# Improving data quality at Europeana

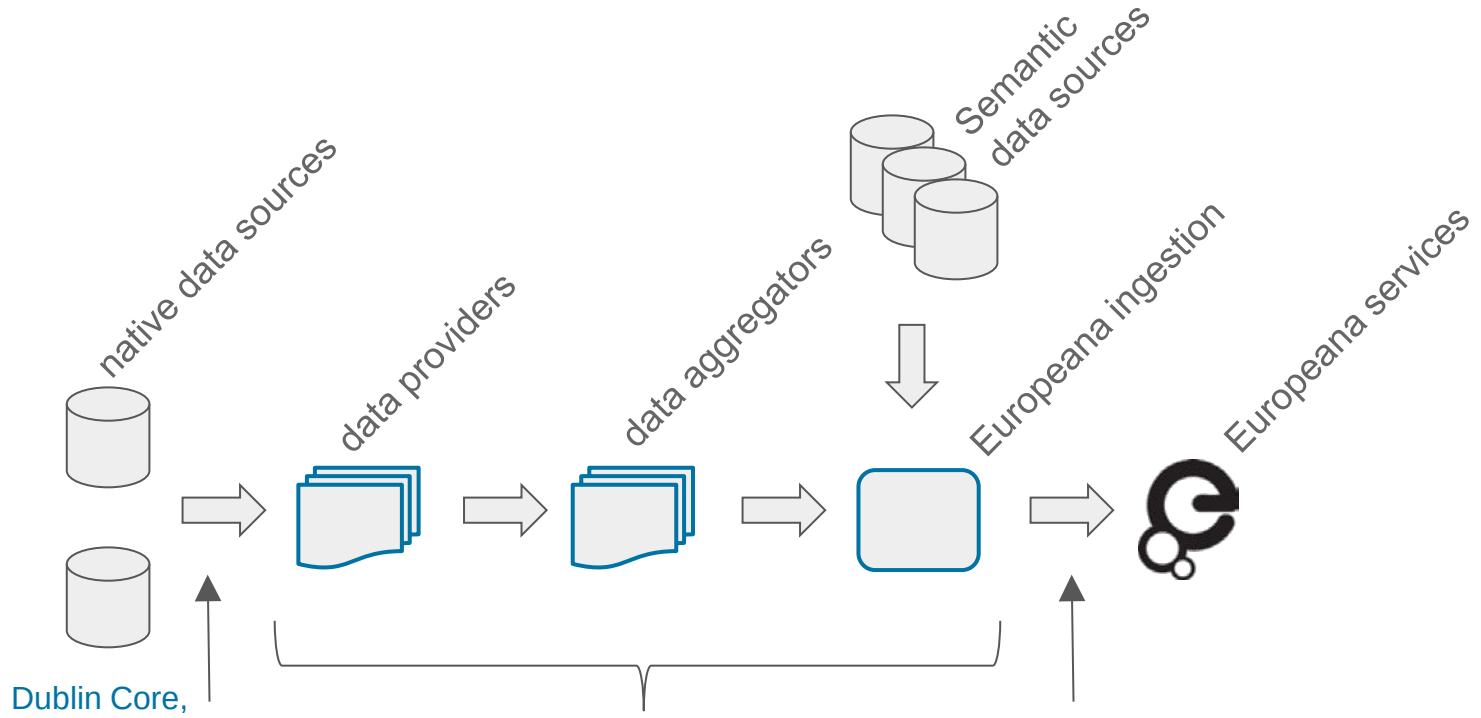
## New requirements and methods for better measuring metadata quality

Péter Király[1], Hugo Manguinhas[2], Valentine Charles[2], Antoine Isaac[2], Timothy Hill[2]

[1]Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

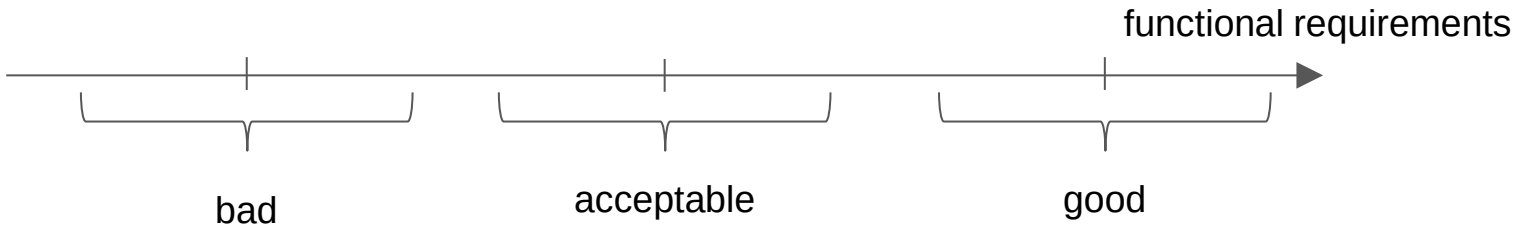
[2]Europeana Foundation, The Netherlands

