.63	5.21	0	177	177	0	14.80	26.56
2	tagged literals	0	62921	62921	0	4.63	24.11	0	289	289	0	18.76	34.33
3	distinct languages	0	179	179	0	3.05	10.69	0	177	177	0	17.26	30.70
4	tagged literals per language	0	62921	62921	0	1.58	20.97	0	94	94	0	0.44	0.55

Figure 5: Screenshot of preliminary results of implemented accessibility scores.[20]

[20] http://144.76.218.178/europeana-qa/multilinguality.php?id=all, please note that the scores as presented in the screenshot are likely to change in future as we are still in the process of testing and adapting the process to quantify language tags correctly.

7. Improvement of Multilinguality in Europeana

This section highlights some of the improvements in Europeana which were executed during DSI-2. The normalization of values in the *dc:language* field is a direct reaction to the high number of diverse language encodings found in this particular field.
Section 7.2 presents the progression of the coverage of the Entity Collection w.

7.1 Normalizing values in the dc:language fields[21]

One of the outcomes of quantifying multilingual information in Europeana is the heterogeneity of the values in the *dc:language* field. To enable a facet that allows to filter results by the language of the object, it is crucial to normalize values in this field. Predominantly, values are normalized in ISO-639-1 or ISO-639-3, but, they also occur in natural language sentences that cannot be processed automatically. In between, we also found the use of language ISO codes (but without explicitly indicating which ISO standard is in use). Quite frequent are also references to languages by their name. Another relevant characteristic of the data is that, in a single data element, references to several languages may be present.

To homogenize the data, a set of rules for languages normalisation in the Europeana dataset was developed. Given the characteristics of the data, the language normalization operation is actually a mix of operations, comprising cleaning, normalization and enrichment of data. The following exemplify some cases of the different types of operations:

Input value	**Language normalization output (in ISO 639-1)**
“English”	“en”
“eng”	“en”
en-GB	“en”
“English and Latin”	“en” “la”
Greek; Latin	“el” “la”

Table 10: Operations conducted to normalize different values in dc:language field.

The following 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

[21] Section taken from Charles et al., 2017.

Total values already normalized (ISO-639-3, three letter codes)	4,831,534

Table 11: Presence of ISO-639 codes in the values of dc:language field.

A thorough description of the operation conducted to normalize this field can be found in Appendix C. The line of work has not been yet included in Europeana's production environment for data management. It is nonetheless included on the roadmap for the V1of Europeana's future ingestion system (Metis)[22].

7.2 Multilingual Entities

The introduction of the Entity Collection leads to an increase of multilingual information in Europeana’s metadata. Figures 6-8 show the amount of Europeana entities, namely places (figure 6), agents (figure 7) and timespan (figure 8) in any given language (blue bar). The more entities are available in a language the more likely it is that they can be suggested to users. The green bar in each figure shows how many of these entities are actually used to enrich Europeana objects making it possible that these enriched objects are retrievable with different language variances of the place names.

