Measuring completeness as metadata quality metric in Europeana

Péter Király
Göttingen eResearch Alliance
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
Göttingen, Germany
peter.kiraly@gwdg.de

Marco Büchler
Institute of Computer Science
Georg-August-Universität Göttingen
Göttingen, Germany
mbuechler@etrap.eu

***Abstract*—Europeana, the European digital platform for cultural heritage, has a heterogeneous collection of metadata records ingested from more than 3200 data providers. The original nature and context of these records were different. In order to create effective services upon them we should know the strength and weakness or in other words the quality of these data. This paper proposes a method and an open source implementation to measure some structural features of these data, such as completeness, multilinguality, uniqueness, record patterns, to reveal quality issues.**

***Keywords*-Big data applications, Data analysis, Data collection, Quality of service, Quality management, Metadata, Data integration**

I. Introduction

> "In the last 24 hours, I wasted a lot of time because I made assumptions about some (meta)data that were just not correct. I spend a long time debugging, but the code was fine, it just couldn't find what's not there. Wrong assumptions are some of the most difficult bugs to catch." – Felix Rau, German linguist on the consequence of metadata issues[1]

Big data applications, Data analysis, Data collection, Quality of service, Quality management

The functionalities of an aggregated metadata collection are dependent on the quality of metadata records. Some examples from Europeana, the European digital platform for cultural heritage[2], illustrate the importance of metadata:

(a) Several thousand records have the title 'Photo' and its synonyms and language variations without further descriptions; how can a user find those objects which depict a particular building on those photos if no or imprecise textual descriptions are available?

(b) Several data providers listed in Europeana's 'Institution' facet under multiple different name variations (e.g. 'Cinecittà Luce S.p.A.' (372,412 records), 'Cinecittà Luce' (2,405 records), 'LUCE' (105 records) refer to the same organization), do we expect that a user is able to select all derivated forms when s/he wants to search objects belong to a particular organization?

(c) Without formalized and unified values in a 'year' facet, we are not able to use the functionality of interactive date range selectors. How can we interpret values such as '13436', or '97500000' when we expect a year?

(d) Some records have only technical identifiers, without any descriptive fields (title, creator, description, subjects, etc.). These records are not interpretable for humans. For this reason they do not support any of the core functionalities of Europeana.

(e) In a multilingual environment the user would expect that s/he get the same result-set when searching for a well-known entity, such as Leonardo's masterpiece 'Mona Lisa' (or 'La Gioconda', 'La Joconde'), however the different language variations return different result-sets: the language variations are not resolved into a common entity.

The question is how to decide which records should be improved, and which are good enough? 'Fitness for purpose' is a well-known slogan of quality assurance, referring to the concept that quality should be defined according to some business purposes. When dealing with the quality of metadata it is relevant to clarify why metadata are important. In Europeana's case it is relatively straightforward: Europeana provides access points to digitized objects. If the features of the record make it impossible to find a record, the intended purpose is not met, the user will not access the object and s/he will not use it. One can argue that the quality of the record is bad. The manual evaluation of each record is not affordable for even a middle-size collection.

This paper proposes a generalized methodology and a scalable software package which can be used in Europeana and elsewhere in the domain of cultural heritage for collections having either small or big data.

II. Background and foundations

Europeana collects and presents cultural heritage metadata records. The database at the time of writing contains more than 58 million records from more than 3200 institutions[3] in the Europeana Data Model (EDM) metadata schema. The

[1] 18 Oct 2018, https://twitter.com/fxru/status/1052838758066868224
[2] http://europana.eu
[3] Extracted from Europeana Search API.

organizations send their data in EDM or in any other metadata standard. Due to the variety of original data formats, cataloguing rules, languages and vocabularies, there are big differences in the quality of the individual records, which heavily affects the functionalities of Europeana's services.

In 2015 a Europeana task force investigated the problem of metadata quality, and published a report (see [1]), however – as stated – 'there was not enough scope ... to investigate ... metrics for metadata quality ....' In 2016 a wider Data Quality Committee[4] (DQC) was founded. In the Committee several experts from different domains (such as metadata theory, cataloguing, academic research, software development) come together to analyse and revise the metadata schema, discuss data normalization, run functional requirements analysis and define 'enabling' elements (answering questions such as 'Which are the core functionalities of Europeana?' and 'Which metadata elements support them?'). DQC also builds a 'problem catalogue', which is a collection of frequently occurred metadata anti-patterns (such as duplicate values, repeated title as description, values for machine consumption in fields which are intended for human consumption) [2]. The questions of multilinguality are given special emphasis.