# Improving data quality at Europeana. **The data workflow**



there are “good” and “bad” metadata records

but we don’t have clear metrics like this:

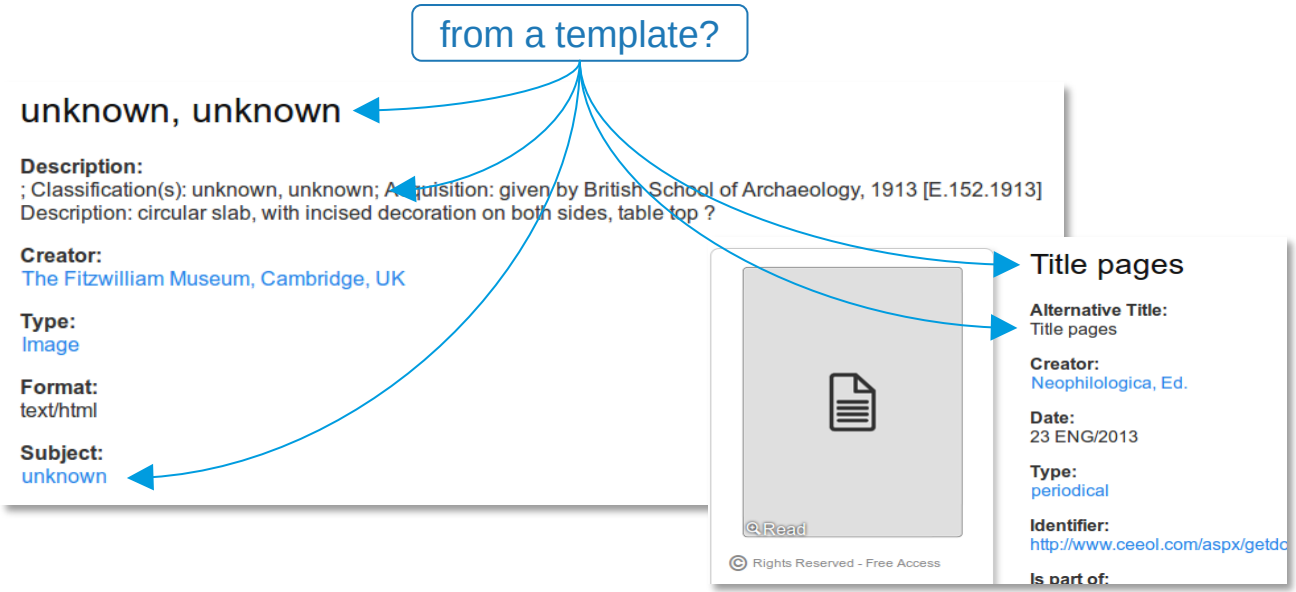


# Improving data quality at Europeana. **Non-informative values**

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**

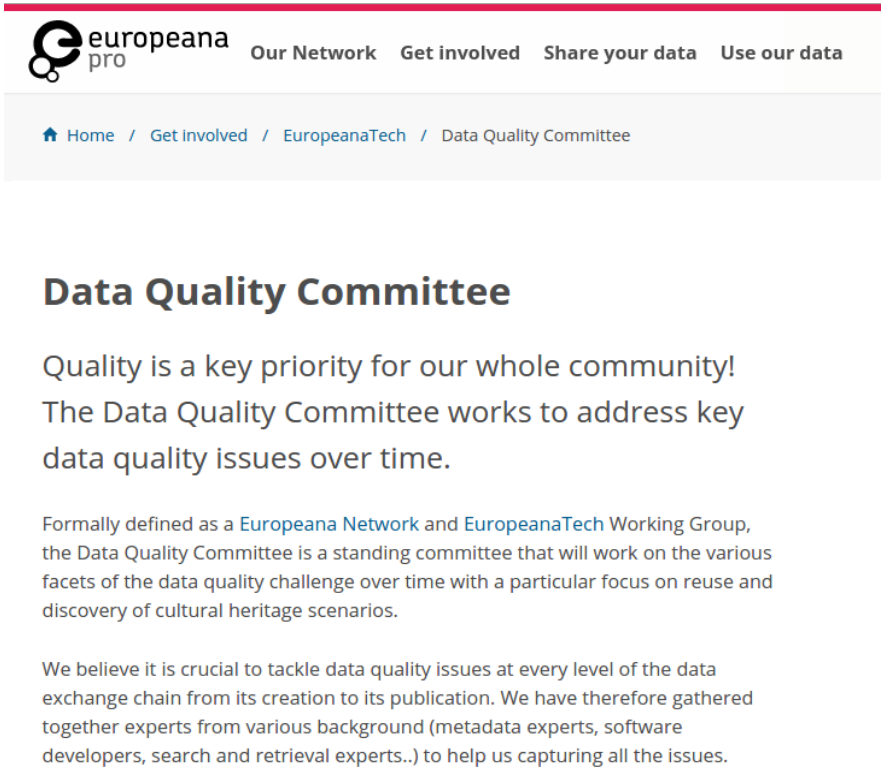



more examples in Report and Recommendations from the Task Force on Metadata Quality (2015)

# “Fitness for purpose” (QA principle)


more explanation:
Data on the Web Best Practices
W3C Working Draft, https://www.w3.org/TR/dwbp/

# Improving data quality at Europeana. **Data Quality Committee**



europeana pro

Our Network Get involved Share your data Use our data

Home / Get involved / EuropeanaTech / 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