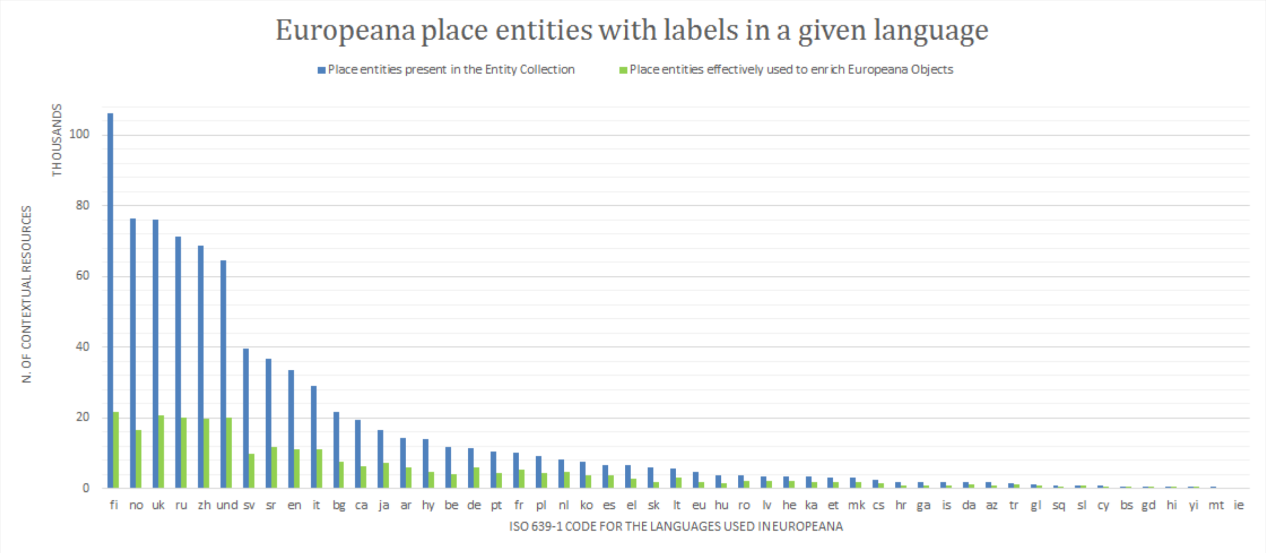


Figure 6: Number of place entities in the Entity Collection (blue bar) and the number of place entities used to enrich Europeana objects (green bar).

[22] http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Milestones/ms1.1-ingestion-workflows-business-requirements-update.pdf

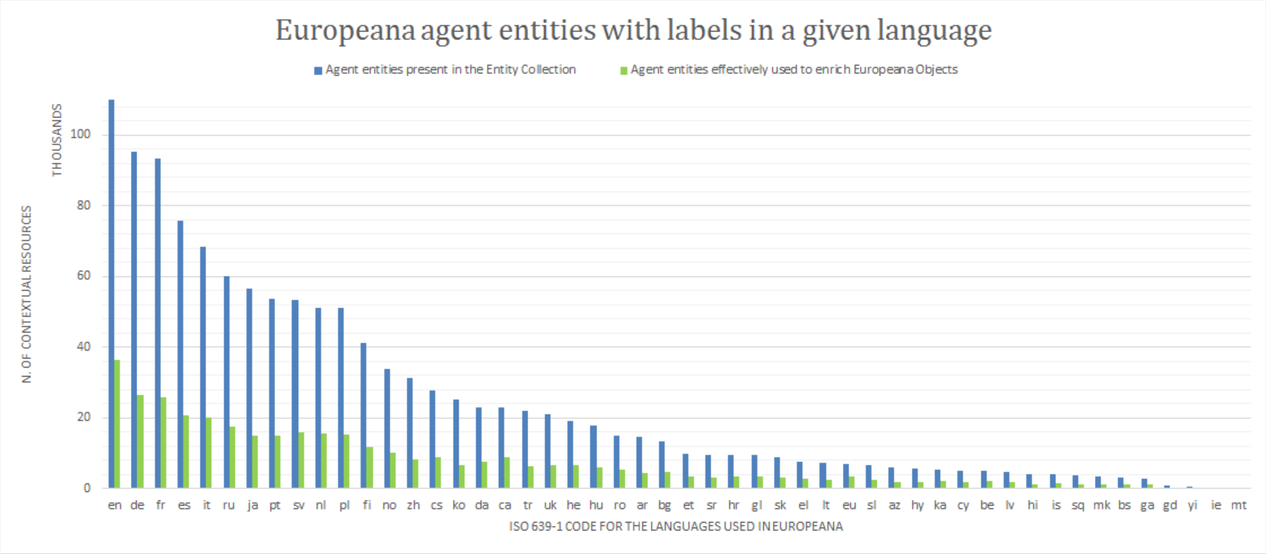


Figure 7: Number of agent entities in the Entity Collection (blue bar) and the number of entities used to enrich Europeana objects.