The current research is conducted in collaboration with the DQC, having the purpose of finding methods, metrics and building an open source tool called 'Metadata Quality Assurance Framework'[5] to measure metadata quality. The proposed method is intended to be a generic tool for measuring metadata quality. It is adaptable to different metadata schemas (planned schemas include – but not limited to – MARC[6] and Encoded Archival Description[7]). The software is scalable to Big Data, as it is built to work together with the distributed file system of Apache Hadoop[8], the general, large-scale data processing engine Apache Spark[9], and Apache Cassandra[10] database. One of the most important features of this approach is the capability of producing understandable reports for data curators, who are not familiar with the language used by software developers, data scientists or statisticians. The reports are generated for those who are able to turn them into actionable plans. The framework is modular: there is a schema-independent core library with schema specific extensions. It is designed for usage in continuous integration for metadata quality assessment.[11]

The research asks the question how the quality of cultural heritage metadata can be best measured. It is generally assumed that quality itself is too complex for a single concept, and it is impossible to measure every aspect of it; also for theoretical reasons (for example current language detection methods do not work well with the short texts typically available in metadata records) and partly for practical reasons (such as limited resources). However a number of structural features of the metadata record are measurable and the outcome provides good approximation in most cases. One could call it 'metadata smells', similar to what is called 'code smells' in software development: 'a surface indication that usually corresponds to a deeper problem in the system'.[12] Approximation means in practice that the outcome should call for further scrunity by metadata experts. It also implies that there is a fair chance that the tool cannot detect variances due to that those errors are not bound to structural features.

The primary purpose of the project is to shed light on improvable metadata records. If we know where the errors are, and we can prioritize them, they can be fixed and the corrections can be planned carefully taking care of the order of importance of the problems. Since Europeana is an aggregator, the corrections should be done at the source of the information, inside the database of the particular data provider. Better data supports more reliable functions, so by fixing the weak records Europeana could build stronger services. Finding typical errors might also lead to improve the underlying metadata schema and its documentation (supposedly some of the errors occurred due to the language used in the schema documentations) and during the measurement process examples can be found for highlighting good and bad practice of certain metadata elements. Lastly high score metadata records could be used to propagate 'good metadata practices' or in the process of prototyping new services.

III. State of the art

The computational methods for metadata quality assessment emerged in the last decade in the cultural heritage domain ([4], [5], [6], [7]). The latest evaluation of the relevant works are conducted by [8]. The applied metrics in the domain of Linked Data (which has an intersection with the cultural heritage domain) are listed in [9]. Papers defined quality metrics and suggested computational implementations. They however mostly analyzed smaller volumes of records, metadata schemas which are less complex than EDM, and usually applied methods to more homogeneous data sets (notable exceptions are [10] investigating 7 million, and [7] investigating 25 million records). The novelty of this

[4]https://pro.europeana.eu/project/data-quality-committee

[5]http://144.76.218.178/europeana-qa/, source code and background information: http://pkiraly.github.io

[6]MAchine Readable Cataloging, https://www.loc.gov/marc/. A MARC assessment tool based on this framework is also created. It is available at https://github.com/pkiraly/metadata-qa-marc. Note that MARC is a much more complex standard than EDM, and the presence of a strict rule-set makes finding individual problems more important than in the case of Europeana records, so there are more emphasis on the "accuracy" and "conformance to expectation" metrics.

[7]http://www.loc.gov/ead/

[8]http://hadoop.apache.org/

[9]http://spark.apache.org/

[10]http://cassandra.apache.org/

[11]See http://pkiraly.github.io/2016/07/02/making-general/ and [3]

[12]The term was coined by Kent Beck and popularized by Martin Fowler in his Refactoring book, see https://martinfowler.com/bliki/CodeSmell.html

research is that it increases the volume of records, introduces new types of data visualizations and quality reports, and provides an open source implementation that is reusable in other collections.

For a comprehensive bibliography of cultural heritage metadata assessment see the Metadata Assessment Zotero library[13] which is maintained by the members of Digital Library Federation's Metadata Assessment group[14] and members DQC including the first author of this paper.

IV. Methodology

A. The EDM schema

An EDM record[15] consists of several entities. The core of the record is called *provider proxy*, it contains the data that the individual organizations (*data providers*) sent to Europeana. The original format of the data might be EDM or a number of different metadata schema used in the cultural heritage domain (such as Dublin Core, EAD, MARC etc.) – in this case the data providers or Europeana transform them to EDM. Other important parts are the *contextual entities*: agents, concepts, places and time spans which contains description of entities (persons, place names, etc.) which are in some relationship with the object. There are two important features of these contextual entities:

(1) They came from multilingual vocabularies, and the instances contain their labels in several languages.

(2) Wherever it is possible the entities have relationships with other entities (the relationships are defined by the SKOS ontology).

The last entity is called *Europeana proxy*. Structurally it is the same as the provider proxy, but it contains only the links between the provider proxy and the contextual entities which are detected by an automatic semantic enrichment process.

Each data element supports or enables one or more functionalities of the services built on top of the data. The Data Quality Committee is working on functional requirement analysis, in which we define the core functions starting from typical user scenarios (how the user interact with the collection), and analyse which metadata elements support them [11]. For the sake of example, see 'Cross-language recall'. Its user scenario is 'As a user, I want to search Europeana collections in the language I am most comfortable with, and feel confident that I will receive relevant results irrespective of language of documents.' These contextual elements are mostly multilingual. The set of enabling elements are defined as 'any element that can be linked to a contextual entity in the Europeana Entity Collection' such as dc:contributor, dc:creator, dc:date, etc.