$\simeq$ 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**

the most important functions

- 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
- ...

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**

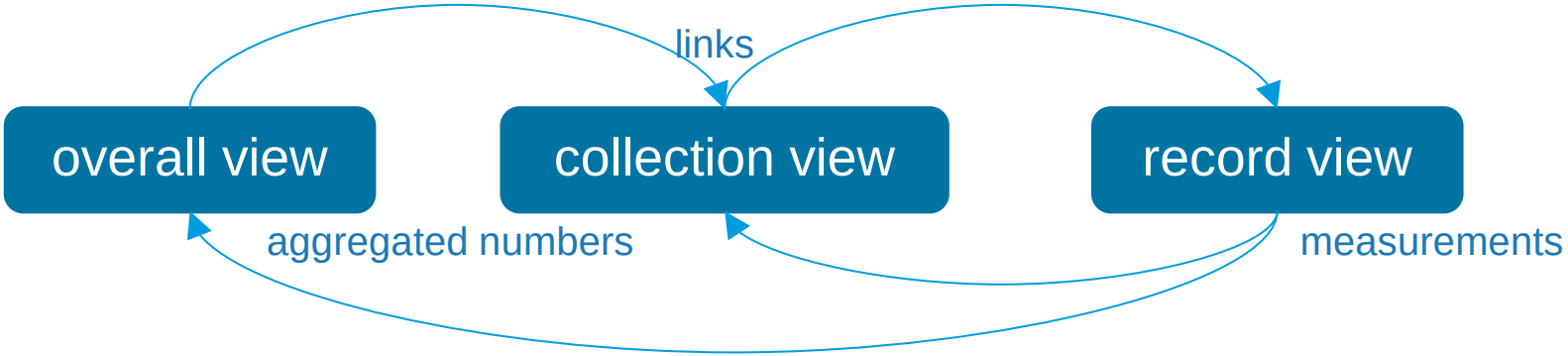
"metadata anti-patterns"

- 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 ()
- Very short description field
- ...

# Improving data quality at Europeana. **Problem definition**

| | |
|---|---|
| Description | Title contents same as description contents |
| Example | /2023702/35D943DF60D779EC9EF31F5DF... |
| Motivation | Distorts search weightings |
| Checking Method | Field comparison |
| Notes | Record display: creator concatenated onto title |
| Metadata Scenario | Basic Retrieval |