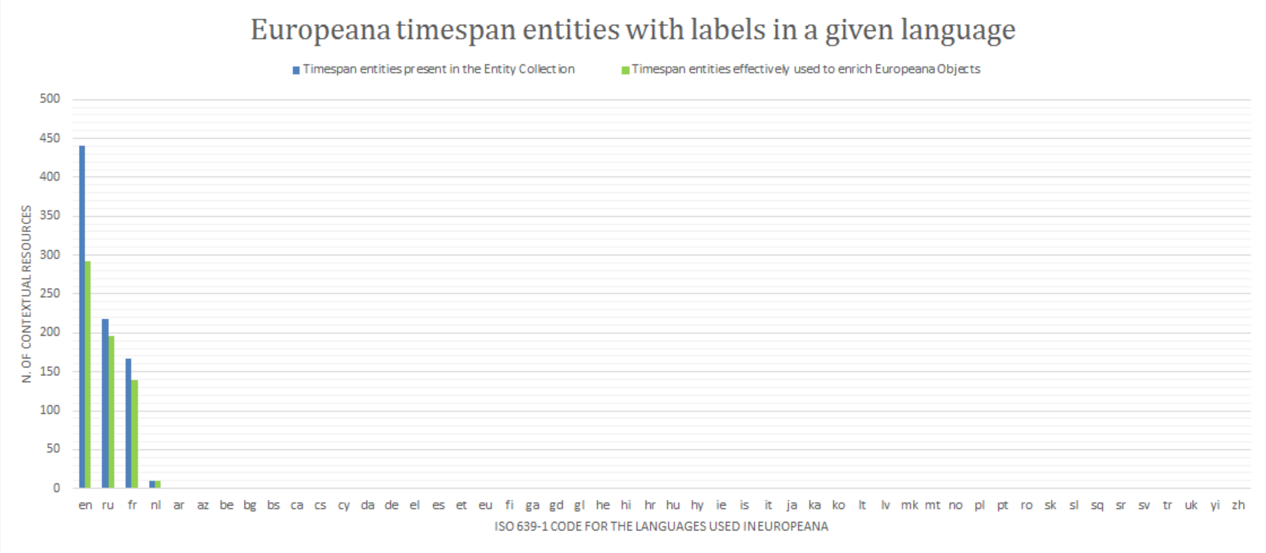


Figure 8: Number of timespan entities in the Entity Collection (blue bar) and the number of entities used to enrich Europeana objects.

The fourth entity, the Entity Collection is covering are concepts. Figure 9 shows the progression of the coverage of entities before and after the update. The update involved the addition of two new thematic vocabularies: 1) a music genres, forms and composition vocabulary obtained from Wikidata; and 2) the Europeana Photography Multilingual Vocabulary[23] developed by the Photoconsortium[24]. Besides the content update, the Semantic Enrichment Framework was also changed for concepts to enrich *dc:format* and *dcterms:medium* besides the already enriched fields of *dc:type* and *dc:subject*. The impact of both changes can be seen on the 4th graph of Figure 9.

The first and the third graph show the number of entities in the Entity Collection, the second and fourth graph show the number of entities used to enrich Europeana objects.

[23] http://www.digitalmeetsculture.net/article/europeanaphotography-multilingual-vocabulary-released-and-disseminated/