[13]http://zotero.org/groups/metadata_assessment

[14]https://dlfmetadataassessment.github.io/

[15]For EDM documentation, guidelines and other materials consult https://pro.europeana.eu/page/edm-documentation

Since the definition of these enabling elements has not been yet harmonized with the purpose of measurement, we started with a simpler model called sub-dimensions. In this model instead of the more complex user scenarios Valentine Charles and Cecile Devarenne defined a matrix of general functionalities and their enabling elements. The sub-dimensions are:

- *Mandatory elements* - fields which should be present in every record. The model also handles group of fields from which at least one should be present, e.g. one from the 'subject heading'-like elements (dc:type, dc:subject, dc:coverage, dcterms:temporal, dcterms:spatial)
- *Descriptiveness* – how much information has the metadata to describe of what the object is about
- *Searchability* – the fields most often used in searches
- *Contextualization* – the bases for finding connected entities (persons, places, times, etc.) in the record
- *Identification* – for unambiguously identifying the object
- *Browsing* – for the browsing features at the portal
- *Viewing* – for the displaying at the portal
- *Re-usability* – for reusing the metadata records in other systems
- *Multilinguality* – for multilingual aspects, to be understandable for all European citizen

At the time of writing this model examines only the existence of the fields, it does not check the match of the content with the expectation – which task will be implemented in the next phase of the research.

B. Measuring

For every record, features are extracted or deducted which somehow related to the quality of the records. The main feature groups are:

- *simple completeness* – ratio of filled fields,
- *completeness of sub-dimensions* – groups of fields to support particular functions, as seen above,
- *existence and cardinality of fields* – which fields are available in a record and how many times,
- *problem catalogue* – existence of known metadata problems[16],
- *uniqueness of the descriptive fields* (title, alternative title, description)[17],
- *multilinguality*[18],
- *record patterns* – which fields form the 'typical record'?

The measurements happen on three levels: on individual records, on subsets (e.g. records of a data provider), and on the whole dataset.

[16]This measurement is an experimental in Europeana context as a proof of the concept. The full problem catalogue will be formally described with the Shapes Constraint Language ([12]).

[17]For the underlying theory see [13]. The applied method is different than it is described in the thesis.

[18]See [14] and [15]

Table I
NORMALIZATION OF CARDINALITY

number of instances	0	1	2-4	5-10	11-
normalized score	0.0	0.25	0.50	0.75	1.0

On the first level the tool iterates on every metadata record. It analyses the records and produces a comma-separated row containing the results of the individual measurements. In total there are more than one thousand numbers extracted from each record, each represents a quality-related feature of a field, a group of fields or the whole record calculated with different scoring algorithms.

The second level is of the subsets. Currently there are three kinds of subsets: datasets that are records ingested together during the same process (they usually handled by the same transformation chain when Europeana received them from the data providers); the records belong to the same data providers, and the intersection of these two: records from the same data provider ingested at the same process. In the future DQC might consider supporting additional facets, such as records ingested from the same country, data aggregator or any other reasonable property of the metadata records.

On the second and third level we calculate aggregated metrics; the completeness of structural entities (such as the main descriptive part and the contextual entities – agent, concept, place, timespan – connecting the description to linked open data vocabularies).

The final completeness score is the combination of two approaches both applying different weighting. In the first one the weighting reflects the sub-dimensions: the 'simple completeness' score's weight is 5 (this score is the proportion of available fields in the record comparing to all the fields in the schema), the mandatory elements' weight is 3, the rest sub-dimensions get 2. The equation is

$$c_{sub-dimensions} = \frac{\sum_{i=1}^{d} score_i \times weight_i}{\sum_{i=1}^{d} weight_i} \quad (1)$$

with d as the number of sub-dimensions, $score_i$ as the proportion of availability of the fields belong to the particular sub-dimension, and $weight_i$ as the weight of a sub-dimension.

In the second approach, the main factor is the normalized version of cardinality to prevent biasing effect of extreme values. Sometimes there are more than one hundred or even thousand field instances in a single record which would have too much effect on the score, so the tool normalizes them according to table I.

The cardinality-based weight is simple: each field equally counts 1, but the rdf:about field (which identifies the individual entities) counts 10 so that the number of entities is taken into account for the weighting. The equation is

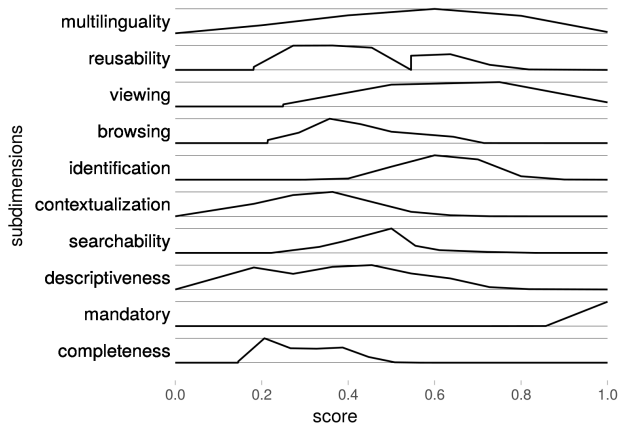

Figure 1. The distribution of sub-dimension and 'simple completeness' scores

$$c_{cardinality} = \frac{\sum_{i=1}^{d} norm(cardinality_i) \times weight_i}{\sum_{i=1}^{d} weight_i} \quad (2)$$

with d as the number of fields, $cardinality_i$ as the cardinality of a fields, $norm()$ as the normalizing function (see table I) and $weight_i$ as the weight of a field in this computation.