# Improving data quality at Europeana. **Measurement**




Completeness – 40 measurements

Field cardinality – 127 measurements

Uniqueness – 6 measurements

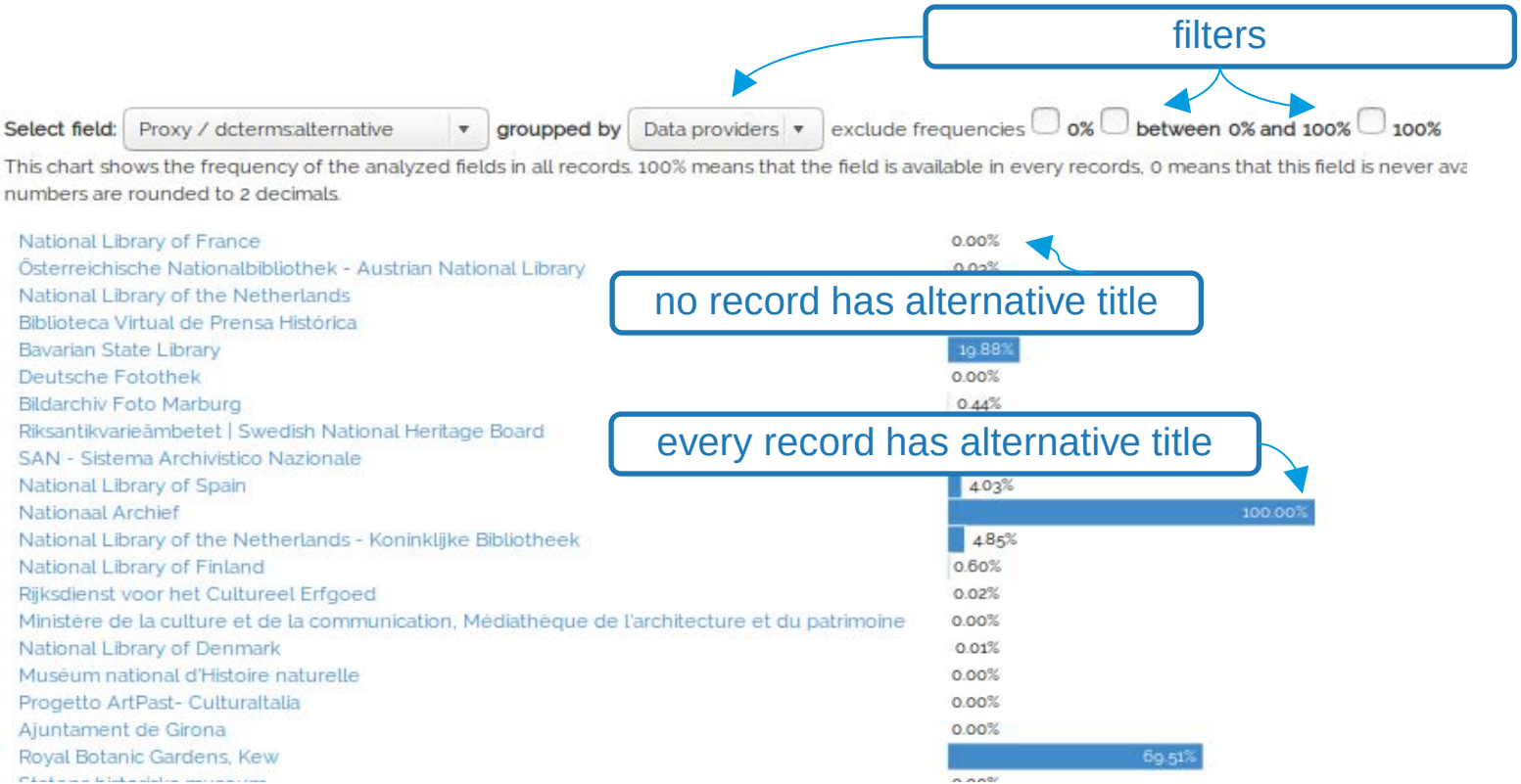
Multilinguality – 300+ measurements

Language specification – 127 measurements

Problem catalog – 3 measurements

etc.

