Measuring Metadata Quality

Thesis
in order to acquire the Doctoral Degree in Philosophy
at the Faculty of Humanities of the Georg-August-Universität Göttingen

submitted by
Péter Király
from Debrecen

Göttingen 2019

Thesis supervisors:
Prof. Dr. Gerhard Lauer
Prof. Dr. Ramin Yahyapour
Dr. Marco Büchler

Abstract

In the last 15 years different aspects of metadata quality have been investigated. Researchers measured the established metrics on a variety of metadata collections. One common aspect of the majority of these research projects is that the tools they produce as a necessary side effect were not intended to be reused in other projects. This research, while focusing mainly on a specific metadata collection, Europeana, investigates practical aspects of metadata quality measurement such as reusability, reproducability, scalability and adaptability.

Europeana.eu – the European digital platform for cultural heritage – aggregates metadata describing 58 million cultural heritage objects from more than 3200 libraries, museums, archives and audiovisual archives across Europe. The collection is heterogeneous with objects in different formats and languages and descriptions that are formed by different indexing practices. Often these records are also taken from their original context. In order to develop effective services for accessing and using the data we should know their strengths and weaknesses or in other words the quality of these data. The need for metadata quality is particularly motivated by its impact on user experience, information retrieval and data re-use in other contexts. In Chapter 2 the author proposes a method and an open source implementation to measure some structural features of these data, such as completeness, multilinguality and uniqueness. The investigation and exposure of record patterns is another aspect to reveal quality issues.

One of the key goals of Europeana is to enable users to retrieve cultural heritage resources irrespective of their origin and the material's metadata language. The presence of multilingual metadata descriptions is therefore essential for successful cross-language retrieval. Quantitatively determining Europeana's crosslingual reach is a prerequisite for enhancing the quality of metadata in various languages. Capturing multilingual aspects of the data requires us to take data aggregation lifecycle into account including data enhancement processes such as automatic data enrichment. In Chapter 3 the author presents an approach developed together with some members of Europeana Data Quality Committee for assessing multilinguality as part of data quality dimensions, namely completeness, consistency, conformity and accessibility. The chapter describes the defined and implemented measures, and provides initial results and recommendations.

The next chapter (Chapter 4) – investigating the applicability of the above mentioned approach – describes the method and results of validation of 16 library catalogues. The format of the catalog record is Machine Readable Cataloging (MARC21) which is the most popular metadata standard for describing books. The research investigates the structural features of the record and as a result finds and classifies

different commonly found issues. The most frequent issues are usage of undocumented schema elements, improper values instead of using terms from controlled vocabulary, or the failure to meet other strict requirements.

The next chapters describe the engineering aspects of the research. First (Chapter 5), a short account of the structure of an extensible metadata quality assessment framework is given, which supports multiple metadata schemas, and is flexible enough to work with new schemas. The software has to be scalable to be able to process huge amount of metadata records within a reasonable time. Fundamental requirements that need to be considered during the design of such a software are i) the abstraction of the metadata schema (in the context of the measurement process), ii) how to address distinct parts within metadata records, iii) the workflow of the measurement, iv) a common and powerful interface for the individual metrics, and v) interoperability with Java and REST APIs. Second (Chapter 6), is an investigation of the optimal parameter settings for a long running, standalone mode Apache Spark based, stateless process. It measures the effects of four different parameters and compares the application's behaviour in two different servers. The most important lessons learned in this experiment is that allocating more resources does not necessary imply better performance. Moreover, what we really need in an environment with limited and shared resources is a 'good enough' state which respectfully let other processes run. To find the optimal settings, it is suggested to pick up a smaller sample, which is similar to the full dataset in important features, and measure performance with different settings. The settings worth to check are number of cores, memory allocation, compression of the source files, and reading from different file systems (if they are available). As a source of ground truth Spark's default log, Spark event log, or measuring points inside the application can be used.

The final chapter explains future plans, the applicability of the method to other subdomains, such as Wikicite (the open citation data collection of Wikidata) and research data, and research collaborations with different cultural heritage institutions.

Zusammenfassung

Übersetzt von Juliane Stiller

In den letzten 15 Jahren wurden verschiedene Aspekte von Metadatenqualität untersucht. In verschiedenen Metadatenkollektionen haben Wissenschaftler und Wissenschaftlerinnen Messwerte für etablierte Kennzahlen erfasst. Gemeinsam ist diesen Forschungsprojekten, dass die für die Messungen benötigten Werkzeuge häufig nicht darauf ausgelegt sind in anderen Projekten wiederverwendet zu werden. Die vorliegende Arbeit beschäftigt sich hauptsächlich mit der speziellen Metadatenkollektion von Europeana und untersucht dabei die praktischen Aspekte von Kriterien zur Messung von Metadatenqualität, wie Wiederverwendung, Reproduzierbarkeit, Skalierbarkeit und Anpassungsfähigkeit.

Europeana.eu, die europäische digitale Plattform für kulturelles Erbe, sammelt Metadaten von 58 Millionen kulturellen Objekten, die aus mehr als 3200 Bibliotheken, Museen, Archiven und audiovisuellen Archiven in Europa stammen. Diese Sammlung ist heterogen und besteht aus Objekten in verschiedenen Formaten und Sprachen, deren Beschreibungen durch unterschiedliche Indexierungspraktiken entstanden sind. Oft wurden die Objekte aus ihrem ursprünglichen Kontext genommen. Um nun Dienstleistungen zu entwickeln, mit denen die Daten zugänglich gemacht und genutzt werden können, muss man die Stärken und Schwächen oder anders ausgedrückt die Qualität der Daten kennen. Der Bedarf an qualitativ hochwertigen Daten ist durch deren Einfluss auf die Nutzererfahrung, das Information Retrieval und die Wiederverwendung von Daten in anderen Zusammenhängen motiviert. Im zweiten Kapitel schlägt der Autor eine Methode sowie eine Open Source Lösung vor, um strukturelle Eigenschaften von Daten, wie Vollständigkeit, Multilingualität und Eindeutigkeit, zu messen. Eine weitere Komponente, um Probleme in Daten aufzudecken, ist die Analyse und Veranschaulichung von Dokumentstrukturen.

Ein zentrales Anliegen von Europeana ist es, Nutzern und Nutzerinnen die Möglichkeit zu bieten Kulturgüter unabhängig ihrer Herkunft und Sprache, in der sie beschrieben sind, zu finden. Für ein erfolgreiches sprachübergreifendes Retrieval sind mehrsprachige Metadatenbeschreibungen unerlässlich. Eine Voraussetzung um überhaupt die Metadatenqualität in verschiedenen Sprachen verbessern zu können, ist die quantitative Bestimmung der sprachlichen Vielfalt der Metadaten in Europeana. Um die Mehrsprachigkeit in den Daten erfassen zu können, müssen der komplette Prozess der Datenaggregation abgebildet und auch Prozesse zur Datenverbesserung, wie bei-

spielsweise automatische Datenanreicherungen, berücksichtigt werden. In Kapitel 3 präsentiert der Autor eine Methode, die er zusammen mit Mitgliedern des Europeana Data Quality Committees entwickelt hat, um Mehrsprachigkeit als Aspekt verschiedener Dimensionen von Datenqualität, wie Vollständigkeit, Konsistenz, Konformität und Zugänglichkeit, messen zu können.

Das nächste Kapitel (Kapitel 4) geht darauf ein, wie das oben beschriebene Konzept skalierbar umgesetzt werden kann und beschreibt die Methode und die Ergebnisse der Validierung von 16 Bibliothekskatalogen. Die Katalogdatensätze liegen in einem maschinenlesbaren Format (MARC21) vor, dem am weitesten verbreiteten Metadatenstandard zur Beschreibung von bibliographischen Einheiten. Die vorliegende Untersuchung ermittelt strukturelle Merkmale der Datensätze und klassifiziert die in diesen häufig auftretenden Probleme. Die häufigsten Probleme sind die Verwendung von undokumentierten Schema-Elementen, falsche Werte an Stellen, an denen ein Wert aus einem kontrollierten Vokabular hätte übernommen werden sollen oder die Missachtung anderer strenger Vorgaben.

Die nächsten Kapitel beschreiben die technischen Aspekte der Forschung. In Kapitel 5 wird ein kurzer Überblick über den Aufbau des erweiterbaren Framework zur Messung von Metadatenqualität gegeben. Dieser unterstützt verschiedene Metadatenschemata und ist flexibel genug, um mit neuen Schemata umgehen zu können. Diese Anwendung muss skalierbar sein, um eine große Anzahl von Metadatensätzen innerhalb einer angemessenen Zeit verarbeiten zu können. Grundlegende Anforderungen, die bei der Entwicklung einer solchen Software berücksichtigt werden müssen, sind i) die Abstraktion des Metadatenschemas (im Rahmen des Messprozesses), ii) der Umgang mit unterschiedlichen Teilen innerhalb von Metadatensätzen, iii) der Messprozess, iv) eine gemeinsame und leistungsfähige Schnittstelle für die einzelnen Metriken und v) die Interoperabilität mit Java- und REST-APIs. In Kapitel 6 wird untersucht welche optimalen Parametereinstellungen für einen lang laufenden Prozess, basierend auf dem Apache Spark Stand-Alone-Modus, nötig sind. Dafür werden die Auswirkungen von vier verschiedenen Parametern gemessen und das Verhalten der Anwendung auf zwei verschiedenen Servern verglichen. Die wichtigste Erkenntnis aus diesem Experiment ist, dass die Zuweisung von mehr Ressourcen nicht unbedingt eine bessere Leistung bedeutet. In einem Umfeld mit begrenzten und geteilten Ressourcen brauchen wir einen Zustand, der "gut genug" ist und anderen Prozessen den Vortritt lässt. Um die optimalen Einstellungen zu finden und die Performance mit verschiedenen Parametern zu messen, sollte ein kleineres Sample herangezogen werden, das in wichtigen Merkmalen dem vollständigen Datensatz ähnelt. Die Einstellungen, die überprüft werden sollten, sind die Anzahl der Rechenkerne, die Speicherzuweisung, die Kompression der Quelldateien und (falls vorhanden) das Auslesen verschiedener Dateisysteme. Als Grundlage der Bewertung können das Standard-Spark-Logging sowie das Event-Logging oder Messpunkte innerhalb der Anwendung verwendet werden.

Das letzte Kapitel (Kapitel 7) erläutert Zukunftspläne, die Anwendbarkeit der Methode auf andere Bereiche wie Wikicite (die offene Datenbank für Zitationsdaten von Wikidata) und Forschungsdaten, sowie Forschungskooperationen mit verschiedenen

Kulturerbeinstitutionen.

Contents

Chapter 1.

Introduction

In the cultural heritage section there is a long tradition of building catalogues. During the centuries museums, archives and libraries developed different systems to record their collections.

There is no good definition for the quality, but much of the literature agrees that quality should somehow be in line with the 'fitness for purpose', i.e. the quality of an object should be measured as how much the object supports a given purpose. The main purposes of the cultural heritage metadata are registering the collection and helping users in discovery. The functional analysis of MARC 21 format (the most popular metadata schema for bibliographic records) goes further and sets up functional groups, such as search, identity, select, manage, process and classifies the underlying schema elements to these categories [27, 16, 49]. So by analysing the fields of the individual records, we can more precisely tell which aspects of the quality are good or bad.

These records are not only for registration and helping discovery of the materials, they are also the sources of additional researches in the Humanities. The catalogue contains lots of factual information, which are not available in other sources (or not in organised way), and therefore before the age of digitisation one could have found the printed catalogues of the most important collections (e.g. British Library, Library of Congress etc.) in the reading rooms of research institutions. In the past two decades several research projects attached existing library metadata to different types of full text datasets (optical character recognised or XML encoded versions), to provide additional facets for the analysis process such as personal or institutional names (creators, publishers), geographical information (places of publication), time span and so on.

Just a few examples: KOLIMO (Corpus of Literary Modernism)[1] uses TEI headers containing catalogue information as well as other metadata, for extracting literature and language features specific to a given time period, or to a particular author. OmniArt [58] is a research project, based on the metadata of Rijksmuseum (Amsterdam), the Metropolitan Museum of Arts (New York) and the Web Gallery of Art[2]. They collected 432,217 digital images with curated metadata (which is the largest collection of that kind) to run categorical analysis. Benjamin Schmidt uses

[1] `https://kolimo.uni-goettingen.de/index.html`
[2] `https://www.wga.hu/`

the HathiTrust[3] digital library and its metadata records to test machine learning classification algorithms, where he can compare the results with the Library of Congress subject headings available in the metadata records [56]. The common features of these project is that they use cultural heritage institutions' catalogue data as primary sources in their own research. It is self evident, that quality of those data might have effect on the conclusions of the research, and on the other hand it is beyond the responsibilities and possibilities of a researcher (or even a research group) to validate the records one by one, and fix them as needed.

This third use case of cultural heritage data become so frequent recently, that two years ago it lead to coining a new phrase: "collections as data". As the Santa Barbara Statement on Collections as Data [14] summarises: "For decades, cultural heritage institutions have been building digital collections. Simultaneously, researchers have drawn upon computational means to ask questions and look for patterns. This work goes under a wide variety of names including but not limited to text mining, data visualisation, mapping, image analysis, audio analysis, and network analysis. With notable exceptions [...], cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use. Thinking about collections as data signals an intention to change that." While collections as data movement emphasises the importance of re-usability of cultural heritage data, and we expect that this great and important movement will help organisations to think more about the scientific usage or their metadata,[4] their principles are focusing on access, and get rid of current barriers, however misses the aspects of quality. The quality assessment aspect we propose in this project would be a complementary element next to the other principles.

1.1. Metadata quality

> "We know it [i.e. metadata quality] when we see it, but conveying the full bundle of assumptions and experience that allow us to identify it is a different matter." (Bruce and Hillmann) [11].
>
> The (US) National Information Standards Organization (NISO) provides a definition for metadata, which is "structured information that describes, explains, locates, or otherwise represents something else." [48] The interesting thing in this definition is the list of verbs: describes, explains, locates, and represents. Metadata is not a static entity, it has multiple different functions and should be in context of other entities. That is in harmony with the famous the quality assurance slogan 'fitness for

[3] `https://www.hathitrust.org/`

[4] A 2016 report which analyses the usage of two important British cultural heritage collections mentions that "The citation evidence that is available shows a growing literature that mentions using EEBO [Early English Books Online] or HCPP [House of Commons Parliamentary Papers]", and "Shifts to humanities data science and data-driven research are of growing interest to scholars". [44]

purpose'. There are different definitions of the slogan, some of them are

- fulfilment of a specification or stated outcomes
- measured against what is seen to be the goal of the unit
- achieving institutional mission and objectives

From these definitions we can draw two important conclusions:

1) an object's quality is not an absolute value, it depends on the context of the object, what goal(s) the agents in the current context would like to achieve with the help of the object

2) the quality is a multi-faceted value. As the object might have different functions, we should evaluate the fulfilment's of them independently.

NISO's definition of metadata nicely fits into this framework, as it highlights the multi-faceted and contextual nature of metadata.

In an aggregated metadata collection such as Europeana, the main purpose of the metadata is to provide access points to the objects which the metadata describe (and stored remotely in the providing cultural heritage institutions, outside of Europeana). If the metadata stored in Europeana is of low quality or missing, the service will not be able to provide access points, and the user will not use the object.

As Bruce and Hillmann states, an expert could recognise if a given metadata record is "good" or "bad". What we would like to achieve is to formalise this knowledge by setting up the dimensions of the quality, and establishing metrics and measurement methods.

1.2. Metrics in the literature

In the literature of metadata quality assessment (see Appendix A) one can find a number of metric definitions. In this section I review some of them which proved to be relevant in my research.

Regarding to the cultural heritage context Bruce and Hillmann's above cited seminal paper ([11]) defines the data quality metrics. Palavitsinis in his PhD thesis [52] summarises them as follows:

Completeness: Number of metadata elements filled out by the annotator in comparison to the total number of elements in the application profile

Accuracy: In an accurate metadata record, the data contained in the fields, correspond to the resource that is being described

Consistency: Consistency measures the degree to which the metadata values provided are compliant to what is defined by the metadata application profile

Objectiveness: Degree in which the metadata values provided, describe the resource in an unbiased way, without undermining or promoting the resource

Appropriateness: Degree to which the metadata values provided are facilitating the deployment of search mechanisms on top of the repositories

Correctness: The degree to which the language used in the metadata is syntactically and grammatically correct

The same author – analysing the metadata quality literature focusing mainly on the Learning Object Repositories metadata – lists the following additional dimensions proposed by different authors: accessibility, conformance, currency, intelligibility, objectiveness, presentation, provenance, relevancy and timeliness. He also repeats the categorisation of Lee et al. [40] regarding to the quality dimensions:

Intrinsic Metadata Quality: represents dimensions that recognise that metadata may have innate correctness regardless of the context in which it is being used. For example, metadata for a digital object may be more or less 'accurate' or 'unbiased' in its own right,

Contextual Metadata Quality: recognises that perceived quality may vary according to the particular task at hand, and that quality must be relevant, timely, complete, and appropriate in terms of amount, so as to add value to the purpose for which the information will be used,

Representational Metadata Quality: addresses the degree to which the metadata being assessed is easy to understand and is presented in a clear manner that is concise and consistent,

Accessibility Metadata Quality: references the ease with which the metadata is obtained, including the availability of the metadata and timeliness of its receipt.

Zaveri Amrapali and her colleagues surveyed the Linked Data Quality literature in 2015 [66]. Their work became the most cited paper regarding to data quality. They investigated what quality dimensions and metrics were suggested by other authors and grouped individual metrics into the following dimensions:

Accessibility dimensions

Availability – the extent to which data (or some portion of it) is present, obtainable, and ready for use. The metrics this dimension are:

- A1 accessibility of the SPARQL endpoint and the server
- A2 accessibility of the RDF dumps
- A3 dereferenceability of the URI
- A4 no misreported content types
- A5 dereferenced forward-links

Licensing – the granting of permission for a customer to reuse a dataset under defined conditions.

- L1 machine-readable indication of a license
- L2 human-readable indication of a license
- L3 specifying the correct license

Interlinking – the degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources.

- I1 detection of good quality interlinks
- I2 existence of links to external data providers
- I3 dereferenced back-links

Security – the extent to which data is protected against alteration and misuse.

- S1 usage of digital signatures
- S2 authenticity of the dataset

Performance – the efficiency of a system that binds to a large dataset.

- P1 usage of slash-URIs
- P2 low latency
- P3 high throughput
- P4 scalability of a data source

Intrinsic dimensions

Syntactic validity – the degree to which an RDF document conforms to the specification of the serialization format

- SV1 no syntax errors of the documents
- SV2 syntactically accurate values
- SV3 no malformed datatype literals

Semantic accuracy – the degree to which data values correctly represent the real-world facts

- SA1 no outliers
- SA2 no inaccurate values
- SA3 no inaccurate annotations, labellings or classifications
- SA4 no misuse of properties
- SA5 detection of valid rules

Consistency – a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms

- CS1 no use of entities as members of disjoint classes
- CS2 no misplaced classes or properties
- CS3 no misuse of owl:DatatypeProperty or owl:ObjectProperty

- CS4 members of owl:DeprecatedClass or owl:DeprecatedProperty not used
- CS5 valid usage of inverse-functional properties
- CS6 absence of ontology hijacking
- CS7 no negative dependencies/correlation among properties
- CS8 no inconsistencies in spatial data
- CS9 correct domain and range definition
- CS10 no inconsistent values

Conciseness – the minimization of redundancy of entities at the schema and the data level

- CN1 high intensional conciseness
- CN2 high extensional conciseness
- CN3 usage of unambiguous annotations/labels

Completeness – the degree to which all required information is present in a particular dataset

- CM1 schema completeness
- CM2 property completeness
- CM3 population completeness
- CM4 interlinking completeness

Contextual dimensions

Relevancy – the provision of information which is in accordance with the task at hand and important to the users' query

- R1 relevant terms within metainformation attributes
- R2 coverage

Trustworthiness – the degree to which the information is accepted to be correct, true, real, and credible

- T1 trustworthiness of statements
- T2 trustworthiness through reasoning
- T3 trustworthiness of statements, datasets and rules
- T4 trustworthiness of a resource
- T5 trustworthiness of the information provider
- T6 trustworthiness of information provided (content trust)
- T7 reputation of the dataset

Understandability – the ease with which data can be comprehended without ambiguity and be used by a human information consumer

- U1 human-readable labelling of classes, properties and entities as well as presence of metadata
- U2 indication of one or more exemplary URIs
- U3 indication of a regular expression that matches the URIs of a dataset

- U4 indication of an exemplary SPARQL query
- U5 indication of the vocabularies used in the dataset
- U6 provision of message boards and mailing lists

Timeliness – how up-to-date data is relative to a specific task

- TI1 freshness of datasets based on currency and volatility
- TI2 freshness of datasets based on their data source

Representational dimensions

Representational conciseness – the representation of the data, which is compact and well formatted on the one hand and clear and complete on the other hand

- RC1 keeping URIs short
- RC2 no use of prolix RDF features

Interoperability – the degree to which the format and structure of the information conform to previously returned information as well as data from other sources

- IO1 re-use of existing terms
- IO2 re-use of existing vocabularies

Interpretability – technical aspects of the data, that is, whether information is represented using an appropriate notation and whether the machine is able to process the data

- IN1 use of self-descriptive formats
- IN2 detecting the interpretability of data
- IN3 invalid usage of undefined classes and properties
- IN4 no misinterpretation of missing values

Versatility – the availability of the data in different representations and in an internationalized way

- V1 provision of the data in different serialization formats
- V2 provision of the data in various languages

Some of these metrics are relevant only in Linked Data context (those which are LD technology specific, such as SPARQL endpoint or RDF dump). On the other hand there are lots of metrics which are useful for non-linked metadata as well. For example we will see in Chapter 2 that there is a tendency to add misinterpretable ad-hoc values into a placeholder ("+++EMPTY+++" to quote an extreme case) when the value is missing. 'V2 provision of the data in various languages' is similar concept than the multilinguality I'll describe in Chapter 3. Downloadable dumps are also very useful even it is not in a specific (e.g. RDF) format.

1.2.1. FAIR metrics

One of the main recent developments regarding to research data management was the formulation of FAIR principles. [64]. "The FAIR Principles provide guidelines for the publication of digital resources such as datasets, code, workflows, and research objects, in a manner that makes them Findable, Accessible, Interoperable, and Reusable." It became the starting point of many different projects which either implement the principles, or investigate further extensions. One of the is FAIRMetrics [65, 20]. It concentrates on the measurement aspects of the FAIR principles: how can we set up metrics upon which we can validate the "fairness" or research data.

The authors suggested, that good metrics in general should have the following properties:

- clear
- realistic
- discriminating
- measurable
- universal

There are 14 FAIR principles, and for each there is a metric. Each metric answers questions, such as 'What is being measured?', 'Why should we measure it?', 'How do we measure it?', 'What is a valid result?', 'For which digital resource(s) is this relevant?' etc.

The creators published the individual metrics as nanopublications and they are working on an implementation. Besides the metrics they defined 'Maturity Indicator tests' which are available as REST API backed by a Ruby based software called FAIR Evaluator. Maturity Indicators are an open set of metrics. Above the core set (which presented by the FAIR-Metrics), the creators invited the research communities to create their own indicators. As they emphasise: "we view FAIR as a continuum of 'behaviors' exhibited by a data resource that increasingly enable machine discoverability and (re)use."

The FAIRmetrics are as follows:

- F1: Identifier Uniqueness (Whether there is a scheme to uniquely identify the digital resource.)
- F1: Identifier persistence (Whether there is a policy that describes what the provider will do in the event an identifier scheme becomes deprecated.)
- F2: Machine-readability of metadata (The availability of machine-readable metadata that describes a digital resource.)

- F3: Resource Identifier in Metadata (Whether the metadata document contains the globally unique and persistent identifier for the digital resource.)
- F4: Indexed in a searchable resource (The degree to which the digital resource can be found using web-based search engines.)
- A1.1: Access Protocol (The nature and use limitations of the access protocol.)
- A1.2: Access authorization (Specification of a protocol to access restricted content.)
- A2: Metadata Longevity (The existence of metadata even in the absence/removal of data.)
- I1: Use a Knowledge Representation Language (Use of a formal, accessible, shared, and broadly applicable language for knowledge representation.)
- I2: Use FAIR Vocabularies (The metadata values and qualified relations should themselves be FAIR, for example, terms from open, community-accepted vocabularies published in an appropriate knowledge-exchange format.)
- I3: Use Qualified References (Relationships within (meta)data, and between local and third-party data, have explicit and 'useful' semantic meaning)
- R1.1: Accessible Usage License (The existence of a license document, for both (independently) the data and its associated metadata, and the ability to retrieve those documents)
- R1.2: Detailed Provenance (There is provenance information associated with the data, covering at least two primary types of provenance information: – Who/what/When produced the data (i.e. for citation); – Why/How was the data produced (i.e. to understand context and relevance of the data))
- R1.3: Meets Community Standards (Certification, from a recognized body, of the resource meeting community standards.)

Most of these metrics rather measure the data repository, than individual research data sets. In this thesis I do not work with research data, it is among my future plans, but it is good to note that FAIRmetrics does not cover classical metadata quality metrics (such as completeness, accuracy etc.), so even if it will have a robust implementation, there will be space left for future research on research (meta)data quality, and on the other hand some of these metrics are applicable for cultural heritage data (e.g. persistent identifiers would help the ingestion process of Europeana, so the *Identifier persistence* metric would be a useful indicator in this workflow).

1.2.2. Vocabularies for validating Linked Data

The domain of Linked Data (or semantic web) is based on 'Open World assumption', which means that objects (entities) and statements about them are separated, different agents could create a statement about an object. Practically it means that there is no concept as "record", since the object does not have clear boundaries. The traditional record based systems have schemas, which describe what kind of statements could be done about an entity. For example the Dublin Core Metadata Element Set consists of 15 metadata element. If we would like to record a colour of a book in this schema, we can not do it directly. Of course we can put this information into a semantically more generic field, such as "format", but then we will loose specificity, and colour will be stored together with other features such as size, dimensions etc. In Linked Data context the situation is different: we can easily introduce a new property, and create a statement, however we loose the control of the schema. We can not tell if the new property is valid or not.

To solve this problem W3C set up RDF Data Shapes working group "to produce a language for defining structural constraints on RDF graphs"[5]. One of the results came from this approach is Shapes Constraint Language (SHACL)[6]

SHACL defined a vocabulary (see Table 1.1) upon which one can create validation rules. It does not set metrics directly, but these constraint definitions are very useful building blocks of a data quality measurement system. The implementation of SHACL is based on Linked Data, but the definitions are meaningful in other contexts as well.

Within Europeana Data Quality Committee we plan to define frequently occurring metadata problems (or 'anti-patterns') with SHACL.

1.2.3. Organising issues per responsible actors

Christopher Groskopf who wrote a guide for data journalists how to recognise data issues [21] followed a different approach. He wrote a practical guide, not an academic paper, so he organised issues based on who could fix them. His main take-away messages are

- be skeptic about the data
- check it with exploratory data analysis
- check it early, check it often

[5] `https://www.w3.org/2014/data-shapes/charter`

[6] `https://www.w3.org/TR/shacl/`. We should note that there is another approach for the same problem: Shape Expressions (ShEx) available at `http://shex.io`.

Table 1.1.: Core constraints in SHACL

category	constrains
Cardinality	minCount, maxCount
Types of values	class, datatype, nodeKind
Shapes	node, property, in, hasValue
Range of values	minInclusive, maxInclusive, minExclusive, maxExclusive
String based	minLength, maxLength, pattern, stem, uniqueLang
Logical constraints	not, and, or, xone
Closed shapes	closed, ignoredProperties
Property pair constraints	equals, disjoint, lessThan, lessThanOrEquals
Non-validating constraints	name, value, defaultValue
Qualified shapes	qualifiedValueShape, qualifiedMinCount, qualifiedMaxCount

His categorisation is the following:

Issues that your source should solve

- Values are missing
- Zeros replace missing values
- Data are missing you know should be there
- Rows or values are duplicated
- Spelling is inconsistent
- Name order is inconsistent
- Date formats are inconsistent
- Units are not specified
- Categories are badly chosen
- Field names are ambiguous
- Provenance is not documented
- Suspicious numbers are present
- Data are too coarse
- Totals differ from published aggregates
- Spreadsheet has 65536 rows
- Spreadsheet has dates in 1900 or 1904
- Text has been converted to numbers

Issues that you should solve

- Text is garbled
- Data are in a PDF
- Data are too granular

- Data was entered by humans
- Aggregations were computed on missing values
- Sample is not random
- Margin-of-error is too large
- Margin-of-error is unknown
- Sample is biased
- Data has been manually edited
- Inflation skews the data
- Natural/seasonal variation skews the data
- Timeframe has been manipulated
- Frame of reference has been manipulated

Issues a third-party expert should help you solve

- Author is untrustworthy
- Collection process is opaque
- Data asserts unrealistic precision
- There are inexplicable outliers
- An index masks underlying variation
- Results have been p-hacked
- Benford's Law fails
- It's too good to be true

Issues a programmer should help you solve

- Data are aggregated to the wrong categories or geographies
- Data are in scanned documents

Groskopf's list is not a definition of general metrics, it is a catalogue of anti-patterns. It was created in reflection to the data journalism context, and it implies that – comparing to cultural heritage data – these project are smaller in both the number of contributors and the number of records. On the other hand, the sole purpose of these data is to be used in data analysis so during the data cleaning process the maintainer has more freedom than that of a librarian, who should keep in mind multiple data reuse scenarios. Despite of these differences cultural heritage projects also get inspirations from Groskopf's list.

1.2.4. Conclusion about the metrics

In the previous section I revised some of the metrics and approaches. This is not a comprehensive overview (for those who would like to read a general review of the metadata quality metrics I suggest the already quoted thesis of Palavitsinis [52]). What I wanted to show is that in different research areas or domains of activities there are quite different approaches for the measurement of metadata quality and detecting individual issues.

There are general metrics, such as completeness, format specific metrics, such as those ones for Linked Data that were collected by Amrapali or those I will discuss in Chapter 4 for MARC records. Some metrics measure data, but there are metrics which focusing on services which helps users to access data (such as existence of different API endpoints, or downloadable data dumps — we could label most of the FAIRmetrics into this category). In one of the early papers in metadata quality [59] Stvilia and his co-authors emphasized that the information quality (IQ) framework they created (which contains "typologies of IQ variance, the activities affected, a comprehensive taxonomy of IQ dimensions along with general metric functions, and methods of framework operationalization"), should be applied to a data source by selecting relevant IQ dimensions. In other words not all metrics are useful in all situation, we should select the appropriate one for each and every use case.

1.3. Research objectives

In this thesis I would like to answer the following questions:

Q1: What kind of quality dimensions are meaningful in the context of two different cultural heritage data sources: the collection of Europeana and MARC 21 format library catalogues.

Q2: How could it be implemented in a flexible way, so the solution should remain easily extensible to measure the same metrics on data sources in other formats.

Since Europeana could be qualified as Big Data (at least in the cultural heritage domain) two more questions arose regarding to scalability:

Q3: How can these measurement be implemented in scalable way?

Q4: How could Big Data analysis be conducted with limited computational resources?

1.3.1. The outline of this thesis

In Chapter 2 I describe the main metrics for Europeana. I also give an overview of the tool I developed for implementing the measurments. Chapter 3 describes a new set of metrics, *multilinguality* which measures how users with different language background can access Europeana's data. Chapter 4 concentrates on traditional library metadata, and shows the results of validation of 16 catalogues. Chapter 5 sheds light on the questions of flexibility: how the tool abstracts measurements in order to support different metadata schemas. Chapter 6 concentrates on resource

optimisation: how the tool (or other tools which uses the same underlying technique, namely Apache Spark) should be optimised for speed in a multi-tenant environment with limited resources. Finally Chapter 7 provides a conclusion and shows future plans.

Chapter 2.

Measuring completeness as metadata quality metric in Europeana[1]

Péter Király and Marco Büchler[2]

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 was different. In order to create effective services upon this data it is important to know the strengths and weaknesses, or in other words, the quality of these data. This chapter proposes a method and an open source implementation to reveal quality issues by measuring some structural features of these data, such as completeness, multilinguality, uniqueness, and record patterns.

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

2.1. 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

[1] This chapter has been first published as extended abstract in Digital Humanities 2017 Conference Abstracts (`https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf`) then as a full paper: [36]

[2] Péter Király created the experiments, the underlying software, and contributed to the text. Marco Büchler contributed to the text.

> the most difficult bugs to catch." – Felix Rau, German linguist on the consequence of metadata issues[3]

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[4], illustrate the importance of metadata:

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

(b) Several data providers are listed in Europeana's 'Institution' facet under multiple name variants (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 a user to select all variant forms when s/he wants to search for objects belonging to a particular organization?

(c) Without formalized and unified values in the '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 human readable and do not support any of the core functionalities of Europeana.

(e) In a multilingual environment the user would expect that s/he would 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 and are not resolved into a common entity.

The question is thus 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 purpose. When dealing with metadata quality it is relevant to clarify why metadata are important. In Europeana's case it is relatively straightforward in that it provides access points to digitized objects. If the features of a record make it impossible to find an object then its intended purpose is not met as the user cannot use an object they cannot access. One could then reasonably argue that the quality

[3] 18 Oct 2018, `https://twitter.com/fxru/status/1052838758066868224`
[4] http://europana.eu

of such a record is insufficient. The manual evaluation of each record, however, is not affordable for even a middle-size collection.

