Improving data quality at Europeana
New requirements and methods for better measuring metadata quality

Péter Király¹, Hugo Manguinhas², Valentine Charles², Antoine Isaac², Timothy Hill²

¹Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
²Europeana Foundation, The Netherlands
Improving data quality at Europeana. **The data workflow**

Native data sources → Data providers → Data aggregators → Semantic data sources → Europeana ingestion → Europeana services

- Dublin Core, LIDO, EAD, MARC, EDM custom, ...
- Data transformations
- Europeana Data Model (EDM)
Improving data quality at Europeana. **The problem**

there are “good” and “bad” metadata records

but we don’t have clear metrics like this:
Improving data quality at Europeana. **Non-informative values**

non informative dc:title:  
“photograph, framed”,  
“group photograph”  
“photograph”

informative dc:title:  
“Photograph of Sir Dugald Clerk”,  
“Photograph of "Puffing Billy"”
Improving data quality at Europeana. **Copy & paste cataloging**

from a template?

unknown, unknown

Description:
; Classification(s): unknown, unknown; Acquisition: given by British School of Archaeology, 1913 [E.152.1913] Description: circular slab, with incised decoration on both sides, table top?

Creator:
The Fitzwilliam Museum, Cambridge, UK

Type:
image

Format:
text/html

Subject:
unknown

Title pages

Alternative Title:
Title pages

Creator:
Neophilologica, Ed.

Date:
23 ENG/2013

Type:
periodical

Identifier:
http://www.ceeol.com/aspx/getdoc

Is part of:

more examples in Report and Recommendations from the Task Force on Metadata Quality (2015);
Improving data quality at Europeana. **Why data quality is important?**

“Fitness for purpose” (QA principle)

- no metadata → no access to data → no data usage

more explanation:
Data on the Web Best Practices
W3C Working Draft, [https://www.w3.org/TR/dwbp/](https://www.w3.org/TR/dwbp/)
Improving data quality at Europeana. **Data Quality Committee**

**Data Quality Committee**

Quality is a key priority for our whole community! The Data Quality Committee works to address key data quality issues over time.

Formally defined as a Europeana Network and EuropeanaTech Working Group, the Data Quality Committee is a standing committee that will work on the various facets of the data quality challenge over time with a particular focus on reuse and discovery of cultural heritage scenarios.

We believe it is crucial to tackle data quality issues at every level of the data exchange chain from its creation to its publication. We have therefore gathered together experts from various background (metadata experts, software developers, search and retrieval experts..) to help us capturing all the issues.
by measuring structural elements we can predict metadata record quality

≃ metadata smell
Improving data quality at Europeana. **Purposes**

- improve the metadata
- services: good data → reliable functions
- better metadata schema & documentation
- propagate “good practice”
Improving data quality at Europeana. **What to measure?**

- Structural and semantic features
  - *Cardinality, uniqueness, length, dictionary entry, data type conformance, multilinguality* (schema-independent measurements)

- Discovery scenarios
  - *Requirements of the most important functions*

- Problem catalog
  - *Known metadata problems*
Improving data quality at Europeana. **Discovery scenarios**

- Basic retrieval with high precision and recall
- Cross-language recall
- Entity-based facets
- Date-based facets
- Improved language facets
- Browse by subjects and resource types
- Browse by agents
- Hierarchical search and facets
- ...

the most important functions
Improving data quality at Europeana. **Metadata requirements**

As a user I want to be able to filter by whether a person is the subject of a book, or its author, engraver, printer etc.

**Metadata analysis**
In each case the underlying requirement is that the relevant EDM fields for objects be populated with URIs rather than free text. These URIs need to be related, at a minimum, to a label for each of the supported languages.

**Measurement rules**
- the relevant field values should be resolvable URI
- each URI should be associated with labels in multiple languages
Improving data quality at Europeana. **Problem catalog**

- Title contents same as description contents
- Systematic use of the same title
- Bad string: “empty” (and variants)
- Shelfmarks and other identifiers in fields
- Creator not an agent name
- Absurd geographical location
- Subject field used as description field
- Unicode U+FFFD (_showing_)
- Very short description field
- ...

"metadata anti-patterns"
Improving data quality at Europeana. **Problem definition**

**Description**

- **contents**
- **Example**

/2023702/35D943DF60D779EC9EF31F5DF...