[24] http://photoconsortium.net/

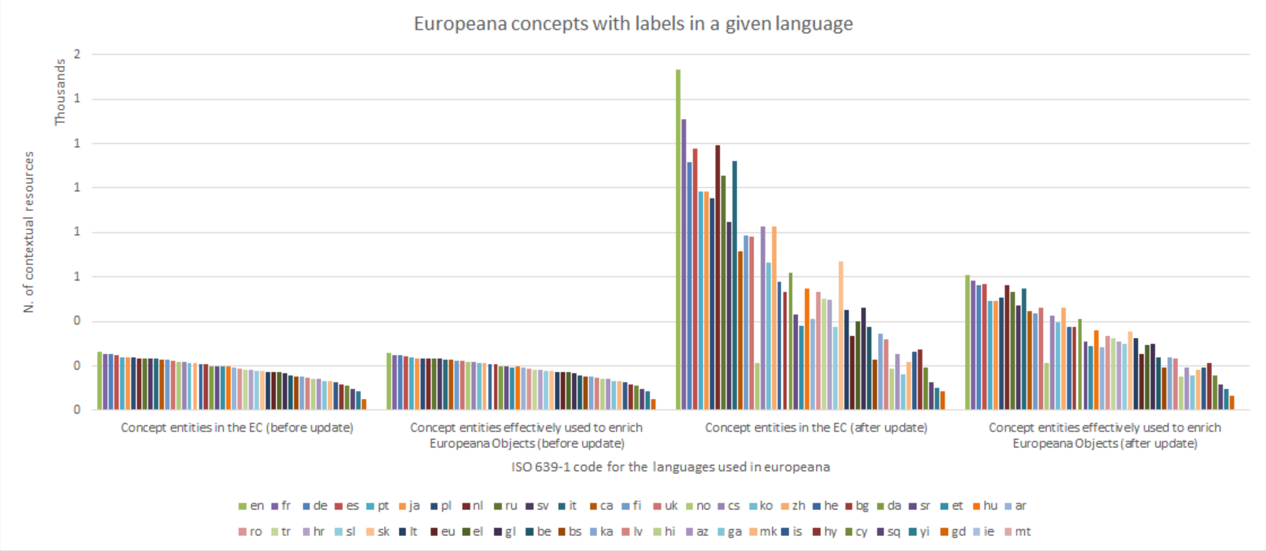


Figure 9: Progression of language coverage of concept entities before and after the update.

Looking at all entities in the Entity Collection, we see that places are the entity types occurring most.[25] Agents and concepts are the other two entity types that can be used to enrich Europeana objects and are beneficial for automatic query suggestion.

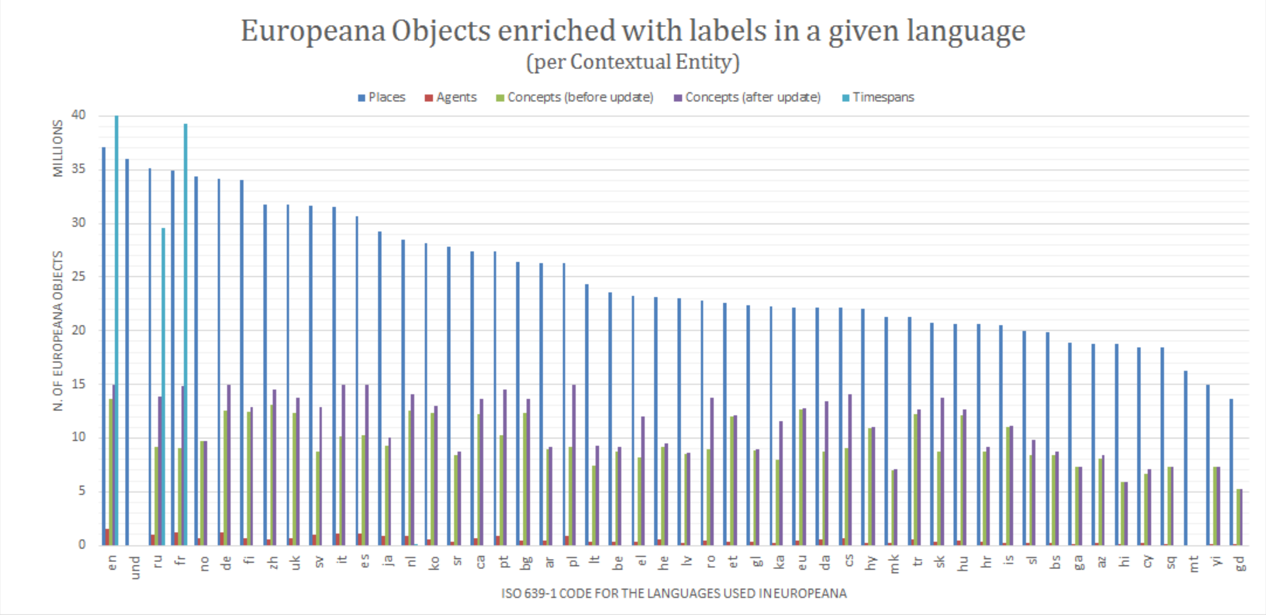


Figure 10: Number of Europeana objects enriched with contextual entities in a given language.