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

2.2. Background and foundations

Europeana collects and presents cultural heritage metadata records. The database at the time of this writing contains more than 58 million records in the Europeana Data Model (EDM) metadata schema from more than 3200 institutions[5] i. The organizations can send their data in EDM or in another metadata standard. Due to the variety of original data formats, cataloguing rules, languages and vocabularies, there are large differences in the quality of individual records, which heavily affects Europeana's service functionalities.

In 2015, a Europeana task force investigated the problem of metadata quality, and published a report (see [15]), however – as stated – 'there was not enough scope ... to investigate ... metrics for metadata quality ....' In 2016, a wider Data Quality Committee[6] (DQC) was founded and several experts on this committee from different domains (such as metadata theory, cataloguing, academic research, software development) came together to analyse and revise the metadata schema, discuss data normalization, run functional requirements analysis and define 'enabling' elements (answering questions such as 'What are the core functionalities of Europeana?' and 'Which metadata elements support them?'). DQC also built a 'problem catalogue', which is a collection of frequently occurring metadata anti-patterns (such as duplicate values, title field repeated as description, values for machine consumption in fields which were intended for human consumption, etc.) [26]. The questions of multilinguality were given special emphasis.

This current research is being conducted in collaboration with the DQC with the purpose of finding methods, defining metrics and building an open source tool called 'Metadata Quality Assurance Framework'[7] 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 are not limited to –

[5] Extracted from Europeana Search API.
[6] https://pro.europeana.eu/project/data-quality-committee
[7] http://144.76.218.178/europeana-qa/, source code and background information: http://pkiraly.github.io

MARC[8] and Encoded Archival Description[9]). The software is scalable to Big Data, as it is built to work together with the distributed file system of Apache Hadoop[10], the general, large-scale data processing engine Apache Spark[11], and the Apache Cassandra[12] database. One of the most important features of this approach is the capability to produce reports understandable to 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 then 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.[13]

The research discussed here questions 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 that it is impossible to measure every aspect of it both for theoretical reasons (for example current language detection methods do not work well with the short texts typically available in metadata records) and for practical reasons (such as limited resources). A number of structural features of the metadata record, however, are measurable and the outcome provides a 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'.[14] Approximation means in practice that the outcome should call for further scrutiny by metadata experts. It also implies that there is a fair chance that the tool cannot detect variances due to those errors that 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, then we can prioritize what needs to be fixed first and corrections to metadata can be planned in order of the importance of the problem. Since Europeana is an aggregator, corrections should be made at the information source itself, inside the database of the particular data provider. Better data supports more reliable functions, so by fixing weak records Europeana could build

[8] 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 ruleset 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.

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

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

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

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

[13] See http://pkiraly.github.io/2016/07/02/making-general/ and [35]

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

stronger services. Finding typical errors might also help improve the underlying metadata schema and its documentation (supposedly some of the errors occurred due to the language used in the schema documentation). In addition, during the measurement process examples of bad and good practice for certain metadata elements could be found and highlighted. Lastly high scoring metadata records could be used to propagate 'good metadata practices' or assist in the process of prototyping new services.

2.3. State of the art

The computational methods for metadata quality assessment emerged in the last decade in the cultural heritage domain ([11], [59], [51], [24]). The latest evaluation of the relevant work was conducted by [52]. The applied metrics in the domain of Linked Data (which has an intersection with the cultural heritage domain) are listed in [66]. While some papers defined quality metrics others suggested computational implementations. Nonetheless, they 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 [50] investigating 7 million, and [24] investigating 25 million records). The novelty of this 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[15] which is maintained by the members of the Digital Library Federation's Metadata Assessment group[16] and members of the DQC including the first author of this chapter.

2.4. Methodology

2.4.1. The EDM schema

An EDM record[17] consists of several entities. The core of the record is called the *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 schemas used

[15] http://zotero.org/groups/metadata_assessment
[16] https://dlfmetadataassessment.github.io/
[17] For EDM documentation, guidelines and other materials consult https://pro.europeana.eu/page/edm-documentation

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 contain descriptions 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 the *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 DQC is working on functional requirement analysis, in which we define the core functions starting from typical user scenarios (how the user interacts with the collection), and analyse which metadata elements support them [25]. For example, consider the user scenario of 'Cross-language recall': 'As a user, I want to search the Europeana collections in the language I am most comfortable with, and feel confident that I will receive relevant results irrespective of document language.' 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.

Since the definition of these enabling elements has not yet been harmonized with the purpose of measurement, DQC 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 'subject heading'-like elements (dc:type, dc:subject, dc:coverage, dcterms:temporal, dcterms:spatial)
- *Descriptiveness* – how well does the metadata describe of what the object is about
- *Searchability* – the fields most often used in searches
- *Contextualization* – the basis 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 displaying results at the portal
- *Re-usability* – for reusing the metadata records in other systems
- *Multilinguality* – for multilingual aspects, to be understandable for all European citizens

At the time of this writing this model examines only the existence of the fields, it does not check if the content matches what type of data is expected – a task which will be implemented during the next research phrase.

2.4.2. Measuring

For every record, features are extracted or deducted which somehow relate 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[18],
- *uniqueness of the descriptive fields* (title, alternative title, description)[19],
- *multilinguality*[20],
- *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.

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 that of the subsets. Currently there are three kinds of subsets: datasets that are records ingested together during the same process (they were usually handled by the same transformation chain when Europeana received them from the data providers); records belonging to

[18] This measurement is experimental in the Europeana context as a proof of concept. The full problem catalogue will be formally described with the Shapes Constraint Language ([38]).

[19] For the underlying theory see [2]. The method applied here is different than as described in the thesis.

[20] See [13] and [57]

Table 2.1.: 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

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 aggregated metrics are calculated including the completeness of structural entities (suchas 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 schemes. In the first approach, 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 of the sub-dimensions get 2. The equation is

$$c_{sub-dimensions} = \frac{\sum_{i=1}^{d} score_i \times weight_i}{\sum_{i=1}^{d} weight_i} \tag{2.1}$$

with d as the number of sub-dimensions, $score_i$ as the proportion of availability of the fields belonging 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 the biasing effect of extreme values. Sometimes there are more than one hundred or even a thousand field instances in a single record which would have too much effect on the score, so the tool normalizes them according to table 2.1.

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

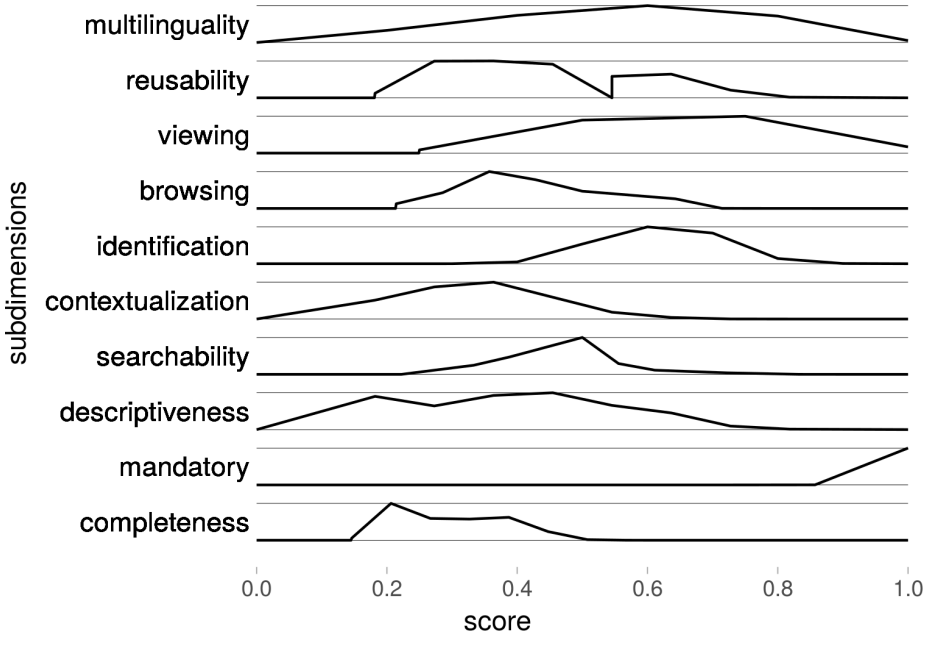

Figure 2.1.: The distribution of sub-dimension and 'simple completeness' scores

equation is

$$c_{cardinality} = \frac{\sum_{i=1}^{d} norm(cardinality_i) \times weight_i}{\sum_{i=1}^{d} weight_i} \tag{2.2}$$

with d as the number of fields, $cardinality_i$ as the cardinality of a field, $norm()$ as the normalizing function (see table 2.1) 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 a higher weight (so it is more important) than the second one:

$$c_{compound} = \frac{c_{sub-dimensions} + (0.4 \times c_{cardinality})}{1.4} \tag{2.3}$$

2.4.3. 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 significant difference in performance between the two, but in other scenarios the Hadoop File System could be a better choice). The record level analyses are written in Java, using the Spark API[21]. It provides automatic and configurable multithreading, so the tool can make use of the available resources of the environment effectively (either if it is a single machine with a multicore processor or a 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 makes use of 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[22] and in the Scala implementation of Spark[23]. It reads the CSV files generated in the previous phase, and produces CSV and JSON files for storing the results 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 the dataset it can process. In terms creating statistics for the whole Europeana dataset this is insufficient. For this reason, Scala on Spark is used for all top level aggregations. Scala's statistical capabilities are not that rich, however, so it does not produce all the metrics that 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.[24] The technical details of the workflow is documented in [34]. 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 were also used at the same time for other research and development projects, so making the calculations resource-effective was

[21] Metadata quality assessment library: https://github.com/pkiraly/ metadata-qa-api, Europeana specific extension: https://github.com/ pkiraly/europeana-qa-api, Apache 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.

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

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

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

Table 2.2.: 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

an important software design constraint.

The data source for this calculation is a snapshot of Europeana data. The first snapshot was created at the end of 2015, which contains 46 million records, 1747 datasets and 3550 data providers[25] (extracted from Europeana's OAI-PMH service). During the project's lifetime additional snapshots have been 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).[26] DQC aims to introduce a monthly update cycle, so the time span between the updates of the Europeana production database and the refreshing of the data quality dashboard should not be more than one month.

2.5. Results

2.5.1. 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. While they correlate by the Pearson's correlation coefficient of 0.59, their shape and ranges are different. Because of the nature of the 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 2.2 and figure 2.2.

There are data providers where all (in some cases more than ten thousand) records have the same scores: they have a uniform structure. Be-

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

[26] 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.

Figure 2.2.: Distribution of completeness calculations

cause one simple score is not enough to establish this, 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 subdimension a data provider has five distinct values (from 0.4 to 0.8) almost evenly distributed while one of the best collections (of this category) is almost homogeneous: 99.7% of the records have the same value: 0.9 (even the remaining 0.3% has 0.8). This means that the corresponding fields[27] 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

[27] 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.

timespan and 40% have *concept* entities). Only the mandatory technical elements appear in every record. 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 the 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.

2.5.2. Multilinguality

DQC has recently published details regarding the results of the multilinguality calculation (see [13] and [37]), so this section presents only 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 a 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

These metrics were 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 2.3 and 2.4 and figure 2.3).

Table 2.4 reflects that only 20% of the records have two or more languages per property in the Provider Proxy. After the enhancement process injects external contextual information (about agents, concepts, places and timespans) from multilingual data sources such as DBpedia and other sources into the Europeana records, the overall multilinguality became higher. Not only are the number of fields with two or more language values increased, but the number of records without any language annotation also decreased.

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,

Table 2.3.: 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

Table 2.4.: 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%)

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 a more thorough picture of language usage.

2.5.3. Uniqueness

One might recall the example of similar titles mentioned at the beginning of this chapter. To find those records we should calculate the uniqueness of the values. Uniqueness is a positive value in those fields which should 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 an ideal case multiple records will 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 have required more resources than were available for this research, three fields were selected for this task: title, alternative title, and description. In calculating this score a modified version of Solr's

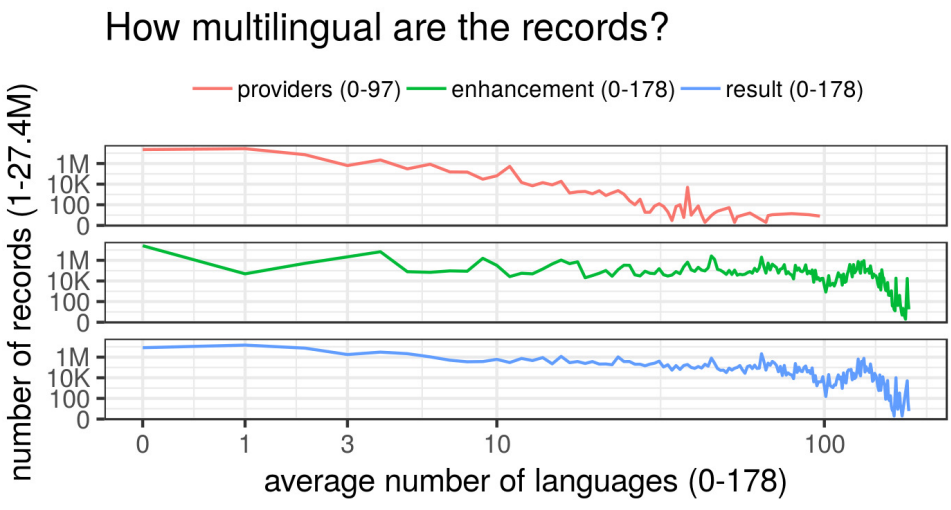

Figure 2.3.: Multilinguality

Table 2.5.: Uniqueness categories by frequency

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

relevancy scoring was applied:

$$score(t_f, v_f) = log\left(1 + \frac{t_f - v_f + 0.5}{v_f + 0.5}\right) \tag{2.4}$$

$$uniqueness_f = \left(\frac{score(t_f, v_f)}{score(t_f, 1.0}\right)^3 \tag{2.5}$$

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

As seen in figure 2.4 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 2.5 displays the category boundaries for these three fields:

Table 2.6.: 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

The result of the categorization is shown in table 2.6. While the absolute majority of the records in regards to all three fields do contain unique values, there are still millions of records with low scores for one or another field, and 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 calculate an average of the result, we find that there are 25 million records with unique values in all available fields while on the other side of the scale only 3.62% of the records are in the lowest category. This 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 a higher chance to be found by a search term.

From the Solr index one could extract the most frequent terms. Along with the "photograph" example in the introduction there are many 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 even a general descriptive term ("Porträtt", "Château", "Plakat", "Rijksmonument"). It would require further investigation to filter out those frequent terms which appear in records especially where the other descriptive fields also lack a necessary level of uniqueness. The tool described here provides a solid basis for such an investigation.

2.5.4. Record patterns

What fields make up a typical record? In other words: what fields do data providers actually use? Record patterns are the typical field collocations. Since the completeness measurement counts the existence of all the fields, a map-reduce based analysis could extract these patterns. 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 were too many similar patterns worthy of grouping together in order to analyze them effectively. A similarity algorithm was therefore applied for clustering the patterns. All patterns were first represented by a string containing zeros

and ones. First, all the fields of a collection were collected and sorted by a standard field order. Each field was then categorized into one of three categories: mandatory fields, important fields (those fields which appeared in a sub-dimension) and non-important 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 fields. The similarity is calculated by the Jaro-Winkler algorithm. In the visualization (as you can see in figure 2.5) the clusters are displayed by default, and the user needs to click to display the patterns belonging to a cluster. The table is ordered by the number of records, so the more typical records are on the top. If the field is only available in some records within the cluster, it is grayed (the color is proportional with the number of records). By default the page does not display patterns occuring in less than 1% of the records.

Thus far two quality problems were revealed by the use of record patterns. The first problem covers those records which had only a 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, and 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 would not be able to discover these records by using facets, due to the lack of descriptive information about the object. The second problem is structural homogeneity: each record in some collection always has the same set of fields. There are 906 such data providers in Europeana, but fortunately most of them are relatively small collections, 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). The problem with such a record is that it contains generic fields instead of specific ones (for example it does not make distinctions among conceptual, spatial and temporal subject headings, and puts different contextual information into dc:type or dc:subject).

2.6. Further work

Europeana is currently working on its new ingestion system called Metis[28], and it will able to integrate the tool described here. It is currently planned that when a new record-set arrives for import the measurement will be

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

launched automatically. The Ingestion Officer can then check the quality report and share both the output and general conclusions with data providers who can then either change their transformation rules or hopefully fix issues with their metadata records if possible.

There are other metrics in addition to the calculation models that were discussed in this chapter, and we are planning to compute them in the near future (e.g. accuracy, information content, timeliness, existence of known metadata anti-patterns). Much of the related literature suggests calculating a top level score, which summarizes all metrics into one final score that characterizes the record's metadata quality. This could be achieved by weighting the metrics or applying machine learning algorithms, such as Principal Component Analysis [31]. It was mentioned previously that the current completeness calculation approach only confirms the existence of a field. The next step on this research front is to extend this model with content evaluation of the relevant fields according to the User Scenarios analysis ([25]).

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 there is 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 conducted by the Digital Public Library of America (which is similar to Europeana regarding to its purpose and its metadata schema). This approach failed partly because there was not enough usage data available at time the research was conducted, however, the proposed method sounds promising, and if Europeana has log files it would be worthwhile to run an experiment.

Other future plans include defining the problem catalogue with W3C's Shapes Constraint Language [38] and publishing the results as linked data fitted to the Data Quality Vocabulary Ontology [63].

The proposed method could also be used in data collections using other metadata schemas, such as MARC based library catalogues[29], EAD-based archival collections,[30] and others.

2.7. Conclusion

This research sought to rethink the relationship between functionality and metadata schema (together with the DQC) and a framework was

[29]Since MARC has lots of strict content related rules, and EDM only has a few, there is a significant distance between the approaches followed in the two projects.

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

implemented that proved successful in measuring structural features that correlate with metadata issues. The user of the framework is able to select between low and high quality records. According to our original hypothesis, structural features such as field existence and cardinality correlated with metadata quality, and this ultimately proved to be true. In addition, this work also extended the volume of records analyzed by introducing big data tools that were not mentioned previously in the literature.

Although this research focused on a particular dataset and metadata schema, the applied method is based on generalized algorithms so it could also be applied to other data schemas. Several Digital Humanities studies (some examples: KOLIMO (Corpus of Literary Modernism)[31], [58], [55]) based on schema defined cultural databases. The research process could also be improved by finding the weak points of the sources, making the conclusions more reliable, and – reflecting on Felix Rau's tweet quoted at the beginning of this chapter – 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 feedback, to Christina Harlow and Zaveri Amrapali for general inspiration, and to Felix Rau for the motto and to GWDG for supporting the research.

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

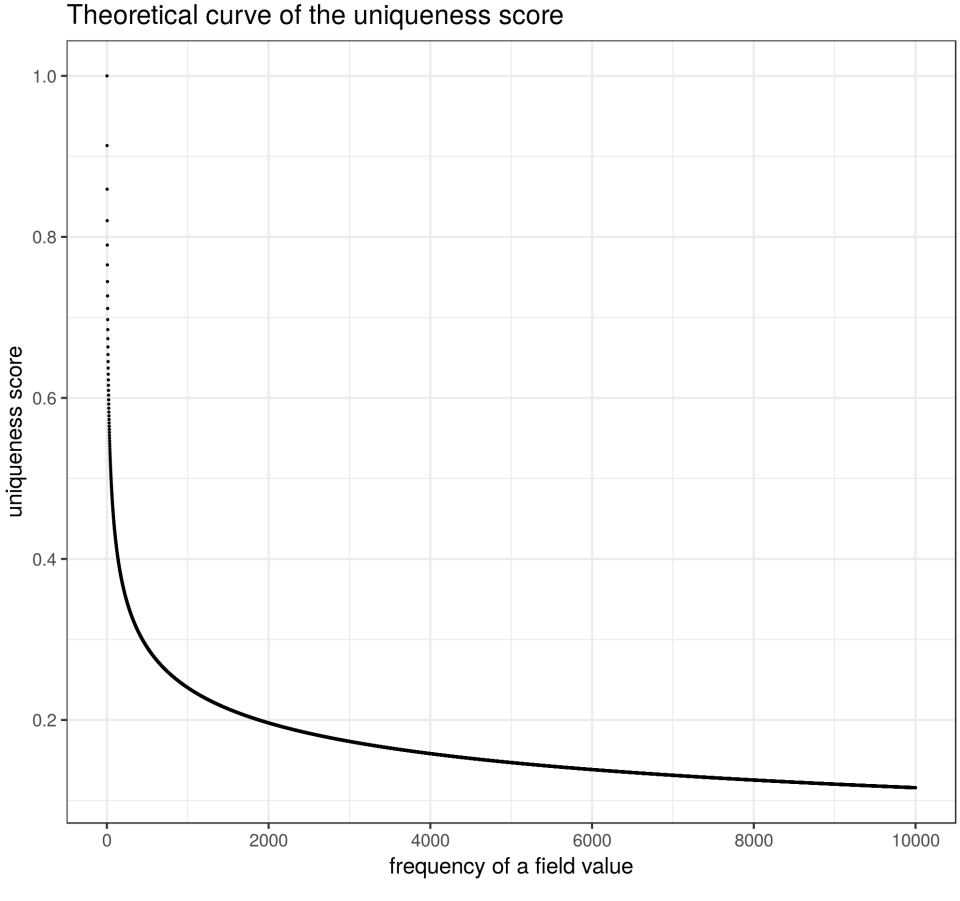

Figure 2.4.: Theoretical curve of uniqueness score. As frequency of terms gets higher, the uniqueness score get radically smaller towards zero.

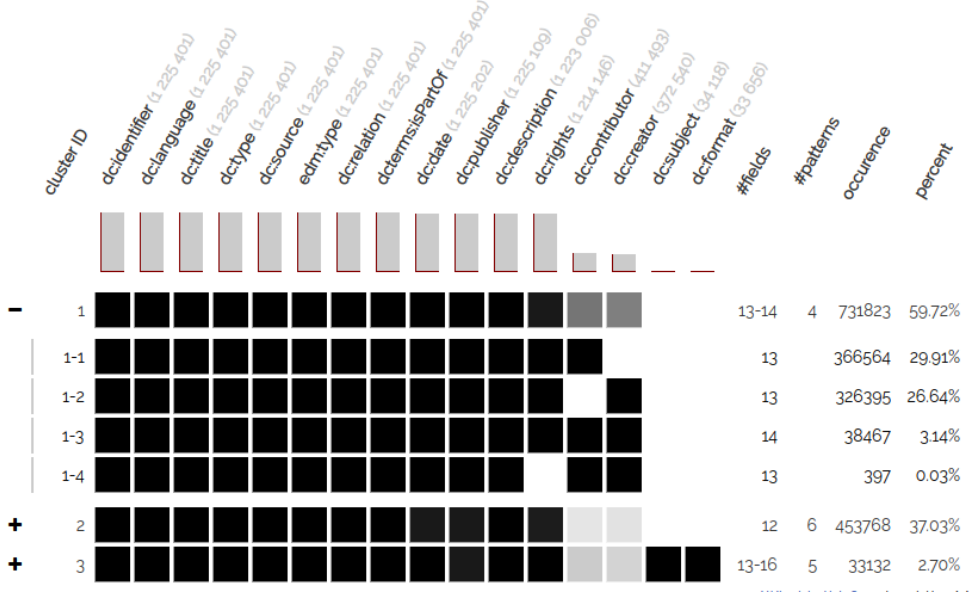

Figure 2.5.: Clustered record patterns. The first line represents a cluster of similar patterns. The next four lines are the patterns belonging to the cluster. The top gray bar represents the frequency of fields in the whole collection.

Chapter 3.

Evaluating Data Quality in Europeana: Metrics for Multilinguality[1]

Péter Király, Juliane Stiller, Charles Valentine, Werner Bailer, and Nuno Freire[2]

Abstract: Europeana.eu aggregates metadata describing more than 50 million cultural heritage objects from libraries, museums, archives and audiovisual archives across Europe. The need for quality of metadata is particularly motivated by its impact on user experience, information retrieval and data re-use in other contexts. One of the key goals of Europeana is to enable users to retrieve cultural heritage resources irrespective of their origin and the material's metadata language. The presence of multilingual metadata description is therefore essential to successful cross-language retrieval. Quantitatively determining Europeana's cross-lingual reach is a prerequisite for enhancing the quality of metadata in various languages. Capturing multilingual aspects of the data requires us to take into account the full lifecycle of data aggregation including data enhancement processes such as automatic data enrichment.The chapter presents an approach for assessing multilinguality as part of data quality dimensions, namely completeness, consistency, conformity and accessibility. We describe the measures defined and implemented, and provide initial results and recommendations.

[1]This chapter has been published as [37]. Previous reports of this research have been published as [57, 13].

[2]Péter Király contributed to algorithms, created the underlying software and contributed to the text. Juliane Stiller, Charles Valentine, Werner Bailer, and Nuno Freire contributed to the algorithms and to the text.

3.1. Introduction

Europeana.eu[3] is Europe's digital platform for cultural heritage. It aggregates metadata describing more than 50 million cultural heritage objects from a wide variety of institutions (libraries, museums, archives and audiovisual archives) across Europe. The need for high-quality metadata is particularly motivated by its impact on search, the overall Europeana user experience, and on data re-use in other contexts such as the creative industries, education and research. One of the key goals of Europeana is to enable users to find the cultural heritage objects that are relevant to their information needs irrespective of their national or institutional origin and the material's metadata language.

As highlighted in the White Paper on Best Practices for Multilingual Access to Digital Libraries [1], most digital cultural heritage objects do not have a specific language, i.e., as they are not in textual form, and can only be searched through their metadata, which is text in a particular language. The presence of multilingual metadata description is therefore essential to improving the retrieval of these objects across language spaces. Quantitatively determining Europeana's cross-lingual reach is a prerequisite for enhancing the quality of metadata in various languages.

In this chapter, we present multilinguality as a measurable component of different data quality dimensions: completeness, consistency, conformity and accessibility. We capture data quality by defining and implementing quality measures along the full data-aggregation lifecycle, taking also into account the impact of data enhancement processes such as semantic enrichment. The model the data is represented in, namely the Europeana Data Model (EDM)[4], is also a key element of our work.

In the next section, we present data quality frameworks, dimensions and criteria that are commonly referred to in the context of data quality measurement. Section 3 describes how multilingual metadata is presented in Europeana's data model and the data quality dimensions we use are also introduced. In Section 4, we describe the implementation of the different measures as well as the calculation of the scores. Section 5 describes first results and measures that were taken to improve metadata along the different quality dimensions and the first recommendations we have been able to identify based on the results from the metrics. We conclude this chapter with an outline of future work.

[3] http://www.europeana.eu/
[4] http://pro.europeana.eu/edm-documentation

3.2. State of the art

Addressing data quality requires the identification of the data features that need to be improved and this is closely linked to the purpose the metadata is serving. Libraries have always highlighted that bibliographic metadata enables users to find material, to identify an item and to select and obtain an entity [27]. Based on this, Park [53] expands functional requirements of bibliographic data to discovery, use, provenance, currency, authentication and administration, and related quality dimensions. The approach to metadata assessment for cultural heritage repositories presented in [9] also starts from use for a specific purpose. While the work mentions the issue of multilinguality, it does not propose specific metrics to measure it.

Different sets of dimensions have been proposed for classifying metadata quality measures. Bruce and Hillmann [11] define the following measures for quality: completeness, accuracy, provenance, conformance, logical consistency and coherence, timeliness and accessibility, multilinguality is not addressed in this work. The existing works that consider multilinguality assign it to different quality dimensions, depending on the purpose of the measurement. Zaveri et al. [66] propose dimensions for quality assessment for linked data. Completeness is listed as an intrinsic criterion, while multilinguality is covered by versatility, which is considered a representational criterion. The ISO/IEC 25012 standard [30] defines a data quality model with 15 characteristics, discriminating between inherent and system-dependent ones, but putting many of the criteria in the overlap between the two classes. Completeness is defined as an inherent criterion, while accessibility and compliance (conformity) are in the overlapping area. Multilinguality could be seen as being compliant to providing a certain number of elements in a certain number of languages, and as enabling access to users who are able to search and understand results in certain languages. Radulović et al. [54] propose a metadata quality model for linked data, and define multiple languages as an indicator for the quality dimension availability, i.e., can it be accessed by users with the requirement to get the data in a specific language. Ellefi et al. [17] propose a taxonomy of features for profiling RDF datasets, of which one part discusses quality. They define representativity as one dimension of quality in their model, under which they see versatility (including multilinguality) as one measure. To the best of the authors' knowledge, the only resource which actually measured multilingual features in metadata is [62], cited in [52]. Albertoni et al. [3] also include multilinguality in a scoring function, although in the context of importing other linked data vocabularies.

It becomes apparent that while multilinguality is considered in some

works, it is usually not treated as a separate quality dimension, but rather as part of other criteria or dimensions existing at quite different levels in different quality models. We use the measures based on the frequency of language tags described in [62] as a basis, and following the conclusion from the literature, consider multilinguality in the context of different quality dimensions. Our work started with the development of metrics to measure the multilingual quality of metadata in Europeana within the EU-funded project Europeana DSI-2[5]. A first iteration of a multilingual saturation score that counted language tags across metadata fields in the Europeana collections as well as the existence of links to multilingual vocabularies was introduced by Stiller and Király [57]. The score was extended in [13] by including measures that define multilinguality as part of different quality dimensions.

3.3. Approach

Firstly, to determine the multilingual degree of metadata across several quality dimensions, we have to understand the different ways multilingual information is expressed in Europeana's data model. Secondly, the structure of multilingual data informs the criteria and metrics that enable us to measure multilinguality across several metadata quality dimensions.

3.3.1. Multilingual information in Europeana's metadata

Multilinguality in Europeana's metadata has two perspectives: concerning the language of the object itself, and the language of the metadata that describes this object. First, the described cultural object, insofar as it is textual, audiovisual or in any other way a linguistic artefact, has a language. The data providers are urged to indicate the language of the object in the *dc:language* field in the Europeana Data Model (EDM) in this way: `<dc:language>de</dc:language>`. If used consistently and in accordance with standards for language codes, this information could then be used to populate a language facet allowing users to filter result-sets by language of objects. The language information is essential for users who want to use objects in their preferred language. Second, the language of metadata is essential for retrieving items and determining their relevance. Metadata descriptions are textual and therefore have a language. Each value in the metadata fields can be provided with a language tag (or language attribute). Ideally, the language is known and indicated by this tag for every literal in each field. If several language

[5] `https://pro.europeana.eu/project/europeana-dsi-2`

tags in different languages exist, the multilingual value can be considered to be higher. For instance, consider as an example this data provided by an institution:

```
<#example> a ore:Proxy ;  # data from provider
  dc:subject "Ballet",    # literal
  dc:subject "Opera"@en . # literal with language tag
```

The first *dc:subject* statement is without language information, whereas the second tells us that the literal is in English. Multilingual information is not only provided by the institutions but can also be introduced by Europeana. Europeana assesses metadata in particular fields to enrich it automatically with controlled and multilingual vocabularies as defined in the Europeana Semantic Enrichment Framework.[6] As shown in the following example, the dereferencing of the link (i.e., retrieving all the multilingual data attached to concepts defined in a linked data service) allows Europeana to add the language variants for this particular keyword to its search index.

```
<#example> a ore:Proxy ; edm:europeanaProxy true ;
  # enrichment by Europeana with multilingual vocabulary
  dc:subject <http://data.europeana.eu/concept/base/264> .

<http://data.europeana.eu/concept/base/264> a skos:Concept ;
  # language variants are added to index
  skos:prefLabel "Ballett"@no, "Ballett"@de, "Balé"@pt,
                 "Baletas"@lt, "Balet"@hr, "Balets"@lv .
```

The record now has more multilingual information than at the time of ingestion into Europeana. The labels are added to the search index and this particular record can be retrieved with various language variants of the term *ballet*. The different language versions from multilingual vocabularies are likely to be translation variants. This distinction between the provided metadata and the metadata created by Europeana needs to be taken into account for measuring multilinguality as defined in Section 4.

3.3.2. Multilinguality as a facet of quality dimensions

For measuring multilinguality, we identify four quality dimensions: completeness, consistency, conformity and accessibility. Each of these dimensions assesses multilinguality from a different perspective.

Completeness. Completeness is a basic quality measure, expressing the number (proportion) of fields present in a dataset, and identifying

[6] https://docs.google.com/document/d/1JvjrWMTpMIH7WnuieNqcTOzpJAXUPo6x4uMBj1pExOY/edit

non-empty values in a record or (sub-)collection. For a fixed set of fields completeness is thus straightforward to measure, and can be expressed as the absolute number or fraction of the fields present and not empty. However, the measure becomes non-trivial when data is represented using a data model with optional fields (that may e.g., only be applicable for certain types of objects), or with certain fields for which the cardinality is unlimited (e.g., allowing zero to many subjects or keywords). These characteristics apply to EDM. In such cases the measure becomes unbounded, and a few fields with high cardinality may outweigh or swamp other fields.

In the context of measuring multilingual completeness, the metric is twofold. First, the concept of completeness can be applied to measuring the presence of fields with language tags. This measure of multilinguality must be seen in relation to the results of measuring completeness. Only fields both present and non-empty can be said to have or lack language tags and translations. A record which is 80% complete can still reach 100% multilingual completeness if all present and non-empty fields have a language tag. Second, the completeness measure can reflect the presence of the *dc:language* field that identifies the language of the described object.