**Motivation**

- Distorts search weightings

**Checking Method**

- Field comparison

**Notes**

- Record display: creator

**Metadata Scenario**

- Basic Retrieval
Improving data quality at Europeana. **Measurement**

- **Completeness** – 40 measurements
- **Field cardinality** – 127 measurements
- **Uniqueness** – 6 measurements
- **Multilinguality** – 300+ measurements
- **Language specification** – 127 measurements
- **Problem catalog** – 3 measurements
- etc.
Improving data quality at Europeana. **Field frequency per collections**

This chart shows the frequency of the analyzed fields in all records. 100% means that the field is available in every record, 0 means that this field is never available. Numbers are rounded to 2 decimals.

**Select field:** Proxy / dcterms:alternative  
**Grouped by:** Data providers

- National Library of France
- Österreichische Nationalbibliothek - Austrian National Library
- National Library of the Netherlands
- Biblioteca Virtual de Prensa Histórica
- Bayerische Staatsbibliothek
- Deutsche Fotothek
- Bildarchiv Foto Marburg
- Riksantikvarieämbetet | Swedish National Heritage Board
- SAN - Sistema Archivistico Nazionale
- National Library of Spain
- Nationaal Archief
- National Library of the Netherlands - Koninklijke Bibliotheek
- National Library of Finland
- Rijksdienst voor het Cultureel Erfgoed
- Ministère de la culture et de la communication, Médiathèque de l'architecture et du patrimoine
- National Library of Denmark
- Museum national d'histoire naturelle
- Progetto ArtPast- Culturalitalia
- Ajuntament de Girona
- Royal Botanic Gardens, Kew
- Universitätsbibliothek Zürich

**Filters**

- 0%
- between 0% and 100%
- 100%

**No record has alternative title**
- National Library of France: 0.00%
- Österreichische Nationalbibliothek: 0.03%
- National Library of the Netherlands: 19.88%
- Bayerische Staatsbibliothek: 0.00%
- Deutsche Fotothek: 0.44%

**Every record has alternative title**
- National Library of Spain: 100.00%
- Nationaal Archief: 4.03%
- National Library of the Netherlands - Koninklijke Bibliotheek: 4.85%
- National Library of Finland: 0.60%
- Rijksdienst voor het Cultureel Erfgoed: 0.02%
- Ministère de la culture et de la communication, Médiathèque de l'architecture et du patrimoine: 0.00%
- National Library of Denmark: 0.01%
- Museum national d'histoire naturelle: 0.00%
- Progetto ArtPast- Culturalitalia: 0.00%
- Ajuntament de Girona: 0.00%
- Royal Botanic Gardens, Kew: 0.00%
- Universitätsbibliothek Zürich: 0.00%
Improving data quality at Europeana. Details of field cardinality

Cardinality of Proxy/dc:subject

Basic statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>0</td>
</tr>
<tr>
<td>max</td>
<td>128</td>
</tr>
<tr>
<td>range</td>
<td>128</td>
</tr>
<tr>
<td>median</td>
<td>0</td>
</tr>
<tr>
<td>mean</td>
<td>2.788</td>
</tr>
<tr>
<td>std.dev</td>
<td>1.846</td>
</tr>
</tbody>
</table>

Median is 0, mean is close to 1

128 subjects in one record

A record with minimal score: 08501/Athena_Update_ProvidedCHO_Bildarchiv_Foto_Marburg_t2_20093960_T_001_LAC_42_589
A record with maximal score: 08501/Athena_Update_ProvidedCHO_Bildarchiv_Foto_Marburg_obj_00011704_C_407_490

link to interesting records

Histogram

<table>
<thead>
<tr>
<th>Range of Values</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 20</td>
<td>830299</td>
<td>99.77%</td>
</tr>
<tr>
<td>20 - 40</td>
<td>1587</td>
<td>0.19%</td>
</tr>
<tr>
<td>40 - 60</td>
<td>211</td>
<td>0.03%</td>
</tr>
<tr>
<td>60 - 80</td>
<td>79</td>
<td>0.01%</td>
</tr>
<tr>
<td>80 - 100</td>
<td>2</td>
<td>0.00%</td>
</tr>
<tr>
<td>100 - 120</td>
<td>3</td>
<td>0.00%</td>
</tr>
<tr>
<td>120 - 140</td>
<td>3</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
Improving data quality at Europeana. **Multilinguality**

