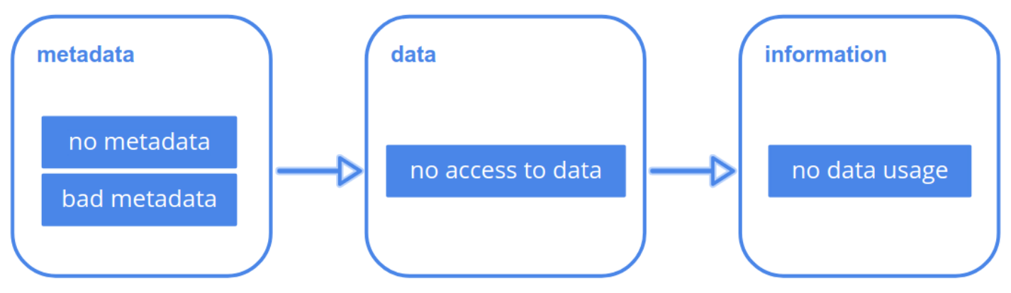
Measuring metadata quality.

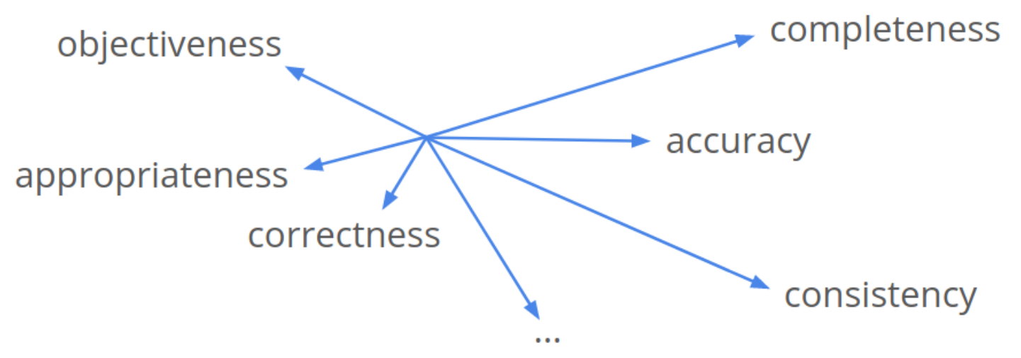
Péter Király
Göttingen eResearch Alliance and GWDG (Göttingen, Germany)
peter.kiraly@gwdg.de

Why to measure metadata quality?

The quality of metadata record is an issue which generated many discussions in different contexts, however it is hard to define and hard to measure. Several papers proposed methods to measure structural elements which could be take as approximations to data quality. In a research project we aimed to create an **open source framework** which refines and implements these metrics, configurable to work with different metadata schemas. **Knowing the issues is the first step in fixing them.** Since quality assessment depends on the purpose of data, process should start with **functional requirement analysis**.


metadata
no metadata
bad metadata
data
no access to data
information
no data usage

Consequence of metadata issues


objectiveness
completeness
accuracy
appropriateness
correctness
consistency
...

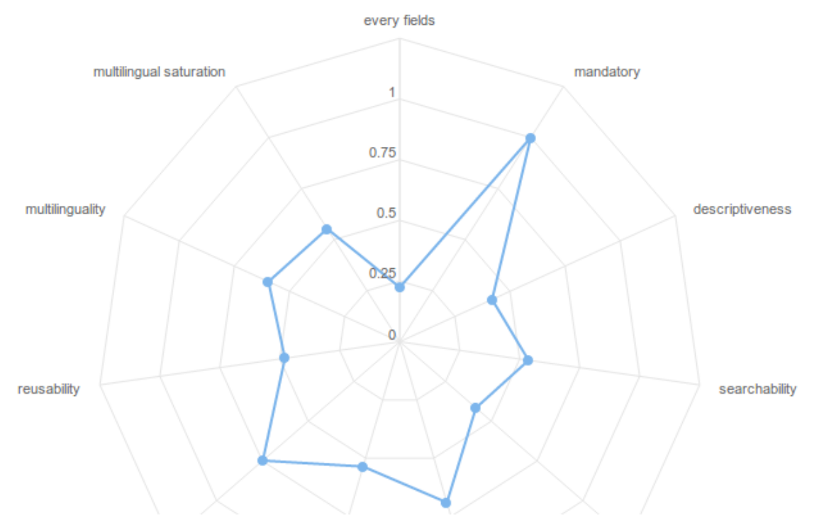
Quality dimensions

The benefits of metadata quality measurement:

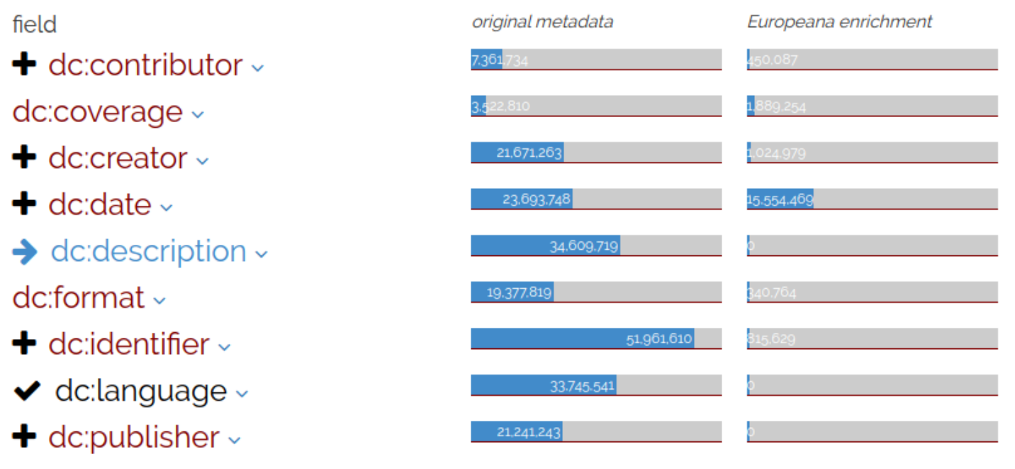
- for researchers, data librarians, data stewards: **check and improve metadata, propagate good use cases** and demonstrate anti-patterns
- for data repository developers, data journal editors: **improve data services**

Reporting

Completeness is based on the existence, cardinality and importance of fields defined in the metadata schema. Fields support one or more different functions, such as searchability, descriptiveness, identification, browsing, or re-usability. We call the completeness sub-dimensions, and they measure how well the functions are supported.