Consistency. Consistency describes the logical coherence of the metadata across fields and within a collection. With regard to multilinguality, the dimension assesses the variety of language values in the *dc:language* field and the language tags that specify the language in a given field. Consistent values should be used to describe the same language.

In Europeana, the consistency measure for the *dc:language* field is mainly relevant for the language based facet. The more consistent languages are expressed, the more useful language facets become. Ideally, inconsistencies in expressing languages through language codes should be fixed through normalization (see Section 3.5).

Conformity. Conformity refers to the accordance of values to a given standard or a set of rules. Here, the language values in the *dc:language* field and the language tags in any given field can be assessed with regard to their conformity to a given standard such as ISO-639-2[7]. The conformity measure for the *dc:language* field influences the usefulness of a language facet.

Accessibility. Accessibility describes the degree to which multilingual information is present in the data, and allows us to understand how easy or hard it is for users with different language backgrounds to access information. So far, Europeana has little knowledge about the distribution of linguistic information in its metadata – especially within single records.

[7] `https://www.loc.gov/standards/iso639-2/php/code_list.php`

Table 3.1.: Dimensions, criteria and measures for assessing multilinguality in metadata.

Dimension	Criteria	Measures
Completeness	Presence or absence of values in fields relating to the language of the object or the metadata	• Share of multilingual fields to overall fields • Presence or absence of *dc:language* field
Consistency	Variance in language notation	• Distinct language notations
Conformity	Compliance to ISO-639-2	• Binary or share of values that comply or not comply
Accessibility	Multilingual Saturation	• Numbers of distinct languages • Number of language tagged literals • Tagged literals per language

To quantify the multilingual degree of data and measure cross-lingual accessibility, the language tag is crucial. The more language tags representing different languages are present, the higher is the multilingual reach. Resulting metrics can be scaled to the field, record and collection levels. In practical terms, the accessibility measure serves to gauge cross-language recall and entity-based facet performance. To summarize: with regard to multilinguality, we identified the dimensions, quality criteria and measures presented in Table 3.1.

3.4. Operationalizing the metrics for multilinguality

The different metrics for the assessment of multilinguality in metadata are implemented in the metadata quality assurance framework of Europeana.[8] Implementation of the metrics requires a good understanding of the data aggregation workflows which can contribute to the increase of multilingual labels (such as machine learning and natural language processing techniques for language detection, automatic tagging, or semantic enrichment) in the metadata. Before being displayed in Europeana, the source data goes through several levels of data aggregation. EDM doesn't represent the different data processes that take place at each of these levels but captures the different data outputs. EDM allows us to distinguish between (a) values provided by the data provider(s) and (b)

[8] `http://144.76.218.178/europeana-qa/multilinguality.php?id=all`

information (automatically) added by Europeana (for instance by semantic enrichment) by leveraging on the proxy mechanism from the Object Re-use and Exchange (ORE) model. The metadata provided to Europeana are captured under a `ore:Proxy` while the metadata created by Europeana are captured under a `edm:EuropeanaProxy`. The examples in Section 3.3.1 demonstrate how the mechanism enables the representation of resources in the context of different aggregations of the same resource [29]. Any implementation of quality measures, and in particular of multilingual ones, needs to take into account this distinction. For instance, the score for accessibility might be higher if we only consider the Europeana proxy where a value was enriched with a multilingual vocabulary (e.g. DBpedia) leading to more language tags than initially provided by an institution.

3.4.1. Measurement workflow

The process for assessing the multilinguality of metadata is based on the metadata quality assurance framework, which has four phases:

1. Data collection and preparation: the EDM records are collected via Europeana's OAI-PMH service[9], transformed to JSON where each record is stored in a separate line, and stored in Hadoop Distributed File System[10].

2. Record-level measurement: the Java applications[11] measuring different features of the records run as Apache Spark jobs, allowing them to scale readily. The process generates CSV files which record the results of the measurements such as the number of field instances, or complex multilingual metrics.

3. Statistical analysis: the CSV files are analyzed using statistical methods implemented in R and Scala. The purpose of this phase is to calculate statistical tendencies on the dataset level and create graphical representations (histograms, boxplots). The results are stored in JSON and PNG files.

4. User interface: interactive HTML and SVG representations of the results such as tables, heat maps, and spider charts. We use PHP, jQuery, d3.js and highchart.js to generate them.

[9] `https://pro.europeana.eu/resources/apis/oai-pmh-service`. Our client library: `https://github.com/pkiraly/europeana-oai-pmh-client/`.

[10] We made two data snapshots available: 2015 December (46 million records, 392 GB): `https://hdl.handle.net/21.11101/EAEA0-826A-2D06-1569-0`, 2018 March (55 million records, 1,1 TB): `http://hdl.handle.net/21.11101/e7cf0a0-1922-401b-a1ae-6ec9261484c0`

[11] Source code and binaries: `http://pkiraly.github.io/about/#source-code`.

Since we intend to measure multilingual saturation of the provided and enriched metadata separately, we perform measurements for the following objects: the provider (source) created proxy S, the Europeana created proxy E (containing enrichments) and the whole EDM record O. Each proxy has several properties, such as *dc:title*, *dc:subject*, etc. These properties might have multiple instances. Each instance might have either a string only, a tagged literal or a URI. We suppose that if the URI is resolvable then a contextual object was created, so we check only whether a contextual entity exists within the same object. If we found one, we use its *skos:prefLabel* property to check whether it is a string or tagged literal.

For each property we define the following quantities: nt_p, the number of tagged literals of a property p, l_p, the list of language tags of p and d_p, the set of distinct language tags of p, thus $|d_p| \leq |l_p|$.

We calculate the basic scores for both proxies. We denote the four resulting values for the proxies as tp_S, tp_E, the number of tagged properties in provider and Europeana proxies, tl_S, tl_E, the number of tagged literals, dl_S, dl_E, the set of distinct language tags, and nl, the number of distinct languages.

On object level, these values are aggregated from the proxies by summation/union, i.e., $tp_O = tp_S + tp_E$, $tl_O = tl_S + tl_E$, $dl_O = dl_S \cup dl_E$, and $nl_O = |dl_O|$.

Note that $l_O \leq (l_S + l_E)$, as the provider and Europeana proxy typically contain overlapping languages. In many practical cases, it is likely that $l_O = \max(l_S, l_E)$.

3.4.2. Deriving metrics from basic scores

In this section, we discuss how we derive metrics from these scores that relate to the different quality dimensions concerning multilingual saturation.

Completeness The number of languages present can be used to measure completeness, in particular, when the resulting score is also checked against a target value. A basic metric is the fraction of properties and literals that have language tags, i.e., $fp_S = \frac{tp_S}{|p \in S|}$ and $fl_S = \frac{tl_S}{\sum_{p \in S} l_p}$, where $p \in S$ is the set of properties of S. The same calculation can be applied to E and O. The languages per property for the proxies and the object are defined as the normalized number of languages, i.e., $lpp_S = \frac{l_S}{tp_S}$ (and analogously for E and O).

Consistency We assess consistency of the language tags used throughout the dataset, such as standard vs. non-standard codes, two vs. three letter codes for the same language, short vs. extended language tags, etc. In order to determine a metric for consistency of language tags, we need external information that groups synonymous language identifications. The Languages Name Authority List (NAL) published in the European Union Open Data Portal[12] provides synonyms for languages. This vocabulary was used for language normalization as reported in Section 3.5.

We denote the set of languages as $L = \{l_1, \ldots, l_n\}$, and the language tag for language l_i in vocabulary v as $t^v_{l_i}$. Examples for v could be the two letter tags from ISO-639-1 or the different three letter tags from ISO-639-2/T and ISO-639-2/B. For each of the languages l_i we can thus define a set of tags T_i. For the standards, it is well defined which tags denote the same language, and using the syntactic rules of extended language tags those can be included as well (e.g., associate "en-gb" with "en"). In addition there may be custom tags, (e.g., "british english") .

We can then determine the consistency as

$$cs_S = \frac{1}{l_S} \sum_{l_i \in dl_S} 1 - \frac{|\{t_{Sj} | j = 1, \ldots, tl_S\} \cup T_i| - 1}{\sum_{k=1}^{|T_i|} |\{t_{Sj} | j = 1, \ldots, tl_S\} \cup t^k_i|}, \tag{3.1}$$

where t_{Sj} is the language tag of literal j in S, and $\{t_{Sj} | j = 1, \ldots, tl_S\}$ is the set of language tags of the literals. This score is 1 if a single language tag is used for all literals, and close to 0 if each literal uses a different language tag. For E and O the score can be determined analogously.

Conformity We assess whether the language tags used are from a standard set of tags, such as one of the parts of ISO-639. Similar as for consistency, we define a set of possible standard tags of a language l_i, denoted as T'_i. We determine a conformity metric as the fraction of language tags from this set.

$$cf_S = \frac{1}{l_S} \sum_{l_i \in dl_S} \frac{\sum_{j=1}^{tl_S} |t_{Sj} \cup T'_i|}{tl_S}, \tag{3.2}$$

where $t_S j$ is the language tag of literal j in S. For E and O the score can be determined analogously.

[12] `https://open-data.europa.eu/en/data/dataset/language`

Table 3.2.: Results for the measures in the different dimensions.

Dimension	Measures	Results
Completeness	• Share of multilingual fields to overall fields • Presence or absence of *dc:language* field	• Measureable for each field per dataset • 25,5% of datasets (35,14% of records) have no *dc:language* field
Consistency	• Distinct language notations	• Over 400 distinct language notation across all fields
Conformity	• Binary or share of values that comply or not comply	• See Table 3.3 for statistics on conformity with ISO-639
Accessibility	• Numbers of distinct languages • Number of language tagged literals • Tagged literals per language	• median of 6.0 (mean 41.2±53.65) per object • median of 15.0 (mean 73.3±111.17) per object • median of 1.6 (mean 2.3±3.46) per object

Accessibility The richness of metadata in a particular language is a metric for how easily the object can be found and interpreted in that language. Next to the number of distinct languages and the number of language tagged literals, we use the average number of tagged literals per language as a metric, and determine it as $tll_S = \frac{tl_S}{l_S}$ (an analogously for E and O).

3.5. Results

The metrics are implemented in the metadata quality assurance framework using a snapshot of the data from March 2018. In table 3.2, we report on some of the results from various dimensions and describe some of the developments they initiated. The data quality issues observed in the results lead to a series of best practices beneficial for further improvement.

Completeness. With regard to the presence of the *dc:language* field, the measure indicates that 905 out of 3,548 datasets have no value in the *dc:language* field, which shows the field is missing. On a record level, 64.86% of the records have a *dc:language* field.[13] Furthermore, we can determine the share of multilingual fields across all records for

[13] http://144.76.218.178/europeana-qa/frequency.php

Table 3.3.: Presence of ISO-639 codes in the values of the dc:language field.

Total values in the Europeana dataset	33,070,941
Total values already normalized (ISO-639-1, 2 letter codes)	23,634,661
Total values already normalized (ISO-639-3, three letter codes)	4,831,534

given fields. For example, 97.05% of all records have a *dc:title* field. The great majority of these fields have no language indicated for their values. Approximately a fourth of the *dc:title* fields have values with a language. Titles in German, English, Dutch, Polish and Italian contribute to more than half of the *dc:title* values that have a language tag. The metric allows to investigate the share of fields with multilingual tags across specific datasets or Europeana as a whole. It is also possible to compare different versions of the Europeana dataset to track progress and improvements. In a multilingual context, the completeness of the metadata is improved by the presence of languages for metadata elements supporting literals (*dc:subject, dc:description, dc:title*), or by the presence of links to contextual entities with multilingual features.

Consistency. Next to measuring the consistency in the language tag notation, we specifically measured the consistency in the *dc:language* field. This revealed that over 400 different language variants are present in the field. To ensure consistent use of language codes over the whole collection, they need to be normalized and standards applied within the *dc:language* field. This element must be provided when a resource is of *edm:type* TEXT and should be provided for these other types (AUDIO, IMAGE, VIDEO, 3D). Identifying the absence of language is also needed to properly assess the degree of multilinguality. We therefore recommend the use of the ISO 639-2 code for non-linguistic content (i.e. "zxx").

Conformity. After determining the heterogeneity of values in *dc:language* (dimension: consistency), we normalize the values in this field. *Dc:language* values are predominantly normalized in ISO-639-1 or ISO-639-3, but, in contrast, values nevertheless sometimes occur in natural language sentences that cannot be processed automatically. We also find language ISO codes without their reference to the ISO standard in use, or references to languages by their name. A language normalization operation was implemented consisting of a mix of operations, comprising cleaning, normalization and enrichment of data. Table 3.3 presents some general statistics about the presence of ISO-639 codes in the values of *dc:language* in the Europeana dataset. The metric helps us to design further language normalization rules which in turn can be used to improve the results of the quality measures. Tackling the heterogeneity of languages tags in other fields is still an open issue that needs to be tackled in future.

Accessibility. As noted earlier, our approach to measuring multilingual saturation in metadata allows us not only to measure the data's quality as it is provided by contributing institutions, but also provides us with insight into the effectiveness of Europeana's data enhancement processes, such as semantic enrichment. The measures for accessibility allow us to determine the number of distinct language tags per dataset or specific fields revealing which languages are covered and can be exploited for display and retrieval. For example, the Europeana collection after applying its automatic data enhancement workflow to its datasets has a median of 6 distinct languages per object where the maximum of distinct languages in an object is 182. Per object, there are 15 language tagged literals (median) with 14.5% or the records do not have any tagged literals and one object having as many as 62997 tagged literals. Delving into datasets, we can determine the amount of objects with particular language tags per field, as well as whether these language tags were coming from providers' data or are added by Europeana automatically. The results enable metadata experts to determine the multilingual reach of a dataset on field level and allow them to develop strategies for increasing the multilingul saturation. Being able to track progression over time by comparing different snapshots of the data is another valuable asset of the framework.

In summary, the results obtained for the dimensions above focus on the multilingual quality of the metadata with the sole objective to improve the accessibility of the cultural heritage objects available in Europeana.

3.6. Conclusion and future work

In this chapter we present our approach for assessing the multilingual quality of data in the context of Europeana. This approach is the result of a long term research activity of Europeana, providing essential conclusions for the establishment of a reliable multilingual quality measurement for its services and data providers. The measures for multilinguality are embedded into the data dimensions of completeness, consistency, conformity, and accessibility. Results of these measures allow Europeana to define and implement language normalization rules and several recommendations for data providers.

We identify several potential improvements on the quality measures, which should be further elaborated in future iterations of this activity at Europeana. We also conclude that improvements of the metrics can be achieved if they consider more the needs of users providing data to Europeana or re-using it for building their own applications. Refining visualization reports will help interpreting the measurements and to adjust our metrics. For instance, in order to get a comprehensive view of the

quality of data, the different metrics will need to be presented together (e.g. multilinguality on top of completeness) so that the interrelation between the different metrics is made visible.

The metrics proposed in this chapter are potentially applicable to a wider range of applications, beyond providing multilingual access to cultural assets, as stated in the Strategic Research Agenda for Multilingual Europe 2020[14]. One other important application is research data, for which multilinguality may also be relevant. The FAIR principles [64] include findability and accessibility by both humans and machines — for which multilinguality is one component. We intend to publish the metrics in a way that can be consumed by third parties interested in the Europeana data, as well as applying them to their data. The recently published W3C Data Quality vocabulary[15] is a good candidate for a machine-readable representation of our metrics and the measurement results.

Acknowledgments.

This work was partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

[14] http://www.meta-net.eu/sra/
[15] `https://www.w3.org/TR/vocab-dqv/`

Chapter 4.

Validating 126 million MARC records

Abstract: The chapter describes the method and results of validation of 14 library catalogues. The format of the catalog record is Machine Readable Catalog (MARC21) which is the most popular metadata standards for describing books. The research investigates the structural features of the record and as a result finds and classifies different commonly found issues. The most frequent issue types are usage of undocumented schema elements, then improper values in places where a value should be taken from a dictionary, or should match to other strict requirements.

4.1. Introduction

How should a book be described properly? This question has a long past (and an even longer future) with several proposed methods which evolved over time. In the current epoch in the history of cataloguing or in other words bibliographic control we see the end of a period, and the start of a new one. There are different conflicting proposals on the table regarding what should be the new big thing, but there is a consensus on that from the middle of 60-es up until now the dominant record format of the book descriptions was MARC, MAchine Readable Cataloging developed and maintained by the Library of Congress [42]. MARC is a format and also a semantic specification. It was invented – after an investigative period started at the end of 1950s – at the middle of the 1960s (at the age of punch cards) in a collaborative effort of different American libraries, led by Henriette Avram.[1] At that time the available information storage space was much less than it is nowadays, so the information should be compressed, therefore one of the main technical features of MARC is that wherever a piece of information could be described by an element of a closed list of terms this path is chosen. The record contains abbreviated

[1] See MARC's early history in [7]

forms, while the standard describes the abbreviated terms in detail. It makes the human understanding of MARC difficult in its native form, but makes the machine readability and thus validation easy. Theoretically at least. The problem is that during the decades while the basic structure of MARC remained the same, MARC continued to grown into a giant standard, with a number of such small or big dictionaries (which sometimes are externally developed and maintained by other organizations, such as the content classification schemes). Roy Tennant, in his famous (infamous for some), manifesto-like article [61] pictures the situation with a colourful sentence „There are only two kinds of people who believe themselves able to read a MARC record without referring to a stack of manuals: a handful of our top catalogers and those on serious drugs." Most of the open source tools for handling MARC concentrate on the structure, and take less care about the semantics, maybe because it would require a huge effort to make the standard itself machine readable in order to make such a tool aware of the meaning of the abbreviations.

Closing into the end of the MARC life cycle, it would be the appropriate time to examine some of the catalogues published under open licenses, and check their quality. This chapter examines 126 million[2] MARC records from 16 different library organizations whether they match the structural requirements of the standard. In order to achieve this goal the author created an open source software application written in Java which implements the structural rules of the MARC21 Bibliographic Description as Java classes.[3] The rule set in a machine readable way could be exported in Avram specification conformant JSON format,[4] so other tools could reuse it.

4.2. Why it important to validate metadata?

I would not claim that metadata plays the most important rule in using the digitised text or audiovisual media. The large text corpus is much more effective than metadata when it comes to searching. However there are some other fields where metadata is very useful. [46] showed that a successful data mining approach on large corpora of digitised journals should use metadata as well. According to [47] one of the current research topics in natural language processing is to reuse the contextual information provided by the metadata of the text. There are several examples when different digital humanities research uses metadata and full text together to reveal new facts (among others [32], [56] and [39]). [10] calls

[2] 126 638 140 to be exact
[3] https://github.com/pkiraly/metadata-qa-marc
[4] http://format.gbv.de/schema/avram/specification

attention to the effects of metadata biases: "because researchers have to rely on metadata to organise and navigate large corpora, there may be a significant number of relevant but essentially 'invisible' documents.". "Collections as data"[5] is a recent LAM movement, it aims to "foster a strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research and teaching". Their statement sheds light to the importance of metadata: "Trustworthy collections as data should include open, robust metadata" [14]. They collect different use cases which includes services for researcher backed by metadata.

Library catalogues became widely reused. MARC is a well documented format, there are different open source and commercial tools to work with, some of the biggest catalogues are openly accessible and reusable, the records are meticulously curated and usually contain contextual information, all of these makes them easily usable data sources for different researches. The reuse aspect makes the effects of metadata issues bigger, since they not just break Ranganathan's rules[6], but they effect scientific conclusions. This research is not looking for the "perfect metadata" [8], but aims to reveal the improvable parts of catalogues.

4.3. Introduction to MARC

While this chapter can not give a full encounter of MARC, it just provides the most important features, which helps in the understanding of the resulting quality measurement.

An example record (excerpt):

[5] https://collectionsasdata.github.io/
[6] https://en.wikipedia.org/wiki/Five_laws_of_library_science

```
01136cnm a2200253ui 4500
001 002032820
005 20150224114135.0
008 031117s2003    gw            000 0 ger d
020   $a3805909810
100 1 $avon Staudinger, Julius,$d1836-1902
      $0(viaf)14846766
245 10$aJ. von Staudingers Kommentar zum ... /
      $cJ. von Staudinger.
250   $aNeubearb. 2003$bvon Jörn Eckert
260   $aBerlin :$bSellier-de Gruyter,$c2003.
300   $a534 p. ;.
500   $aCiteertitel: BGB.
500   $aBandtitel: Staudinger BGB.
700 1 $aEckert, Jörn
852 4 $xRE$bRE55$cRBIB$jRBIB.BUR 011 DE 021
      $p000000800147
```

The above example is a widely accepted representation of a MARC record however MARC is stored differently in MARC files. It has a semi-binary format, it uses some delimiter characters and fixed length fields to separate content parts.

The first line is the leader (sometimes abbreviated as LDR), it is a fixed length field and a component of individual "portions". One should split the content this way (here '|' characters represent the boundaries between the portions, the first two lines denote character positions):

```
0               1               2
01234 5 6 7 8 9 0 1 2345 6 7 8 9 0 1 2 3
01136|c|n|m| |a|2|2|0025|3|u|i| |4|5|0|0
```

Using the positions as the key, the meaning of the individual portions should be read as

- `LDR/0-4`: Record length. '01136' – is a number padding with zeros (max. value: 99999) denoting the length of the record, so this record is 1136 byte long.
- `LDR/5`: Record status. 'c' is a dictionary term, means "Corrected or revised"
- `LDR/6`: Type of record: 'n' is not among the defined types, this is an error
- `LDR/7`: Bibliographic level. 'm' means "Monograph/Item"
- ...

In this document I will use the term 'subfield' for these information (the standard does not have a term for this).

Different library materials require different descriptive element sets, however in MARC – maybe due to the necessity of compressing information –

there is not a single "record type" field. Instead its value comes from the combination of two subfields of the Leader: the Type of Record, and the Bibliographic level (see Table 4.1). Several other elements of the schema depend on the value of this combination, however – as we will see – there are examples where the combination returns an undefined value.

Leader and fields 001-008 are so called control fields. Most of them are structurally similar to the Leader. The interpretation of 008 is dependent from the record type as defined by the combination mentioned above. Let's see the split of 008.

```
0           1           2                3
012345 6 7890 1234 567 8901 2 3 4567 8 9 0 1 2 3 4 567 8 9
031117 s 2003      gw                   0 0 0   0   ger   d
......|.|....|....|...|____|_|_|____|_|_|_|_|_|_|_|...|.|.
```

Here the first five (in position 0-17: different dates and place of publication) and the last three subfields (35-39: language, is the record modified?, cataloging source) are common for all types (denoted by dots), but the in-between part (18-34, denoted by underscores) are type-specific (for a book: illustrations, target audience, form of item, nature of contents, government publication, conference publication, Festschrift, index, literary form, biography – all encoded) and does not only provide the meaning, but the internal structure as well (the number of portions and their positions)! This structure is quite fragile, one single character deletion from these fields would break the structure and mix the meaning of a position.

From 010 up until the end of the record are 'datafields'. Their internal structure is different from that of the control fields. They have two indicators and one or more subfields. The indicators contain a single character to represent dictionary terms – usually they are qualifiers of the subfields. The subfields start with a '$' character, then comes a character long subfield code, and finally the value of the subfield. The value could be

- a code, or
- a literal value, which might be
  - a free text, or
  - a dictionary term, or
  - a string match to a fixed format (e.g. yymmdd), or
  - a combination of fixed format string and dictionary terms (e.g. for 045's subfield 'a' – conventionally abbreviated as 045$a in the string "d7n6" 'd7' means B.C. 299-200, 'n' means A.D. '900-999', and 6 stands for the 60-es of that century)[7], or

[7] http://www.loc.gov/marc/bibliographic/bd045.html

Table 4.1.: Record type

	Type of record (LDR/05)	Bibliographic level (LDR/06)		type	Form of material (006/00)	
a	Language material	a c d m	Monographic component part Collection Subunit Monograph/Item	Books	a	Language material
a	Language material	b i s	Serial component part Integrating resource Serial	Continuing Resources	s	Serial/Integrating resource
t	Manuscript language material			Books	t	Manuscript language material
c d i j	Notated music Manuscript notated music Nonmusical sound recording Musical sound recording			Music	c d i j	Notated music Manuscript notated music Nonmusical sound recording Musical sound recording
e f	Cartographic material Manuscript cartographic material			Maps	e f	Cartographic material Manuscript cartographic material
g k o r	Projected medium Two-dimensional nonprojectable graphic Kit Three-dimensional artifact or naturally occurring object			Visual Materials	g k o r	Projected medium Two-dimensional nonprojectable graphic Kit Three-dimensional artifact or naturally occurring object
m	Computer file			Computer Files	m	Computer file/Electronic resource
p	Mixed materials			Mixed Materials	p	Mixed materials

– combination of fixed positions and dictionary terms

Both the field and the subfield can be repeatable or non-repeatable.

4.3.1. The validation tool

In order to validate MARC records I have developed a software. The core of the software is a data model, which records the whole MARC 21 Bibliographic documentation as a set of Java classes. In the model the main units are the fields (*DataFieldDefinition* class), which have the code of the tag (such as 245), it's label ("Title Statement"), cardinality (repeatable or non repeatable), the URL of the definition at the Library of Congress page (`https://www.loc.gov/marc/bibliographic/bd245.html`). Then come the definition of the indicators, each having label, list of possible codes and their label. The last part of the field definition is the list of subfields: code, label, cardinality. The standard provides notes for obsolete elements, the Java class stores them as historical codes and subfields. Whenever it was possible I also recorded the BIBFRAME 2.0 equivalences based on the MARC 21 to BIBFRAME 2.0 Conversion Specifications[8]. BIBFRAME names holds meaning (e.g. "responsibilityStatement" versus MARC's 245$c, which is a language neutral notation) which could be uses in exporting data, because they don't contain spaces, so could be processed without problems in different software environments and self-describing. Since naturally there is no BIBFRAME equivalent for all MARC element, I have created a similar machine-readable tag. Some elements in MARC has a specially encoded value (e.g. the $6 subfield in lots of field, which records linkage between parallel elements), so the model let us to attach content parsers. Several subfields should contain a term from a controlled dictionary and some of the subfields have formal rules to check whether they are valid or not. The tool contains all these dictionaries and validator classes have been implemented to check against these rules. Some rules are outside of MARC such as ISBN and ISSN rules. These identifiers are composite of a sequence of numbers, where the last one is a check digit, it should be equal of a result of a sequence of computations with all of the remaining numbers. It is also possible to connect external online tools to validate specific MARC element. I made some experiments with the automatic Universal Decimal Classification (UDC) analyser service developed by Attila Piros[9]. The experience showed, that these kind of tools could be part of the validation process if they are fast enough, otherwise it is more advisable to extract data from the records (in this case the subfields containing UDC

[8] https://www.loc.gov/bibframe/mtbf/
[9] `http://piros.udc-interpreter.hu/`

classification numbers) and run the UDC analysis asynchronously. As part of the FRBR works Tom Delsey created a mapping between the 12 functions and the MARC elements [16]. The Java model also built in this mapping.

The tool works together with another Java library, Marc4j[10]. With the help of this library my tool can read from binary MARC and from MARCXML formats. In regards to parallel processing with Apache Spark, the tool can read from a special MARC binary format where the records are separated with line breaks.[11]

As a side effect, those knowledge built into the tool makes it useful for different other tasks, not just validation. One can use for formatting records, extracting information, index with Apache Solr, etc.

Since it took long time to build this model, I thought it would be useful to make it exportable, so other MARC related projects could use it as a machine-readable MARC specification. Jakob Voß introduced Avram JSON schema[12] to provide a language for creation of machine readable metadata specification. Due to the implementation of this schema the tool can export MARC into Avram schema.

4.3.2. Addressing elements - MARCspec

In the process of MARC validation it is important that one should be able to address specific parts of the record. For XML format this purpose is fulfilled by XPath, a W3C standard. For JSON there is not such a standard, but Stefan Gössner proposed JSONPath[13] for this purpose. We saw that `245$a` is a conventional way to address subfield 'a' of field 245, however for the less trivial uses cases (see them below) there are no similar conventions. Carsten Klee, the librarian of Zeitschriftendatenbank (Berlin) proposed MARCspec, a common MARC record path language[14], and that is what the tool implemented. Here are some MARCspec expressions:

- `260` – field
- `245^2` – the second indicator of a field
- `700[0]` – the first instance of a field
- `245$c` – a subfield
- `245$b{007/0=\a|007/0=\t}` – subfield 'b' of field '245', if character with position '0' of field 007 equals 'a' OR 't'.

[10] `https://github.com/marc4j/marc4j`
[11] See details at `http://pkiraly.github.io/2018/01/18/marc21-in-spark/`.
[12] `http://format.gbv.de/schema/avram/specification`
[13] http://goessner.net/articles/JsonPath/
[14] http://marcspec.github.io/MARCspec/marc-spec.html

- `020$c{$q=paperback}` – subfield 'c' if subfield 'q' equals to 'paperback'.

The tool extends this concept with two more things. We saw that most of the positions of control fields are type specific. Karen Coyle suggested a naming convention[15] handling this situation, so following that instead of '008/33' (which has 5 different type specific definitions) the software displays '008/33 (tag008book33)' which locates a single definition. The other convention is to map all fields and subfields to self-descriptive labels which could be used in displaying records or indexing with Solr.[16]

4.3.3. Versions

A very peculiar feature makes the interpretation of MARC difficult. As we saw there are competing proposals and practices for the bibliographical description, and MARC by design was created to support different content standards. The current version supports Anglo-American Cataloging Rule 2 (AACR2) [4] and different versions of International Standard Bibliographic Description (ISBD) [28]. In some aspects these are top level standards, and different countries adapt them to their local customs and practices. One consequence was that MARC itself was also localized and now there are about 50 different (international, national, and consortial) MARC versions. The different versions introduce new fields, delete or overwrite existing fields, or change the semantic granularity (e.g. in MARC21 the author's name should be recorded in one schema element, while the Hungarian HUNMARC distinguishes family and given names). On the other hand MARC itself is an evolving standard, there are new, deleted and changed elements every year. Finally, fields and subfields with 9 in their code are reserved for local usage, i.e. every library might have its own locally defined field/subfield set which are not part of the standard (with the exception of field 490[17]).

There are two big problems with versions

1. There is no schema element in the standard to record the MARC version
2. The MARC versions and local extensions are not always properly documented.

Without proper documentation these fields could not be understood, and thus validated. The validation tool should report it, but can not decide

[15] http://kcoyle.net/rda/elementslist.txt
[16] http://pkiraly.github.io/2017/09/24/mapping/
[17] https://www.loc.gov/marc/bibliographic/bd490.html

Listing 4.1: Subfield definition in Java

```
// core MARC21
setSubfieldsWithCardinality(
  "a", "International␣Standard␣Book␣Number", "NR",
  "c", "Terms␣of␣availability", "NR",
  "q", "Qualifying␣information", "R",
  ...
);
// obsolete
setHistoricalSubfields(
  "b", "Binding␣information␣(BK,␣MP,␣MU)␣[OBSOLETE]"
);
// version specific extension
putVersionSpecificSubfields(
  MarcVersion.DNB,
  Arrays.asList(
    new SubfieldDefinition(
      "9", "ISBN␣mit␣Bindestrichen", "R"
)));
```

if an undocumented element is correct or not, because the requirements are not clear.

The software has the following approaches to handle documented versions. The individual schema elements are represented by Java classes which have similar properties to their schema element pairs, e.g. a *DataFieldDefinition* has *Indicators* with *Codes*, and *Subfields* such as the standard's data fields. The MARC21 standard has special notes about the changing of the standard, and the model records them as *historicalCodes* or *historicalSubfields*. The same technique works for version specific codes and subfields. If a version does not extend a field, but introduces a new one or overwrites an existing one, a new *DataFieldDefinition* class could be created in the version's dedicated name space, so it could not be mixed with the core MARC21 implementation. The user can specify the supposed version with *–marcVersion [version]* parameter. The software tries to find the definition in its name space, and if does not find it (which means that particular field was not overwritten in that version), it checks if the core field definition has any version specific definition. If such a definition is found the software validates the element against that, otherwise it uses the default MARC21 definition. To illustrate the definition part, Listing 4.1 is a short example for 020 (ISBN) field definition.

First *setSubfieldsWithCardinality()* defines the core subfields. The parameters of it are set of triplets, in which the first element is the code of the subfield, the second is the description, the third is the cardinality, where "NR" denotes non-repeatable, "R" denotes repeatable sub-

Listing 4.2: Detailed report

```
./validator [parameters] [file]

recordId,MarcPath,type,message,url
010000178,900,field: undefined field,900,""
010000178,008/33 (tag008book33),invalid value, \
  "␣",https://www.loc.gov/marc/.../bd008b.html
```

fields. *setHistoricalSubfields()* defines the obsolete subfield 'b' (the first parameter is the subfield code, the second is the note MARC21 standard provides). Finally *putVersionSpecificSubfields()* defines '9' as a locally defined subfield for the German national bibliography. In this example it is a definition of a new field, which is not defined in MARC21, but this method can be used to overwrite an existing subfield.

At time of writing this chapter the following versions are defined (fully or partially)

- *MARC21*, Library of Congress MARC21
- *DNB*, the Deutsche Nationalbibliothek's MARC version
- *OCLC*, the OCLC's MARC version (partially implemented)
- *GENT*, fields available in the catalog of Gent University (Belgium)
- *SZTE*, fields available in the catalog of Szegedi Tudományegyetem (Hungary)
- *FENNICA*, fields available in the Fennica catalog of Finnish National Library

4.4. Record validation

4.4.1. Validating individual records

The tool has a command line interface (*./validator*) which iterates over one or more MARC or MARCXML records. There are two kinds of output: one which reports all issues with its record identifier (see Listing 4.2), and a summary, which reports similar issue types together without individual ids (Listing 4.3). Both contain URL of element definition if available.