| Proxy/dc:description | Al-Qur‘ān
| Contient : Coran ; Coran ; Coran ; Coran
| Provenant de la collection Asselin de Cherville. |
|----------------------|----------------------------------|
| Proxy/dc:type        | ‘manuscrit’@fr
| ‘manuscript’@en      |
| Proxy/dc:identifier  | http://gallica.bnf.fr/ark:/12148/btv1b85508577 |
| Proxy/dc:language    | ar |
| Proxy/dcterm:isPartOf | http://data.theeuropeanlibrary.org/Collection/a0142 (@resource) |
| Proxy/dc:source      | Bibliothèque nationale de France, Département des manuscrits, Arabe 383 |
| Proxy/dc:rights      | ‘domaine public’@fr
| ‘public domain’@en   |
Improving data quality at Europeana. **Language frequency**
Improving data quality at Europeana. **Encoding problems**

*same language, different encodings*
Improving data quality at Europeana. **Multilingual saturation**

<table>
<thead>
<tr>
<th>Levels of Multilinguality per field</th>
<th>Expressed in numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing field</td>
<td>NA</td>
</tr>
<tr>
<td>Text string without language tag</td>
<td>0</td>
</tr>
<tr>
<td>Text string with language tag</td>
<td>1</td>
</tr>
<tr>
<td>Text string with 2-3 different language tags</td>
<td>2</td>
</tr>
<tr>
<td>Text string with 4-9 different language tags</td>
<td>2.3</td>
</tr>
<tr>
<td>Text string with 10+ different language tags</td>
<td>2.6</td>
</tr>
<tr>
<td>Link to controlled vocabulary</td>
<td>3</td>
</tr>
<tr>
<td>Penalty for strings mixed with translations with no language tag</td>
<td>-0.2</td>
</tr>
</tbody>
</table>
Improving data quality at Europeana. **Multilingual saturation**
Improving data quality at Europeana. **Information content**

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Range</th>
<th>Median</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1215</td>
<td>1215</td>
<td>0.0052</td>
<td>0.8521</td>
<td>31.21</td>
</tr>
<tr>
<td>0</td>
<td>283.0541</td>
<td>283.0541</td>
<td>0.2403</td>
<td>0.4998</td>
<td>1.584</td>
</tr>
<tr>
<td>0</td>
<td>79.2039</td>
<td>79.2039</td>
<td>0.6333</td>
<td>1.9368</td>
<td>4.7845</td>
</tr>
<tr>
<td>0</td>
<td>76.036</td>
<td>76.036</td>
<td>0.2719</td>
<td>0.5632</td>
<td>1.1319</td>
</tr>
<tr>
<td>0.0046</td>
<td>69.8988</td>
<td>69.8942</td>
<td>5.3988</td>
<td>19.0286</td>
<td>20.2707</td>
</tr>
<tr>
<td>65.0993</td>
<td>0.0806</td>
<td>62.7452</td>
<td>0.1362</td>
<td>1.362</td>
<td>2.4574</td>
</tr>
</tbody>
</table>

1 means a unique term
0.0000x means a very frequent term

These are cumulative numbers

\[
\text{entropy}_{\text{cumulative}} = \text{term}_1 + \ldots + \text{term}_n
\]
Improving data quality at Europeana. **Outliers**

**Graphs**

- Bulk of records are close to zero.
- Although 25% are between 0.05 and 1.25.
Improving data quality at Europeana. **Architecture**

- **OAI-PMH client (PHP)**
- **Hadoop File System**
- **Apache Spark**
- **Apache Solr**
- **NoSQL datastore**
- **Analysis with Spark (Scala)**
- **Analysis with R**
- **Web interface (php, d3.js)**

Files and formats involved:
- JSON files
- CSV files
- Image files

Workflow types:
- Recent workflow
- Planned workflow
Improving data quality at Europeana. **Further steps**

**human analysis**
- Translate the results into documentation, recommendations
- Communication with data providers
- Human evaluation of metadata quality
- Cooperation with other projects

**technical**
- Incorporating into Europeana’s new ingestion tool
- Shape Constraint Language (SHACL) for defining patterns
- Process usage statistics
- Measuring changes of scores
- Machine learning based classification & clustering
Improving data quality at Europeana. **Links**

- site: [http://144.76.218.178/europeana-qa/](http://144.76.218.178/europeana-qa/)
- codes: [http://pkiraly.github.io/about/#source-codes](http://pkiraly.github.io/about/#source-codes)