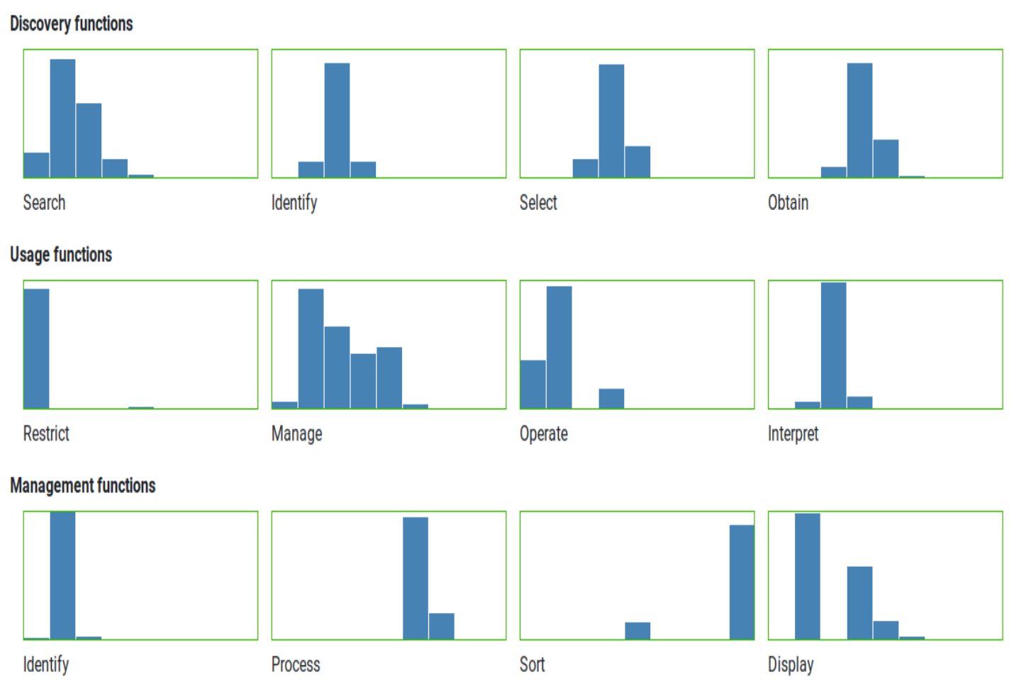


Spider chart of sub-dimensions

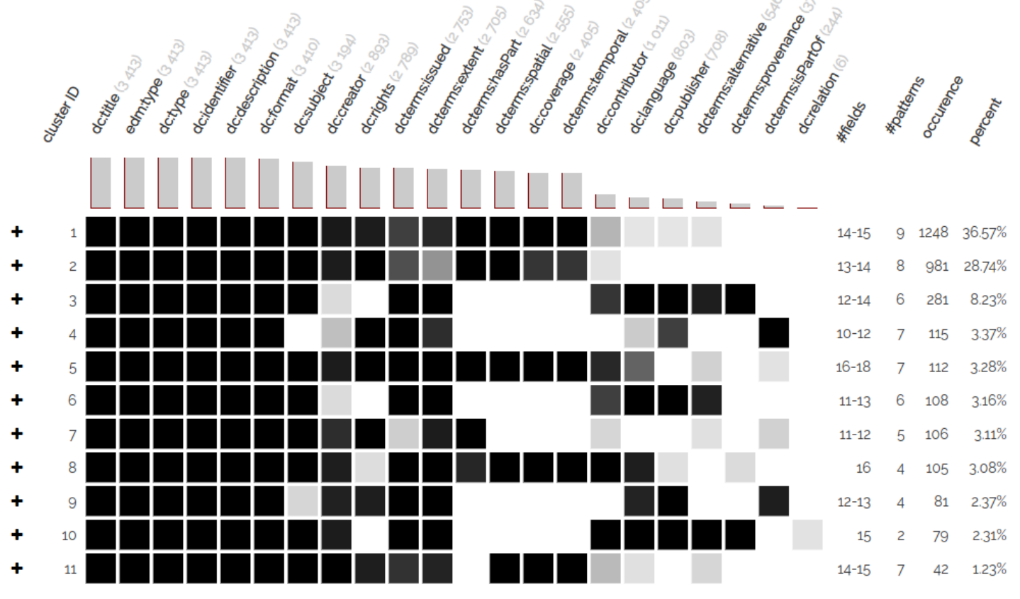

field
original metadata
Europeana enrichment

field	original metadata	Europeana enrichment
+ dc:contributor	7,361,734	150,087
dc:coverage	3,522,810	1,889,254
+ dc:creator	21,671,263	1,024,979
+ dc:date	23,693,748	15,554,469
→ dc:description	34,609,719	0
dc:format	19,377,819	340,764
+ dc:identifier	51,961,610	315,629
✓ dc:language	33,745,541	0
+ dc:publisher	21,241,243	0

Completeness (Europeana)



Functional analysis (MARC)