We already observed that with *–marcVersion* parameter the user can specify the MARC version, while *–defaultRecordType* could be used to step in as a substitution if the record type is unknown.

Listing 4.3: Summarized report

```
./validator --summary [file]

MarcPath,type,message,url,count
900,field: undefined field,900,"",3
```

4.4.2. Results

This research covered the evaluation of 14 catalogues. They are[18] (with their abbreviation, download location, formats and license):

- *bay*: Bibliotheksverbundes Bayern[19], a union catalog of Bavarian libraries
- *bzb*: Bibliotheksservice-Zentrum Baden Würtemberg[20], a union calatogue of Baden-Würtemberg libraries
- *col*: Columbia University Library[21]
- *cer*: Heritage of the Printed Book Database of Consortium of European Research Libraries (CERL)[22]
- *dnb*: Deutsche Nationalbibliothek[23]
- *gen*: Universiteitsbibliotheek Gent[24]
- *har*: Harvard University Library[25]
- *loc*: Library of Congress[26]
- *mic*: University of Michigan Library[27]
- *nfi*: Fennica – the Finnish National Bibliography provided by the Finnish National Library[28]
- *ris*: Répertoire International des Sources Musicales[29]
- *sfp*: San Francisco Public Library[30]

[18] There are some more downloadable catalogues listed at https://github.com/pkiraly/metadata-qa-marc#datasources.)

[19] https://www.bib-bvb.de/web/b3kat/open-data MARCXML format, CC0 license.

[20] https://wiki.bsz-bw.de/doku.php?id=v-team:daten:openaccess:swb. MARCXML format, CC0.

[21] https://library.columbia.edu/bts/clio-data.html MARC21 and MARCXML format, CC0 license.

[22] There is no public download link. I received the catalog from the courtesy of Marian Lefferts (CERL), Alex Jahnke and Maike Kittelmann (SUB).

[23] http://www.dnb.de/EN/Service/DigitaleDienste/Datendienst/datendienst_node.html (note: it is not a direct link, you have to register and contact with librarians to get access to the downloadable dataset). MARC21 and MARCXML format, CC0 license.

[24] https://lib.ugent.be/info/exports Aleph Sequential format, ODC ODbL license.

[25] https://library.harvard.edu/open-metadata MARC21 format, CC0 license.

[26] https://www.loc.gov/cds/products/marcDist.php. MARC21 (UTF-8 and MARC8 encoding), MARCXML formats, open access.

[27] https://www.lib.umich.edu/open-access-bibliographic-records. MARC21 and MARCXML formats, CC0 license.

[28] http://data.nationallibrary.fi/download/. MARCXML, CC0 license.

[29] https://opac.rism.info/index.php?id=8&id=8&L=1. MARCXML, RDF/XML, CC-BY license.

[30] https://archive.org/ MARC format, CC0 for Public Domain Dedication

- *sta*: Stanford University[31]
- *szt*: Szegedi Tudományegyetem Klebelsberg Kuno Könyvtára[32]
- *tib*: Leibniz-Informationszentrum Technik und Naturwissenschaften Universitätsbibliothek (TIB)[33]
- *tor*: Toronto Public Library[34]

Table 4.2.: Number of records in the catalogs (in millions)

bay	bzb	cer	col	dnb	gen	har	loc	mic	nfi	ris	sfp	sta	szt	tib	tor
27.3	23.1	6.0	6.7	16.7	1.8	13.7	10.1	1.3	1.0	1.3	0.9	9.4	1.2	3.5	2.5

4.4.3. Validation

The following issue types were detected in each of the catalogues (see Table 4.3 and 4.4 for detailed distribution).

On the record level it appeared that the 'linkage' between fields could be invalid or ambiguous. Linkage is a special MARC feature, using subfield $6, containing a "data that links fields that are different script representations of each other", mainly used for transcription of foreign language titles.[35]. *Ambiguous linkage* (R1) occurs when the link's target is unclear, *invalid linkage* (R2) occurs when the link itself is missing or its target is not existing in the record. Sometimes *type error* (R3) occurs: the values mentioned in Table 4.1 are missing, invalid or their combination is invalid.

Control subfield's *invalid code* (C1) denotes the case when a control field's code is outside of the provided dictionary, while an *invalid value* (C2) occurs when the code is not a dictionary term, but should match some rule. For example 008/00-05 represents the date the record was created as six digits matching the "yymmdd" (year, month, day) pattern. "993006" is an invalid value, because the middle part "30" could not be interpreted as a month. The software reports "Invalid content: '993006'. Text '993006' could not be parsed: Invalid value for MonthOfYear (valid values 1 - 12): 30".

[31] There is no public download link. I received the catalog from the courtesy of Philip E. Schreur. MARC21 format.

[32] There is no public download link. I received the catalog from the courtesy of Károly Kokas. MARCXML format, CC0.

[33] https://www.tib.eu/de/die-tib/bereitstellung-von-daten/katalogdaten-als-open-data/. (no download link, use OAI-PMH instead) Dublin Core, MARC21, MARCXML, CC0.

[34] https://opendata.tplcs.ca/. 2.5 million MARC21 records, Open Data Policy

[35] https://www.loc.gov/marc/bibliographic/ecbdcntf.html

Table 4.3.: The percentages of records with issues

	bay	bzb	cer	col	dnb	gen	har	loc	mic	nfi	ris	sfp	sta	szt	tib	tor
all issues	100.0	100.0	2.8	90.4	13.9	40.8	100.0	30.5	80.8	62.1	99.7	82.7	92.7	30.8	100.0	100.0
filtered issues	18.8	76.1	2.8	66.0	0.2	27.3	97.3	29.3	67.5	58.1	57.1	60.4	92.5	30.6	100.0	74.2

Note: 'filtered issues': issues excluding the undocumented tags and subfields

Table 4.4.: The percentages of typical structural issues

type	bay	bzb	cer	col	dnb	gen	har	loc	mic	nfi	ris	sfp	sta	szt	tib	tor
issues on record level																
R1 ambig. link	0.0	–	–	–	–	–	0.0	–	0.0	–	–	–	0.0	–	–	0.0
R2 invalid link	0.0	–	0.1	0.0	0.0	0.0	0.0	2.1	0.0	0.0	–	0.0	0.0	–	–	0.0
R3 type error	0.0	–	–	0.0	–	0.0	0.0	–	0.0	0.0	–	0.0	0.8	–	–	0.0
control fields issues																
C1 invalid code	0.0	0.0	25.9	0.1	0.1	0.0	0.0	0.0	0.1	38.4	–	0.0	1.5	0.1	0.7	0.5
C2 invalid value	3.2	8.4	33.1	38.9	0.2	61.6	13.4	2.0	58.8	47.3	–	5.3	14.6	98.7	29.5	16.4
field issues																
F1 missing ref	–	–	–	0.0	–	0.3	0.0	0.0	–	0.0	–	0.0	0.0	0.0	–	0.0
F2 non-repeatable	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.8	0.0	0.0	0.0	0.0	–	0.0
F3 undefined	86.9	85.9	–	26.7	–	0.0	54.2	3.2	34.2	8.4	69.8	90.6	35.8	0.1	30.2	55.3
indicator issues																
I1 invalid	0.4	0.6	23.8	1.4	2.1	0.1	0.8	19.5	1.5	0.2	13.8	0.1	0.6	0.1	29.7	4.9
I2 non-empty	0.0	0.0	9.5	0.4	0.1	0.2	24.5	22.0	1.1	0.0	2.9	0.4	0.3	0.0	–	8.5
I3 obsolete	–	–	–	11.6	–	0.0	6.9	50.3	2.3	0.0	0.1	3.3	2.2	0.0	–	12.7
subfield issues																
S1 classification	–	–	–	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	–	0.0
S2 ISBN	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.4	0.1	0.1
S3 ISSN	0.3	0.0	0.4	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0
S4 length	0.0	–	–	0.0	–	0.0	0.0	0.0	0.0	0.0	–	0.0	0.0	0.0	–	0.0
S5 invalid value	–	–	–	0.0	0.0	0.0	0.0	0.0	0.0	0.0	–	0.0	0.0	0.0	–	0.0
S6 repetition	0.1	0.0	0.1	0.2	0.1	0.5	0.1	0.4	0.2	0.3	–	0.1	0.0	0.1	–	0.1
S7 undefined	9.0	5.1	6.9	20.5	97.3	37.1	0.1	0.3	1.6	4.4	11.9	0.1	44.0	0.4	9.8	1.5
S8 format	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	–	0.0	0.1	0.0	–	0.0

Note: The numbers in the columns represents the percentage of a given type for all issues in the catalog. Character '–' means that a given type does not occur in the catalog, while '0.0' means a percentage close to zero.

For the field the *missing reference subfield (880$6)* (F2) refers to a special linkage issue, when the '880' field does not have the mandatory subfield '$6'. Fields could be repeatable or non-repeatable. *Non-repeatable* (F2) denotes the case when a non-repeatable field is available more than once in a record. An *undefined field* (F3) represents the problem when the documentation of the field is missing. If the field name contains somewhere a digit 9, one could suppose that it is a locally defined, but undocumented field (with the exception of 490, which is defined in the standard). It can not be validated, because as it lacks proper documentation the validator does not know the requirements. The validator could not be sure that the field usage was intentional or not.

The indicators have 3 types of issues. *Invalid value* (I1) means that the value is not a term from the dictionary, a *non-empty value* (I2) occurs when the indicator should be empty, but it holds some non-space value, while *obsolete value* (I3) occurs when the field contains a value which was valid in the past, but not any more.

Finally come the issues with the data subfields. *Classification* (S1) is the problem of specifying an information source (typically a classification scheme). In several fields if the second indicator contains '7', subfield $2 should point to a dictionary term. If the subfield is missing this issue is reported. *invalid ISBN* (S2) and *invalid ISSN* (S3) occurs if the ISBN or ISSN field does not contain any string which looks like an ISBN or ISSN identifier, or the found string doesn't fit the rules (the last character of these identifiers is a check value, it should match the result of some calculations on all previous characters). *Invalid length* (S4) issue occurs when the value is shorter or longer than a specified length, *invalid value* (S5) happens when the value is not a dictionary term, *non-repeatable* (S6) happens when a non-repeatable subfield occurs more than once, while *undefined subfield* (S7) refers to unavailable subfield definition. *Non well-formatted field* (S8) is a formatting issue and is similar to what we have seen at the date parsing: the content does not match a predefined format.

From Table 4.4 it became clear that the most frequent issues are the usage of undocumented schema elements. The next large source of issues are the invalid codes and values in the control fields. One can think it might be due to the fragility of those fields I discussed earlier, but there might be other sources. Several libraries uses MARC only as a data exchange format, they export MARC from converting some other format. It must be a deeper investigation to separate the transformation and the original issues, which already exist in the source record. The indicator-issues also represents a surprisingly large proportion, while there are relatively less issues in data subfields.

Table 4.5.: Percentage of records where different metadata types are available

	bay	bzb	col	cer	dnb	gen	har	loc	mic	nfi	ris	sfp	sta	szt	tib	tor
01x	100.0	100.0	99.9	100.0	100.0	100.0	98.7	100.0	100.0	100.0	94.3	82.7	92.6	100.0	100.0	98.6
1xx	69.1	66.6	81.4	75.3	59.0	66.1	80.1	81.3	84.6	65.4	97.8	69.5	69.0	69.8	81.7	82.4
20x	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	85.6	82.7	92.7	100.0	99.7	100.0
25x	99.2	98.7	99.6	95.5	75.2	100.0	96.9	99.9	97.3	99.6	41.5	82.7	92.0	100.0	100.0	95.4
3xx	80.3	100.0	98.8	89.4	95.0	92.5	95.3	99.9	92.5	100.0	78.8	82.6	89.2	73.4	96.5	95.0
4xx	30.6	26.7	31.1	2.1	23.8	31.8	27.4	32.5	23.1	37.3	12.2	22.6	29.6	45.5	–	26.0
5xx	36.8	37.3	81.3	58.2	42.2	59.7	73.9	75.3	100.0	57.4	60.1	61.1	75.3	87.4	100.0	74.0
6xx	45.0	34.7	84.3	–	41.4	49.6	74.3	86.2	77.4	42.9	70.7	72.7	81.4	58.8	58.0	87.3
70x	37.5	45.2	42.4	57.3	34.6	47.6	47.3	43.8	37.1	61.4	45.6	35.5	50.3	44.2	46.5	47.5
76x	25.2	37.3	14.8	18.8	42.2	1.9	15.5	0.3	6.2	6.9	53.2	2.3	9.8	18.6	53.5	5.2
80x	16.0	16.5	30.7	1.2	16.8	2.8	27.5	9.3	6.3	36.0	–	5.5	28.3	45.0	–	6.7
84x	17.1	17.6	100.0	99.2	91.2	97.9	9.9	16.7	12.9	7.7	83.3	9.8	39.3	25.7	100.0	15.3
hol	–	0.1	6.9	–	–	–	0.0	–	0.0	0.1	–	–	7.2	0.9	–	0.0
oth	0.0	0.0	0.0	0.0	71.1	100.0	0.0	0.0	0.0	59.7	0.0	0.0	0.0	38.6	0.0	0.0

01X-09X: Numbers and code, 1XX: Main entry, 20X: Title, 25X: Edition and imprint, 3XX: Physical description, 4XX: Series statement, 5XX: Note, 6XX: Subject access, 70X: Added entry, 76X: Linking entry, 80X: Series added entry, 84X: Holdings & location & alternate graphics, hol: holdings, oth: localized fields.

4.4.4. Completeness

The completeness of the catalogues has been also analyzed, the result is shown at Table 4.5. As in other metadata standards, there are different kind of information. Some fields contain technical information (such as identifiers, creation and modification dates of the MARC record), descriptive information (e.g. title, author, publisher, publishing date, dimensions), and contextual information, such as normalized name forms or subject headings. MARC groups individual fields into categories, the research followed it to show the completeness of them.

01X-09X: Numbers and code (standard numbers, classification numbers, codes)[36]. 1XX: Main entry, name or a uniform title heading used as main entry[37], 20X-24X: Title and title-related fields (variant and former titles, uniform title)[38], 25X-28X: Edition, imprint, etc. (descriptive fields other than titles)[39], 3xx: Physical description (physical characteristics, graphic representation, physical arrangement, publication frequency, and security information),[40] 4XX: Series Statement (information about series the publication is part of),[41] 5XX: bibliographic notes,[42] 6xx: Subject access field, description of the (topical, geographical, chronological etc.) subjects typically terms coming from subject heading systems/thesauri[43], 70X-75X: Added entry (additional name or a uniform title headings),[44] 76X-78X: Linking entry ("information that identifies other related bibliographic items"),[45] 80X-83X: Series added entry (normalized names relating to the series described in 4XX),[46] 841-88X: Holdings, location, alternate graphics,[47] are one of the place for information about the storage location of the physical object the record describes. An alternative method is to create "holdings records" separate from the bibliographical description [41]. MARC lets cataloguers to incorporate fields defined for the holdings record into the bibliographical record, this is shown in the *hol* row. The *hld* row refers to these fields. The last row *oth* refers to fields defined locally or in other MARC versions.

As expected the catalogues usually has high coverage in technical and de-

[36] https://www.loc.gov/marc/bibliographic/bd01x09x.html
[37] https://www.loc.gov/marc/bibliographic/bd1xx.html
[38] https://www.loc.gov/marc/bibliographic/bd20x24x.html
[39] https://www.loc.gov/marc/bibliographic/bd25x28x.html
[40] https://www.loc.gov/marc/bibliographic/bd3xx.html
[41] https://www.loc.gov/marc/bibliographic/bd4xx.html
[42] https://www.loc.gov/marc/bibliographic/bd5xx.html
[43] https://www.loc.gov/marc/bibliographic/bd6xx.html
[44] 1XX records the agents chiefly responsible for the work, while 70X records contributors. https://www.loc.gov/marc/bibliographic/bd70x75x.html
[45] https://www.loc.gov/marc/bibliographic/bd76x78x.html
[46] https://www.loc.gov/marc/bibliographic/bd80x83x.html
[47] https://www.loc.gov/marc/bibliographic/bd84188x.html

scriptive metadata (01X, 20X, 25X, 3XX), since they are semi-automatically created, or their creation require less resources than that of the contextual metadata. The series statement, and the contextual entities belong to them (4XX, 80X) should be present only if the described item is series ("continuing resource"), so similar values means, that serial statements are mostly contextualized. The evaluation of notes (5XX) is really difficult, because their existence is dependent on information which may or may not available in the described work. This explains why this value shows great variety over the catalogues. The authorized name forms (1XX, 7XX) are kind of information duplication, they are normalized forms of entities occurred in the descriptive fields. To create them requires intellectual efforts, not rarely distinct investigations. The hardest part of the cataloguers' work is classification (6XX). Ideally every work should belong to at least one conceptual class, but due to the limitation of resources it can not be reach. The automatic classification has been a popular research topic inside machine learning, and there were experiences with library catalogues, and one could expect that in the future it will help librarians, but according to these numbers it is not there. A deeper investigation should reveal which parts of the catalogue have classification.

4.4.5. Functional analysis

In 2003 Tom Delsey created a comparative analysis of MARC, FRBR and AACR (Functional Analysis of the MARC 21 Bibliographic and Holdings Formats [16], which has been revised by the Library of Congress in 2006 [49]). An interesting part of this work is a mapping of MARC data elements to user tasks. There are 12 such tasks defined, groupped into three main categories. The definitions of the tasks are the following:

Resource Discovery

- Search – Search for a resource corresponding to stated criteria (i.e., to search either a single entity or a set of entities using an attribute or relationship of the entity as the search criteria).
- Identify – Identify a resource (i.e., to confirm that the entity described or located corresponds to the entity sought, or to distinguish between two or more entities with similar characteristics).
- Select – Select a resource that is appropriate to the user's needs (i.e., to choose an entity that meets the user's requirements with respect to content, physical format, etc., or to reject an entity as being inappropriate to the user's needs).
- Obtain – Access a resource either physically or electronically through an online connection to a remote computer, and/or acquire a resource through purchase, licence, loan, etc.

Resource Use

- Restrict – Control access to or use of a resource (i.e., to restrict access to and/or use of an entity on the basis of proprietary rights, administrative policy, etc.).
- Manage – Manage a resource in the course of acquisition, circulation, preservation, etc.
- Operate – Operate a resource (i.e., to open, display, play, activate, run, etc. an entity that requires specialized equipment, software, etc. for its operation).
- Interpret – Interpret or assess the information contained in a resource.

Data Management

- IdentifyIdentify a record, segment, field, or data element (i.e., to differentiate one logical data component from another).
- Process – Process a record, segment, field, or data element (i.e., to add, delete, replace, output, etc. a logical data component by means of an automated process).
- Sort – Sort a field for purposes of alphabetic or numeric arrangement.
- Display – Display a field or data element (i.e., to display a field or data element with the appropriate print constant or as a tracing).

The software's above mentioned data model provides a field to register these user tasks or functions to each type of MARC elements. Following the 2006 revision of the mapping the majority of the default MARC 21 element registered one or more functions, and a separate analysis method has been written to calculate the coverage of the individual functions per catalogs. The average scores can be find in Table 4.6, and its visualization in Figure 4.1 and 4.2. The numbers should be interpret in the scale of 0 to 100 and shows the proprtion of fields available per record from the totality of fields supporting a given function. In her 2007 paper Miksa [45] evaluated four functions in OCLC catalog containing 50 million records. She applied a threshold to filter results, and get significantly larger numbers. I showed previously that MARC 21 is a quite fine grained metadata schema, there are more than 3400 data elements defined, and more than 1800 of them has a function attached to it (number of data elements supporting individual functions: *resource discovery*: search–464, identify–976, select–360, obtain–466, *resource use*: restrict–24, manage–107, operate–67, interpret–118, *data management*: identify–491, process–529, sort–26, display–80). On the other hand the average record contains about 100 of such elements. It is not a big surprise that on a scale which makes the bar high the average score is very low. It would make sense to set some kind of filtering mechanism to normalise these result. The critics of these functional mapping [45, 23] mention,

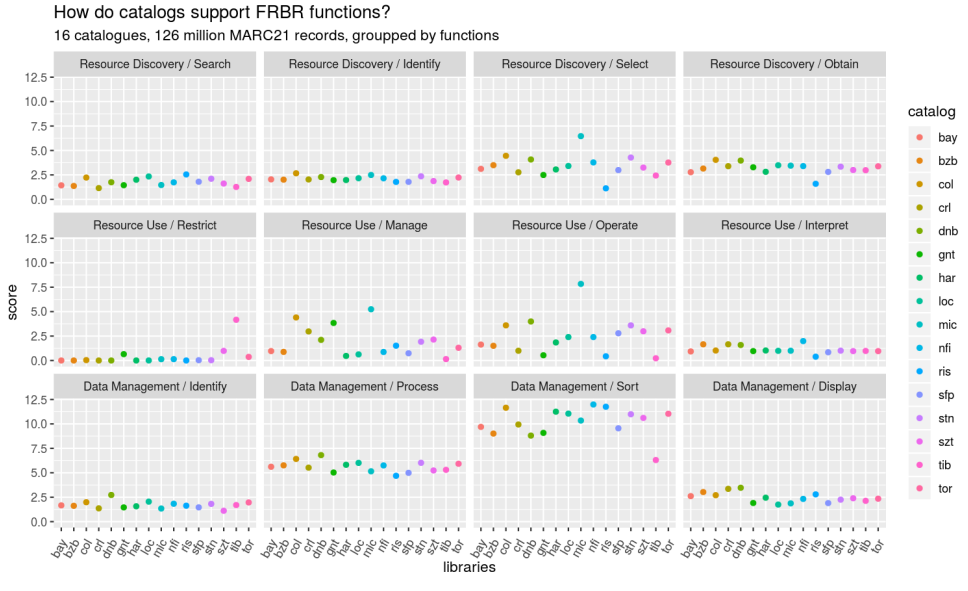

Figure 4.1.: Support of user tasks per catalogues I. Comparison per catalog

that instead of individual elements it would be better to define a combination of elements which together support a function. The current result shows that four functions have real discriminative effects, so they reveal distinctions between catalogues: resource discovery/select, resource use/manage and operate, and data management/sort. From these manage and sort are among those functions which are supported by a low number of fields, however resource use/restrict is also low, however it is not very discriminative. These numbers require additional investigation to draw important conclusions from them.

4.5. Future work

MARC is an evolving standard, its new rules should be implemented in the newer versions of the software. MARC has a number of strict syntactic rules, but there are also semantic rules, which are not as easy to validate. Content wise there are external rule sets such as the ISBD [28], AACR2 [4] or RDA[48] which were not explored in this research. I would

[48] http://rda-rsc.org/content/rda_faq

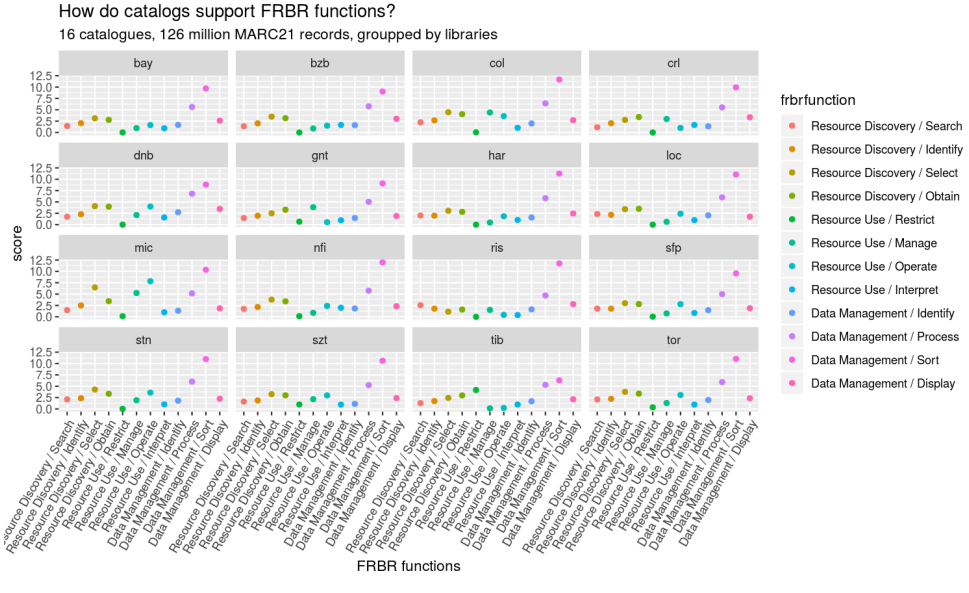

Figure 4.2.: Support of user tasks per catalogs II. Comparison per function

Table 4.6.: Support of user tasks (average score, scale: 0-100)

function	bay	bzb	cer	col	dnb	gen	har	loc	mic	nfi	ris	sfp	sta	szt	tib	tor
Resource discovery																
Search	1.4	1.4	2.2	1.2	1.7	1.4	2.0	2.3	1.5	1.7	2.5	1.8	2.1	1.6	1.3	2.1
Identify	1.6	1.5	2.3	1.8	1.7	1.7	1.7	1.9	1.4	1.8	2.0	1.6	1.9	1.5	1.3	1.8
Select	2.0	2.5	4.0	2.6	2.6	2.2	3.1	3.6	2.9	3.8	1.8	3.0	3.3	2.4	1.3	3.1
Obtain	2.1	2.3	3.4	3.0	3.1	2.9	2.3	3.1	1.8	2.9	1.9	2.4	2.5	2.3	2.3	2.7
Resource use																
Restrict	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.1	0.1	0.0	0.0	0.0	1.0	4.2	0.4
Manage	0.6	0.6	5.9	5.0	1.0	6.5	0.8	0.9	2.4	1.1	2.6	0.9	1.2	1.8	0.1	0.9
Operate	2.1	2.0	6.8	3.1	4.5	1.6	5.8	7.2	8.3	6.3	1.4	6.4	5.5	5.0	0.0	4.5
Interpret	0.1	0.9	0.2	0.9	0.8	0.1	0.2	0.2	0.2	1.2	0.4	0.2	0.2	0.1	0.1	0.1
Data management																
Identify	1.3	1.2	1.8	1.0	2.3	1.2	1.4	1.6	0.9	1.6	1.2	1.2	1.4	0.9	1.3	1.6
Process	2.0	2.2	2.8	1.9	3.2	1.6	2.2	2.4	1.6	2.2	1.9	2.0	2.4	1.7	1.7	2.3
Sort	9.7	9.0	11.7	9.9	8.8	9.1	11.3	11.1	10.3	12.0	11.8	9.6	11.0	10.6	6.3	11.0
Display	2.6	3.0	2.7	3.4	3.5	1.9	2.5	1.7	1.9	2.3	2.8	1.9	2.2	2.4	2.1	2.3

like to encourage the libraries to publish their local MARC element definitions and to update the software with these rules. A web based user interface with faceted search and data visualization is under construction[49]. Data science methods provide us with the possibility of deeper analysis. A Gent catalogue record has the following responsibility statement: "Herr Seele (tekeningen); Toon Coussement (foto's); Peter Claes, Kris Coremans en Hera Van Sande, vakgroep architectuur en stedenbouw Universiteit Gent (vormgeving)." The same record lists four authority entries: "Herr Seele", "Coussement, Toon", "Claes, Peter", and "Van Sande, Hera" while "Kris Coremans" is missing. A comparision of the authority entries and a list extracted by named entity detection would highlight the missing name.

4.6. Note about reproducibility

This research analyzed mostly freely available data sources. The download links, format, and license information of each are provided in section 3.2. The analysis software is an open source tool that is available under the GPL-3.0 license. The binary versions are distributed via Maven Central, a repository for Java libraries. The software is properly documented, and provides helper scripts. Continuous integration (via Travis CI) and automatically generated transparent code coverage reports help to maintain the quality of the software. For every research software project it is a crucial point whether the tool could escape the confines of the laboratory walls. The author is happy to cooperate with libraries to improve the software, and thus the quality of the catalogues. The generated reports behind Table 4.3 and 4.4 are available as supplemental materials[50].

4.7. Acknowledgement

Thanks for Johann Rolschewski and Phú for their help in collecting the list of published library catalog, Jakob Voß for the Avram specification and for his help in exporting MARC schema to Avram, Carsten Klee for the MARCspec. I would like to thank the early users of the software, Patrick Hochstenbach (Gent), Osma Suominen and Tuomo Virolainen (FNL), Kokas Károly and Bernátsky László (SZTE), Sören Auer and Berrit Genat (TIB), Shelley Doljack, Darsi L Rueda, and Philip E. Schreur (Stanford), Marian Lefferts (CERL), Alex Jahnke and Maike Kittelmann (SUB) who provided data, suggestions or other kinds of feedback,

[49] https://github.com/pkiraly/metadata-qa-marc-web
[50] https://doi.org/10.25625/AMF8JC

Justin Christoffersen for language assistance. Special thanks to Reinhold Heuvelmann (DNB) for terminological and language suggestions.

Chapter 5.

Towards an extensible measurement of metadata quality[1]

5.1. Introduction

In the literature about metadata quality measurement[2] it is rare that the authors give any hint about their implementation methods or give a reference to their source codes. Thus – since as readers of these papers we lack the opportunity of studying the implementation details –, we are forced to assume that these projects aimed to measure one specific dataset based on a particular metadata schema. When I started my research, it was an important aspect that the tool I design (called as the time being "Metadata Quality Assurance Framework"[3]) should work together with different metadata schemas and formats.

The quality assessment process has three phases:

1. Measuring individual metadata records
2. Analyzing the results with statistical methods to get metrics of the dataset
3. Reporting results

The results of the first phase are table-like data structures containing mostly numerical values. Since these type of results are very similar for different metadata schemas, the second phase can already be handled in general way with statistical and data science methods. The first step

[1] This chapter is an extended version of the paper [35].

[2] The most current bibliography of the topic is the collaborative Metadata Assesment Zotero group library created by the DLF AIG Metadata Working Group (`http://dlfmetadataassessment.github.io/`). Available at `https://www.zotero.org/groups/metadata_assessment`.

[3] Background information, documentation and source codes are available from `http://pkiraly.github.io/`.

should however be investigates further. The basic process of measuring is the following: the tool takes a record as a formatted string (in JSON or XML format), addresses individual parts to check their existence, cardinality or other properties (different features of their content) and returns the results of these checks (typically in numerical form). If one would like to abstract this process she should figure it out how to list and access individual parts of the metadata record, and she also should provide schema-independent measurement methods.

This chapter will describe the steps I took so far to support this abstraction, which enables us to measure different metadata collections with a single tool.

5.2. Types of measurement

The main types of data quality measurement the tool can run on metadata records are the following:

1. General structural and semantic metrics. These measurements are the most well known in the literature, and following the seminal articles [11, 51] several projects reported to measure the metrics of completeness (the existence of the defined fields in the records), accuracy (comparison of a full data object and its metadata), conformance to expectations (schema rule validation and information value), logical consistency and coherence, accessibility (how easy is to understand the text of the record), timeliness (the metadata quality change over time) and provenance (the relationship between other metrics and the creator of the data).

2. Support of functional requirements. Each data schema is created for supporting a set of functionalities, such as searching, identifying or describing objects. The data elements support one or more of these functionalities, and their existence and content has an impact of these functionalities. An example: a timeline widget expects a specific date format; if the field value is in another format the widget will ignore it. This family of metrics gives measures the scale of support of the functional requirement. To apply these metrics a functional requirement analysis of the data schema should be conducted, and mapping the individual data elements (classes and properties) to the functionalities. The result will be a report which tells how the data support the intended functions. Following the terminology established in [19] I call these scores 'sub-dimensions'. In the Europeana Data Quality Committee Valentine Charles and Cecile Devarenne defined a number of sub-dimensions (such as searchability, descriptiveness, identification, contextualization, browsing etc.) which could be re-used in other metadata domains.

3. Existence of known data patterns. These are schema- and domain-specific patterns which occurs frequently in the datasets. There are good patterns which detect good data creation practices, and anti-patterns, which should be avoided (such as data repetition, meaningless data etc.). For some domains there are existing pattern catalogs (e.g. the Europeana Data Quality Committee works on a Europeana specific pattern catalog, while [60] examined three SKOS validation criteria catalogs).

4. Multilinguality. RDF provides an easily adaptable technique to add a language tag to literal values, and multilinguality has become a key aspect in the Linked Open Data world. In cultural heritage databases the translation of the descriptive fields (such as title, description) might be quite a resource-intensive task. On the other hand reusing existing multilingual dictionaries for subject headings is a relatively simple and cheap process. On the measurement side the nice thing is that generally the multilingual layer in metadata schemas (even in those not built on top of RDF) are similar, so the implementation can be abstracted. The big problem is how to handle the biases generated by the different cardinality and importance of the data elements. Imagine that we have a subject heading which is accessible in several language, but it is attached to a great portion of records, so its information value or distinctive power is low. See more about multilinguality in [57].

The common point in these metrics is that they can be implemented as generic functions where input parameters are specific elements of a data schema. The functions themselves should not know about the details of the schema; that is to say they should be schema-independent. In other words: the only thing need to be created on a schema by schema basis is a method which takes care of mapping the schema elements and measurement functions and feeds these generic functions with the appropriate metadata elements.

Note: based on these metrics on the following, analytic phase a mathematical model have to be created that generates one or more top level data quality score for the record; this chapter does not discuss this phase.