# Improving data quality at Europeana. **Field frequency per collections**



# Improving data quality at Europeana. **Details of field cardinality**

## Cardinality of Proxy/dc:subject

### Basic statistics

| | |
|---|---|
| min | 0 |
| max | 128 |
| range | 128 |
| median | 0 |
| mean | 1.2789 |
| SE.mean | 0.0031 |
| CI.mean.0.95 | 0.006 |
| var | 7.8056 |
| std.dev | 2.7939 |
| coef.var | 2.1846 |

128 subjects in one record

median is 0, mean is close to 1

link to interesting records

A record with minimal score 08501/Athena_Update_ProvidedCHO_Bildarchiv_Foto_Marburg_t2_20093960_T_001_LAC_42_589

A record with maximal score08501/Athena_Update_ProvidedCHO_Bildarchiv_Foto_Marburg_obj_00011704_C_407_490

### Histogram

| range of values | 0 - 20 | 20 - 40 | 40 - 60 | 60 - 80 | 80 - 100 | 100 - 120 | 120 - 140 |
|---|---|---|---|---|---|---|---|
| count | 830299 | 1587 | 211 | 79 | 2 | 3 | 3 |
| percentage | 99.77% | 0.19% | 0.03% | 0.01% | 0.00% | 0.00% | 0.00% |

# Improving data quality at Europeana. **Multilinguality**

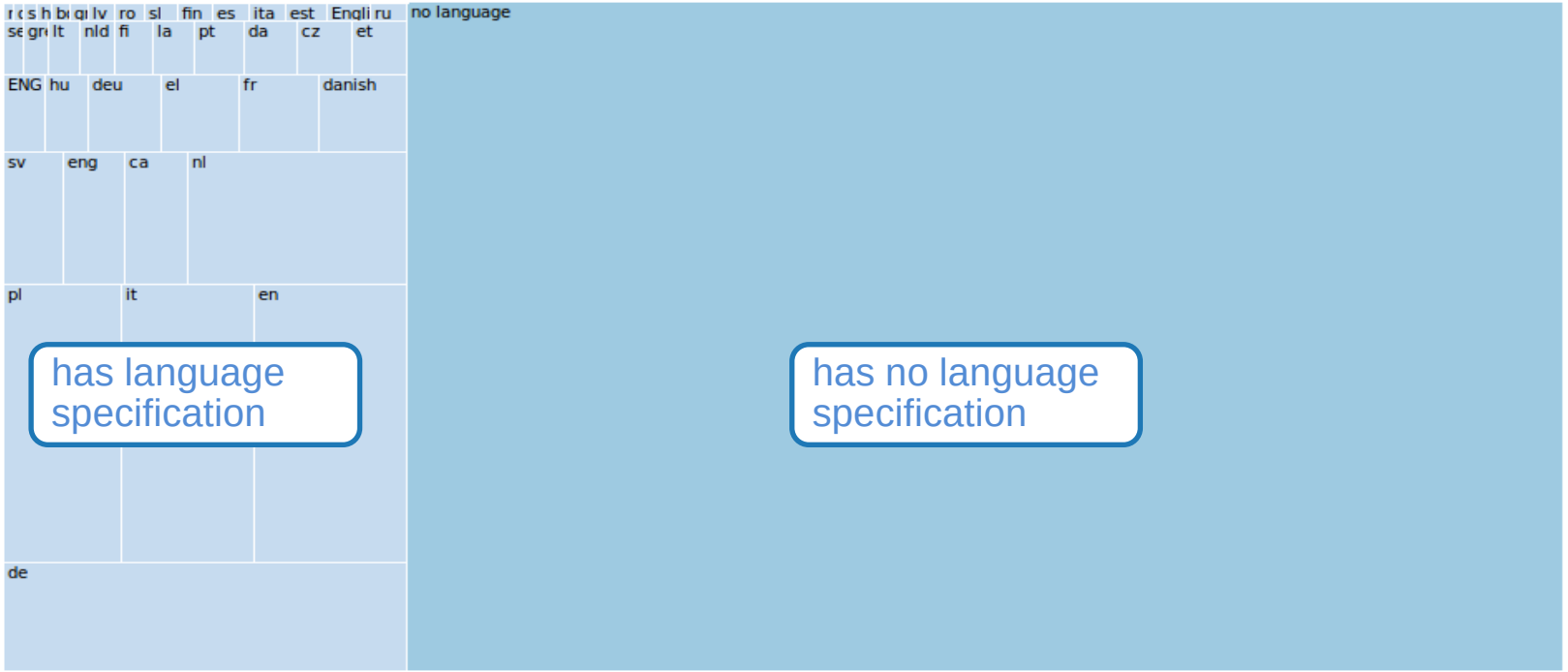
| | |
|---|---|
| Proxy/dc:description | ▪ Al-Qur'ān<br>▪ Contient : Coran ; Coran ; Coran ; Coran<br>▪ Provient de la collection Asselin de Cherville. |
| Proxy/dc:type | ▪ "manuscrit"@fr<br>▪ "manuscript"@en |
| Proxy/dc:identifier | http://gallica.bnf.fr/ark:/12148/btv1b85508577 |
| Proxy/dc:language | ar |
| Proxy/dcterms: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<br>▪ "public domain"@en |