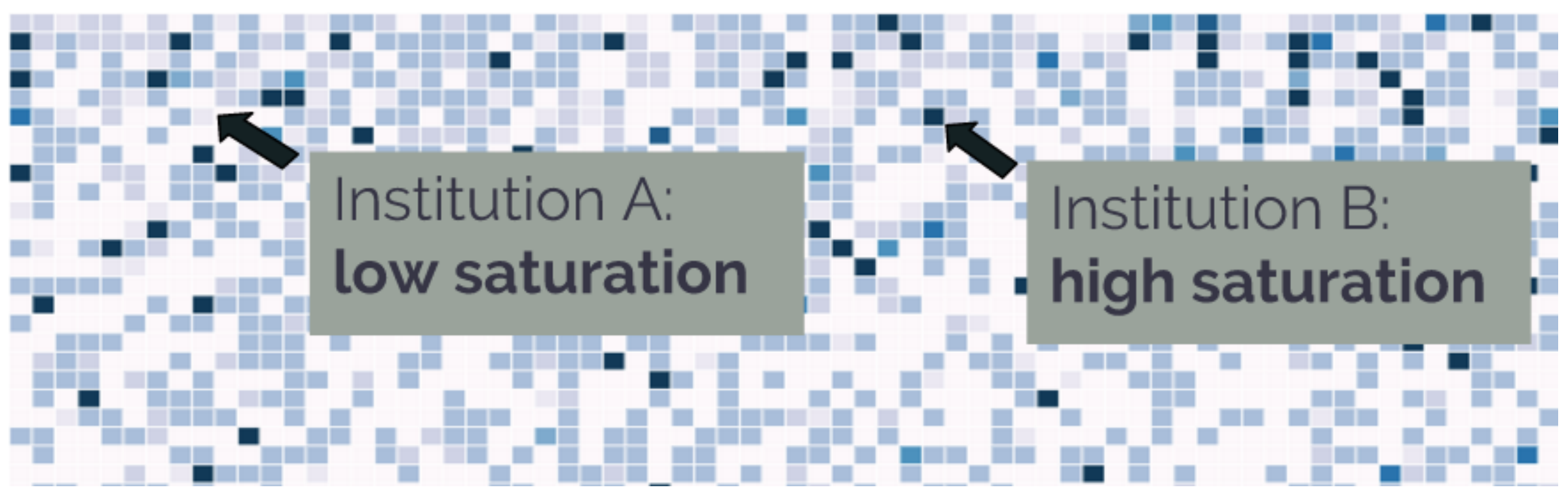


cluster ID	#fields	#patterns	occurence	percent
1	14-15	9	1248	36.57%
2	13-14	8	981	28.74%
3	12-14	6	281	8.23%
4	10-12	7	115	3.37%
5	16-18	7	112	3.28%
6	11-13	6	108	3.16%
7	11-12	5	106	3.11%
8	16	4	105	3.08%
9	12-13	4	81	2.37%
10	15	2	79	2.31%
11	14-15	7	42	1.23%

Record patterns (Europeana)

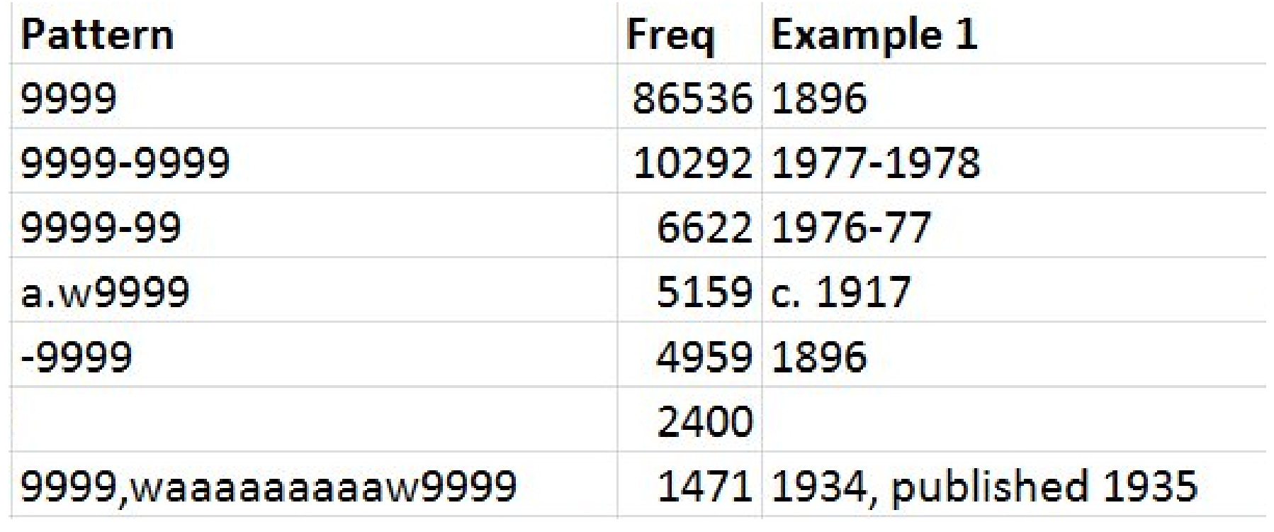
Multilinguality

Text w/o language annotation	"Germany"
Texts w/ language annotation	"Germany"@en, "Deutschland"@de
Link to multilingual vocabulary	<http://www.geonames.org/2921044/>



Heat-map of multilingual saturation. Squares represent collections.