5.3. Mapping schema and measurements

A metadata schema describes the structure of the record, and optionally gives us constraints upon the values of the fields. In order to measure the metadata quality various pieces of information are needed about the schema:

- What are the fields to be analysed?

- Are there any special properties of a field (e.g. mandatory or optional, repeatable, content-related constraints, special format)?
- Are there fields the tool has to extract to identify the record or a subset of the collection?
- Are there field groups which can behave in special way? (For example: if a record must have at least a title or an alternative title, the tool should group these together, and when checking the mandatory elements it should return true if at least one element of the group exists. Conversely, disjoint fields (which are mutually exclusive) should be checked that only one one of them is available in the record, but never more.)

In the previous section I mentioned that I had to create a mapping mechanism which dispatches the elements of the metadata record to different measurement functions. In the first iteration of the Metadata Quality Assurance Framework I have created an abstract concept of the schema, which lists the metadata elements needed for the quality analyses. Each element has a name, an address with which its occurrences can be found within the record, and different properties which denote its role in a particular function (for example the list of sub-dimensions it is part of, the list of field groups, whether it might have language annotation etc.). The mapping supports parent-children relationship, so the functions can recurse down the hierarchy.

A prerequisite for measuring information value is a searchable index, which I implemented with Apache Solr. In the framework, metadata fields are accordingly mapped to Solr field names.

In this mapping the tool records all information about the metadata schema which is necessary for the measuring. At time of writing the manifestation of this mapping is a Java class.[4] To run the measurement on a new schema, one should create first this mapping object. In the future I will create user friendly interfaces (web-based editorial form and XML/RDF annotations) which are more familiar tools for the intended audience, the community of metadata experts.

5.3.1. Addressing elements

An important part of schema handling is how it addresses the particular parts of the record. In the XML world XPath[5] provides with a standard way to solve this problem. The current version of the Metadata Quality

[4]Such as `https://github.com/pkiraly/metadata-qa-api/blob/master/src/main/java/de/gwdg/metadataqa/api/schema/EdmOaiPmhXmlSchema.java`

[5]XML Path Language (XPath). Version 1.0. W3C Recommendation 16 November 1999 (Status updated October 2016). `https://www.w3.org/TR/xpath/`

Assurance Framework (version 0.4) supports only JSON records, so it makes use of a similar tool, JsonPath[6] which has a different syntax, but offers the same functionality. To illustrate it here is an example for the Europeana Data Model (EDM) metadata schema[7]:

Listing 5.1: A JSON path example

```
$.['ore:Proxy']
  [?(@['edm:europeanaProxy'][0] == 'false')]
  ['dc:title']
```

This expression addresses the dc:title field instances of the ore:Proxy part in which the value of the first edm:europeanaProxy instance is 'false'.

Listing 5.2: An excerpt of an EDM metadata record

```
{
  "ore:Proxy": [
    {
      "edm:europeanaProxy": ["false"],
      "dc:title": [
        {
          "@lang": "de",
          "#value": "Pyrker-Oberwart,␣Johann␣Ladislaus"
        }
      ],
      ...
    },
    {
      "edm:europeanaProxy": ["true"],
      ...
    }
  ]
}
```

Here you can see that the ore:Proxy is a list of two objects. The first one's edm:europeanaProxy is 'falue', the second one's is 'true'. Since we are looking for the the object with the 'false' value, we get the first. It has a 'dc:title' property. The return value will be the Java representation of the following JSON string:

Listing 5.3: The selected part of the record

```
[
  {
    "@lang": "de",
    "#value": "Pyrker-Oberwart,␣Johann␣Ladislaus"
  }
]
```

[6]Stefan Goessner: JSONPath — XPath for JSON. http://goessner.net/articles/JsonPath/. Actually it uses its Java port, the Jayway JsonPath available at https://github.com/jayway/JsonPath

[7]The Europeana Data Model Documentation is available at http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation

In this example an absolute path is shown, which searches from the root of the record, but since the schema abstraction supports parent-child relations, relative paths can be used as well.

5.3.2. Flexible and configurable measurements

The overall picture

The system's central entry point is the CalculatorFacade (or its extension). It provides us with a number of configuration options. The most important one is the registration of the schema mapping. Based on the settingsin the schema mapping it prepares the measurement classes for running. When it is ready to run the process passes every record to the measure() method. It accepts the metadata record as a JSON string, runs all the measurements which have been configured, and returns a CSV representation of the result of the metrics. As indicated by the name it is just a Facade object[8]: it only coordinates the process, the actual measurements are done by the individual 'calculators' (the generic schema-agnostic functions I mentioned above). These all implement the Calculator interface, and they have three important methods. The 'measure()' method accepts a cacheable representation of the metadata record and performs the measurement. The 'getCsv()' method returns the result of the measurement, while 'getHeader()' method returns a list of the column names for the CSV row. The cacheable representation is a special object. It contains the full record or a part thereof, applies the JsonPath expressions, and transforms the resulting object into a uniform Java object, which provides a simplified DOM-like interface. In this fashion the Calculators can access the values in a generalized way, and can reuse the already-processed parts from the cache. Retrieving the parts of the record is computationally expensive: one part might participate in multiple measurements, but this way the tool has to retrieve it only once.

Each calculator implements one or more metrics. Since at run-time (in the measuring phase) they can not get extra arguments, if a calculator has conditional steps depending on properties which are not part of the schema, they should be configured ahead of the measuring phase via the CalculatorFacade's 'configure()' method. For example the 'MultilingualitySaturationCalculator' can emit its results in either simple and complex form, depending on the 'resultType' setting. The client should decide the format before running the measurement, and set the appropriate one in the Facade class.

[8] See `https://en.wikipedia.org/wiki/Facade_pattern`

This way the CalculatorFacade are extensible, and one can write additional Calculators. The tool currently provides all Calculators required for our current research.

Metadata problem patterns

Problem patterns are known issues in the metadata record instances. In case of Europeana Timothy Hill and Hugo Manguinhas have led the initiative to collect all those problems[9]. They categorized the problems into several types, such as duplicate or redundant information, irrelevant information (such as non-meaningful titles, e.g. "unknown title"), missing or incomplete information, misuse of fields, just to name a few.

The current implementation of the problem catalog measurement has been done based on the Observer design pattern [10]. There is a central class — ProblemCatalog[11] — which implements the Calculator interface, so it has a 'measure()' method. This class acts as the observable subject, and it notifies its subscribers (the observers) when they have to measure a new record. Each individual problem is associated with a distinct class (a ProblemDetector[12]), which implements the Observer interface, and accordingly has a method called 'update()'. It has two parameters: the metadata record, and a variable which is a collector of the measurement results. In this case the result should be a number: how many times the pattern occurs in the record. When the measure is started, the ProblemCatalog class creates a collector for these results, and the client (the facade class) will retrieve it when the measurement is done.

The ProblemCatalog class has 'addObserver()' method to register the subscribers, which should be done at the central facade class at configuration time.

For detecting new problems one has to create new ProblemDetectors and register them in the central CalculatorFacade.

5.3.3. Extensions and APIs

The current Java APIs are defined in a core library (metadata-qa-api) which contains the workflow governing mechanism, and the general schema-agnostic functionalities. There is an extension of this API, called 'europeana-

[9] Data Quality Committee: Problem Patterns. Available at `http://bit.ly/2jIXQGU`

[10] `https://en.wikipedia.org/wiki/Observer_pattern`

[11] `https://github.com/pkiraly/metadata-qa-api/blob/master/src/main/java/de/gwdg/metadataqa/api/problemcatalog/ProblemCatalog.java`

[12] See for example the 'EmptyStrings' class which detects empty strings in field instances: `https://github.com/pkiraly/metadata-qa-api/blob/master/src/main/java/de/gwdg/metadataqa/api/problemcatalog/EmptyStrings.java`

qa-api' that contains the Europeana-specific measurement and facade. These libraries are published in the central Maven repository[13], so any further extension can add these into its dependency tree in the local Maven configuration file (pom.xml) as

Listing 5.4: Including the Java libraries into other project

```
<dependencies>
  <dependency>
    <groupId>de.gwdg.metadataqa</groupId>
    <artifactId>metadata-qa-api</artifactId>
    <version>0.4</version>
  </dependency>
  <dependency>
    <groupId>de.gwdg.metadataqa</groupId>
    <artifactId>europeana-qa-api</artifactId>
    <version>0.4</version>
  </dependency>
  ...
</dependencies>
```

I have created two clients for the Java APIs: the first enables measurement with the popular Big Data analytics tool Apache Spark; the second one provides us with a REST interface.

Big Data analysis

Apache Spark is an extremely efficient tool for batch processing of huge datasets (the Europeana dataset is more than 400 GB). Combined with Apache Hadoop's distributed file system it can be run either on a single machine or in a distributed computing environment. The users can specify the details of allowable resource usage (number of CPUs, memory usage, etc.) and it comes with a web based monitoring tool. In the file-based workflow, Spark reads and processes lines one by one; it requires us to store one record per line (however Spark also supports different other data sources such as NoSQL databases, where this constraint does not exist). In our Spark based client the result is stored as CSV files.

Data analysis with REST APIs

The REST interface provides two kinds of API: a simple Record API, which runs the measurement on individual records and returns a CSV or JSON response, and another one which is called Workflow API and enables a full measurement workflow.

[13] http://mvnrepository.com/artifact/de.gwdg.metadataqa

Listing 5.5: Quality measurement of a single record REST API response

```
GET /07602/696BE60475DFAF290B7CD7759E840CD4FFF86E24.json

{
  "existingFields":[
    "edm:ProvidedCHO/@about", "Proxy/dc:title", "Proxy/dc:creator",
    "Proxy/dc:publisher", "Proxy/dc:type", "Proxy/dcterms:spatial",
    ...
  ]
  "emptyFields":[],
  "labelledResults":{
    "fields":{
      "recordId":"/07602/696BE60475DFAF290B7CD7759E840CD4FFF86E24",
      "dataset":"07602_Ag_IT_Culturalitalia_RegioneMarche",
      "dataProvider":"Regione␣Marche␣/␣SchedeS:AN" },
      "completeness":{
        "TOTAL":0.542857, "MANDATORY":1.0, "DESCRIPTIVENESS":0.454545,
        ...
      },
      "existence":{
        "edm:ProvidedCHO/@about":true, "Proxy/dc:title":true,
        "Proxy/dcterms:alternative":false, "Proxy/dc:description":false,
        ...
      },
      "cardinality":{
        "edm:ProvidedCHO/@about":1, "Proxy/dc:title":1,
        "Proxy/dcterms:alternative":0, "Proxy/dc:description":0,
        ...
      },
      "uniqueness":{
        "dc:title:sum":0.001084, "dc:title:avg":1.807687E-4,
        "dcterms:alternative:sum":0.0, "dcterms:alternative:avg":0.0,
        "dc:description:sum":0.0, "dc:description:avg":0.0
      },
      "problemCatalog":{
        "LongSubject":1.0, "TitleAndDescriptionAreSame":0.0,
        "EmptyStrings":0.0
      },
      "languages":{
        "Proxy/dc:title":{ "_0":1 },
        "Proxy/dcterms:alternative":{ "_1":1 },
        ...
      }
    },
    "termsCollection":{
      "dc:title":[
        {"term":"ancona", "tf":1, "df":12388, "tfIdf":8.072328-5},
        {"term":"carta", "tf":1, "df":87048, "tfIdf":1.148791-5},
        ...
      ],
      "dcterms:alternative":[],
      "dc:description":[]
    }
  }
}
```

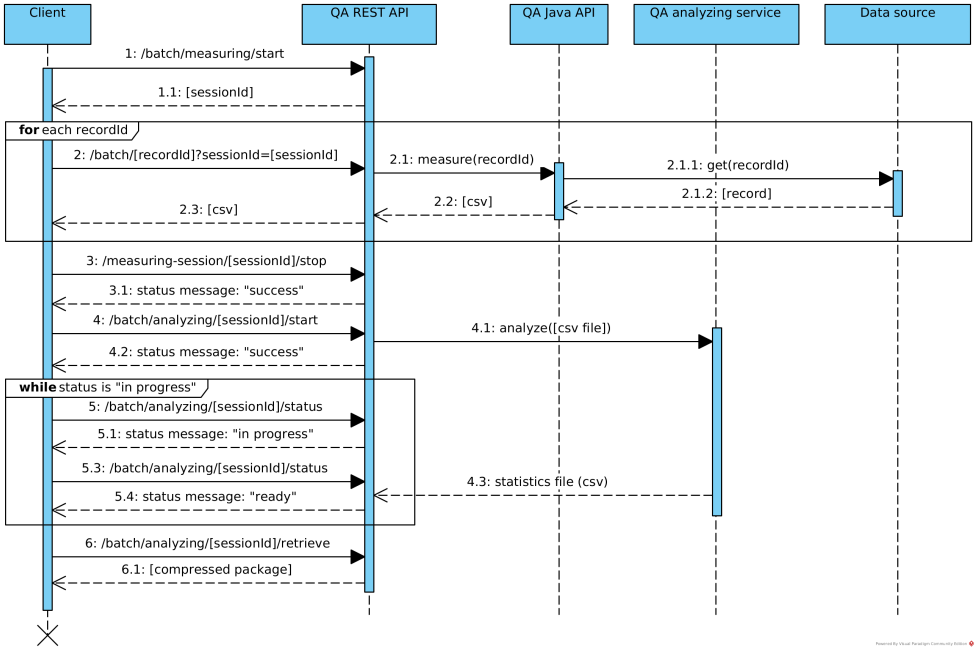

Figure 5.1.: Processing time per records with different settings

Listing 5.5 shows a formatted representation of the Record API call. It accepts two parameters, the record ID, and the format (either as JSON or CSV) as file extension. The result contains all the measurement results for an individual record. At time of writing the measurements are existence, cardinality, uniqueness, problem catalog and languages. The languages part contains pairs of languages and their count for each fields. There are three special codes to encode cases where there is no language annotation:

- _0: no language tag specified
- _1: the field is missing (the very same information that of field existence metric)
- _2: the field is a resource (it contains a URL or tagged as resource)

The user interface's record level display is based on the Record API.

The Workflow API supports the whole life cycle of the metadata quality measurement process, so it measures individual records, then analyze them by running statistical functions on aggregated metrics. The full workflow is depicted by Figure 5.1.

First, the client initializes a session, to which the server returns a session

identifier which should be used during the whole workflow. Then client submits record IDs one by one. The server is responsible for storing the results in a CSV file, but it also returns a CSV row for each record call, same way as the Record API does. Once all records were measured, the client stops the measuring part, and starts the analysis part, in which the server calls the external analyzer (either a R script or a Spark based analysis process) to produce a report containing the result of the statistical analysis and (optionally) generated images files. This process is time consuming, and the client can repeatedly ask for the status of the process. The server reports if the process is "in progress" or "ready". When it is finished, the client can retrieve the result as a zipped package. This process is very useful if the client would like to check only a part of the whole collection. Combined with other tools (e.g. with a search API) it is possible to select the IDs of a subset – for example the newly ingested records, or a specific result set.

The main reason for creating REST APIs is that this way the back end functionalities of the framework became interoperable: the metrics could be reused in the client's own existing systems.

5.4. Conclusions and future works

In this chapter I showed the most important technical requirements of an extensible metadata quality assessment framework. I discussed the first phase of the metadata quality analytics workflow only, the measurement part. It can ingests different metadata schemas, but emits numerical, tabular data in a standardised form. The statistical analysis, and the reporting based on this, and do not require the same abstracting approach as the first step.

I showed the main characteristics of the metadata quality metrics and the metadata schema, then how to map schema features to measurement functions and how to address internal parts of a metadata record. I discussed the relevant design decisions of our implementation, and highlighted the parts which should be extended if one adapts the method for another metadata schema. Finally I described the APIs of the system.

By the time of writing no other projects started to use this framework, so I do not have real external feedback about the flexibility of the framework. Our own experiments were based on two metadata schema: Europeana Data Model (EDM) and MAchine Readable Catalog (MARC21) — both of which have relatively simple structures. I expect that every new metadata schema will raise at least some new requirements, and only after having successfully dealt with a variety of schemas I will be in a position

to release a 1.0 version of the tool. The level of schema abstraction is adequate for current purposes, but I anticipate the addition of several new features in the future. I am constantly looking for collaborating partners to try this approach on new schemas and in different IT environments. Starting and continuing discussions with organizations having similar approaches and requirements in the realms of digital libraries, the semantic web, digital humanities, and learning objects, and learning from each other will be a crucial aspect of creating a truly interoperable framework.

5.5. Acknowledgments

I would like to thank Europeana and GWDG for providing computational environment for this research. For this research Cecile Devarenne and Yorgos Mamakis (Europeana) for giving suggestions regarding to the Workflow API, and all the members of the Data Quality Committee[14] for their work, and Timothy Hill for language-related suggestions.

[14] `http://pro.europeana.eu/page/data-quality-committee`

Chapter 6.

Predicting optimal Spark settings in standalone mode

Abtsract What are the optimal parameter settings for a long running, standalone mode, Spark-based stateless process? The chapter investigates the effects of four different Spark parameters, and compares the application's behaviour in two different servers. Two important lessons learned from this experiment are: i) allocating more resources (e.g. memory, CPU) does not necessary imply better performance of the process, ii) in an environment with limited and shared resources instead of maximising the performance one should rather tune the system to be 'good enough' in terms of performance and at the same time is respectful with other running processes. To discover the optimal settings, it is suggested to pick a small sample that shares important features with the full dataset, and measure performance of the process against different Spark parameter settings. The settings to check are: the number of cores, memory allocation, compression of the source files, and reading data stored on different file systems (if they are available). As a source of ground truth, Spark log, Spark event log, or measuring points inside the application, can be used.

6.1. Introduction

Measuring the quality of cultural heritage metadata is a task which has two important features. First, in Digital Humanities (DH) context the size of source data (catalogues of libraries, archives, and museums – commonly abbreviated as LAM; or aggregated catalogues and datasets such as Europeana[1], Wikidata[2] or Archives Portal Europe[3]) could be regarded as Big Data. Big Data is a relative concept, usually referring to data

[1] `https://europeana.eu`
[2] `https://wikidata.org`
[3] `http://www.archivesportaleurope.net/`

larger than can be processed by traditional methods within the available infrastructure of a LAM organisation. DH and cultural heritage contexts have not traditionally been equipped with high performance computing tools, so the volume of Big Data managed in these contexts has been generally smaller than, for example, in the fields of astrophysics or medicine. On the other hand, the variety of data in DH and cultural heritage contexts is large. Second, even after a decade of research in metadata quality ([43]), the LAM community has not yet reach a clear consensus on the exact meaning of 'quality' – however different quality dimensions and metrics have been established

Current research (Measuring Metadata Quality) is still experimental, and based (to an extent) on trial and error workflow, in which metrics are selected from the literature (or new metrics are invented); their measurements are implemented and tried on the (meta)data; and metadata experts evaluate the result and suggest changes on the measurement. A consequence of this research cycle is that it necessitates the execution of multiple long running measurements of the same Big Data set. It takes considerable time, and requires a significant allocation of computer resources (memory, CPU, disk capacity), with a tendency to block other tasks running concurrently on the same computer. Apache Spark[4] among other tools, decreases the duration of this research cycle by letting the existing process run in parallel fashion. While Spark could be run in a cluster, clusters are rarely available to DH/LAM organizations, where a more typical use is to run Spark in 'standalone mode' utilising the multicore architecture of a single machine and simplifying the writing of multi-threaded software code. This chapter suggests easy preliminary measurements for a Spark-based process to identify its optimal settings in each context. Another aim of this chapter to fulfil a gap in the literature concerning Spark's usage and performance in standalone contexts, as commercial and scientific conference presentations concerning Spark's performance concentrate upon its work in clustered environments.

6.2. Measuring completeness of Europeana records

Europeana is a digital platform of the European cultural heritage, which aggregates catalogue records from European libraries, archives, museums, and other cultural organisations (called data providers). This experimental research uses a snapshot of Europeana records created during 2018 August containing approximately 62 million records. Each record are in

[4] `http://spark.apache.org/`

Europeana Data Model (EDM)[5] metadata schema, which has the several parts or 'entities': a descriptive metadata part ('proxy'), and optional contextual entities that describe the agents, concepts, places and time spans mentioned in the proxy. Moreover Europeana not only aggregates these records, it enhances them as well. With the help of semantic web technologies and linked open data, Europeana attempts to detect entities in the data provider's proxy, and to save them as additional contextual entities.

The snapshot of Europeana's data used in this research is stored in a MongoDB database. Accessing MongoDB from Spark has limitations as reading from MongoDB is a time-consuming process, and Spark's Mongo connector does not support a specific reference type which is heavily used in the database. As such, the author chose to export the data in text files in which every line is an individual, normalised record.[6] This process was built upon Spark's Mongo connector, enhanced with some extra API calls[7]. Spark's Mongo connector partitions the database in order to run the processing tasks (reading and and exporting records to text files) in parallel. At the end of the process 1740 files – each with 35.6 thousand records – were created. The size of the files varies, averaging 0.47GB in size; the bulk of files are between 0.23 GB and 0.9 GB.

This preparation phase was followed by a record-level-measuring process using Spark's Java API. Within the broader research project (Measuring Metadata Quality) the author conducted multiple measurements. For this experiment, one of them, the *completeness measurement* was selected. The completeness measurement takes a JSON string and checks every field in the schema that is available in that record, and returns an integer (zero or more) denoting the number of available field instances. The result is serialised as a CSV formatted string containing record identifiers, some metadata (the identifiers of sources of the records), and the cardinalities of the fields. Spark takes care of reading the input, writing the output, and distribution of the processing over the available CPUs. The part of the process that interacts with Spark is provided below in Listing 6.1. The next step is the statistical analyses of this CSV file. This produces, in turn, a set of CSV files with statistical description of the whole collection and its 20 thousand subcollections (that is, records harvested in the same dataset, created by the same data provider, aggregated by the same provider, derived from particular countries, or written in the same language – or the combination of these), an their visualisation on a web based user interface.

[5] `https://pro.europeana.eu/resources/standardization-tools/edm-documentation`

[6] Here, 'normalisation' refers to the record packaging together the proxies and all linked contextual entities, instead of merely keeping references to them.

[7] `https://github.com/pkiraly/europeana-qa-spark`

Listing 6.1: The part of Java client code which interacts with Spark API

```
// initialize Spark
SparkConf conf = new SparkConf().setAppName("CompletenessCount");
JavaSparkContext context = new JavaSparkContext(conf);

// initialize the processing class
final EdmCalculatorFacade facade = ...

// read input file
JavaRDD<String> inputFile = context.textFile(parameters.getInputFileName());

// definition of a method which process a lines
Function<String, String> baseCounts = new Function<String, String>() {
  @Override
  public String call(String jsonString) throws Exception {
    String result = "";
    try {
      result = facade.measure(jsonString);
    } catch (InvalidJsonException e) {
      // error reporting
    }
    return result;
  }
};

// processing every lines of input files
JavaRDD<String> baseCountsRDD = inputFile.map(baseCounts);

// save result
baseCountsRDD.saveAsTextFile(parameters.getOutputFileName());
```

The standalone Spark process is a minimalistic use of Spark API's capabilities – above its mandatory input and output calls only one extra method is used: map(). It is important to note that Spark API is similar to SQL queries: it is a high level API, and its engine optimises it and creates a low level implementation. When the Spark process is started, it analyses the input to calculate the number of 'tasks' it needs to complete. Each task takes an input, runs the code, and saves the output. Outside of these tasks Spark runs a monitoring web server, which is started before reading the first file, and shut down after producing the last output. Spark also runs some file managing processes in the background when merging and renaming output files. The final output will be a set of files, which need to be merged into a single file outside of the standalone Spark process.

6.3. Tuning Spark and measuring performance

There are several settings in standalone Spark process which are deserving of further experimentation:

- number of cores (CPUs) of the processing computer;
- memory allocation
- for file inputs:
  - whether they are stored in the operating-system's file system, or in a Hadoop File System
  - whether they are compressed of not

In the experiment the author ran the same measurement on two different machines using two input sets. The first input set is the full corpus, while the second input set contains only ten files. The average size of these files are smaller than the average size of files in the whole collection (mean is 0.3 GB, while it is 0.47 GB for the full set). The first machine ('europeana') has a CPU 2.4 times faster than the second one ('roedel'). 'Europeana' has eight cores, 'roedel' has 16 cores. The data and the processing code were identical, and both machines used Spark version 2.4.0.

The numbers measured are recorded in different sources. Spark log provides information about 'stage' and 'job' duration. In the Spark processing hierarchy the process might be split into multiple jobs, each having multiple stages; in this experiment there is only one job with a single stage. The process is launched by a bash script, which also measures the overall duration. Spark has a special event log (see later), which provides other important metrics, such as executors' run-time and CPU-time. Finally the author put a time counter into the client source code, to record how much time the map() function takes to complete the process (see Listing 6.2). Bash scripts read these information, and the charts were created with R.[8]

6.3.1. Number of cores and compression

In next experiment, the effect of the number of cores used, and whether the files were compressed, has been measured. The appropriate parameters are listed in Listing 6.3. The results (the processing time per record under given number of cores) are displayed in Fig. 6.1. A number of conclusions could be drawn from the chart.

1) The per record processing times for the small set is faster. It is not surprising given, that processing time depends on the complexity of the

[8]The source files of the experiment are available at `https://github.com/pkiraly/euro-par`.

Listing 6.2: Use of Spark accumulator to measure duration

```
// accomulators are special, thread-safe, distributed Spark variables
LongAccumulator accum = context.sc().longAccumulator();

...
Function<String, String> baseCounts = new Function<String, String>() {
  @Override
  public String call(String jsonString) throws Exception {
    long start = System.nanoTime();
    ...
        accum.add(System.nanoTime() - start);
    return result;
  }
};
...
logger.info(formatDurationInfo(accum));
accum.reset();
```

Listing 6.3: Spark settings for number of cores and input specification

```
spark-submit --master local[<number-of-cores>] --inputFileName file:///...
```

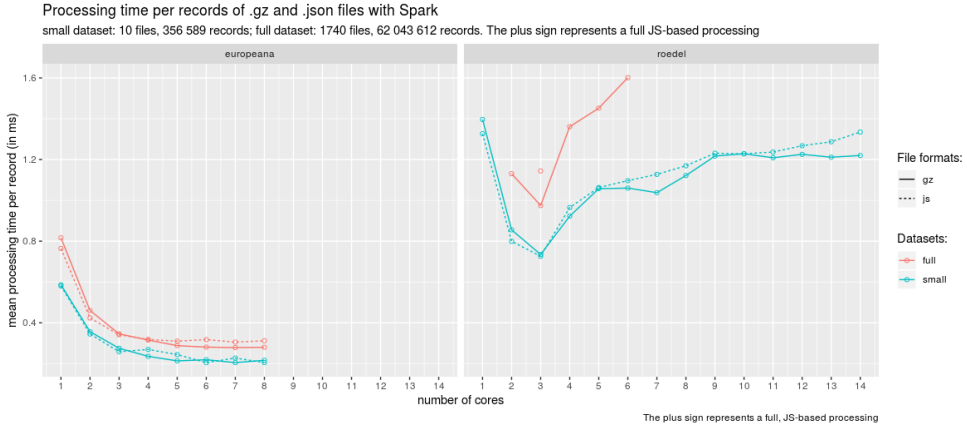

Figure 6.1.: Processing time per records with different settings

record structure which correlates with the size of the record. It is more important that the shape of the small set and the full set are close to each other; correspondingly, running a small set for this kind of application might predict the running of the full set. It is important that the processing function – that is, 'measure()' – is stateless, ensuring the incorporating classes do not collect data (unless in exceptional cases), and the measurement undertaken does not depend upon previous records. Not all Spark client code works this way: the prediction holds only for stateless ones.

2) Gzip compressed files are usually slightly faster than uncompressed files. Processing compressed files has two side effects. First, uncompressed files are partitioned, and (as we saw above) each partition is paired with a distinct task with its own overhead. Second, decompression also has its own overhead. The evident advantage of using compressed files is sparing of disk space.[9]

3) The shape of lines on the two machines are significantly different. On 'euroepana', as more cores are engaged performance continually increases until seven cores, however after a given number of cores (3-4) the improvement is not significant. To determine the optimal settings for the number of cores is not easy here, because there is no clearly identifiable optimal core number. There are two discernible factors at play: first, the examined speed might be already 'good enough'; second, while using more cores does slightly improve the performance of the current process, it takes resources away from other processes running on the same system, which is an impolite behaviour. On 'roedel' the situation is radically different; it has a clear peak at 3 cores. The reason is that the system throughput – that is, the combination of CPU, memory and I/O operation speed and other factors – has a maximum. If more parallel processes are running, some processes will consume and lock all available resources thus creating a bottleneck because others will wait for moments until the locked resources have been released. To detect the actual bottleneck is not easy; some techniques for this task are discussed below.

6.3.2. Memory allocation

The next setting tested was the amount of allocated memory. The process was run on the small, compressed set using different number of cores and allocating 1, 2, 3, and 4 GB memory for both Spark driver (the central controller) and executors (the parts which execute client code). The Spark settings are available in Listing 6.4.

[9]Since the full process took more than one day on 'roedel' machine which exceeded the author's limited machine time, the fastest predictably setting was run on the uncompressed files.

Listing 6.4: Spark's memory allocation options

```
spark-submit --driver-memory <memory> --executor-memory <memory> ...
```

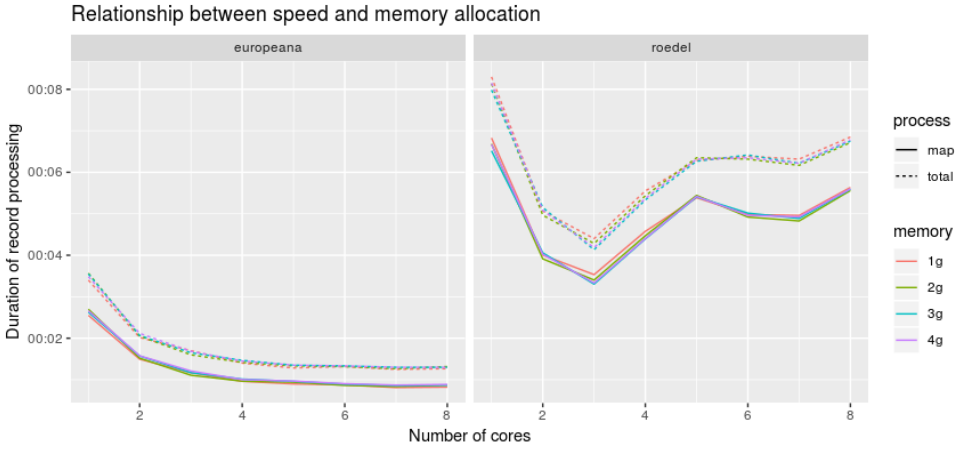

Figure 6.2.: Relationship between speed and memory allocation.

Fig. 6.2 shows the result of different memory allocations. Succinctly, differences in allocated memory do not produce any clear effect. This is not a surprise, as the application is virtually stateless. If it were to accumulate large (or lots of) variables in memory – which happens in several Spark SQL or Spark ML based applications – we would see a significantly different graph. The conclusion is that it is not worth allocating more memory than the default for this application: which is 1 GB for the current Spark version.

6.3.3. HDFS or normal FS?

While Spark works well with Hadoop Distributed File System (HDFS)[10], this chapter questions whether Spark performs better in a standalone setup, where HDFS runs on a single machine, and is not distributed over nodes? The Spark settings is available in Listing 6.5. The result is shown in Fig. 6.3.

In most cases, the results are slightly better for operating system's file

[10] http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

Listing 6.5: Reading from HDFS

```
spark-submit --inputFileName hdfs://localhost:9000/europeana/*.gz ...
```

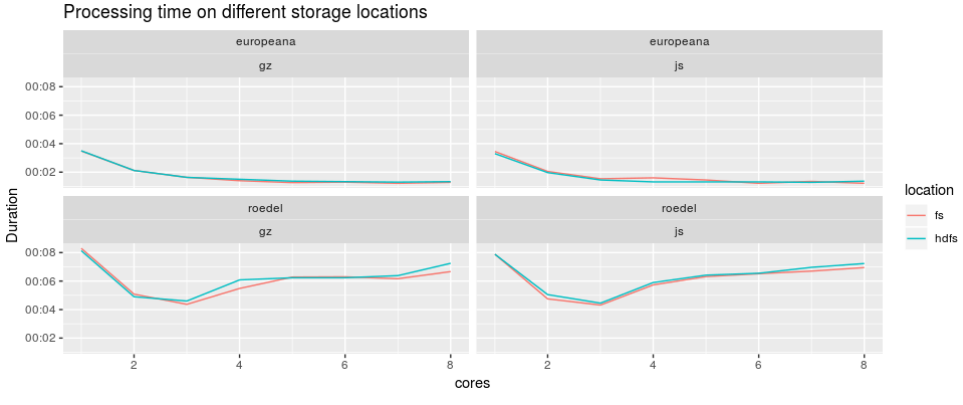

Figure 6.3.: Processing time on different storage locations

system, but not always, and usually the difference between the results is small. The question of whether it is worth to use HDFS or not depends on the speed gain and the cost of the setup of HDFS. The experiment shows there is little advantage in storing files in HDFS in these two test systems.