The final equation is the combination of these two approaches where the first approach has higher weight (so more important) than the second one:

$$c_{compound} = \frac{c_{sub-dimensions} + (0.4 \times c_{cardinality})}{1.4} \quad (3)$$

C. Implementation

The data processing workflow has four phases. The current workflow ingests data from a MongoDB database, and stores the extracted records in line-oriented JSON files either in a Linux file system or in a Hadoop File System (using the available resources there is no big difference in performance between the two, but in other scenarios Hadoop File System could be a better choice). The record level analyses are written in Java, using the Spark API[19]. It provides automatic and configurable multithreading, so the tool can make use the available resources of the environment effectively (either

[19]Core library: https://github.com/pkiraly/metadata-qa-api, Europeana specific extension: https://github.com/pkiraly/europeana-qa-api, Spark-interface: https://github.com/pkiraly/europeana-qa-spark. The APIs (and the MARC assessment tool) are available as compiled Java libraries within Maven Central Repository: https://mvnrepository.com/artifact/de.gwdg.metadataqa, so one could use it in 3rd party Java or Scala projects.

if it is a single machine with multicore processor or high performance computing cluster with several nodes). The output of these calculations are CSV files, which are also indexed by Apache Solr for occasional record based retrieval. The tool's quality dashboard make use the search and retrieval functionalities in displaying the results, and finding records with given quality metrics.

The third phase is a statistical analysis of the record level metrics. For datasets and data providers the software is written in R[20] and in the Scala implementation of Spark[21]. It reads the CSV files generated in the previous phase, and produces CSV and JSON files for storing the result of the calculations and image files for graphs, visualizing central tendencies or other statistical features of the data. R however has a weak point: it works exclusively in memory, so the size of memory limits the size of dataset it can process. For creating statistics for the whole Europeana dataset this is not enough. For this reason, for top level aggregations Scala on Spark is used. Scala's statistical capabilities are not that rich, so it does not produce all the metrics as R does.

The last phase is an online statistical dashboard, a light-weighted, PHP and JavaScript based website which displays the output of the previous phases.[22] The technical details of the workflow is documented in [16]. All phases are run in a single commodity hardware (Intel Core i7-4770 Quad-Core processor with 32 GB DDR3 RAM, with Ubuntu 16.04 operating system) which at the same time were also used for other research and development projects, so making the calculation resource-effective was an important constrain in the software design.

The data source of the calculation is a snapshot of Europeana data. The first snapshot were created at the end of 2015, which contains 46 million records, 1747 datasets and 3550 data providers[23] (extracted from Europeana's OAI-PMH service). In the project's lifetime additional snapshots were created, the latest one is from August 2018 (62 million records, 1.27 TB in total, the data source is a replica of Europeana's MongoDB database).[24] DQC aims to introduce a monthly update cycle, so the time span between the updates of Europeana production database and the refreshing of the data quality dashboard should not be more than one month.

[20]source code: https://github.com/pkiraly/europeana-qa-r

[21]https://github.com/pkiraly/europeana-qa-spark/tree/master/scala

[22]source code: https://github.com/pkiraly/europeana-qa-web

[23]the name of data providers has not been not normalized so far, some organizations have several different names.

[24]In order to make the research repeatable, three full data snapshots are available for download at http://hdl.handle.net/21.11101/0000-0001-781F-7 and the first one is archived for long term preservation at the Humanities Data Center, Göttingen: https://hdl.handle.net/21.11101/EAEA0-826A-2D06-1569-0. The format of these snapshot is JSON, one record per line.

Table II
BASIC STATISTICS OF COMPLETENESS CALCULATIONS

metric	mean	std.dev.	min.	max.
sub-dimension-based	0.50	0.07	0.22	0.93
cardinality-based	0.12	0.05	0.05	0.48
compound	0.39	0.06	0.17	0.78

V. RESULTS

A. Completeness

A comparison of the scores of sub-dimension-based (where the field importance counts) and the field-cardinality-based approaches (where the number of field instances counts) reveals that they give different results. They correlate by the Pearson's correlation coefficient of 0.59, however their shape and ranges are different. Because of the nature of calculation the compound score is quite close to the first approach and the cardinality-based calculation has smaller effect on the final score. The sub-dimension-based scores are in the range of 0.22 and 0.92 while cardinality based scores are in the range of 0.05 and 0.48. The details of the distribution are shown in table II and figure 2.