Problem patterns



Pattern	Freq	Example 1
9999	86536	1896
9999-9999	10292	1977-1978
9999-99	6622	1976-77
a.w9999	5159	c. 1917
-9999	4959	1896
	2400	
9999,waaaaaaaaaw9999	1471	1934, published 1935

Figure 1: Variant date representations in MoMe collection (from H. Klinke, bit.ly/date-paterns)

Problem patterns are known, frequently occurred issues in the metadata record instances. They could be categorized into several types, such as duplicate or redundant information, irrelevant information, missing or incomplete information, misuse of fields.

Implementation

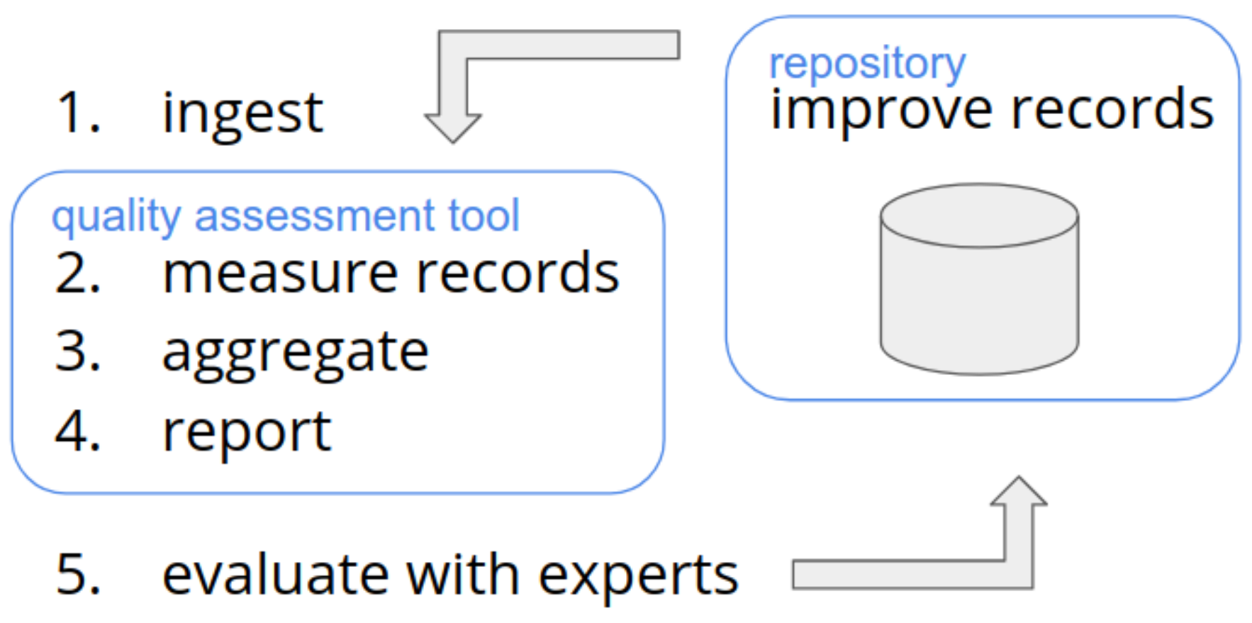

1. ingest
quality assessment tool
2. measure records
3. aggregate
4. report
5. evaluate with experts
repository
improve records

Figure 2: Data workflow

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 includes **Java and REST APIs**.

Metadata quality communities

- Europeana Data Quality Commission **http://bit.ly/europeana-dqc**
- DLF Metadata Assessment WG **http://dlfmetadataassessment.github.io**
- ADOCHS (Belgium) **http://adochs.be**

References & credits

Codes and data sources are freely available and reusable in the spirit of open science.

- Doctoral thesis **doi:10.13140/RG.2.2.33177.77920**
- Source code **http://pkiraly.github.io/about/#source-codes**
- Demo site: **http://144.76.218.178/europeana-qa**
- Project updates and documentation: **http://pkiraly.github.io/**
- Zotero bibliography **http://bit.ly/qa-bibliography**

Thanks to the participants of the Europeana Data Quality Commission for their contribution to this common effort. Special thanks to Gerhard Lauer, Ramin Yahyapour, Marco Büchler, and Juliane Stiller.

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 via pkiraly@gwdg.de or @kiru!