6.4. Event log and history server – to measure performance

It was mentioned earlier that Spark launches a monitoring web service (unless it is disabled), that provides useful information about the running process. Although he application is shut down when the process ends, Spark provides a useful tool to keep historical information. To enable this tool requires the setting up a directory which stores the events (and their metrics). This history server is similar to the monitoring server, yet it processes data from the saved event log and not from the live process (as per the monitoring service). As the event log is a simple text file containing JSON formatted lines, it is possible to move the event log and

Listing 6.6: Spark history server setting and launch

```
> cd $SPARK_HOME
> more conf/spark-defaults.conf
...
spark.eventLog.enabled           true
spark.eventLog.dir               file:/path/to/spark-event-log
spark.history.fs.logDirectory    file:/path/to/spark-event-log
> sbin/start-history-server.sh
```

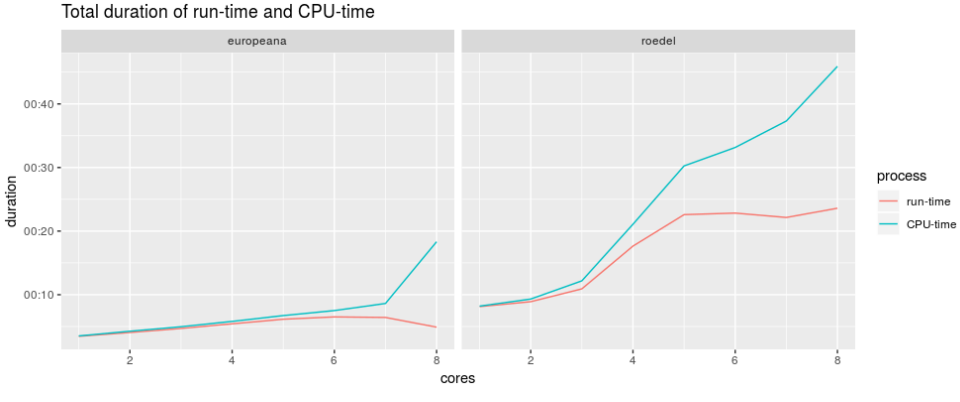

Figure 6.4.: Processing time on different storage locations

display its content on a different machine.[11] It is important to note that the web interface does not show all the information from the event log. The history server provides an API, enabling data to be programmatically read from it.

The most important information for this experiment are 'executorRunTime' and 'executorCpuTime' variables. According to Spark API documentation [6] 'executorRunTime' is the "time the executor spends actually running the task (including fetching shuffle data)", while 'executorCpuTime' is the "CPU Time the executor spends actually running the task (including fetching shuffle data) in nanoseconds". Since the experiment process did not have shuffle steps, it can be supposed that most of the time happens inside the map() function. According to [12] the differ-

[11] The event log file name consists of the master name and the timestamp of the start (such as 'local-1550822115584'). This name is also used as an identifier inside the file. If required for use in another machine, file name should be changed, and its occurrences should be also changed inside the file content (e.g. in this experiment the author used names such as 'europeana-c4', where 'europeana' reflected to the machine name, while 'c4' stands for experiment run with 4 cores setting).

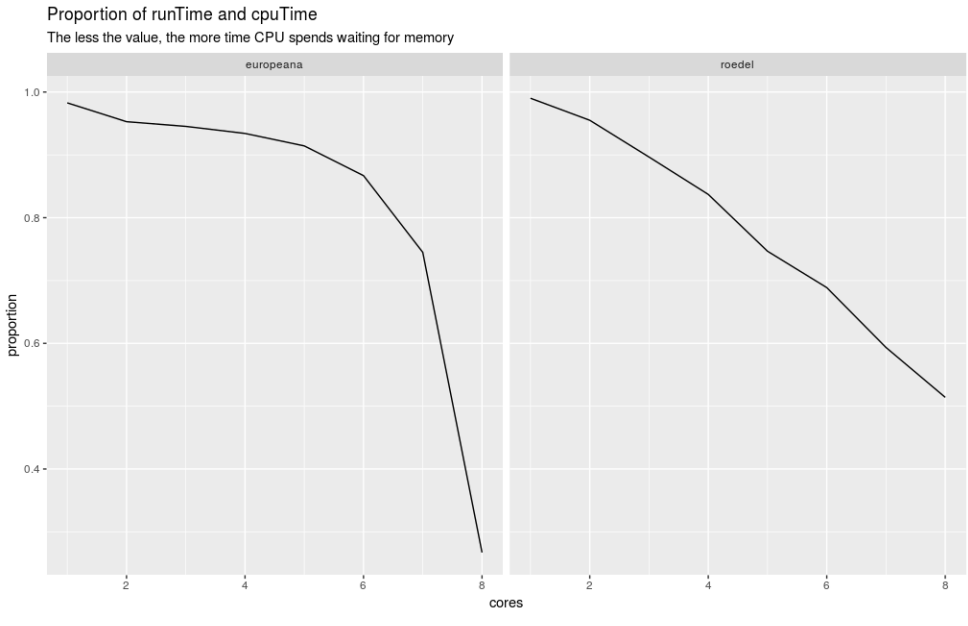

Figure 6.5.: Proportion of run-time and CPU-time. The larger the distance, the more time CPU is waiting.

ence between run-time and CPU-time is the time the CPU waits for the memory. Fig. 6.4 reveals that that the time spent on the map() function grows increasingly larger in both machines as the number of cores utilised are increased. It is natural, because individual processes will increasingly compete for resources. The two machines used in this experiment behave differently due to the increases in difference between times.

If the time consumption were displayed differently, highlighting the relative numbers (i.e. the proportion of run-time and CPU-time in Fig. 6.5), an interesting pattern can be observed. While in 'europeana' the degradation is moderated up until the utilisation of 6 cores, and from then on is progressive, on 'roedel' the degradation is linear. This means that the CPU spends a lot of time waiting, including when performing well utilising only a small number of cores.

In Figure 6.6 run-time and CPU-time are displayed alongside the duration of the map method and total processing time. Figure 6.6 reveals the reason of the performance peak: the quickest performance happens before the duration of the map method exceeds the CPU-time. It is not clear exactly what other processes are involved in the task, yet it is clear,

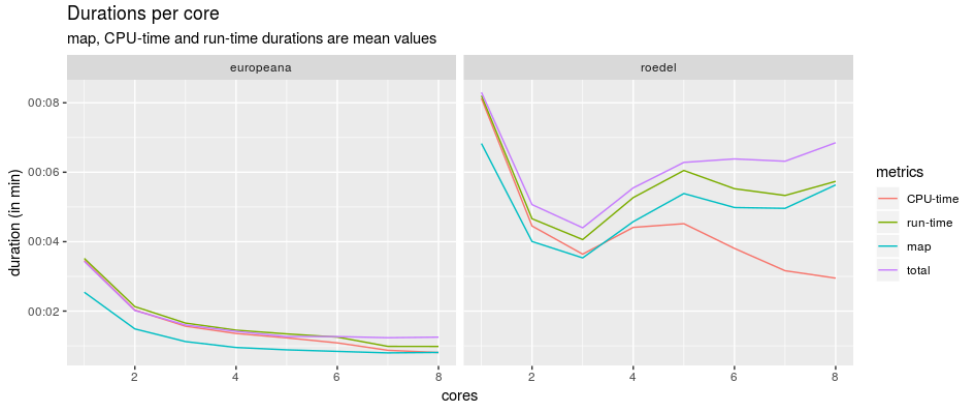

Figure 6.6.: Duration of different components per CPU.

that the CPU is undertaking other processes concurrently as the crossing CPU-time is larger then the duration of the map method. When the duration of the map method gets bigger than the CPU-time, it is clear that map() function – and the other confounded processes – has to wait for free memory. Notably, this crossing occured on 'europeana' between 7 and 8 cores, and on 'roedel' between 3 and 4 cores. It is also interesting that at 8 cores on 'roedel' the map-time almost reaches run-time, and almost half of its time is spent waiting.

One last note about the hidden processes. As mentioned earlier, Spark optimises its code, and this code is not the high level API which runs in the Java Virtual Machine (JVM). Another important feature of this API is that its methods are classified either as transformations or as actions. In Spark, nothing occurs until an action triggers the launch of the processing workflow in which the transformations happen in a pipeline manner. This has two important consequences. First, the duration of some methods can not be measured in the client code, as the methods only assemble the building blocks but do not run them. Second, there is no clear mapping between client code and what a Java profiler shows. This is not only due to the optimisation process, but also because of the intensive usage of language constructions (such as lambda functions). For example, the textFile() method – which reads the file – does not show in the Java stack trace.

6.5. Conclusion

The most important lesson learned in this experiment is that allocating more resources (CPU, memory, IO) to a Spark-based process does not necessary imply better performance. In an environment with limited and shared resources, what is required is a 'good enough' state which respectfully lets other processes run concurrently. To find the optimal parameter settings for a Spark process, this chapter recommends first selecting a small sample from the full dataset (the subject of analysis) that demonstrates similarities – in important characteristics and measuring speed with different settings – to the whole, such as the number of cores, memory allocation, compression of the source files, and different file systems (if they are available). As a source of ground truth one can use Spark log (which contains some performance indicators), Spark event log (disabled by default), or measuring points in your application (via Spark accumulators [5]).

Chapter 7.

Conclusion

7.1. Results

It is time to revisit the questions I aimed to answer in the introductory chapter.

Q1: *What kind of quality dimensions are meaningful in the context of two different cultural heritage data sources: the collection of Europeana and MARC 21 format library catalogues?*

What I found useful – following the suggestion of Stvilia et al. [59] – is a mix of different quality dimensions, metrics and approaches. The main types of data quality measurement in this dissertation were the following:

1. *General structural and semantic metrics.* These measurements are the most well known in the literature. Following the seminal articles of this research domain [11, 51] they are:

- *completeness* (the existence of the defined fields in the records)
- *conformance to expectations* (schema rule validation and information value),
- *accessibility* (how easy is to understand the text of the record),
- *logical consistency and coherence*,
- *provenance* (the relationship between other metrics and the creator of the data)

I did not examined the *accuracy* dimension (comparison of a full data object and its metadata), because it requires comparison of metadata and the subject of it – i.e. full text of books – which were not available.

In Chapter 2 and 4 I showed different variations of completeness, which add a weighting factor to fields reflecting their cardinality (number of field instances) and importance (which come from functional requirement analysis).

2. *Support of functional requirements.* This dimension is a variation of completeness. Each data schema is created for supporting a set of functions, such as searching, identifying or describing objects. The data elements support one or more of these functions, and their existence and content has an impact of these functions. An example: a timeline widget expects a specific date format; if the field value is in another format the widget will ignore it. This family of metrics gives measures the scale of support of the functional requirement. To apply these metrics we should take the functional requirement analysis of the data schema and map the individual data elements (classes and properties) to the functions. The result will be a report which tells how the data supports the intended functions. Following the terminology established in [19] we call these scores 'sub-dimensions'. The Europeana Data Quality Committee defined a number of sub-dimensions (such as searchability, descriptiveness, identification, contextualisation, browsing etc.) which could be reused in other metadata domains. Regarding to MARC 21 schema the Library of Congress defined 12 user tasks, and created a mapping between them and the schema's data elements [16, 49]. It turned out that the approach to measure of functional support is closely bound to completeness, and since the total number of data elements in MARC are much higher than the actually available fields in the records, not just the completeness is low, but the functional support is low as well.

3. *Existence of known data patterns.* These are schema- and domain-specific patterns which occur frequently in the datasets. There are good patterns which detect good data creation practices, and anti-patterns, which should be avoided (such as data repetition, meaningless data etc.). For some domains there are existing pattern catalogues (e.g. the Europeana Data Quality Committee works on a Europeana specific pattern catalogue, while [60] examined three SKOS validation criteria catalogues). In Chapter 4 I showed some of the anti-patterns in MARC 21 records. These measurements could be categorized under *conformance to expectations.*

4. *Multilinguality.* Resource Description Framework (RDF)[1] provides an easily adaptable technique to add a language tag to literal values, which makes multilinguality an important aspect in the Linked Open Data world. In cultural heritage databases the translation of the descriptive fields (such as title, description) might be quite a human resource intensive task. On the other hand reusing existing multilingual dictionaries for subject headings is a relatively simple and cheap process. On the measurement side the nice thing is that generally the multilingual layer in metadata schemas (even in those not built on top of RDF) are similar, so the implementation can be abstracted. The big problem is how to

[1] `https://www.w3.org/RDF/`

handle the biases generated by the different cardinality and importance of the data elements. For example Europeana has "document" as a subject heading which is accessible in more than seventy languages, but it is attached to a great portion of records (more than 20%), so its information value or distinctive power is low – if the user searches for documents she receives millions of records. Chapter 3 explained multilinguality in details. This measurement could be categorised under *conformance to expectations* and *accessibility*.

The common point in these metrics is that they can be implemented as generic functions where input parameters are specific elements of a data schema. The functions themselves should not know about the details of the schema; that is to say they should be schema-independent. In other words: the only thing we should create on a schema by schema basis is a method which takes care of mapping the schema elements and measurement functions and feeds these generic functions with the appropriate metadata elements.

The measurement process has the following phases:

1. data ingestion
2. measuring individual records
3. analysing the the results of the measurement to get an aggregate view for the whole or a subset of the collection
4. reporting the results
5. discussing the results within an expert community

These phases create a loop; after phase 5 the process either ends, or goes back to phase 2, 3, or 4.

This thesis concentrated on the second and third phases, of course it uses the discussions in phase 5, but only touches fourth phase. Data ingestion is a data source specific technical task, while research on reporting has not yet been finished.

Q2: *How could it be implemented in a flexible way, so the solution should remain easily extensible to measure the same metrics on data sources in other formats?*

In Chapter 5 I described the levels of abstractions which needs in order to make measurement flexible. Its basic blocks are:

- The metadata schema should be mapped into a class which implement an abstract Schema interface, and where each measurable data element is addressed in a standard addressing language (JsonPath). The interface provides methods to access subsets of fields, individual fields, the properties of fields etc.

- The measurement methods access schema through a (Java) interface. It practically means that the methods should not know the details of the schema, it is enough if they can retrieve the data elements to analyse.
- Each measurement implements the same methods to launch the action of measurement and retrieve the result(s) of it.

This abstraction – following Liskov's substitution principle[2] – guarantees that the process of general measurements can be conducted smoothly on records of different metadata schemas.

I should note that at the beginning of the research I thought that the schema abstraction could be easily implementable with simply analysing an XML schema, but it turned out, that there are several layers which requires knowledge not recorded in a schema file, such as mapping with functions, or which fields should be covered in or left out from distinct analyses. For this phase of the research I did not solved this problem, but I could imagine a tool which reads an XML schema, and then guide the user through a "wizzard" to add the missing information by hand.

Q3: *How can these measurement be implemented in scalable way?*

In this current status the measurement process is based on Apache Spark, which could be run on a single machine or in a cluster of machines. I tested it within GWDG HPC cluster and a master student, Al-Gumaei [2] tested it a small cluster of 3 virtual machines. Both scaling tests worked, and shortened the processing time. Apache Spark not just run on multiple nodes, but can use the multicore CPUs, and its API hides the details of multithread management, so the developer can focus on task itself, while Spark focuses on scalability. I should note, that in the Europeana environment HPC or VM cluster is not available, the tool runs on a single machine which should be shared with other applications. I had to respect this constraint in the development process, and it made the tool more flexible I believe.

A historical note: at the beginning of the research it became evident, that Europeana's data size is too big to analyse with traditional methods. Traditional approaches has different limitations. For example in R language (which provides the best statistical feature set among the open source tools) a developers can read as much data as the size of the memory. Studying the state-of-the-arts solutions, I found first Apache Hadoop, then Apache Spark. Hadoop implemented the map-reduce paradigm, which makes parallelisation of taks quite simple (keeping the details of Java multithreading out of the picture). Apache Spark moves this concept even further with a richer API, which covers statistical functions,

[2] `https://stackify.com/solid-design-liskov-substitution-principle/`

and machine learning algorithms as well. The measurement process has a record level stage, and an aggregation/statistical analysis stage. At first I implemented only the first stage based on Spark, while the statistical analysis has been done with R. The limitation for R made me introduce different trick and hacks to overcome it, but on the long run it does not proved to be maintainable: these tricks broke the workflow of the process, and required manual interventions. So in the last phase of the research I rewrote these analysis in Scala based on Spark API, so now this phase is also run with Spark. At the end the whole measurement process (including data ingestion) takes 4 days (at the beginning it took 3 months). The MARC validation is also can be run with Apache Spark,[3] however since they are not qualified as big data it has less relevancy there.

Q4: *How could Big Data analysis be conducted with limited computational resources?*

It sounds as a poor man's Big Data analysis. My subjective impression is that those organisations I met during this research usually do not have enough financial background to invest in high capacity computational infrastructure. There are exceptions of course, but my intention was that the measurement should not require expensive hardware. It should run on environment with limited resources, even if multiple applications use its resources at the same time. In Chapter 6 I did experiences with four parameter settings (number of CPU, allocated memory, file system type, and data compression), and showed that maximal resource allocation doesn't maximise the performance (measured in speed), but running simple tests with a sample of the data source could predict parameter settings which results a 'good enough' performance (close to the maximum), and respects the resource requirements of other applications.

7.2. Deliverables

During the research I had three tracks of activities

1. Studying the metadata quality domain
2. Engineering the tool
3. Communicating about metadata quality and dissemination of the results

In the first track I studied the metadata quality literature (including a portion of data and information quality literature). Since this domain is not a purely theoretical field, looking for current practice – as far as it was possible – I participated in or at least followed the activities

[3]Regarding to the details, consult with my blog entry `http://pkiraly.github.io/2018/01/18/marc21-in-spark/`.

of metadata quality projects and metadata quality related activities of digital libraries (among others Europeana, Deutsche Digitale Bibliothek, Digital Public Library of America). My interest was rather practical, and less theoretical. My special point of view was the question how the results of these papers could be turned into an Open Source software which could be used in different contexts, however during the process – together with members of Europeana Data Quality Committee – we also introduced a new metric we haven't found in the literature, which measures the multilingual aspects of metadata.

The main purpose of the engineering track was to create a general metadata quality measuring framework. It became a set of modules, such as a core Java API, an Europeana specific Java API, clients for them such as a REST API, a Spark interface, a web user interface, then a MARC 21 tool, a Wikidata-centric tool etc. each adding a specific purpose layer to the lower levels. The tool – tested through different experiments – proved to be both flexible (adaptable to different metadata schemas) and scalable.

The software packages created as part of the dissertation are:

- Harvester client for Europeana's OAI-PMH server (`https://github.com/pkiraly/europeana-oai-pmh-client`). It turned out, that Europeana's OAI-PMH implementation is not robust enough to let clients to harvest the whole dataset. Later I was given access to a MongoDB server, and in most of the time I used this directly instead of using the OAI-PMH protocol.[4]
- General Metadata QA API. Source code: `https://github.com/pkiraly/metadata-qa-api`, Maven[5] artifact (binary Java library): `http://mvnrepository.com/artifact/de.gwdg.metadataqa/metadata-qa-api`
- The Europeana-specific Europeana QA Java API: Source code: `https://github.com/pkiraly/europeana-qa-api`, Maven artifact: `http://mvnrepository.com/artifact/de.gwdg.metadataqa/europeana-qa-api`
- Apache Spark interface of Europeana QA Java API: `https://github.com/pkiraly/europeana-qa-spark`
- Analysis with R: `https://github.com/pkiraly/europeana-qa-r`. By the end of 2018 it became clear that the limitations of R as a language (mainly that it could process as much data as the size of the available memory) blocks the targetted monthly cycle of data

[4]Later the issues I found were reportedly fixed, however I haven't tested the service again.

[5]Maven is a Java building tool and a network of distributed Java library repositories. As a Java or Scala developer you can reuse others' libraries. These libraries has versions. The Maven repositories contain signed binary files and their metadata, which can be used in different Java projects independent from the building tool.

analysis, so the whole analysis was rewritten in Scala with Spark API, which became part of the europeana-qa-spark codebase, and this R based code base will not have been updated in the future.
- Web interface: `https://github.com/pkiraly/europeana-qa-web`
- REST and command line interface: `https://github.com/pkiraly/europeana-qa-client`
- Apache Solr connector: `https://github.com/pkiraly/europeana-qa-solr`
- Cassandra connector: `https://github.com/pkiraly/europeana-qa-cassandra`
- Metadata assessment for MARC records. Source code: `https://github.com/pkiraly/metadata-qa-marc`, Maven artifacts: `https://mvnrepository.com/artifact/de.gwdg.metadataqa/metadata-qa-marc`
- Web interface for displaying the result of the quality assessment of MARC records `https://github.com/pkiraly/metadata-qa-marc-web`
- Quality assessment for the bibliographic records of Wikidata `https://github.com/pkiraly/metadata-qa-wikidata`
- Supplementary materials for paper submitted to EURO-PAR conference `https://github.com/pkiraly/euro-par`

The third aspect of the research was the communication and dissemination of the results which activity accumulated in the current dissertation. The list of presentations and papers published in the context of this presentation is available in Appendix C.

As we saw, metadata quality has multiple dimensions. For each data source we should select those which fit to them both theoretically and practically. The measures have their "computational footprints": the calculation require a given amounts of human and IT resources (and they are not always foreseeable) we should take it into account both in research and non-research projects. Another important aspect is the human condition: the metrics should not be something which is meaningful only from statistical viewpoint, they should be meaningful for the maintainer of data as well. The metrics should help a decision making process about the modification of the data. During the research it was the hardest point: to find the intersection of the interests with metadata experts. It repeatedly happened that what I provided as a result was not useful from the cataloguers' perspective, so I had to improve it based on the feedbacks. It was a pleasant situation, that during the research I worked together with an expert group, the Europeana Data Quality Committee, whose members provided me constant feedback (and requests too, naturally). This collaboration started in 2016, but the work hasn't yet

finished, since we haven't reached a clear consensus on what information are useful.

I plan to continue this kind of research in the future, however a similar collaboration with institutions or experts is a strong necessity for any further research.

7.3. Future work

7.3.1. Research data

CoreTrustSeal is a certification for research data repositories, based on the DSA-WDS Core Trustworthy Data Repositories Requirements[6]. The certification is a successor of Data Seal of Approval. Its purpose is prove that the certified repositories are following best practices of research data management. Organizations should explain their activities in 15 areas, such as data access, licences, workflow, data integrity etc. There are two areas which are interesting from the aspect of metadata quality measurement: appraisal and data quality. The certificates contain the organization's answer and the certifying institution's notes, and they are publicly available[7]. At the time of writing there are 54 CoreTrustSeal certified repositories. The certifications are quite interesting documents, and together they provide a kind of cross section of the state of the art in the 15 areas of data repositories. It seems that their data and metadata quality activities concentrate on the following topics:

- setting the list of recommended and accepted file formats, and checking incoming files against it
- documentation efforts on different levels (general, domain specific, national) creating manuals and guides both for the users and the maintainers of the repository
- data curation by experts – most of these repositories are not self-service, the deposited materials are carefully checked by human experts. They check both the archival aspects (formats, metadata) and domain aspects (content relevancy)
- management of sensitive data (secure data management or excluding non anonymized data)
- setting mandatory, recommended and optional fields regarding to the metadata records

[6] see Core Trustworthy Data Repositories Extended Guidance v1.1 (June, 2018) https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf

[7] https://www.coretrustseal.org/why-certification/certified-repositories/

- online form validation – for the metadata created via an online user interface
- some repositories apply XML validators when the metadata record is expected to be available in XML format

Among the traditional metadata quality dimensions only completeness is mentioned, and it is used as a synonym of the case when all of the mandatory fields are available in the metadata record ("Ensuring DDI fields are completed in the metadata ensures quality control of completeness." wrote the Australian Data Archive[8]). Only a small portion of repositories mentioned usage of controlled vocabulary, and in this preliminary research I found only one repository which named an independent tool used for automating the metadata quality check[9]. The Worldwide Protein Data Bank[10] mentioned that they created two kinds of representations of data quality assessment: one for specialists, and another for non-specialists. The later contains simple graphical depiction that "highlights a small number of essential quality metrics". Different repositories mentioned that they reuse good quality metadata records as examples in the documentation.

It is worth to quote the checklist of Digital Repository of Ireland[11] in which they describe the recommended steps to conduct regular metadata quality assessments:

- Designate one or a small team of information professionals to take responsibility for the audit.
- Decide to what extent any mistakes found during the audit will be fixed within the live database.
- On a quarterly or biannual basis, upload a sample set of records to the software application OpenRefine.
- Use the Faceting and Cluster tools in OpenRefine to identify and record errors, such as misspellings, inconsistent use of capitalisation or blank cells.
- Compile the documentation so that any changes in quality can be noted over a period of time. This will be particularly useful if the organisation has recently started using new cataloguing methods.

The most widely used general metadata schemas are the elements of Data Documentation Initiative (DDI)[12] framework,[13] and The Dublin Core

[8] https://assessment.datasealofapproval.org/assessment_245/seal/html/

[9] FDAT, Tübingen uses docuteam packer see https://wiki.docuteam.ch/doku.php?id=docuteam:packer

[10] https://assessment.datasealofapproval.org/assessment_281/seal/html/

[11] https://repository.dri.ie/catalog/sj13pg68d

[12] http://www.ddialliance.org/

[13] DDI Lifecycle (http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/) and DDI Codebook (http://www.ddialliance.org/

Metadata Initiative's DCMI Metadata Terms[14]. Regarding to domain specific metadata schemas CLARIN's Component Metadata[15] could be viewed as a domain specific standard in lingistic data repositories.

An important conclusion from this preliminary survey is that there is a kind of "market gap" both in research and tool development in the domain of research data management. The elements of data quality mentioned in the certificates (completeness, format consistency, content relevancy, checking facets for errors etc.) are not different than those elements one can find in other metadata domains. There are elements which exist, but apparently did not get so far the popularity they deserve, e.g. the "frictionaless data" data description metadata format [18] or FAIRmetrics [20] I discussed in Chapter 1. Not to mention general elements of the metadata quality research (dimensions, metrics and tools), which could be introduced into this domain, for the satisfaction of both parties.

7.3.2. Citation data

Citation data, or bibliographic data of scholarly articles is a neuralgic point for the libraries. In the "Western World" and for large languages, the publishers are those players which traditionally built databases for the scholarly articles (such as Web of Knowledge, Scopus) instead of libraries. By and large there has been exceptions even in Western European countries. In case of smaller languages and for poorer countries the large publishers does not see the market value to publish scientific journals in vernacular languages therefore those journals are not covered in their databases. In the last two decades several different projects have been launched to make these metadata out of "paywalls". The largest of these project is the DOI database, but the larger part of DOI metadata is also not freely available, however the Initiative for Open Citations[16] works on making the citation data open. Recently WikiCite[17] is the largest freely available citation database based on the bibliographic data imported into Wikidata[18]. It provides query interface and database dumps[19]. Together with Jakob Voß, a volunteer of WikiCite and Wikidata we started a research project[20] to analyze the data. This research is in a preliminary stage. Now I highlight only one feature of the citation data namely *page*

Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html)

[14] http://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[15] https://www.clarin.eu/content/component-metadata

[16] https://i4oc.org/

[17] http://wikicite.org/

[18] https://www.wikidata.org/wiki/Wikidata:Main_Page

[19] http://wikicite.org/access.html

[20] Its code is en extension of the same codebase I wrote for the Europeana analysis. It is available at https://github.com/pkiraly/metadata-qa-wikidata/wiki.

numbers, which seems to be simple, but reveals some complex problems. One can expect that page numbers are arabic or roman numbers separated by dashes and commas (sometimes with some text before or after the numbers). I found however several hundred patterns, which do not fit this expectation. Here I show three issues with "strange" page numbers.

A note before the examples. Wikidata uses a language neutral notation for describing its semantic structure, the entities are denoted by 'Q' and a number, while properties are denoted by 'P' and a number. For example: P304 is the property of the page numbers. Its human readable label in English is "page(s)"[21].

1. Using article identifier as page number

Example #1. Q40154916[22]: P304 = "e0179574"

This article was published in PLoS ONE[23]. The publisher provides citation text, and metadata in RIS and BibTeX format. The citation contains 'e0179574', however it does not explain what exactly it means (as it neither explains any other elements):

Citation on the journal's page:

Vincent WJB, Harvie EA, Sauer J-D, Huttenlocher A (2017) *Neutrophil derived LTB4 induces macrophage aggregation in response to encapsulated Streptococcus iniae infection.* PLoS ONE 12(6): e0179574. `https://doi.org/10.1371/journal.pone.0179574`

In the PDF version of the paper[24] there are page numbers. It also does not explain what 'e0179574' means.

Metadata in RIS[25] (only the relevant part):

```
SP  - e0179574
EP  -
```

SP stands for starting page, EP stands for ending page. It is evident here, that 'e0179574' is not a starting page.