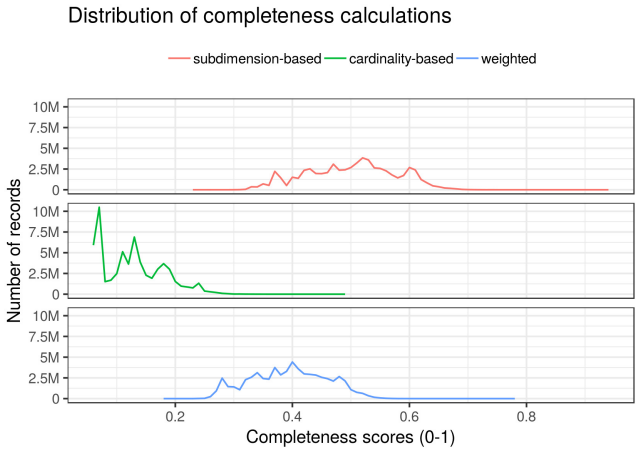

Figure 2. Distribution of completeness calculations

There are data providers for which all (in some cases more than ten thousand) records have the same scores: they have a uniform structure. Because one simple score is not able to testify it, the field-level analysis shows that in these collections all the records have the very same (Dublin Core based) field set. On the other end there are collections where both scores diverge a lot. For example in the identification of sub-dimension a data provider has five distinct values (from 0.4 to 0.8) almost evenly distributed while one of the best collection (of the category) is almost homogeneous: 99.7% or the records have the same value: 0.9 (even the rest 0.3%

has 0.8). It means that the corresponding fields[25] are usually not available in the records of the first dataset, while they are almost always there in the second dataset. The tool provides different graphs and tables to visualize the distribution of the scores.

From the distribution of the fields the first conclusion is that lots of records miss contextual entities, and only a couple of data providers have 100% coverage (6% of the records have *agent*, 28% have *place*, 32% have *timespan* and 40% have *concept* entities). Only the mandatory technical elements appear in every records. There are fields, which are defined in the schema, but not filled in the records and there are overused fields – e.g. dc:description is frequently used instead of more specific fields (such as table of contents, subject related fields or alternative title).

Users can check all the features on top, collection, and records level on the quality dashboard. Data providers get a clear view of their data, and based on this analysis they can design a data cleaning or data improvement plan.

B. Multilinguality

DQC has recently published the details and the results of the multilinguality calculation (see [14] and [17]), so this section is a very short summary of the outcome. EDM follows the RDF model for language annotation, so data creators could denote that a string is written in a particular language (e.g. *"Brandenburg Gate"@en*, where 'Brandenburg Gate' is the value of the field, and 'en' denotes English language). This construct is called tagged literal. DQC found four relevant record-level metrics.

- number of tagged literals
- number of distinct language tags
- number of tagged literals per language tags
- average number of languages per property for which there is at least one language-tagged

It was calculated for the Provider Proxy (which is the original data the organizations submit), the Europeana Proxy (which contains enhancements, typically from multilingual vocabularies), and finally for the whole object. The output is summarized in tables III and IV and figure 3).

Table IV reflects that only 20% of the records have two or more languages per property in the provider proxy. Since the enhancement process which inject external contextual information (about agents, concepts, places and timespans) from multilingual data sources, such as DBpedia and other into the Europeana records, the overall multilinguality became higher. Not just the number of fields with two or more language values are increased, but the number of records without any language annotation decreased.

[25]dc:title, dcterms:alternative, dc:description, dc:type, dc:identifier, dcterms:created, dc:date and dcterms:issued in the Provider Proxy and edm:provider and edm:dataProvider in the Aggregation.

Table III
METRICS OF MULTILINGUALITY (MEANS)

metric	provider	europeana	whole object
number of tagged literals	5.44	64.34	69.79
number of distinct language tags	1.67	37.92	38.79
number of tagged literals per language tags	2.64	0.95	2.17
average number of languages per property for which there is at least one language-tagged literal	1.10	28.10	20.21

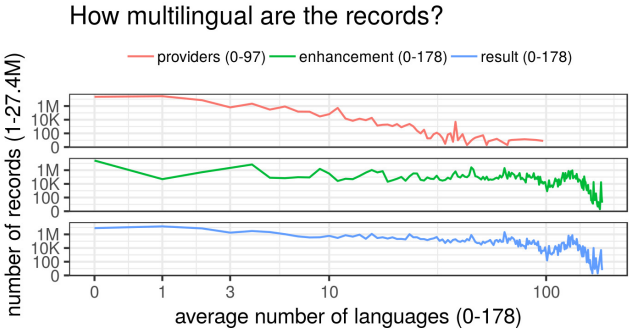

Figure 3. Multilinguality

Another finding is that the language tags are not always standardized. Different data providers follow different standards, or use ad-hoc tags. In the whole dataset there are more than 400 different language tags, but several tags denote the same language (e.g. "en", "eng", "Eng" etc. refer to English). A further investigation should analyze records with normalized language tags, to get proper picture of language usage.

