Measuring metadata quality.  
A quick overview in the context of Europeana metadata

Peter Király
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Germany
peter.kiraly@gwdg.de

Why to measure metadata quality?
The quality of metadata record is an issue which generated many discussions in digital library context, however it is hard to define and hard to measure. Several papers proposed methods to measure structural elements such as completeness, descriptiveness, identification, browsing, or re-usability. We call the completeness the number of records defined in the metadata schema and implemented these metrics, configurable to work with different metadata schemas (such as MARC21, CIDOC-CRM), and could be used by cultural heritage institutions. Knowing the issues in the first step in fixing them. In the first phase we use a snapshot of Europeana’s metadata. It contains 46 million records from 3500 data providers 400 GB in JSON format.

Problem patterns
In every metadata collection there are known issues which occur in several records. Some examples:
1. Equal title and description fields
2. Inappropriate field values
    <Issue2> sh:property [sh:predicate dc:title; sh:hasValue "Unknown"] .
    <Issue3> dc:title: "unknown"
    <Issue4> dc:type: "EMPTY"
We can formalize these issues with Shapes Constraint Language (see http://www.w3.org/TR/shacl/) to make it machine readable.

Implementation
Both the volume of data and the complexity of the schema make us to take care of robustness and reliability aspects of the tool. We used Big Data analytic and data science tools, such as Hadoop, Apache Spark, and R. The tool is modular, integrates feature extraction, statistical analysis, and data visualization features. The users – data providers and aggregators – gets a detailed overview of their records. The tool’s Java and REST APIs let the collections be built in the analysis into the ingestion process, so they can detect in an early phase if there are problems with the records.

Metadata quality communities
Two active communities discuss recently about the topic:
1. Europeana Data Quality Commission
2. Digital Library Federation (DLF) Metadata Assessment Working Group

References & credits
Codes and data sources are freely available and reusable in the spirit of open science.
1. Source codes (Java, R, PHP, Scala, JavaScript) – GPL 3.0 licence
2. Downloadable Europeana Data Quality Commission (March 2015) – CCO licence
3. Digital Library Federation (DLF) Metadata Assessment Working Group

Call for collaboration!
I am looking for opportunities to measure other metadata collections. If you, as a researcher or (meta)data curator are open for collaborating, contact me!

Heat-map of multilingual saturation of dc:subject. Squares represent collections. The darker the square the higher the score.

The basis of scoring follows the RDF data model.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Example</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>bare language annotation</td>
<td>&quot;German&quot;</td>
<td>0</td>
</tr>
<tr>
<td>text with language annotation</td>
<td>&quot;Germany/g&quot;</td>
<td>1</td>
</tr>
<tr>
<td>linked to multilingual vocabulary</td>
<td><a href="http://www.gironames.org/g/2921041/">http://www.gironames.org/g/2921041/</a></td>
<td>3</td>
</tr>
</tbody>
</table>

Additional metrics implemented in progress: distinct languages per object, language tags per object, literals per language, number of multilingual fields (properties), number of multilingual field instances (statements), average number of languages per property with language, average number of languages per original and enriched parts.

Multi-lingual saturation

The basis of scoring follows the RDF data model.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Example</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>text w/ language annotation</td>
<td>&quot;German&quot;</td>
<td>0</td>
</tr>
<tr>
<td>text with language annotation</td>
<td>&quot;Germany/g&quot;</td>
<td>1</td>
</tr>
<tr>
<td>linked to multilingual vocabulary</td>
<td><a href="http://www.gironames.org/g/2921041/">http://www.gironames.org/g/2921041/</a></td>
<td>3</td>
</tr>
</tbody>
</table>

Additional metrics implemented in progress: distinct languages per object, language tags per object, literals per language, number of multilingual fields (properties), number of multilingual field instances (statements), average number of languages per property with language, average number of languages per original and enriched parts.

Completeness
Completeness is based on a Boolean value: whether a field defined in the metadata schema is existing in the record. Fields support one or more different functions, such as searchability, descriptiveness, identification, browsing, or re-usability. We call the completeness the number of records defined in the metadata schema and implemented these metrics, configurable to work with different metadata schemas (such as MARC21, CIDOC-CRM), and could be used by cultural heritage institutions. Knowing the issues in the first step in fixing them. In the first phase we use a snapshot of Europeana’s metadata. It contains 46 million records from 3500 data providers 400 GB in JSON format.

Implementation
Both the volume of data and the complexity of the schema make us to take care of robustness and reliability aspects of the tool. We used Big Data analytic and data science tools, such as Hadoop, Apache Spark, and R. The tool is modular, integrates feature extraction, statistical analysis, and data visualization features. The users – data providers and aggregators – gets a detailed overview of their records. The tool’s Java and REST APIs let the collections be built in the analysis into the ingestion process, so they can detect in an early phase if there are problems with the records.