The BibTeX version[26] contains page number, however does not contain the article identifier at all:

[21] `https://www.wikidata.org/wiki/Property:P304`
[22] `https://www.wikidata.org/wiki/Q40154916`
[23] `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0179574`
[24] `https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0179574&type=printable`
[25] https://journals.plos.org/plosone/article/citation/ris?id=10.1371/journal.pone.0179574
[26] `https://journals.plos.org/plosone/article/citation/bibtex?id=10.1371/journal.pone.0179574`

```
@article{10.1371/journal.pone.0179574,
  year = {2017},
  month = {06},
  volume = {12},
  pages = {1-16},
  ...
}
```

The DOI database follows the content of RIS metadata file instead of BibTeX[27]:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:j.1="http://prismstandard.org/namespaces/basic/2.1/"
    xmlns:j.2="http://purl.org/ontology/bibo/" ...>
  <rdf:Description rdf:about=".../10.1371/journal.pone.0179574">
    <j.2:pageStart>e0179574</j.2:pageStart>
    <j.1:startingPage>e0179574</j.1:startingPage>
    ...
  </rdf:Description>
</rdf:RDF>
```

Example #2. Q21820630[28]: P304 = "c181"

This paper was published in the British Medical Journal[29]. It does not have a PDF version, the only available online version is HTML.

Hrynaszkiewicz Iain, Norton Melissa L, Vickers Andrew J, Altman Douglas G. *Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers* BMJ 2010; 340 :c181 `https://www.bmj.com/content/340/bmj.c181` (DOI: 10.1136/bmj.c181)

The RIS metadata[30] contains bad page number, uses the article identifier:

```
SP  - c181
```

In BibTex metadata[31] there is no page number, but an 'elocation-id' field is available:

```
@article{Hrynaszkiewiczc181,
  elocation-id = {c181},
  ...
}
```

The DOI database follows again the RIS metadata file instead of BibTeX[32], however it repeats the same string in ending page fields:

[27] I used the Crossref API with the following command to retrieve a DOI metadata `curl -H "Accept: application/rdf+xml" http://data.crossref.org/[DOI]`, such as http://data.crossref.org/10.1371/journal.pone.0179574

[28] `https://www.wikidata.org/wiki/Q21820630`

[29] `https://www.bmj.com/content/340/bmj.c181`

[30] `https://www.bmj.com/highwire/citation/194003/ris`

[31] `https://www.bmj.com/highwire/citation/194003/bibtext`

[32] `http://data.crossref.org/10.1136/bmj.c181`

```
<j.1:startingPage>c181</j.1:startingPage>
<j.2:pageStart>c181</j.2:pageStart>
<j.1:endingPage>c181</j.1:endingPage>
<j.2:pageEnd>c181</j.2:pageEnd>
```

A possible conclusion would be that if there is a BibTeX source available, and it doesn't have page number element, but elocation-id, it is a better source than the RIS version.

JATS (Journal Archiving and Interchange Tag Library)[33] defines the `<elocation-id>` element as "replaces the start and end page elements just described for electronic-only publications;"[34].

In this journal (British Medical Journal) some article has a PDF version, (e.g. `https://doi.org/10.1136/bmj.d1584`, `https://doi.org/10.1136/bmj.e1454`, `https://doi.org/10.1136/bmj.a494`) where there are clearly page numbers. The journal provided RIS and BibTeX metadata do not contain page numbers in those cases.

2. Wikidata contains extra info, which is not available elsewhere

Q39877401[35]: P304 = "108-17; quiz 118-9"

The article has been published in *Orthopaedic Nursing*[36]

Schroeder, Diana L.; Hoffman, Leslie A.; Fioravanti, Marie; Medley, Deborah Poskus; Zullo, Thomas G.; Tuite, Patricia K. *Enhancing Nurses' Pain Assessment to Improve Patient Satisfaction* Orthopaedic Nursing: March/April 2016 - Volume 35 - Issue 2 - p 108–117. DOI: 10.1097/NOR.0000000000000226.

Page number in DOI[37] repeats the information provided at the journal:

```
<bibo:pageStart>108</bibo:pageStart>
<prism:startingPage>108</prism:startingPage>

<bibo:pageEnd>117</bibo:pageEnd>
<prism:endingPage>117</prism:endingPage>
```

In the publisher's citation the page number contains the first part of the page number string of the Wikidata value, but not the "; quiz 118-9" part. In the table of contents there is another article[38] at page 118-119 of the

[33] NISO JATS Version 1.1 (ANSI/NISO Z39.96-2015) `https://jats.nlm.nih.gov/archiving/tag-library/1.1/`

[34] `https://jats.nlm.nih.gov/archiving/tag-library/1.1/element/elocation-id.html`

[35] `https://www.wikidata.org/wiki/Q39877401`

[36] `https://journals.lww.com/orthopaedicnursing/pages/articleviewer.aspx?year=2016&issue=03000&article=00010&type=abstract`

[37] `http://data.crossref.org/10.1097/NOR.0000000000000226`

[38] `https://journals.lww.com/orthopaedicnursing/Citation/2016/03000/Enhancing_Nurses__Pain_Assessment_to_Improve.11.aspx#print-article-link`, DOI: 10.1097/NOR.0000000000000236

same issue. It's title is the same, but it does not have authors recorded. It is categorised under "CE Tests" (where CE means continuing education). The two articles don't link directly to each other. They are different, and have different DOIs. The DOI database neither contains any link between them.

Wikidata should keep them separated, however it would require rather long time to investigate cases like this.

3. Wikidata uses page number field to add comment

Q28710224[39]: P304 = "E3523; author reply E3524–5"

Alain Pierret, Valéry Zeitoun, Hubert Forestier: *Irreconcilable differences between stratigraphy and direct dating cast doubts upon the status of Tam Pa Ling fossil.* Proceedings of the National Academy of Sciences Dec 2012, 109 (51) E3523; DOI: `http://doi.org/10.1073/PNAS.1216774109`, URL in journal: `https://www.pnas.org/content/109/51/E3523`, URL in Wikidata: `https://www.wikidata.org/wiki/Q28710224`.

E3524–5 is not part of the article. It is a related article, which is also available in Wikidata (as Q28710226[40]):

Fabrice Demeter, Laura L. Shackelford, Kira E. Westaway, Philippe Duringer, Thongsa Sayavongkhamdy, and Anne-Marie Bacon: *Reply to Pierret et al.: Stratigraphic and dating consistency reinforces the status of Tam Pa Ling fossil* PNAS December 18, 2012 109 (51) E3524-E3525; `https://doi.org/10.1073/pnas.1217629109`, URL in journal: `https://www.pnas.org/content/109/51/E3524`.

These two articles are not interlinked with distinct properties. One could suppose, that the occurrence 'author reply' in other Wikidata records' page number could reveal similar hidden links.

7.3.3. Fixing issues – is that possible?

The Swedish National Heritage Board just launched a project called *Wikimedia Commons Data Roundtripping*[41]. *Roundtripping* is the name of the workflow in which a cultural heritage institution publish their data in Wikimedia Commons, the users enrich these openly available media (such as adding translations of descriptive texts into other languages,

[39] `https://www.wikidata.org/wiki/Q28710224`
[40] `https://www.wikidata.org/wiki/Q28710226`
[41] `https://outreach.wikimedia.org/wiki/GLAM/Newsletter/February_2019/Contents/Special_story`

identifying people, names and aliases, locations and subject matter or linking to authority data and using it to retrieve third party contributions from other memory organisations), then institutions ingest these data and update their original database.

Wikidata doesn't seem to be the platform where the data are generated, but rather it is a place where data created elsewhere are imported to and maybe enriched. Additionally, the bibliographical data in general seems to have their own data flow: from different sources they are duplicated into different targets, such as DOI database(s), commercial or community funded discovery interfaces (Scopus[42], Web of Science[43], Google Scholar[44], Microsoft Academic[45], SpringerNature SciGraph[46], ORCID[47]), institutional repositories, open data platforms (Wikidata, Wikipedia, DBPedia) etc. It is clear, that it would be impossible to fix the data in each platform, and probably the best place to fix them in their origin.[48] It seems that DOI database is a kind of central hub in this network.

Who are responsible for metadata published in DOI database?

According to Göttingen eResearch Alliance's DOI experts Timo Gnadt and Sven Bingert DOI data is created and updated by the data providers, i.e. the institutions which own the data (usually they are the institutions behind the landing page of the metadata records), thus in our case they are the publishers of the journals.

In one case I wrote a report to Springer Nature regarding to the page numbers for Wilkinson et al, *The FAIR Guiding Principles for scientific data management and stewardship*,[49] so far the discussion is about to clarify the problem, I do not have the publisher's opinion about the issue (and if they consider it as an issue).

This research could provide suggestions for Wikidata to fix the data in a format which could be used by bots to update the content. If such a fixing process is implemented, Wikidata's ingestion process should be aware of the updated information, to be sure that they are not overwritten by a next ingestion process. Also these corrections should be sent to the publishers.

[42] `https://www.scopus.com/`
[43] `https://www.webofknowledge.com/`
[44] `https://scholar.google.de/`
[45] `https://academic.microsoft.com/`
[46] `https://scigraph.springernature.com/`
[47] `https://orcid.org/`
[48] Ruben Verborgh, in his keynote speech at ELAG 2018 conference ("The delicate dance of decentralization and aggregation" `http://slides.verborgh.org/ELAG-2018/`) suggested a researcher centric scholarly communication where the origin of the citation data is the researchers' own home pages, and aggregation services like those mentioned above harvest data from there.
[49] Scientific Data volume 3, Article number: 160018 (2016) `https://www.nature.com/articles/sdata201618`

7.3.4. Participation in metadata quality activities

In 2016 two important groups formed in the Cultural Heritage sector which started a deep investigation of data quality in particular segments: the Europeana Data Quality Committee[50] (DQC) and DLF Metadata Assessment Working Group[51] (MAWG). DQC examines the metadata issues specific to the Europeana collection, and involved in creating the measuring framework which is the main subject of this dissertation. The MAWG does not focus on a particular service and metadata schema, they collect relevant literature, use cases, and aims to form a set of recommendations on metadata assessment. In 2017 ADOCHS[52] (Auditing Digitalization Outputs in the Cultural Heritage Sector, Belgium) launched aiming to improve the quality control process concerning the digitized heritage collections of the Belgian national library and national archives[53]. Similar activities of the Digital Public Library of America (DPLA) is described in [22].

I participated or followed the work of these groups. I am a member of DQC and contributed to the environmental scan task of MAWG. I presented my results in a MAWG virtual seminar, and I was a member of ADOCHS follow up committee. I should note that I was successful in initialising communications with a number of other actors, such as Deutsche Digitale Bibliothek, Bibliothèque National de France, DPLA, University of North Texas libraries, and others, but rarely I was able to turn this initial sympathy into a fruitful collaboration. Although it is not a research purpose, but in the future I would like to improve those skills which are required for taking this additional step.

7.4. Acknowledgement

I would like to thank the help of Juliane Stiller, Zaveri Amrapali, Christina Harlow, Valentine Charles, and Antoine Isaac, and Stefan Gradmann. Thanks to the anonymous conference or journal reviewers who gave me important feedback.

Thanks to Philipp Wieder who encouraged me to conduct a PhD research, and to Gerhard Lauer, Ramin Yahyapour and Marco Büchler for supervising this research.

Thanks to GWDG[54] for supporting my research in different ways, to

[50] `http://pro.europeana.eu/page/data-quality-committee`
[51] `https://dlfmetadataassessment.github.io/`
[52] `http://adochs.be/`
[53] The results of ADOCHS project could be found in the publications of Anne Chardonnens and Ettore Rizza – see Appendix A.
[54] `http://gwdg.de`

Europeana and eTRAP[55] research group for using their computers, to JetBrains s.r.o. for IntelliJ IDEA[56] community licence, to developers of Open Source software packages, and infrastructure services I used in the research, and to Open Data publishers for their data.

Köszönet Ildikónak, Zsófinak, Verának és Lucának szeretetükért és türelmükért. Köszönöm továbbá családomnak és barátaimnak – kinek-kinek mást és mást.

„Lesznie kell –” (“It must to be –”, Péter Esterházy)[57]

[55] `https://www.etrap.eu/`

[56] `https://www.jetbrains.com/idea/`

[57] A little Hungarian pornography. Translated by Judith Sollosy (Northwestern University Press, 1997.). Chapter 3 „?” p. 153.

Appendix A.

Metadata assessment bibliography

Péter Király, Sarah Potvin, Nikos Palavitsinis, Anna Neatrour, Ayla Stein, Sara R., Juliane Stiller, Kevin Clair, Pru Mitchell, Corey Harper, Christina Harlow, Laura Akerman, Laura J. Smart, David Maus, jenyoung, Kate Flynn, leonardaa, Cassandra Baker

This bibliography is the result of a community effort. It was built on Zotero platform as Metadata Assessment Group Library[1]. The bibliography was initialized by Corey Harper in February 2016, and DFL Metadata Assessment working group used for recording items found during their environmental scan. Soon Europeana Data Quality Committee also started contributing to it. During my PhD research I intensively used it as well. According to the Zotero API[2] the creators of the bibliography entries are Péter Király (94 items), Sarah Potvin (32), Nikos Palavitsinis (18), Anna Neatrour, Ayla Stein, Sara R. (7), Juliane Stiller, Kevin Clair (4), Pru Mitchell, Corey Harper, Christina Harlow, Laura Akerman, Laura J. Smart (2), David Maus, jenyoung, Kate Flynn, leonardaa. Cassandra Baker contributed with improving of existing bibliography items.

*

Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., and Lehmann, J. 2013. Crowdsourcing Linked Data Quality Assessment. *The Semantic Web – ISWC 2013*, Lecture Notes in Computer

[1] `https://www.zotero.org/groups/488224/metadata_assessment`

[2] `https://api.zotero.org/groups/488224/items`. In order to retrieve every items, one should have an API key, and should use parameters 'start' and 'limit' (see Zotero API documentation at `https://www.zotero.org/support/dev/web_api/v3/basics`.

Science 8219. Heidelberg: Springer, 260–276. https://doi.org/10.1007/978-3-642-41338-4_17

Albertoni, R., De Martino, M., and Podestà, P. 2018. Quality measures for skos: ExactMatch linksets: an application to the thesaurus framework LusTRE. *Data Technologies and Applications* https://doi.org/10.1108/DTA-05-2017-0037.

Alemneh, D.G. 2009. Metadata Quality Assessment: A Phased Approach to Ensuring Long-term Access to Digital Resources. *Proceedings of the American Society for Information Science and Technology 46*, 1, 1–8 https://doi.org/10.1002/meet.2009.1450460380.

Askham, N., Cook, D., Doyle, M., et al. 2013. *The Six Primary Dimensions For Data Quality Assessment. Defining Data Quality Dimensions*. DAMA UK. https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf.

Bade, D. 2008. The Perfect Bibliographic Record: Platonic Ideal, Rhetorical Strategy or Nonsense? *Cataloging & Classification Quarterly 46*, 1, 109–133 https://doi.org/10.1080/01639370802183081.

Barton, J., Currier, S., and Hey, J.M.N. 2003. Building Quality Assurance into Metadata Creation: An Analysis based on the Learning Objects and e-Prints Communities of Practice. *Papers and Project Reports for DC-2003 in Seattle, 28 September - 2 October 2003. Supporting Communities of Discourse and Practice*, 39–48 http://dcpapers.dublincore.org/pubs/article/view/732.

Beall, J. 2005. Metadata and Data Quality Problems in the Digital Library. *Journal of Digital Information 6*, 3, 20.

Bellini, E. and Nesi, P. 2013. Metadata Quality Assessment Tool for Open Access Cultural Heritage Institutional Repositories. *Information Technologies for Performing Arts, Media Access, and Entertainment*, Springer, 90–103. https://doi.org/10.1007/978-3-642-40050-6_9.

Bellini, P., Bruno, I., Nesi, P., and Paolucci, M. 2015. IPR Centered Institutional Service and Tools for Content and Metadata Management. *International Journal of Software Engineering and Knowledge Engineering 25*, 08, 1237–1270. https://doi.org/10.1142/S0218194015500242.

Bruce, T.R. and Hillmann, D.I. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: D. Hillman and E. Westbrooks, eds., *Metadata in practice*. ALA Editions, Chicago, IL, 238–256. http://ecommons.cornell.edu/handle/1813/7895.

Bruce, T.R. and Hillmann, D.I. 2013. Metadata Quality in a Linked Data Context. https://blog.law.cornell.edu/voxpop/2013/01/24/metadata-quality-in-a-linked-data-context.

Cai, L. and Zhu, Y. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal 14*, 2 https://doi.org/10.5334/dsj-2015-002.

Cechinel, C., da Silva Camargo, S., Sánchez-Alonso, S., and Sicilia, M.-Á. 2012. On the Search for Intrinsic Quality Metrics of Learning Objects. *Metadata and Semantics Research*, Springer, 49–60 https://doi.org/10/gfphbr.

Cechinel, C., da Silva Camargo, S., Sicilia, M.-Á., and Sánchez-Alonso, S. 2016. Mining Models for Automated Quality Assessment of Learning Objects. *Journal of Universal Computer Science 22*, 1, 94–113. https://doi.org/10.3217/jucs-022-01-0094.

Charbonneau, M. 2005. Production benchmarks for catalogers in academic libraries: are we there yet? *Library Resources and Technical Services 49*, 1, 40–48.

Charles, V., Stiller, J., Király, P., Bailer, W., and Freire, N. 2018. Data Quality Assessment in Europeana: Metrics for Multilinguality. *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) (TDDL_MDQual_Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Grand Hotel Palace, Thessaloniki, Greece, 21 September 2017*, CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2038/paper6.pdf.

Chen, H.-H., Wu, J., and Giles, C.L. 2018. Compiling Keyphrase Candidates for Scientific Literature Based on Wikipedia. *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) (TDDL_MDQual_Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Grand Hotel Palace, Thessaloniki, Greece, 21 September 2017*, CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2038/paper6.pdf.

Chen, Y.-N., Wen, C.-Y., Chen, H.-P., Lin, Y.-H., and Sum, H.-C. 2011. Metrics for metadata quality assurance and their implications for digital libraries. *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation (Proceedings of the 13th International Conference on Asia-Pacific Digital Libraries, 24–27 October 2011, Beijing, China.)*, Springer, 138–147. http://ceur-ws.org/Vol-2038/paper4.pdf.

Clair, K. 2016. Technical Debt as an Indicator of Library Metadata Quality. *D-Lib Magazine 22*, 11/12. https://doi.org/10.1007/978-3-642-24826-9_19.

Clements, K., Pawlowski, J., and Manouselis, N. 2014. Why Open Educational Resources Repositories fail - Review of Quality Assurance Approaches. *EDULEARN14 Proceedings. 6th International Conference on Education and New Learning Technologies Barcelona, Spain, 2014*, IATED, International Association of Technology, Education and Development, 929–939. https://doi.org/10.1016/j.chb.2015.03.026.

Clements, K., Pawlowski, J., and Manouselis, N. 2015. Open educational resources repositories literature review – Towards a comprehensive quality approaches framework. *Computers in Human Behavior 51, Part B*, 1098–1106. https://jyx.jyu.fi/dspace/handle/123456789/44031.

Conrad, S. 2015. Using Google Tag Manager and Google Analytics to track DSpace metadata fields as custom dimensions. *The Code4Lib Journal* 27. http://journal.code4lib.org/articles/10311.

Debattista, J., Auer, S., and Lange, C. 2016. Luzzu - A methodology and framework for linked data quality assessment. *Journal of Data and Information Quality 8*, 1, 4:1-4:32. https://doi.org/10.1145/2992786.

Debattista, J., Lange, C., and Auer, S. 2014. Representing Dataset Quality Metadata using Multi-Dimensional Views. *SEM '14. Proceedings of the 10th International Conference on Semantic Systems*, ACM, 92–99. https://doi.org/10.1145/2660517.2660525.

Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., and Lumineau, N. 2016a. Open datasets for evaluating the interpretation of bibliographic records. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, ACM, 253–254. https://doi.org/10.1145/2910896.2925457.

Decourselle, J., Duchateau, F., Aalberg, T., Takhirov, N., and Lumineau, N. 2016b. BIB-R: A Benchmark for the Interpretation of Bibliographic Records. *Research and Advanced Technology for Digital Libraries*, Springer International Publishing, 163–174. https://doi.org/10.1007/978-3-319-43997-6.

Degerstedt, S. and Philipson, J. 2016. Lessons Learned from the First Year of E-Legal Deposit in Sweden: Ensuring Metadata Quality in an Ever-Changing Environment. *Cataloging & Classification Quarterly 54*, 7, 468–482. https://doi.org/10.1080/01639374.2016.1197170.

Diakopoulos, N., Friedler, S., Arenas, M., et al. 2016. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. http://www.fatml.org/resources/principles-for-accountable-algorithms.

DLF Aquifer Metadata Working Group and Digital Library Federation. 2009. *Digital Library Federation / Aquifer Implementation Guidelines for Shareable MODS Records, version 1.1.* Digital Library Federation. https://wiki.dlib.indiana.edu/display/DLFAquifer/DLF+Aquifer+Public+Metadata+Documents?preview=/28330/120160257/DLFMODS_ImplementationGuidelines.pdf.

DLF/NSDL Working Group on OAI PMH Best Practices. 2007. *Best Practices for OAI PMH DataProvider Implementations and Shareable Metadata.* Digital Library Federation, Washington, D.C. https://old.diglib.org/pubs/dlf108.pdf.

Dodds, L. Quality Indicators for Linked Data Datasets. http://answers.semanticweb.com/questions/1072/quality-indicators-for-linked-data-datasets.

Doorn, P. and Tsoupra, E. 2018. A Simple Approach to Assessing the FAIRness of Data in Trusted Digital Repositories. *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) (TDDL_MDQual_Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Grand Hotel Palace, Thessaloniki, Greece, 21 September 2017*, CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2038/invited2.pdf.

DPLAFest Participants. 2015. Metadata Quality Research. https://docs.google.com/document/d/15pmA276_fxShkCEagoloJwCXH89PhrF3qWBgB8xSrag/edit#.

Dublin Core Metadata Initiative. 2014. DCMI Task Group RDF Application Profiles. http://wiki.dublincore.org/index.php/RDF_Application_Profiles.

Durco, M. and Windhouwer, M. 2014. The CMD Cloud. *Proceedings of LREC 2014*, 687–690. http://www.lrec-conf.org/proceedings/lrec2014/pdf/156_Paper.pdf.

Dushay, N. and Hillmann, D.I. 2003. Analyzing Metadata for Effective Use and Re-Use. *Proceedings, Dublin Core Metadata Conference, DC-2003*, Dublin Core Metadata Initiative. http://dcpapers.dublincore.org/pubs/article/view/744.

Efron, M. 2007. Metadata Use in OAI-Compliant Institutional Repositories. *Journal of Digital Information 8*, 2. http://people.ischool.illinois.edu/~mefron/papers/efron-metadatause.pdf.

Ellefi, M.B., Bellahsene, Z., Breslin, J., et al. 2017. RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications. *Semantic Web Preprint*, Preprint. http://www.semantic-

web-journal.net/content/rdf-dataset-profiling-survey-features-methods-vocabularies-and-applications.

Europeana Tech. 2015. Evaluation and Enrichments Task Report Outcomes. http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments.

Färber, M., Bartscherer, F., Menne, C., and Rettinger, A. 2017. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web Preprint*, Preprint, 1–53. https://doi.org/10.3233/SW-170275.

Fischer, K.S. 2005. Critical Views of LCSH, 1990–2001: The Third Bibliographic Essay. *Cataloging & Classification Quarterly 41*, 1, 63–109. https://doi.org/10.1300/J104v41n01_05.

Foulonneau, M. 2007. Information redundancy across metadata collections. *Information Processing & Management 43*, 3, 740–751. http://dx.doi.org/10.1016/j.ipm.2006.06.004.

Foulonneau, M. and Cole, T.W. 2005. Strategies for Reprocessing Aggregated Metadata. *Research and Advanced Technology for Digital Libraries*, Springer Berlin Heidelberg, 290–301. https://www.researchgate.net/profile/Muriel_Foulonneau/publication/221176072_Strategies_for_Reprocessing_Aggregated_Metadata/links/54173a450cf2f48c74a403d1.pdf.

Foulonneau, M. and Riley, J. 2008. *Metadata for Digital Resources: Implementation, Systems Design and Interoperability*. Chandos Publishing, Oxford.

Fowler, D., Barratt, J., and Walsh, P. 2018. Frictionless Data: Making Research Data Quality Visible. *International Journal of Digital Curation 12*, 2, 274–285. https://doi.org/10.2218/ijdc.v12i2.577.

Friesen, N. 2004. *International LOM Survey: Report (Draft)*. http://arizona.openrepository.com/arizona/bitstream/10150/106473/1/LOM_Survey_Report2.doc.

Fürber, C. and Hepp, M. 2011. Towards a Vocabulary for Data Quality Management in Semantic Web Architectures. http://www.slideshare.net/cfuerber/towards-a-vocabulary-for-data-quality-management-in-semantic-web-architectures.

Gavrilis, D., Makri, D.-N., Papachristopoulos, L., et al. 2015. Measuring Quality in Metadata Repositories. *Proceedings from the 19th International Conference on Theory and Practice of Digital Libraries*, Springer, 56–67. https://doi.org/10.1007/978-3-319-24592-8_5.

Gendered Expectations for Leadership in Libraries – In the Library with the Lead Pipe. http://www.inthelibrarywiththeleadpipe.org/2015/libleadgender/.

GO FAIR Metrics Group. FAIR Metrics. http://fairmetrics.org/.

Gonçalves, M.A., Moreira, B.L., Fox, E.A., and Watson, L.T. 2007. "What is a good digital library?" – A quality model for digital libraries. *Information Processing & Management 43*, 5, 1416–1437. https://doi.org/10.1016/j.ipm.2006.11.010.

Goovaerts, M. and Leinders, D. 2012. Metadata quality evaluation of a repository based on a sample technique. *Metadata and Semantics Research*, Springer, 181–189. https://doi.org/10/gfphbs.

Greenberg, J., Spurgin, K., and Crystal, A. 2005. *Final Report for the AMeGA (Automatic Metadata Generation Applications) Project.* http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf.

Groskopf, C. 2015. The Quartz guide to bad data. *Quartz.* https://qz.com/572338/the-quartz-guide-to-bad-data/.

Gueguen, G. 2019. Metadata quality at scale: Metadata quality control at the Digital Public Library of America. *Journal of Digital Media Management 7*, 2, 115–126.

Guinchard, C. 2006. Dublin Core use in libraries: a survey. *OCLC Systems & Services: International digital library perspectives 18*, 1, 40–50. https://doi.org/10.1108/10650750210418190.

Guy, M., Powell, A., and Day, M. 2004. Improving the Quality of Metadata in Eprint Archives. *Ariadne* 38. http://www.ariadne.ac.uk/issue38/guy/.

Harper, C. 2016. Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA). *The Code4Lib Journal* 33. http://journal.code4lib.org/articles/11752.

Haslhofer, B. and Klas, W. 2010. A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys 42*, 2, 1–37. https://doi.org/10.1145/1667062.1667064. https://doi.org/10.1145/1667062.1667064.

Hillmann, D.I. 2008. Metadata Quality: From Evaluation to Augmentation. *Cataloging & Classification Quarterly 46*, 1, 65–80. https://doi.org/10.1080/01639370802183008.

Hillmann, D.I., Dushay, N., and Phipps, J. 2004. Improving Metadata Quality: Augmentation and Recombination. http://www.cs.cornell.edu/naomi/DC2004/MetadataAugmentation--DC2004.pdf.

Höffernig, M., Orgel, T., Russegger, S., and Bailer, W. 2015. Assessing Quality in Automated Metadata Aggregation and Mapping

Services. EEXCEES, 1–6. http://eexcess.eu/wp-content/uploads/2013/03/JoanneumResearch_Assessing_Quality.pdf.

Höffernig, M., Orgel, T., Russegger, S., and Bailer, W. 2013. Assessing Quality in Automated Metadata Aggregation and Mapping Services. http://eexcess.eu/wp-content/uploads/2013/03/JoanneumResearch_Assessing_Quality.pdf.

van Hooland, S. 2009. Metadata Quality in the Cultural Heritage Sector: Stakes, Problems and Solutions. http://homepages.ulb.ac.be/~svhoolan/these.pdf.

Huang, Y. and Chiang, F. 2018. Refining Duplicate Detection for Improved Data Quality. *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) (TDDL_MDQual_Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Grand Hotel Palace, Thessaloniki, Greece, 21 September 2017*, CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2038/paper3.pdf.

Hughes, B. 2004. Metadata Quality Evaluation: Experience from the Open Language Archives Community. *Digital Libraries: International Collaboration and Cross-Fertilization*, Springer, 320–329. https://doi.org/10.1007/978-3-540-30544-6_34.

Jackson, A., Han, M.-J., Groetsch, K., Mustafoff, M., and Cole, T.W. 2008. Dublin Core Metadata Harvested Through OAI-PMH. *Journal of Library Metadata 8*, 1, 5–21. https://doi.org/10.1080/10911360802076682.

Jay, M., Simpson, B., and Smith, D. 2009. CatQC and Shelf-Ready Material: Speeding Collections to Users While Preserving Data Quality. *Information Technology and Libraries 28*, 1, 41–48. https://doi.org/10.6017/ital.v28i1.3171.

Kaffee, L.-A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L., and Pintscher, L. 2017. A glimpse into babel: An analysis of multilinguality in wikidata. *OpenSym '17 Proceedings of the 13th International Symposium on Open Collaboration*, ACM, 14:1–14:5. https://doi.org/10.1145/3125433.3125465.

Kaffee, L.-A. and Simperl, E. 2018. The human face of the web of data: a cross-sectional study of labels. 66–77. https://doi.org/10.1016/j.procs.2018.09.007.

Kapidakis, S. 2012. Comparing metadata quality in the Europeana context. *PETRA '12 Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments. Heraklion,*

Crete, Greece — June 06 - 08, 2012, ACM, 25:1-25:8. https://doi.org/10.1145/2413097.2413129.

Kapidakis, S. 2015. Rating quality in metadata harvesting. *PETRA '15 Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments. Article No. 65. Corfu, Greece — July 01 - 03, 2015.*, ACM, 65:1-65:8. https://doi.org/10.1145/2769493.2769512.

Kelly, B., Closier, A., and Hiom, D. 2005. Gateway Standardization: A Quality Assurance Framework for Metadata. *Library Trends 53*, 4, 637–650.

Király, P. 2015a. Metadata Quality Assurance Framework. http://pkiraly.github.io/.

Király, P. 2015b. A Metadata Quality Assurance Framework. Research plan. http://pkiraly.github.io/metadata-quality-project-plan.pdf.

Király, P. 2017. Towards an extensible measurement of metadata quality. *DATeCH2017: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, ACM, 111–115. https://doi.org/10.1145/3078081.3078109.

Király, P. and Büchler, M. 2018. Measuring completeness as metadata quality metric in Europeana. *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2711–2720. https://doi.org/10.1109/BigData.2018.8622487.

Király, P., Stiller, J., Charles, V., Bailer, W., and Freire, N. 2019. Evaluating Data Quality in Europeana: Metrics for Multilinguality. *Metadata and Semantic Research*, Springer International Publishing, 199–211. https://doi.org/10.1007/978-3-030-14401-2_19.

Kontokostas, D., Mader, C., Dirschl, C., et al. 2016. Semantically Enhanced Quality Assurance in the JURION Business Use Case. *The Semantic Web. Latest Advances and New Domains. ESWC 2016*, Springer, 661–676. https://doi.org/10.1007/978-3-319-34129-3_40.

Kontokostas, D., Zaveri, A., Auer, S., and Lehmann, J. 2013. TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data. Springer, 265–272. https://doi.org/10.1007/978-3-642-41360-5_22.

Koster, L. 2014. Analysing library data flows for efficient innovation. *Commonplace.net*. http://commonplace.net/2014/11/library-data-flows/.

Lagoze, C., Krafft, D.B., Jesuroga, S., Cornwell, T., Cramer, E.J., and Shin, E. 2005. An Information Network Overlay Architecture

for the NSDL. *JCDL '05 Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, ACM, 384–384. https://doi.org/10.1145/1065385.1065487.

Lei, Y., Sabou, M., López, V., Zhu, J., Uren, V.S., and Motta, E. 2006. An Infrastructure for Acquiring High Quality Semantic Metadata. *The Semantic Web: Research and Applications*, Springer, 230–244. https://doi.org/10.1007/11762256_19.

Lim, S. and Li Liew, C. 2011. Metadata quality and interoperability of GLAM digital images. *Aslib Proceedings 63*, 5, 484–498. https://doi.org/10.1108/00012531111164978.

Lopatin, L. Metadata Practices in Academic and Non-Academic Libraries for Digital Projects: A Survey. *Cataloging & Classification Quarterly 48*, 8, 716–742. http://dx.doi.org/10.1080/01639374.2010.509029.

Loshin, D. 2013. Building a Data Quality Scorecard for Operational Data Governance. http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/building-data-quality-scorecard-for-operational-data-governance-106025.pdf.

Lovins, D. and Hillmann, D.I. 2017. Broken-World Vocabularies. *D-Lib Magazine 23*, 3/4 (March/April). https://doi.org/10.1045/march2017-lovins.

Ma, S., Lu, C., Lin, X., and Galloway, M. 2009. Evaluating the metadata quality of the IPL. *Proceedings of the American Society for Information Science and Technology 46*, 1, 1–17. https://doi.org/10.1002/meet.2009.1450460249.

Mader, C., Haslhofer, B., and Isaac, A. 2012. Finding Quality Issues in SKOS Vocabularies. *Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings*, Springer, 222–233. https://doi.org/10.1007/978-3-642-33290-6_25.

Madnick, S.E., Wang, R.Y., Lee, Y.W., and Zhu, H. 2009. Overview and Framework for Data and Information Quality Research. *ACM Journal of Data and Information Quality 1*, 1, 1–22. https://doi.org/10.1145/1515693.1516680.

Margaritopoulos, M., Margaritopoulos, T., Mavridis, I., and Manitsaris, A. 2012. Quantifying and Measuring Metadata Completeness. *Journal of the American Society for Information Science and Technology 63*, 4, 724–737. https://doi.org/10.1002/asi.21706.

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., and Manitsaris, A. 2008. A Conceptual Framework for Metadata Quality

Assessment. http://dcpapers.dublincore.org/pubs/article/view/923.

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., and Manitsaris, A. 2009. A Fine-Grained Metric System for the Completeness of Metadata. *Conference Paper in Communications in Computer and Information Science*, Springer, 83–94. https://doi.org/10.1007/978-3-642-04590-5_8.

Matienzo, M.A. and Rudersdorf, A. 2014. The Digital Public Library of America Ingestion Ecosystem: Lessons Learned After One Year of Large-Scale Collaborative Metadata Aggregation. 12–23. http://dcpapers.dublincore.org/pubs/article/view/3700.

Micic, N., Neagu, D., Campean, I.F., and Habib Zadeh, E. 2017. Towards a Data Quality Framework for Heterogeneous Data. https://bradscholars.brad.ac.uk/bitstream/handle/10454/12323/PID4808071.pdf?sequence=3&isAllowed=y.

MODS Guidelines Levels of Adoption. 2009. *MODS Guidelines Levels of Adoption - American Social History Online*. https://wiki.dlib.indiana.edu/display/DLFAquifer/MODS+Guidelines+Levels+of+Adoption.

Moen, W.E., Stewart, E.L., and McClure, C.R. 1997. *The Role of Content Analysis in Evaluating Metadata for the U.S. Government Information Locator Service (GILS): Results from an Exploratory Study Citations, Rights, Re-Use*. University of North Texas Libraries, Digital Library. https://digital.library.unt.edu/ark:/67531/metadc36312/.

Najjar, J. and Duval, E. 2006. Actual Use of Learning Objects and Metadata: An Empirical Analysis. *TCDL Bulletin 2*, 2. http://www.ieee-tcdl.org/Bulletin/v2n2/najjar/najjar.html.

Najjar, J., Ternier, S., and Duval, E. 2003. The actual use of metadata in Ariadne: an empirical analysis. *Proc. ARIADNE 3rd International Conference (2003)*, 6. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3666&rep=rep1&type=pdf.

Neto, C.B., Kontokostas, D., Hellmann, S., Müller, K., and Brümmer, M. 2016. Assessing Quantity and Quality of Links Between Linked Data Datasets. *Proceedings of the Workshop on Linked Data on the Web co-located with the 25th International World Wide Web Conference (WWW 2016)*, CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1593/#article-07.

Newman, D., Hagedorn, K., Chemudugunta, C., and Smyth, P. 2007. Subject Metadata Enrichment Using Statistical Topic Models. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, 366–375. https://doi.org/10.1145/1255175.1255248.

Ngomo, A.-C.N., Auer, S., Lehmann, J., and Zaveri, A. 2014. Introduction to Linked Data and Its Lifecycle on the Web. In: *Reasoning Web. Reasoning on the Web in the Big Data Era: 10th International Summer School 2014, Athens, Greece, September 8-13, 2014. Proceedings*. Springer, Heidelberg, 1–99. http://jens-lehmann.org/files/2013/reasoning_web_linked_data.pdf.

Nichols, D.M., McKay, D., and Twidale, M.B. 2008. A lightweight metadata quality tool. *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries - JCDL '08*, ACM Press, 385–388. https://doi.org/10.1145/1378889.1378957.

Noh, Y. 2011. A study on metadata elements for web-based reference resources system developed through usability testing. *Library Hi Tech 29*, 2, 242–265. https://doi.org/10.1108/07378831111138161.

Nomikos, V. 2018. Repolytics: Identifying Measurable Insights for Digital Repositories. *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) (TDDL_MDQual_Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Grand Hotel Palace, Thessaloniki, Greece, 21 September 2017*, CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2038/paper7.pdf.

Nwala, A.C., Weigle, M.C., and Nelson, M.L. 2018. Bootstrapping Web Archive Collections from Social Media. *Proceedings of the 29th on Hypertext and Social Media*, ACM, 64–72. https://doi.org/10.1145/3209542.3209560.

Ochoa, X. and Duval, E. 2009. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries 10*, 2–3, 67–91. https://doi.org/10.1007/s00799-009-0054-4.

Olson, J.E. 2003. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann. https://books.google.com/books/about/Data_Quality.html?id=x8ahL57VOtcC.

Palavitsinis, N. 2014. Metadata Quality Issues in Learning Repositories. https://www.researchgate.net/publication/260424499_Metadata_Quality_Issues_in_Learning_Repositories.

Palavitsinis, N., Manouselis, N., and Sanchez-Alonso, S. 2014a. Metadata quality in learning object repositories: A case study. *The Electronic Library 32*, 1, 62–82. https://doi.org/10.1108/EL-12-2011-0175.

Palavitsinis, N., Manouselis, N., and Sanchez-Alonso, S. 2014b. Metadata quality in digital repositories: Empirical results from the cross-

domain transfer of a quality assurance process: Journal of the American Society for Information Science and Technology. *Journal of the Association for Information Science and Technology 65*, 6, 1202–1216. http://dx.doi.org/10.1002/asi.23045.

Palavitsinis, N., Manouselis, N., and Sanchez-Alonso, S. 2017. Metadata and Quality in Digital Repositories and Libraries from 1995 to 2015: A Literature Analysis and Classification. *International Information & Library Review*, 11. http://dx.doi.org/10.1080/10572317.2016.1278194.

Park, E.G. 2007. Building interoperable Canadian architecture collections: initial metadata assessment. *The Electronic Library 25*, 2, 207–218. https://doi.org/10.1108/02640470710741331.

Park, E.G. and Richard, M. 2011. Metadata assessment in e-theses and dissertations of Canadian institutional repositories. *The Electronic Library 29*, 3, 394–407. https://doi.org/10.1108/02640471111141124.

Park, J. 2006. Semantic interoperability and metadata quality: an analysis of metadata item records of digital image collections. *Knowledge Organization 33*, 1, 20–34.

Park, J. 2009. Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly 47*, 3–4, 213–228. https://doi.org/10.1080/01639370902737240.

Park, J. and Yuji, T. 2010. Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. *Cataloging & Classification Quarterly 48*, 8, 696–715. https://doi.org/10.1080/01639374.2010.508711.

Phillips, M. 2015. Metadata Quality, Completeness, and Minimally Viable RecordsMetadata Quality, Completeness, and Minimally Viable Records. *Mark E. Phillips Journal*. http://vphill.com/journal/post/4075/.

Phipps, J., Hillmann, D.I., and Paynter, G. 2005. Orchestrating metadata enhancement services: Introducing Lenny. *Proceedings from the International Conference on Dublin Core and Metadata Applications, 2005*, 49–58. http://dcpapers.dublincore.org/pubs/article/view/803.

Pirmann, C. 2009. Alternative Subject Languages for Cataloging. http://courseweb.lis.illinois.edu/~pirmann2/LIS577/toolbox/langhead.html.

Radulovic, F., Mihindukulasooriya, N., García-Castro, R., and Gómez-Pérez, A. 2017. A comprehensive quality model for Linked Data. *Semantic Web Preprint*, Preprint, 1–22. https://doi.org/10.3233/SW-170267.

Rasaiah, B.A., Jones, Simon.D., Bellman, C., Malthus, T.J., and Hueni, A. 2015. Assessing field spectroscopy metadata quality. *Remote Sensing 7*, 4, 4499–4526. https://doi.org/10/gfphbq.

Reiche, K. and Höfig, E. 2013. Implementation of Metadata Quality Metrics and Application on Public Government Data. Institute of Electrical and Electronics Engineers (IEEE), 236–241. https://doi.org/10.1109/COMPSACW.2013.32.

Reiche, K., Höfig, E., and Schieferdecker, I. 2014. Assessment and Visualization of Metadata Quality for Open Government Data. *CeDEM14. Conference for E-Democracy and Open Governement*, Edition Donau-Universität Krems, 335–346. http://www.donau-uni.ac.at/imperia/md/content/department/gpa/zeg/bilder/cedem/cedem14/cedem14_proceedings_1st_edition.pdf.

Reiche, K.J. 2013. Assessment and Visualization of Metadata Quality for Open Government Data. http://www.inf.fu-berlin.de/inst/ag-se/theses/Reiche13-metadata-quality.pdf.

Rennau, H.-J. 2017. Location trees enable XSD based tool development. *XML London 2017 Conference Proceedings*, XML London, 20–37. https://doi.org/10.14337/XMLLondon17.Rennau01.

Rizza, E., Chardonnens, A., and van Hooland, S. 2019. Close-reading of Linked Data: a case study in regards to the quality of online authority files. *arXiv e-prints*. https://arxiv.org/abs/1902.02140.

Robertson, J.R. 2005. Metadata quality: implications for library and information science professionals. *Library Review 54*, 5, 295–300. https://doi.org/10.1108/00242530510600543.

Rula, A. and Zaveri, A. 2014. Methodology for Assessment of Linked Data Quality. *Proceedings of the 1st Workshop on Linked Data Quality*, CEUR. http://ceur-ws.org/Vol-1215/paper-04.pdf.

Shreeves, S.L., Kaczmarek, J.S., and Cole, T.W. 2003. Harvesting cultural heritage metadata using the OAI Protocol. *Library Hi Tech 21*, 2, 159–169. https://doi.org/10.1108/07378830310479802.

Shreeves, S.L., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B., and Cole, T.W. 2005. Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practices for Federated Collections. *Currents and Convergence: Navigating the Rivers of Change. ACRL Twelfth National Conference.*, Association of College & Research Libraries, 223–237. https://www.ideals.illinois.edu/bitstream/handle/2142/145/shreeves05.pdf.

Sicilia, M.A., Garcia, E., Pages, C., Martinez, J.J., and Gutierrez, J.M. 2005. Complete metadata records in learning object repositories: some evidence and requirements. *International Journal of Learn-*

ing Technology 1, 4, 411–424. https://doi.org/10.1504/IJLT.2005.007152.

Simon, A., Vila Suero, D., Hyvönen, E., et al. 2014. *EuropeanaTech Task Force on a Multilingual and Semantic Enrichment Strategy: final report*. Europeana. http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/multilingual-and-semantic-enrichment-strategy.

Simons, G., ed. 2009. Open Language Archives Community (OLAC) Metadata Metrics. http://www.language-archives.org/NOTE/metrics.html.

Snow, K. 2017. Defining, Assessing, and Rethinking Quality Cataloging. *Cataloging & Classification Quarterly*, 1–18. https://doi.org/10.1080/01639374.2017.1350774.

Soto-Hernández, S. and Naumis-Peña, C. Metadata in Mexican Television News Broadcasts on the Web. *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) (TDDL_MDQual_Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Grand Hotel Palace, Thessaloniki, Greece, 21 September 2017*, CEUR Workshop Proceedings. http://ceur-ws.org/Vol-2038/paper5.pdf.

Stiller, J. and Király, P. 2017. Multilinguality of Metadata. Measuring the Multilingual Degree of Europeana's Metadata. *Everything Changes, Everything Stays the Same? Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017)*, Verlag Werner Hülsbusch, 164–176. https://www.researchgate.net/publication/314879735_Multilinguality_of_Metadata_Measuring_the_Multilingual_Degree_of_Europeana%27s_Metadata.

Stvilia, B. and Gasser, L. 2008. Value-based metadata quality assessment. *Library & Information Science Research 30*, 1, 67–74. https://doi.org/10.1016/j.lisr.2007.06.006.

Stvilia, B., Gasser, L., Twidale, M.B., Shreeves, S.L., and Cole, T.W. 2004. Metadata quality for federated collections. *Proceedings of the Ninth International Conference on Information Quality (ICIQ-04)*, 111–125. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.552.1921&rep=rep1&type=pdf.

Stvilia, B., Gasser, L., Twidale, M.B., and Smith, L.C. 2007. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology 58*, 12, 1720–1733. https://doi.org/10.1002/asi.20652.

Stvilia, B., Hinnant, C., Wu, S., et al. 2015. Research project tasks, data, and perceptions of data quality in a condensed matter physics community. *Journal of the Association for Information Science and Technology 66*, 2, 246–263. https://doi.org/10.1002/asi.23177.

Suominen, O. and Hyvönen, E. 2012. Improving the Quality of SKOS Vocabularies with Skosify. *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012.*, Springer, 383–397. https://doi.org/10.1007/978-3-642-33876-2_34.

Suominen, O. and Mader, C. 2014. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics 3*, 1, 47–73. https://doi.org/10.1007/s13740-013-0026-0.

Szostak, R., Scharnhorst, A., Beek, W., and Smiraglia, R.P. 2018. Connecting KOSs and the LOD Cloud. *arXiv:1802.08141 [cs]*. http://arxiv.org/abs/1802.08141.

Tallerås, K. 2017. Quality of Linked Bibliographic Data: The Models, Vocabularies, and Links of Data Sets Published by Four National Libraries. *Journal of Library Metadata 17*, 2, 126–155. https://doi.org/10.1080/19386389.2017.1355166.

Tallerås, K., Dahl, J.H.B., and Pharo, N. 2018. User conceptualizations of derivative relationships in the bibliographic universe. *Journal of Documentation 74*, 4, 894–916. https://doi.org/10.1108/JD-10-2017-0139.

Tallerås, K., Massey, D., Dahl, J.H.B., and Pharo, N. 2013. Ordo ad chaos – Linking Norwegian black metal. In: S.K. Nilsson and A. Frenander, eds., *Libraries, black metal and corporate finance: Current research in Nordic Library and Information Science*. University of Borås, Borås, 136–150. https://www.diva-portal.org/smash/get/diva2:883968/FULLTEXT01.pdf.

Tani, A., Candela, L., and Castelli, D. 2013. Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management 49*, 6, 1194–1205. https://doi.org/10.1016/j.ipm.2013.05.003.

Tarver, H., Phillips, M., Zavalina, O., and Kizhakkethil, P. 2015. An Exploratory Analysis of Subject Metadata in the Digital Public Library of America. *Proceedings from the International Conference on Dublin Core and Metadata Applications 2015*, Dublin Core Metadata Initiative, 30–40. https://digital.library.unt.edu/ark:/67531/metadc725779/.

Tarver, H., Zavalina, O., Phillips, M., Alemneh, D.G., and Shakeri, S. 2014. How Descriptive Metadata Changes in the UNT Libraries' Collection: A Case Study. *Proceedings of the International Con-*

ference on Dublin Core and Metadata Applications, Dublin Core Metadata Initiative, 43–52. http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/235.

Tcholtchev, N. 2014. Visualization of Metadata Quality for Open Government Data. http://www.slideshare.net/dgpazegovzpi/konrad-cedem-praesi.

Thomas, S.E. 1996. Quality in Bibliographic Control. *Library Trends 44*, no.3 (winter 1996), 491–505.

Tönnies, S. 2012. Quality Control using Semantic Technologies in Digital Libraries. https://publikationsserver.tu-braunschweig.de/receive/dbbs_mods_00046510.

Tönnies, S. and Balke, W.-T. 2009. Using Semantic Technologies in Digital Libraries - A Roadmap to Quality Evaluation. *Research and Advanced Technology for Digital Libraries. ECDL 2009.*, Springer, 168–179. http://ceur-ws.org/Vol-581/gvd2010_6_3.pdf.

Tönnies, S. and Balke, W.-T. 2010. Quality Assessment in Digital Libraries - Challenges and Chances. *Proceedings of the 22. GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken)*, CEUR Workshop Proceedings. https://doi.org/10.1007/978-3-642-04346-8_18.

Trippel, T., Broeder, D., Durco, M., and Ohren, O. 2014. Towards automatic quality assessment of component metadata. *Proceedings of LREC 2014*, 3851–3856. http://hdl.handle.net/11858/00-001M-0000-0024-3233-2.

Tsiflidou, E. and Manouselis, N. 2013. Tools and Techniques for Assessing Metadata Quality. *Metadata and Semantics Research: Proceedings of the 7th Research Conference, MTSR 2013, Thessaloniki, Greece, November 19-22, 2013.*, Springer, 99–110. https://doi.org/10.1007/978-3-319-03437-9_11.

Van Kleeck, D., Langford, G., Lundgren, J., Nakano, H., O'Dell, A.J., and Shelton, T. 2016. Managing Bibliographic Data Quality in a Consortial Academic Library: A Case Study. *Cataloging & Classification Quarterly 54*, 7, 452–467. https://doi.org/10.1080/01639374.2016.1210709.

Vassilakaki, E. and Garoufallou, E. 2013. Multilingual Digital Libraries: A review of issues in system-centered and user-centered studies, information retrieval and user behavior. *International Information & Library Review 45*, 1–2, 3–19. https://doi.org/10.1016/j.iilr.2013.07.002.

Ward, J.H. 2002. A Quantitative Analysis of Dublin Core Metadata Element Set (DCMES) Usage in Data Providers Registered with the Open

Archives Initiative (OAI). http://ils.unc.edu/MSpapers/2816.pdf.

Welsh, A. 2016. The Rare Books Catalog and the Scholarly Database. *Cataloging & Classification Quarterly 54*, 5–6, 317–337. https://doi.org/10.1080/01639374.2016.1188433.

Westbrook, R.N., Johnson, D., Carter, K., and Lockwood, A. 2012. Metadata Clean Sweep: A Digital Library Audit Project. *D-Lib Magazine 18*, 5/6. https://doi.org/10.1045/may2012-westbrook.

Young, J.Y., Westbrook, J.D., Feng, Z., et al. 2017. OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive. *Structure 25*, 3, 536–545. https://doi.org/10.1016/j.str.2017.01.004.

Zavalina, O.L. 2014. Complementarity in Subject Metadata in Large-Scale Digital Libraries: A Comparative Analysis. *Cataloging & Classification Quarterly 52*, 1, 77–89. https://doi.org/10.1080/01639374.2013.848316.

Zavalina, O.L., Kizhakkethil, P., Alemneh, D.G., Phillips, M.E., and Tarver, H. 2015a. Building a Framework of Metadata Change to Support Knowledge Management. *Journal of Information & Knowledge Management 14*, 01, 1550005-1-1550005–16. https://doi.org/10.1142/S0219649215500057.

Zavalina, O.L., Kizhakkethil, P., and Shakeri, S. 2015b. Metadata change in traditional library collections and digital repositories: Exploratory comparative analysis. *Proceedings of the Association for Information Science and Technology 52*, 1, 1–5. https://doi.org/10.1002/pra2.2015.1450520100146.

Zavalina, O.L., Shakeri, S., Kizhakkethil, P., and Phillips, M.E. 2018. Uncovering Hidden Insights for Information Management: Examination and Modeling of Change in Digital Collection Metadata. *Transforming Digital Worlds*, Springer International Publishing, 645–651. https://doi.org/10/gfphbt.

Zavalina, O.L., Zavalin, V., Shakeri, S., and Kizhakkethil, P. 2016. Developing an Empirically-based Framework of Metadata Change and Exploring Relation between Metadata Change and Metadata Quality in MARC Library Metadata. *Procedia Computer Science 99*, 50–63. https://doi.org/10.1016/j.procs.2016.09.100.

Zaveri, A., Kontokostas, D., Sherif, M.A., Bühmann, L., Morsey, M., and Auer, S. 2013. User-driven quality evaluation of DBpedia. *Proceedings of the 9th International Conference on Semantic Systems*, ACM, 97–104. https://doi.org/10.1145/2506182.2506195.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. 2016. Quality assessment for Linked Data: A Survey.

Semantic Web 7, 1, 63–93. https://doi.org/10.3233/SW-150175.

Zeitlyn, D. and Beardmore-Herd, M. 2018. Testing Google Scholar bibliographic data: Estimating error rates for Google Scholar citation parsing. *First Monday 23*, Number 11-5 November 2018. https://doi.org/10.5210/fm.v23i11.8658.

Zeng, M.L., Subrahmanyam, B., and Shreve, G.M. 2004. Metadata Quality Study for the National Science Digital Library (NSDL) Metadata Repository. *Digital Libraries: International Collaboration and Cross-Fertilization*, Springer, 339–340. https://doi.org/10.1007/978-3-540-30544-6_36.

Appendix B.

Representing MARC 21 in Avram JSON schema

This listing shows fragments of the full export of MARC 21 Format for Bibliographic Data into Avram JSON schema described in Chapter 4. It contains examples for a simple control field (001), a complex control field (008) and a data field (245). The reader can access he full Avram export at `http://pkiraly.github.io/2018/01/28/marc21-in-json/`.