no language specification

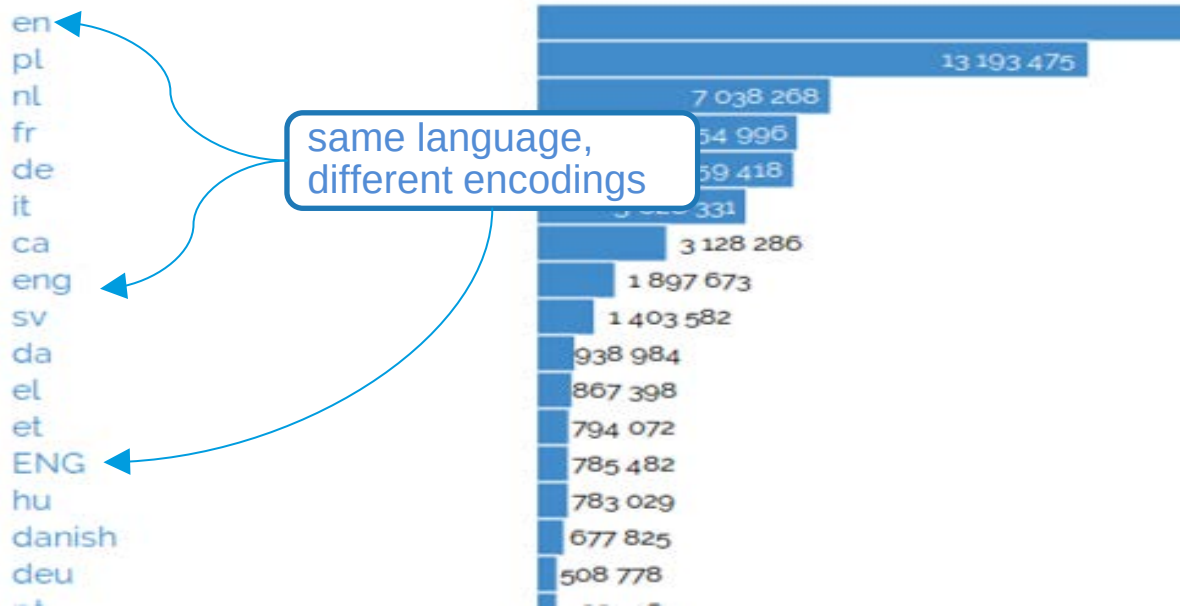
@ = language notation in RDF

@resource is a URI

# Improving data quality at Europeana. **Language frequency**



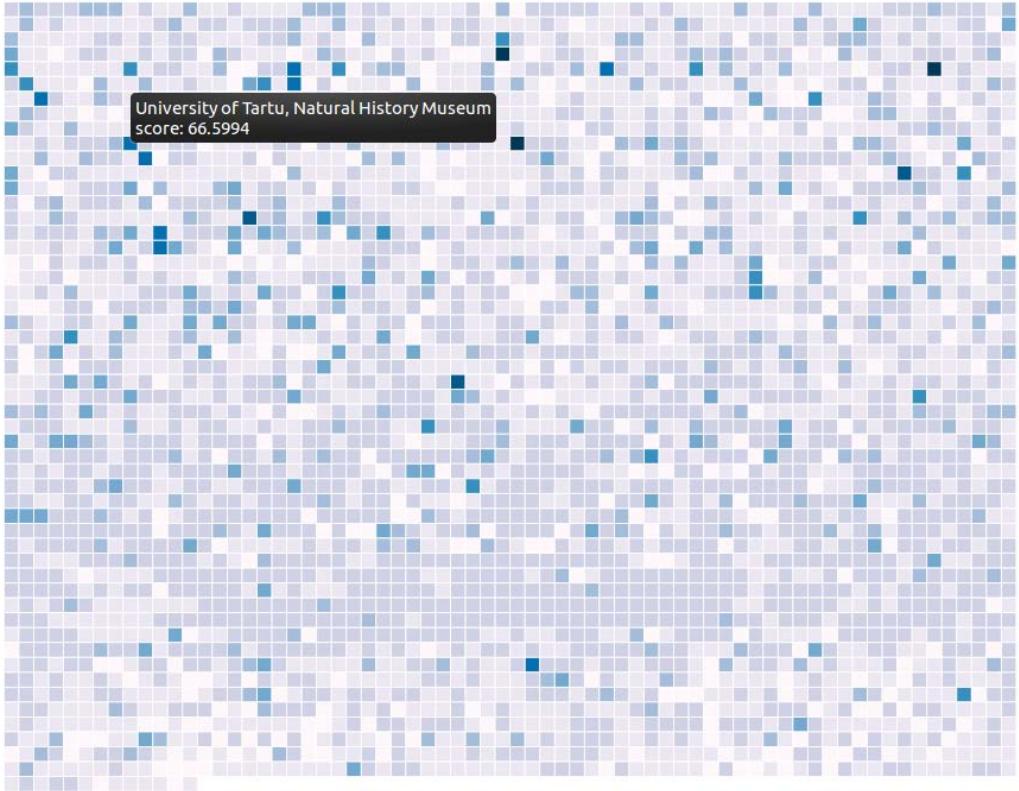
# Improving data quality at Europeana. **Encoding problems**