[25] For recent numbers on the vocabularies used and objects enriched, please refer to the document "Europeana Semantic Enrichment Framework" available here: https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcT0zpJAXUPo6x4uMBj1pEx0Y/edit#heading=h.so4ujx9oye9f

8. Conclusion & Future Work

Quantifying the multilingual degree of Europeana’s data is an ongoing endeavour. First results have shown that it is possible to count the diversity and the amount of language tags as well as the existence of fields that indicate the language of the objects. These preliminary results help Europeana to not only control the data’s quality with regard to completeness, consistency conformity and accessibility but also make more precise plans and recommendations to increase multilingual data quality. Tracking the progress over time is another benefit the developed metrics bring.

In this report, we also report on efforts that the Europeana DSI2 project has undertaken to improve multilingual metadata quality on two of the aspects we've identified to be relevant: normalization of the field indicating the language of the Cultural Heritage Object, and the automatic enrichment of metadata using multilingual sources integrated in the Europeana Collection. The former effort has not been yet deployed in Europeana's production is, but the latter has been, leading to gains for the user's multilingual experience (better matching between queries and objects, and automatic query suggestion UI features) - and for re-users of Europeana's services eager to develop their own multilingual features.

In the future, the impact of multilingual data on user interactions needs to be further assessed. One has to find out if multilingual data with good quality reflecting many different languages really fulfils its expectations: users search, browse and retrieve material independent of their preferred language. This deliverable offers some quantifiable metrics to assess language crossing within search and browsing tasks and provides means to determine the degree of understanding of material in a foreign language. But to really know how well users deal with the provided multilingual information, user-centred evaluations need to be designed.

References

Albertoni, Riccardo, Monica De Martino, and Paola Podesta. "A Linkset Quality Metric Measuring Multilingual Gain in SKOS Thesauri." *LDQ@ ESWC*. 2015. Available online: http://ceur-ws.org/Vol-1376/LDQ2015_paper_01.pdf

Bruce, T. R., & Hillmann, D. I. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In D. Hillman & E. Westbrooks (Hrsg.), *Metadata in practice* (S. 238–256). Chicago, IL: ALA Editions. Abgerufen von http://ecommons.cornell.edu/handle/1813/7895

Charles, V., Stiller, J., Király, P., Bailer, W., & Freire, N. (2017). Data Quality Assessment in Europeana: Metrics for Multilinguality. (Meta)-data quality workshop at TPDL 2017, Springer Berlin / Heidelberg. (forthcoming)

Functional requirements for Bibliographic records: final report / IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). München: K.G. Saur.

Harlow, C. (2017) Nailing Jello to a Wall: Metrics, Frameworks, & Existing Work for Metadata Assessment. DCMI/ASIS&T Joint Webinar, 27 April 2017. Available at http://dublincore.org/resources/training/ASIST_Webinar-20170427/Harlow_Webinar.pdf

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

Mader, C., Haslhofer, B., & Isaac, A. (2012). Finding quality issues in SKOS vocabularies (pp. 222–233). Presented at the TPDL - Theory and Practice of Digital Libraries.

Park, J. (2009). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly*, *47*(3–4), 213–228. https://doi.org/10.1080/01639370902737240

Petras, V., Hill, T., Stiller, J., & Gäde, M. (2017). Europeana – a Search Engine for Digitised Cultural Heritage Material. *Datenbank-Spektrum*, 1–6. https://doi.org/10.1007/s13222-016-0238-1

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is “Quality” Metadata “Shareable” Metadata? The Implications of Local Metadata Practices for Federated Collections. In *Currents and Convergence: Navigating the Rivers of Change. ACRL Twelfth National Conference.* (pp. 223–237). Minneapolis, Minnesota: Association of College & Research Libraries. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/145/shreeves05.pdf

Stiller, J., & Király, P. (2017). Multilinguality of Metadata. Measuring the Multilingual Degree of Europeana’s Metadata. In M. Gäde, V. Trkulja, & V. Petras (Hrsg.), *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017)* (S. 164–176). Glückstadt, Germany: Verlag Werner Hülsbusch. Available here: https://edoc.hu-berlin.de/bitstream/handle/18452/2109/stiller.pdf

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

Stvilia, B., & Gasser, L. (2008). Value-based metadata quality assessment. *Library & Information Science Research*, *30*(1), 67–74.