Metadata quality communities
Two active communities discuss recently about the topic:
1. Europeana Data Quality Commission
2. Digital Library Federation (DLF) Metadata Assessment Working Group
3. Metadata Assessment Working Group

References & credits
Codes and data sources are freely available and reusable in the spirit of open science.
1. Source codes (Java, R, PHP, Scala, JavaScript) – GPL 3.0 licence
2. Downloadable Europeana Data Quality Commission (March 2015) – CCO licence
3. Digital Library Federation (DLF) Metadata Assessment Working Group

Call for collaboration!
I am looking for opportunities to measure other metadata collections. If you, as a researcher or (meta)data curator are open for collaborating, contact me!

Heat-map of multilingual saturation of dc:subject. Squares represent collections. The darker the square the higher the score.

The basis of scoring follows the RDF data model.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Example</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>text w/ language annotation</td>
<td>&quot;German&quot;</td>
<td>0</td>
</tr>
<tr>
<td>text with language annotation</td>
<td>&quot;Germany/g&quot;</td>
<td>1</td>
</tr>
<tr>
<td>linked to multilingual vocabulary</td>
<td><a href="http://www.gironames.org/g/2921041/">http://www.gironames.org/g/2921041/</a></td>
<td>3</td>
</tr>
</tbody>
</table>

Additional metrics implemented in progress: distinct languages per object, language tags per object, literals per language, number of multilingual fields (properties), number of multilingual field instances (statements), average number of languages per property with language, average number of languages per original and enriched parts.

Multilingual saturation

The basis of scoring follows the RDF data model.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Example</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>text w/ language annotation</td>
<td>&quot;German&quot;</td>
<td>0</td>
</tr>
<tr>
<td>text with language annotation</td>
<td>&quot;Germany/g&quot;</td>
<td>1</td>
</tr>
<tr>
<td>linked to multilingual vocabulary</td>
<td><a href="http://www.gironames.org/g/2921041/">http://www.gironames.org/g/2921041/</a></td>
<td>3</td>
</tr>
</tbody>
</table>

Additional metrics implemented in progress: distinct languages per object, language tags per object, literals per language, number of multilingual fields (properties), number of multilingual field instances (statements), average number of languages per property with language, average number of languages per original and enriched parts.

Completeness
Completeness is based on a Boolean value: whether a field defined in the metadata schema is existing in the record. Fields support one or more different functions, such as searchability, descriptiveness, identification, browsing, or re-usability. We call the completeness the number of records defined in the metadata schema and implemented these metrics, configurable to work with different metadata schemas (such as MARC21, CIDOC-CRM), and could be used by cultural heritage institutions. Knowing the issues in the first step in fixing them. In the first phase we use a snapshot of Europeana’s metadata. It contains 46 million records from 3500 data providers 400 GB in JSON format.

Implementation
Both the volume of data and the complexity of the schema make us to take care of robustness and reliability aspects of the tool. We used Big Data analytic and data science tools, such as Hadoop, Apache Spark, and R. The tool is modular, integrates feature extraction, statistical analysis, and data visualization features. The users – data providers and aggregators – gets a detailed overview of their records. The tool’s Java and REST APIs let the collections be built in the analysis into the ingestion process, so they can detect in an early phase if there are problems with the records.

Metadata quality communities
Two active communities discuss recently about the topic:
1. Europeana Data Quality Commission
2. Digital Library Federation (DLF) Metadata Assessment Working Group
3. Metadata Assessment Working Group

References & credits
Codes and data sources are freely available and reusable in the spirit of open science.
1. Source codes (Java, R, PHP, Scala, JavaScript) – GPL 3.0 licence
2. Downloadable Europeana Data Quality Commission (March 2015) – CCO licence
3. Digital Library Federation (DLF) Metadata Assessment Working Group

Call for collaboration!
I am looking for opportunities to measure other metadata collections. If you, as a researcher or (meta)data curator are open for collaborating, contact me!

Heat-map of multilingual saturation of dc:subject. Squares represent collections. The darker the square the higher the score.

The basis of scoring follows the RDF data model.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Example</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>text w/ language annotation</td>
<td>&quot;German&quot;</td>
<td>0</td>
</tr>
<tr>
<td>text with language annotation</td>
<td>&quot;Germany/g&quot;</td>
<td>1</td>
</tr>
<tr>
<td>linked to multilingual vocabulary</td>
<td><a href="http://www.gironames.org/g/2921041/">http://www.gironames.org/g/2921041/</a></td>
<td>3</td>
</tr>
</tbody>
</table>

Additional metrics implemented in progress: distinct languages per object, language tags per object, literals per language, number of multilingual fields (properties), number of multilingual field instances (statements), average number of languages per property with language, average number of languages per original and enriched parts.