C. Uniqueness

One might recall the example of similar titles mentioned at the beginning of this paper. To find those records we should calculate the uniqueness of the values. Uniqueness is a positive value in those fields which describe unique properties of an object, and less positive (or even negative) in those fields which connects records to contextual information where the values should come from a controlled vocabulary, and thus in ideal case multiple records share the same terms. In order to effectively establish the uniqueness of a value, one should be able to check a search index with the special requirement, that it should index and store field values as a phrase. Since building such an index for the whole dataset would require

Table IV
DISTRIBUTION OF AVERAGE NUMBER OF LANGUAGES PER PROPERTY

entity	0	1	2 or more
Provider Proxy	22.4M (36.2%)	27.3M (44.1%)	12.1M (19.6%)
Europeana Proxy	25.8M (41.7%)	49K (0.07%)	36.1M (58.2%)
Object	8.2M (13.3%)	14.6M (23.7%)	39.1M (63.0%)

Table V
UNIQUENESS CATEGORIES BY FREQUENCY

field	*****	****	***	**	*
title	2-	8-	37-	293-	5226-
alternative	2-	6-	23-	132-	1514-
description	2-	7-	34-	252-	4128-

Table VI
HOW UNIQUE ARE EUROPEANA RECORDS?

field	unique	*****	****	***	**	*
title	59.4	9.5	8.3	8.7	7.1	6.6
alternative	62.4	11.2	7.1	3.6	2.7	12.7
description	54.6	9.0	7.3	10.2	6.7	11.9
together	45.4	10.8	15.6	18.2	6.3	3.62

more resources than what were available for this research, three fields were selected for this task: title, alternative title, and description. Calculating the score we applied a modified version of Solr's relevancy scoring:

$$score(t_f, v_f) = log\left(1 + \frac{t_f - v_f + 0.5}{v_f + 0.5}\right) \quad (4)$$

$$uniqueness_f = \left(\frac{score(t_f, v_f)}{score(t_f, 1.0}\right)^3 \quad (5)$$

t_f is the number of records field f is available, v_f is the frequency of a value.

As seen in figure 4) the score decreases radically as the field value became more frequent. On the user interface there is a categorization: besides the unique values, there are 5 categories denoted with stars. Table V) displays the category boundaries for this three fields:

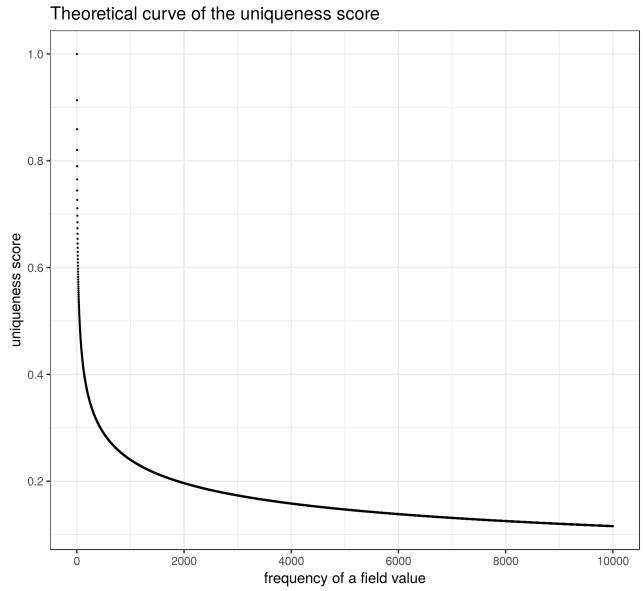

Figure 4. Theoretical curve of uniqueness score. As frequency of terms gets higher, the uniqueness score get radically smaller towards zero.

The result of the categorization is shown in table VI. The absolute majority of the records regarding to all three fields contains unique values, however still there are millions of records with low scores for one or another field, moreover there are almost ten thousand records where none of these fields are available. When we examine the three values together (see the last row of the table), and calculating an average of the result is that there are 25 million records with unique values in all available fields and on the other side of the scale only 3.62% of the records are in the lowest category. It means, that even if some values are low, most of the time there is at least one field with a less frequent value, so the record has higher chance to be found by a search term.

From the Solr index we could extract the most frequent terms. Above the "photograph" example there are frequent phrases in the title field denoting missing information (e.g. "Unbekannt", "Onbekend" or "+++EMPTY+++"), collection, journal or institution names ("Journal des débats politiques et littéraires", "ROMAN COIN") or a general descriptive term ("Porträtt", "Château", "Plakat", "Rijksmonument"). It would require further investigation to filter out those frequent terms, which stand in records, where the other descriptive fields also lack a necessary level of uniqueness. The tool provides solid basic for such an investigation.

D. Record patterns

What fields make typical records? In other words: what fields do the data providers make use? Record patterns are the typical field collocations. Since the completeness measurement collects the existence of all the fields, a map-reduce based analysis could extract the pattern. In this case the mapping function creates the patterns (each pattern is a list of field names available in a particular record) while the reduce-function counts them. In the first iteration it turned out that there are too many similar patterns which would be worth to group together in order to analyze effectively. In clustering patterns a similarity algorithm was applied. All patterns are first represented by a string containing of zeros and ones. First, all the fields of a collection was collected and sorted by a standard field order. Each fields are categorized into one of three categories: mandatory fields, important fields (those fields which appeared in a sub-dimension) and non-distinguished fields. If the field exists in the pattern it is represented by one or more ones otherwise one or more zeros. The mandatory fields get three characters, the important fields get two, and others gets only one character. This way the patterns having the same important fields and different unimportant fields are closer to each other than patterns sharing the non-important field. The similarity is calculated by the Jaro-Winkler algorithm. In the visualization (as you can see in figure 5) the clusters are displayed by default, and the user should click to unhide

the patterns belong to this cluster. The table is ordered by the number of records, so the more typical records are on the top. If the field is not available in all records within the cluster only in some, it is grayed (the color is proportional with the number of record). By default the page does not display patterns occur in less than 1% of the records.