Taylor, K., & White, J. (1998). Predicting What MT Is Good for: User Judgments and Task Performance. In D. Farwell, L. Gerber, & E. Hovy (Eds.), *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas AMTA'98 Langhorne, PA, USA, October 28–31, 1998 Proceedings* (pp. 364–373). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/3-540-49478-2_33

Appendix A: EDM fields where use of language tag and/or multilingual resources is encouraged

Field	**Ideal Values**
edm:ProvidedCHO/@about	n.a.
Proxy/dc:contributor	Linked to controlled vocabulary
Proxy/dc:coverage	Linked to controlled vocabulary
Proxy/dc:creator	Linked to controlled vocabulary
Proxy/dc:description	translations with language tag
Proxy/dc:format	Linked to controlled vocabulary
Proxy/dc:rights	translations with language tag
Proxy/dc:source	translations with language tag
Proxy/dc:subject	Linked to controlled vocabulary
Proxy/dc:title	with language tag, if appropriate translation
Proxy/dc:type	Linked to controlled vocabulary
Proxy/dcterms:alternative	translations with language tag
Proxy/dcterms:created	should be handled as date
Proxy/dcterms:extent	translations with language tag
Proxy/dcterms:isReferencedBy	translations with language tag
Proxy/dcterms:issued	handed as date field

Proxy/dcterms:medium	Linked to controlled vocabulary
Proxy/dcterms:provenance	translations with language tag
Proxy/dcterms:spatial	Linked to controlled vocabulary
Proxy/dcterms:temporal	Linked to controlled vocabulary
Proxy/edm:currentLocation	Linked to controlled vocabulary
Proxy/edm:hasMet	Linked to controlled vocabulary
Proxy/edm:isRelatedTo	Linked to controlled vocabulary
ore:Aggregation	
Aggregation/edm:dataProvider	Linked to controlled vocabulary
Aggregation/edm:provider	Linked to controlled vocabulary
Aggregation/edm:intermediateProvider	Linked to controlled vocabulary
Aggregation/dc:rights	translations with language tag

Appendix B: requirements for multilingual logging in Europeana

Logging	**Comment**
Number of different languages occurring in SERP	Distinct languages from facet (populates by edm:language)
Number of document per language in SERP	Can be taken from the facet
Interface language (change)	The interface language or the interface language change from the default setting to a preferred language via a drop down menu, cookie or link to the appropriate language version.
Language and country of external referrer	Language and country version of external links (e.g. from Google).

Language facet usage	User refines result lists according to the language facet
Country facet usage	User refines results by providing country fact
Language of the user’s operating system / browser language	Information about the browsers can include the language version.
Language of an object	If there are language tags
Language of collection of clicked object	As indicated by language of provider
IP address	To infer country of origin of user
Query language	Language of search terms
Language of the full results	Language of result clicked by the user

Appendix C: Language Normalization Evaluation Report

Editor:	Nuno Freire
Contributors:	Antoine Isaac, Cécile Devarenne, Hugo Manguinhas, Karl Pineau, Timothy Hill, Valentine Charles
Revisions:	● 03/02/2017 (first draft) ● 11/08/2017 (v1.0)

Introduction

The data cleaning and normalization plugin for Metis, being developed in DSI-2, includes a specific operation for the normalization of values from data elements and attributes that refer to languages. The specific EDM parts that contain this type of data, and are potential objects of normalization, are the following:

- The Dublin Core element dc:language - a property of edm:ProvidedCHO and ore:Proxy;
- The xml:lang attributes that may be present in every element of EDM containing textual data[26].