# Improving data quality at Europeana. **Multilingual saturation**

| Levels of Multilinguality per field | Expressed in numbers |
|---|---|
| Missing field | NA |
| Text string without language tag | 0 |
| Text string with language tag | 1 |
| Text string with 2-3 different language tags | 2 |
| Text string with 4-9 different language tags | 2.3 |
| Text string with 10+ different language tags | 2.6 |
| Link to controlled vocabulary | 3 |
| Penalty for strings mixed with translations with no language tag | -0.2 |

# Improving data quality at Europeana. **Multilingual saturation**



# Improving data quality at Europeana. **Information content**

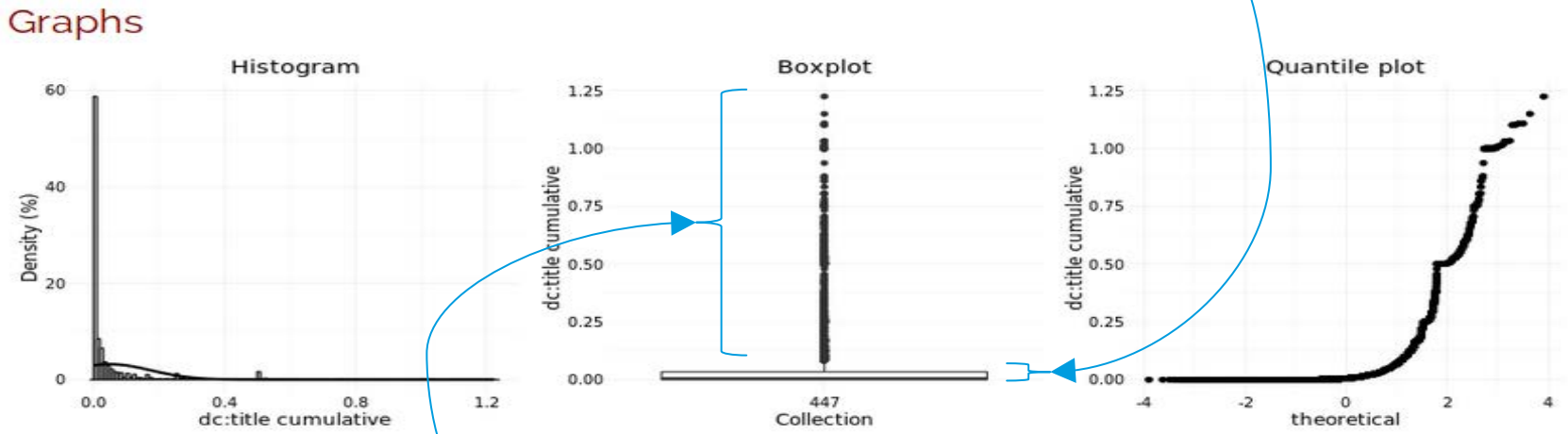
| | Minimum | Maximum | Range | Median | Mean | Standard deviation |
|---|---|---|---|---|---|---|
| | 0 | 1215 | 1215 | 0.0052 | 0.8521 | 31.21 |
| | 0 | 283.0541 | 283.0541 | 0.2403 | 0.4998 | 1.584 |
| | 0 | 79.2039 | 79.2039 | 0.6333 | 1.9368 | 4.7845 |
| | 0 | 76.036 | 76.036 | 0.2719 | 0.5632 | 1.1319 |
| | 0.0046 | 69.8988 | 69.8942 | 5.3988 | 19.0286 | 20.2707 |
| | | | | 65.0993 | 0.0806 | |
| | | | | 62.7452 | 0.1362 | |