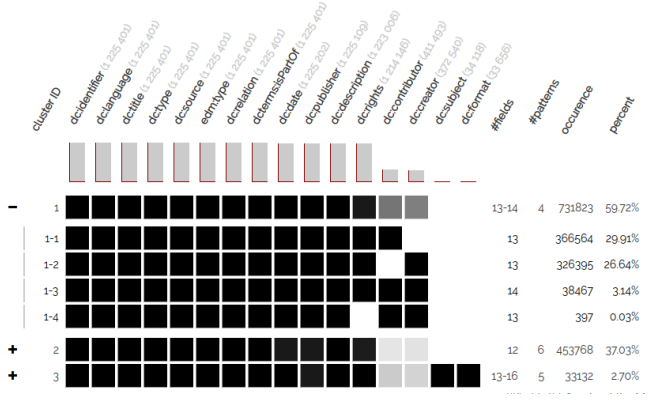

Figure 5. Clustered record patterns. The first line represents a cluster of similar patterns. The next four lines are the patterns belong to the cluster. The top gray bar represents the frequency of fields in the whole collection.

Thus far two quality problems were revealed by the use of record patterns. The first problem covers those records which has only small number of fields. There were more than $150,000$ records having only the following four fields in the Provider Proxy entity: dc:title, dc:type, dc:rights, edm:type of which only the first two might contain descriptive information about the object. It is evident, that there is a high chance that users are not able to discover them by using facets, because those objects do not provide enough information. The second problem is a structural homogeneity: each record in some collection has always the same set of fields. There are 906 such data providers in Europeana, but fortunately most of them are relatively small collection, only 26 have more than a thousand records. The biggest homogeneous collection (with over $500,000$ records) however contains only 5 fields of which 3 are descriptive fields. The problem with such a record is, that it contains generic fields instead of specific ones (for example do not make distinction among conceptual, spatial and temporal subject headings, and putting different contextual information into dc:type or dc:subject).

VI. Further work

Europeana works on its new ingestion system called Metis[26] which is able to integrate the tool. According to the plans when a new record-set arrives for the import, the measurement is launched automatically, the Ingestion Officer can check the quality report and share the output and the conclusions with the data providers who can react either by changing the transformation rules or fixing the issues if it is possible.

Beside the discussed calculation models there are other metrics which are planned to be computed in the near future (e.g. accuracy, information content, timeliness, existence of known metadata anti-patterns). The relevant literature suggests a top level score, which summarizes all metrics into one score, which finally characterizes the record's metadata quality. This could be achieved by weighting the metrics or applying machine learning algorithms, such as Principal Component Analysis [18]. It was mentioned previously that current completeness calculation approaches examine the existence of a field. The next step in this front is to extend this model with the evaluation of the content of the relevant fields according to the User Scenarios analysis ([11]).

In DQC we also plan to compare the scores with experts' evaluation and with usage data (log files). Harper ran a test to reveal whether is there a correlation between the usage of an object (the frequency of access via their portal and API) and the scores calculated by a quality assessment in Digital Public Library of America (which is similar to Europeana regarding to its purpose and its metadata schema). This approach failed partly because there were not enough usage data available at time of the research, however the proposed method sounds promising, and if Europeana has its log files, it would worth to run an experience.

To define the problem catalogue with W3C's Shapes Constraint Language [12] is planned. Another plan is to publish the results as linked data fit to the ontology of Data Quality Vocabulary [19].

The proposed method can be used in collections of other metadata schemas, such as MARC based library catalogues[27], EAD-based archival collections,[28] and others.

VII. Conclusion

In the research the relationship between functionality and the metadata schema (together with DQC) are rethought, and a framework is implemented which proved to be successful in measuring structural features which correlate with metadata issues. The user of the framework is able to select low and high quality records. According to the hypothesis structural features such as existence and cardinality of fields correlate with metadata quality, and it proved to be true. The research extended the volume of the analyzed records by introducing big data tools that were not mentioned previously in the literature.

In this research a particular dataset and metadata schema were covered, however the applied method is based on

[26]https://github.com/europeana/metis-framework

[27]Since MARC has lots of strict content related rules, and EDM does has only a few, there is significant distance between the approach followed in the two project.

[28]The biggest European archival collection Archives Portal Europe (http://www.archivesportaleurope.net/) published their data via a REST API under CC0 license.

generalized algorithms, so it is applicable to other data schema. Several Digital Humanities studies (some examples: KOLIMO (Corpus of Literary Modernism)[29], [20], [21]) based on schema defined cultural databases. The research process could be improved by finding the weak points of the sources, making the conclusions more reliable, and – reflecting to Felix Rau's tweet quoted at the beginning of this paper – by forming more realistic assumptions about the data.