This document presents the functionality of the language normalization operation, its underlying algorithm, and results of evaluation tasks that have been performed. Although this normalization operation can be applied to both cases, the evaluation was performed only on the dc:language values. We finish with conclusions from the evaluation, focusing on assessing feasibility and ensuring a good reliability of its output.

Functionality

In the Europeana dataset, dc:language values are currently quite heterogeneous. dc:language values are predominantly normalized in ISO-639-1 or ISO-639-3, but, in contrast, values sometimes consist in natural language sentences that cannot be processed automatically. In between, we find also the use of language ISO codes (but without explicitly indicating which ISO standard is in use). Quite frequent are also references to languages by their name. Another relevant characteristic of the data is that, in a single data element, references to several languages may be present.

Given the characteristics of the data, the language normalization operation is actually a mix of operations, comprising cleaning, normalization and enrichment of data. The following exemplify some cases of the different types of operations:

Input value	**Language normalization**

[26] While this tags are expected to be much more controlled than the values for dc:language, investigation shows that a small proportion of them are filled with free text, against the formal rules for language tags (fixing them would thus be validation/correction rather than normalization). See https://europeanadev.assembla.com/spaces/europeana-ingestion/tickets/1040, https://europeanadev.assembla.com/spaces/europeana-ingestion/tickets/1247 and https://europeanadev.assembla.com/spaces/europeana-npc/tickets/927 .

	output (in ISO 639-1)
“English”	“en”
“eng”	“en”
en-GB	“en”
“English and Latin”	“en” “la”
Greek; Latin	“el” “la”

The language normalization algorithm

The language normalization algorithm is constituted by several sub-algorithms. It applies the sub-algorithms, in a specified order, until one of the sub-algorithms is capable of outputting a normalized result. If none of the algorithms is capable of providing a normalized value, then no normalization is done.

The output of the normalization plugin is a value from a language vocabulary. It can be configured to normalize against one of several language vocabularies, as described in the following subsection.

Language vocabularies

The algorithms work based on a core vocabulary. Internally all normalization operations are performed using the core vocabulary. The core vocabulary contains alignments with several other vocabularies, allowing the final output to be given according to any of the aligned vocabularies.

The core vocabulary is the Languages Name Authority List (NAL) published in the European Union Open Data Portal - https://open-data.europa.eu/en/data/dataset/language. NAL is aligned with ISO 639-1 (the current vocabulary in use at Europeana), ISO 639-2b, , ISO 639-2t and ISO 639-3. In addition, it provides human readable labels for all languages in all european languages.

The output of the normalization plugin can be configured for URIs of the NAL vocabulary or to any of the aligned ISO sets of codes. When the normalization results in a language that is not included in the output vocabulary, the output is empty, thus no normalization is done.

The vocabulary to be applied is a parameter of the algorithm, which is defined through a parameter of the data normalization plugin for Metis.

Matching algorithms

Before applying the sub-algorithms, the normalization operation starts by tokenizing the field value. The result is a sequence of tokens which may consist of individual words or punctuation marks.

The sub-algorithms applied are described next, in the order they are applied (starting from the most reliable):

1. Matching with a code from ISO 639-1 - this algorithm checks for the current ISO 639 code list in use at Europeana, since this is the most reliable match. The algorithm detects only

when a single language code is present in the value. If extra punctuation marks are in the value, the punctuation is removed.

2. Matching with any ISO code - this algorithm has the same behaviour as the previous one, but matches against any of the ISO code lists supported in NAL.
3. Match with a human readable label - this algorithm checks if the value matches a language label existing in NAL. A match is considered only when the full value matches a label, excluding punctuation marks. When a match is found, but the labels is ambiguous within NAL, the match is not considered, since it could lead to unreliable results.
4. Match with multiple codes - as in the second algorithm, it matches against any of the ISO code lists supported in NAL, but at this stage it also allows several codes to be present in the value (for example, “eng; fra” results in a normalization to two languages). All non-punctuation tokens must match a code, otherwise the algorithm does not consider any match.
5. Match with multiple labels - as in the third algorithm, it matches against any non-ambiguous label in NAL, but at this stage it also allows several labels to be matched in the value (for example, “English, French” results in a normalization to two languages). All non-punctuation tokens must match a label, otherwise the algorithm does not consider any match.
6. Partial value matches with ISO codes or language labels - this algorithm allows only parts of the values to match on either codes or labels. Although this algorithm is implemented, it is currently not in use, since our inspection of the results has detected several false matches in its results. Some examples of incorrect matches are in the following table:

Input value	**Language normalization output (in ISO 639-1 or ISO 639-2)**
“Escrito **en** Flandes”	“en”
“cifrada, **con** cifra interlineal”	“con”
“Letra **de** varias manos”	“de”

Evaluation procedure

An evaluation of the normalization results was conducted in December 2016. The evaluation was focused on the algorithms applied in the steps 3 to 5 described in the previous section. The results of step 6 were observed but not formerly evaluated, since the observations were enough to conclude that the results of step 6 are not reliable enough for application in the *real-world* scenario of Europeana.

The evaluation was done with the normalized output vocabulary set for ISO-639-3. It included the manual inspection of 993 distinct values of dc:language data, which were subject of the normalization conditions of steps 3 to 5 of the normalization algorithm. The manual inspection of the results was conducted by five participants from the Europeana R&D and Ingestion teams, which classified the results as *correct*, *incorrect* or *unknown*.

Results

To measure the results of the normalization of language values, the normalization algorithm was applied to the complete Europeana data set, and the estimated number of correct/incorrect normalizations is calculated based on the following:

- Values that were already in normal form in the original metadata (that is, in ISO-639-1) were not considered for the calculation.
- Whenever the normalization algorithm matched a value with a code from any other ISO-639 standard (steps 1 and 2 of the matching algorithm), it was considered correct.
- Values manually classified as *unknown* were not considered for the calculation.
- Whenever the normalization algorithm processed a value included in the evaluation the result was compared with the classification done manually, and considered correct or incorrect for the calculation.

General statistics

The following 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

Results of normalization using ISO-639-1

Given the results of the manual evaluation, the results obtained from the application of the normalization algorithm to the complete Europeana dataset are would be the following:

Total values requiring normalization to ISO639-1	9,436,280
Correct normalizations	8,108,044
Incorrect normalizations	44
Not normalized	1,328,192

The final result of normalizing the complete dataset of Europeana using ISO-639-1 as the target vocabulary would be:

Total values in the Europeana dataset	33,070,941
Total values in ISO-639-1	31,803,048 (96,17%)
Total values non-normalized	1,267,893

	(3,83%)
Error rate of the normalization (approx.)	1 / 212,766

Results of normalization using ISO-639-3

In the case Europeana decides to adopt three letter codes as the target vocabulary for dc:language values, additional values can be normalized. In such case, the final result of normalizing the complete dataset of Europeana using ISO-639-3 as the target vocabulary would be:

Total values in the Europeana dataset	33,070,941
Total values in ISO-639-3	32,823,935 (99,25%)
Total values non-normalized	247,006 (0,75%)
Error rate of the normalization (approx.)	1 / 454,545

Note that the error rate is lower than with ISO-639-2 normalization because the target vocabulary is bigger and allows to detect more matches - and, it seems, with more precision.

Conclusion

A significative amount of dc:language values exist in the Europeana dataset, which can be normalized. The results of the evaluation indicate that the language normalization operation can be performed reliably by Europeana on its complete dataset and ingestion of additional data in the future.

The reliability achieved also indicates that the same operation may be reliably applied to the wrong xml:lang attributes present in other elements of EDM (see earlier footnote).

It was also observed that the dataset contains a significative amount of language values that are not supported by ISO-639-1. The evaluation also allowed to measure the results that would be obtained if ISO-639-3 is adopted.

Summarizing, with ISO-639-1 as the target vocabulary, 96.27% of the dc:language values can become normalized in the dataset. With the adoption of ISO-639-3, it would be possible to increase the percentage of normalized values to 99.25%.