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

These are cumulative numbers
$entropy_{cumulative} = term_1 + ... + term_n$

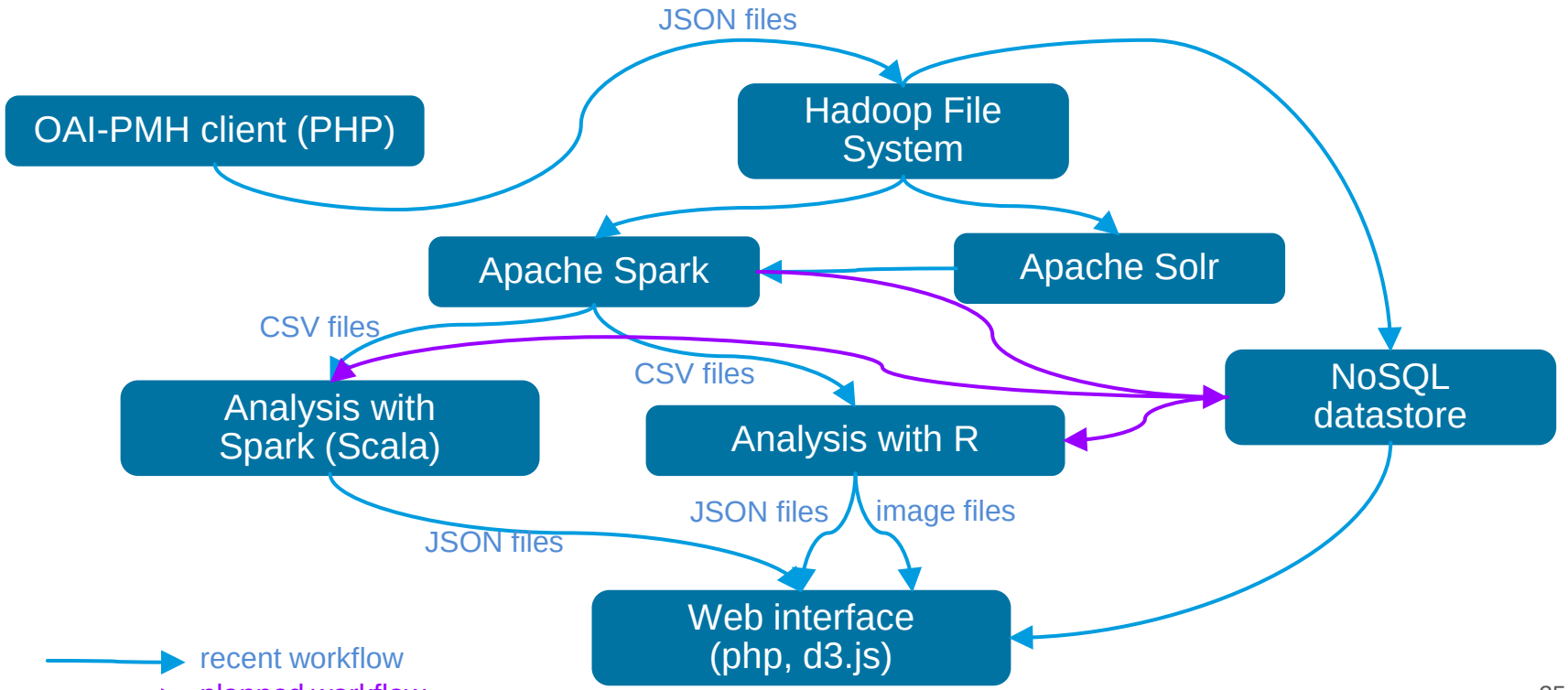
# Improving data quality at Europeana. **Outliers**

bulk of records are close to zero




although 25% are between 0.05 and 1.25

# Improving data quality at Europeana. **Architecture**



# 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

- Europeana Data Quality Committee: http://pro.europeana.eu/europeana-tech/data-quality-committee
- site: http://144.76.218.178/europeana-qa/
- codes: http://pkiraly.github.io/about/#source-codes