ACKNOWLEDGMENT

The first author would like to thank to all the past and current members of the Europeana Data Quality Committee, to the supervisors of his PhD research, Gerhard Lauer, and Ramin Yahyapour, to Jakob Voß, Juliane Stiller, Mark Phillips for providing feedbacks, to Christina Harlow and Zaveri Amrapali for general inspirations, and to Felix Rau for the motto and to GWDG for supporting the research.

REFERENCES

[1] M.-C. Dangerfield *et al.*, "Report and recommendations from the task force on metadata quality." Europeana, Tech. Rep., 2016. [Online]. Available: https://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf

[2] T. Hill and H. Manguinhas, "Internal dqc problem patterns," Europeana, Tech. Rep., 2016. [Online]. Available: http://bit.ly/2jIXQGU

[3] P. Király, "Towards an extensible measurement of metadata quality," in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017*. ACM Press, 2017. [Online]. Available: https://doi.org/10.1145/3078081.3078109

[4] T. R. Bruce and D. I. Hillmann, "The continuum of metadata quality: Defining, expressing, exploiting," in *Metadata in practice*, D. Hillman and E. Westbrooks, Eds. ALA Editions, 2004, pp. 238–256. [Online]. Available: http://ecommons.cornell.edu/handle/1813/7895

[5] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith, "A framework for information quality assessment," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 12, pp. 1720–1733, 2007. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.20652/full

[6] X. Ochoa and E. Duval, "Automatic evaluation of metadata quality in digital repositories," *International Journal on Digital Libraries*, vol. 10, no. 2, pp. 67–91, 2009.

[7] C. Harper, "Metadata analytics, visualization, and optimization: Experiments in statistical analysis of the digital public library of america (DPLA)," *The Code4Lib Journal*, no. 33, 2016. [Online]. Available: http://journal.code4lib.org/articles/11752

[29] https://kolimo.uni-goettingen.de/

[8] N. Palavitsinis, "Metadata quality issues in learning repositories," Ph.D. dissertation, Alcala de Henares, 2014. [Online]. Available: https://www.researchgate.net/publication/260424499_Metadata-_Quality_Issues_in_Learning_Repositories

[9] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web*, vol. 7, no. 1, pp. 63–93, 2015. [Online]. Available: http://content.iospress.com/articles/semantic-web/sw175

[10] D. Newman, K. Hagedorn, C. Chemudugunta, and P. Smyth, "Subject metadata enrichment using statistical topic models," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '07. ACM, 2015, pp. 366–375. [Online]. Available: http://doi.acm.org/10.1145/1255175.1255248

[11] T. Hill, V. Charles, and A. Isaac, "Discovery - user scenarios and their metadata requirements - v.3," Europeana, Tech. Rep., 2016. [Online]. Available: https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_WG/DataQualityCommittee/DQC_DiscoveryUserScenarios_v3.pdf

[12] H. Knublauch and D. Kontokostas, "Shapes constraint language (SHACL)," W3C, W3C Recommendation, jul 2017, https://www.w3.org/TR/2017/REC-shacl-20170720/.

[13] K. Al-Gumaei, "Scalable measurement of the information content of the metadata instances using big data framework europeana metadata as case study (master's thesis)," Master's thesis, Georg-August-Universität Göttingen, 2016.

[14] V. Charles, J. Stiller, W. Bailer, N. Freire, and P. Király, "Evaluating data quality in europeana: Metrics for multilinguality," 2017. [Online]. Available: https://www.researchgate.net/publication/319956489_Data_Quality_Assessment_in_Europeana_Metrics_for_Multilinguality

[15] J. Stiller and P. Király, "Multilinguality of metadata. measuring the multilingual degree of europeana's metadata," in *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017)*, ser. Schriften zur Informationswissenschaft, M. Gäde, V. Trkulja, and V. Petras, Eds. Verlag Werner Hülsbusch, 2017, pp. 164–176. [Online]. Available: https://www.researchgate.net/publication/314879735_Multilinguality_of_Metadata_Measuring_the_Multilingual_Degree_of_Europeana's_Metadata

[16] P. Király, "How to run the analysis? a cheat sheet," Tech. Rep., 2015. [Online]. Available: http://pkiraly.github.io/cheatsheet/

[17] P. Király, J. Stiller, V. Charles, W. Bailer, and N. Freire, "Evaluating data quality in europeana: Metrics for multilinguality," in *Proceedings of the 12th Metadata and Semantic Research Conference - MTSR2018*, 2018.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer New York, 2013. [Online]. Available: https://doi.org/10.1007/978-1-4614-7138-7

[19] R. Albertoni and A. Isaac, "Data on the web best practices data quality vocabulary," W3C, W3C Note, dec 2016, https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/.

[20] G. Strezoski and M. Worring, "Omniart: Multi-task deep learning for artistic data analysis," *CoRR*, vol. abs/1708.00684, 2017. [Online]. Available: http://arxiv.org/abs/1708.00684

[21] B. Schmidt, "Stable random projection: Standardized universal dimensionality reduction for library-scale data," in *Digital Humanities 2017. Conference Abstracts*, R. Lewis, C. Raynor, D. Forest, M. Sinatra, and S. Sinclair, Eds., 2017, pp. 340–342. [Online]. Available: https://dh2017.adho.org/abstracts/497/497.pdf