```
{
 "$schema": "https://format.gbv.de/schema/avram/schema.json",
 "title": "MARC 21 Format for Bibliographic Data.",
 "description": "MARC 21 Format for Bibliographic Data.",
 "url": "https://www.loc.gov/marc/bibliographic/",
 "fields": {
  ...
  "001": {
   "tag": "001",
   "label": "Control Number",
   "repeatable": false
  },
  ...
  "008": {
   "tag": "008",
   "label": "General Information",
   "repeatable": false,
   "types": {
    "All Materials": {
     "positions": {
      "00-05": {
       "label": "Date entered on file",
       "url": "https://www.loc.gov/marc/bibliographic/bd008a.html"
      },
      "06": {
       "label": "Type of date/Publication status",
       "url": "https://www.loc.gov/marc/bibliographic/bd008a.html",
       "codes": {
        "b": {
         "label": "No dates given; B.C. date involved"
        },
        "c": {
         "label": "Continuing resource currently published"
        },
        "d": {
         "label": "Continuing resource ceased publication"
        },
        "e": {
```

```
    "label": "Detailed date"
   },
   "i": {
    "label": "Inclusive dates of collection"
   },
   "k": {
    "label": "Range of years of bulk of collection"
   },
   "m": {
    "label": "Multiple dates"
   },
   "n": {
    "label": "Dates unknown"
   },
   "p": {
    "label": "Date of distribution/release/issue and production/
        recording session when different"
   },
   "q": {
    "label": "Questionable date"
   },
   "r": {
    "label": "Reprint/reissue date and original date"
   },
   "s": {
    "label": "Single known date/probable date"
   },
   "t": {
    "label": "Publication date and copyright date"
   },
   "u": {
    "label": "Continuing resource status unknown"
   },
   "|": {
    "label": "No attempt to code"
   }
  }
 },
 "07-10": {
  "label": "Date 1",
  "url": "https://www.loc.gov/marc/bibliographic/bd008a.html"
 },
 "11-14": {
  "label": "Date 2",
  "url": "https://www.loc.gov/marc/bibliographic/bd008a.html"
 },
 "15-17": {
  "label": "Place of publication, production, or execution",
  "url": "https://www.loc.gov/marc/bibliographic/bd008a.html"
 },
 "35-37": {
  "label": "Language",
  "url": "https://www.loc.gov/marc/bibliographic/bd008a.html"
 },
 "38": {
  "label": "Modified record",
  "url": "https://www.loc.gov/marc/bibliographic/bd008a.html",
  "codes": {
   " ": {
    "label": "Not modified"
   },
   "d": {
    "label": "Dashed-on information omitted"
   },
   "o": {
```

```
    "label": "Completely romanized/printed cards romanized"
   },
   "r": {
    "label": "Completely romanized/printed cards in script"
   },
   "s": {
    "label": "Shortened"
   },
   "x": {
    "label": "Missing characters"
   },
   "|": {
    "label": "No attempt to code"
   }
  },
  "historical-codes": {
   "u": {
    "label": "Unknown [OBSOLETE] [CAN/MARC only]"
   }
  }
 },
 "39": {
  "label": "Cataloging source",
  "url": "https://www.loc.gov/marc/bibliographic/bd008a.html",
  "codes": {
   " ": {
    "label": "National bibliographic agency"
   },
   "c": {
    "label": "Cooperative cataloging program"
   },
   "d": {
    "label": "Other"
   },
   "u": {
    "label": "Unknown"
   },
   "|": {
    "label": "No attempt to code"
   }
  },
  "historical-codes": {
   "a": {
    "label": "National Agricultural Library [OBSOLETE, 1997] [
        USMARC only]"
   },
   "b": {
    "label": "National Library of Medicine [OBSOLETE, 1997] [
        USMARC only]"
   },
   "l": {
    "label": "Library of Congress cataloguing [OBSOLETE, 1997] [
        CAN/MARC only]"
   },
   "o": {
    "label": "Other institution cataloguing [OBSOLETE, 1997] [CAN
        /MARC only]"
   },
   "n": {
    "label": "Report to New serials titles [OBSOLETE, 1997] [
        USMARC only]"
   },
   "r": {
    "label": "Reporting library [OBSOLETE, 1997] [CAN/MARC only]"
   }
```

```
    }
   }
  }
 },
 "Books": {
  "positions": {
   "18-21": {
    "label": "Illustrations",
    "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
    "codes": {
     " ": {
      "label": "No illustrations"
     },
     "a": {
      "label": "Illustrations"
     },
     "b": {
      "label": "Maps"
     },
     "c": {
      "label": "Portraits"
     },
     "d": {
      "label": "Charts"
     },
     "e": {
      "label": "Plans"
     },
     "f": {
      "label": "Plates"
     },
     "g": {
      "label": "Music"
     },
     "h": {
      "label": "Facsimiles"
     },
     "i": {
      "label": "Coats of arms"
     },
     "j": {
      "label": "Genealogical tables"
     },
     "k": {
      "label": "Forms"
     },
     "l": {
      "label": "Samples"
     },
     "m": {
      "label": "Phonodisc, phonowire, etc."
     },
     "o": {
      "label": "Photographs"
     },
     "p": {
      "label": "Illuminations"
     },
     "|": {
      "label": "No attempt to code"
     }
    }
   },
   "22": {
    "label": "Target audience",
```

```
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   " ": {
    "label": "Unknown or not specified"
   },
   "a": {
    "label": "Preschool"
   },
   "b": {
    "label": "Primary"
   },
   "c": {
    "label": "Pre-adolescent"
   },
   "d": {
    "label": "Adolescent"
   },
   "e": {
    "label": "Adult"
   },
   "f": {
    "label": "Specialized"
   },
   "g": {
    "label": "General"
   },
   "j": {
    "label": "Juvenile"
   },
   "|": {
    "label": "No attempt to code"
   }
  },
  "historical-codes": {
   "u": {
    "label": "School material at first level [OBSOLETE]"
   },
   "v": {
    "label": "School material at second level [OBSOLETE]"
   }
  }
 },
 "23": {
  "label": "Form of item",
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   " ": {
    "label": "None of the following"
   },
   "a": {
    "label": "Microfilm"
   },
   "b": {
    "label": "Microfiche"
   },
   "c": {
    "label": "Microopaque"
   },
   "d": {
    "label": "Large print"
   },
   "f": {
    "label": "Braille"
   },
   "o": {
```

```
    "label": "Online"
   },
   "q": {
    "label": "Direct electronic"
   },
   "r": {
    "label": "Regular print reproduction"
   },
   "s": {
    "label": "Electronic"
   },
   "|": {
    "label": "No attempt to code"
   }
  },
  "historical-codes": {
   "g": {
    "label": "Punched paper tape [OBSOLETE, 1987]"
   },
   "h": {
    "label": "Magnetic tape [OBSOLETE, 1987]"
   },
   "i": {
    "label": "Multimedia [OBSOLETE, 1987]"
   },
   "z": {
    "label": "Other form of reproduction [OBSOLETE, 1987]"
   }
  }
 },
 "24-27": {
  "label": "Nature of contents",
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   " ": {
    "label": "No specified nature of contents"
   },
   "a": {
    "label": "Abstracts/summaries"
   },
   "b": {
    "label": "Bibliographies"
   },
   "c": {
    "label": "Catalogs"
   },
   "d": {
    "label": "Dictionaries"
   },
   "e": {
    "label": "Encyclopedias"
   },
   "f": {
    "label": "Handbooks"
   },
   "g": {
    "label": "Legal articles"
   },
   "i": {
    "label": "Indexes"
   },
   "j": {
    "label": "Patent document"
   },
   "k": {
```

```
  "label": "Discographies"
 },
 "l": {
  "label": "Legislation"
 },
 "m": {
  "label": "Theses"
 },
 "n": {
  "label": "Surveys of literature in a subject area"
 },
 "o": {
  "label": "Reviews"
 },
 "p": {
  "label": "Programmed texts"
 },
 "q": {
  "label": "Filmographies"
 },
 "r": {
  "label": "Directories"
 },
 "s": {
  "label": "Statistics"
 },
 "t": {
  "label": "Technical reports"
 },
 "u": {
  "label": "Standards/specifications"
 },
 "v": {
  "label": "Legal cases and case notes"
 },
 "w": {
  "label": "Law reports and digests"
 },
 "y": {
  "label": "Yearbooks"
 },
 "z": {
  "label": "Treaties"
 },
 "2": {
  "label": "Offprints"
 },
 "5": {
  "label": "Calendars"
 },
 "6": {
  "label": "Comics/graphic novels"
 },
 "|": {
  "label": "No attempt to code"
 }
},
"historical-codes": {
 "h": {
  "label": "Handbooks [OBSOLETE]"
 },
 "x": {
  "label": "Technical reports [OBSOLETE, 1997]"
 },
 "3": {
```

```
    "label": "Discographies [OBSOLETE, 1997]"
   },
   "4": {
    "label": "Filmographies [OBSOLETE, 1997]"
   }
  }
 },
 "28": {
  "label": "Government publication",
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   " ": {
    "label": "Not a government publication"
   },
   "a": {
    "label": "Autonomous or semi-autonomous component"
   },
   "c": {
    "label": "Multilocal"
   },
   "f": {
    "label": "Federal/national"
   },
   "i": {
    "label": "International intergovernmental"
   },
   "l": {
    "label": "Local"
   },
   "m": {
    "label": "Multistate"
   },
   "o": {
    "label": "Government publication-level undetermined"
   },
   "s": {
    "label": "State, provincial, territorial, dependent, etc."
   },
   "u": {
    "label": "Unknown if item is government publication"
   },
   "z": {
    "label": "Other"
   },
   "|": {
    "label": "No attempt to code"
   }
  },
  "historical-codes": {
   "n": {
    "label": "Government publication-level undetermined [OBSOLETE
        ]"
   }
  }
 },
 "29": {
  "label": "Conference publication",
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   "0": {
    "label": "Not a conference publication"
   },
   "1": {
    "label": "Conference publication"
   },
```

```
   "|": {
    "label": "No attempt to code"
   }
  }
 },
 "30": {
  "label": "Festschrift",
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   "0": {
    "label": "Not a festschrift"
   },
   "1": {
    "label": "Festschrift"
   },
   "|": {
    "label": "No attempt to code"
   }
  }
 },
 "31": {
  "label": "Index",
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   "0": {
    "label": "No index"
   },
   "1": {
    "label": "Index present"
   },
   "|": {
    "label": "No attempt to code"
   }
  }
 },
 "33": {
  "label": "Literary form",
  "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
  "codes": {
   "0": {
    "label": "Not fiction (not further specified)"
   },
   "1": {
    "label": "Fiction (not further specified)"
   },
   "d": {
    "label": "Dramas"
   },
   "e": {
    "label": "Essays"
   },
   "f": {
    "label": "Novels"
   },
   "h": {
    "label": "Humor, satires, etc."
   },
   "i": {
    "label": "Letters"
   },
   "j": {
    "label": "Short stories"
   },
   "m": {
    "label": "Mixed forms"
```

```
      },
      "p": {
       "label": "Poetry"
      },
      "s": {
       "label": "Speeches"
      },
      "u": {
       "label": "Unknown"
      },
      "|": {
       "label": "No attempt to code"
      }
     },
     "historical-codes": {
      " ": {
       "label": "Non-fiction [OBSOLETE, 1997]"
      },
      "c": {
       "label": "Comic strips [OBSOLETE, 2008]"
      }
     }
    },
    "34": {
     "label": "Biography",
     "url": "https://www.loc.gov/marc/bibliographic/bd008b.html",
     "codes": {
      " ": {
       "label": "No biographical material"
      },
      "a": {
       "label": "Autobiography"
      },
      "b": {
       "label": "Individual biography"
      },
      "c": {
       "label": "Collective biography"
      },
      "d": {
       "label": "Contains biographical information"
      },
      "|": {
       "label": "No attempt to code"
      }
     }
    }
   }
  },
  "Computer Files": {
   "positions": {
    "22": {
     "label": "Target audience",
     "url": "https://www.loc.gov/marc/bibliographic/bd008c.html",
     "codes": {
      " ": {
       "label": "Unknown or not specified"
      },
      "a": {
       "label": "Preschool"
      },
      "b": {
       "label": "Primary"
      },
      "c": {
```

```
    "label": "Pre-adolescent"
   },
   "d": {
    "label": "Adolescent"
   },
   "e": {
    "label": "Adult"
   },
   "f": {
    "label": "Specialized"
   },
   "g": {
    "label": "General"
   },
   "j": {
    "label": "Juvenile"
   },
   "|": {
    "label": "No attempt to code"
   }
  }
 },
 "23": {
  "label": "Form of item",
  "url": "https://www.loc.gov/marc/bibliographic/bd008c.html",
  "codes": {
   " ": {
    "label": "Unknown or not specified"
   },
   "o": {
    "label": "Online"
   },
   "q": {
    "label": "Direct electronic"
   },
   "|": {
    "label": "No attempt to code"
   }
  }
 },
 "26": {
  "label": "Type of computer file",
  "url": "https://www.loc.gov/marc/bibliographic/bd008c.html",
  "codes": {
   "a": {
    "label": "Numeric data"
   },
   "b": {
    "label": "Computer program"
   },
   "c": {
    "label": "Representational"
   },
   "d": {
    "label": "Document"
   },
   "e": {
    "label": "Bibliographic data"
   },
   "f": {
    "label": "Font"
   },
   "g": {
    "label": "Game"
   },
```

```
   "h": {
    "label": "Sound"
   },
   "i": {
    "label": "Interactive multimedia"
   },
   "j": {
    "label": "Online system or service"
   },
   "m": {
    "label": "Combination"
   },
   "u": {
    "label": "Unknown"
   },
   "z": {
    "label": "Other"
   },
   "|": {
    "label": "No attempt to code"
   }
  }
 },
 "28": {
  "label": "Government publication",
  "url": "https://www.loc.gov/marc/bibliographic/bd008c.html",
  "codes": {
   " ": {
    "label": "Not a government publication"
   },
   "a": {
    "label": "Autonomous or semi-autonomous component"
   },
   "c": {
    "label": "Multilocal"
   },
   "f": {
    "label": "Federal/national"
   },
   "i": {
    "label": "International intergovernmental"
   },
   "l": {
    "label": "Local"
   },
   "m": {
    "label": "Multistate"
   },
   "o": {
    "label": "Government publication-level undetermined"
   },
   "s": {
    "label": "State, provincial, territorial, dependent, etc."
   },
   "u": {
    "label": "Unknown if item is government publication"
   },
   "z": {
    "label": "Other"
   },
   "|": {
    "label": "No attempt to code"
   }
  }
 }
```

```
    }
   },
   ...
  }
 },
 "245": {
  "tag": "245",
  "label": "Title Statement",
  "url": "https://www.loc.gov/marc/bibliographic/bd245.html",
  "repeatable": false,
  "indicator1": {
   "label": "Title added entry",
   "codes": {
    "0": {
     "label": "No added entry"
    },
    "1": {
     "label": "Added entry"
    }
   }
  },
  "indicator2": {
   "label": "Nonfiling characters",
   "codes": {
    "0": {
     "label": "No nonfiling characters"
    },
    "1-9": {
     "label": "Number of nonfiling characters"
    }
   }
  },
  "subfields": {
   "a": {
    "label": "Title",
    "repeatable": false
   },
   "b": {
    "label": "Remainder of title",
    "repeatable": false
   },
   "c": {
    "label": "Statement of responsibility, etc.",
    "repeatable": false
   },
   "f": {
    "label": "Inclusive dates",
    "repeatable": false
   },
   "g": {
    "label": "Bulk dates",
    "repeatable": false
   },
   "h": {
    "label": "Medium",
    "repeatable": false
   },
   "k": {
    "label": "Form",
    "repeatable": true
   },
   "n": {
    "label": "Number of part/section of a work",
    "repeatable": true
   },
```

```
      "p": {
       "label": "Name of part/section of a work",
       "repeatable": true
      },
      "s": {
       "label": "Version",
       "repeatable": false
      },
      "6": {
       "label": "Linkage",
       "repeatable": false
      },
      "8": {
       "label": "Field link and sequence number",
       "repeatable": true
      }
     },
     "historical-subfields": {
      "d": {
       "label": "Designation of section/part/series (SE) [OBSOLETE,
           1979]"
      },
      "e": {
       "label": "Name of part/section/series (SE) [OBSOLETE, 1979]"
      }
     }
    },
    ...
   }
  }
```

Appendix C.

Curriculum Vitae

Surname, given name: Király, Péter
E-mail: peter.kiraly@gwdg.de

Profiles on the web: GitHub[1], Google Scholar[2], ORCID[3], ResearchGate[4], Academia.edu[5], LinkedIn[6], SlideShare[7], Twitter[8]

University education

1990-1996: Diploma in History and Textual Studies, University of Miskolc, Hungary

Skills

- languages: Hungarian (native), English (working level proficiency), Latin (intermediate), German (beginner), Dutch (beginner)
- Librarian and Archival skills (focusing on descriptive standards)
- Researching historical documents (including textual studies)
- Computer Science (Web, Programming, Data Science)

Employment history

- 2014 October-: Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG, Germany) as researcher, software developer
- 2012-2014 September: Europeana.eu (The Netherlands) as Portal Backend Developer
- 2008-2013: eXtensible Catalog Project (USA, from 2010 eXtensible Catalog Organisation) as software developer (from 2012 in part time)

[1] https://github.com/pkiraly/
[2] https://scholar.google.de/citations?user=hvoZFusAAAAJ&hl=en
[3] https://orcid.org/0000-0002-8749-4597
[4] https://www.researchgate.net/profile/Peter_Kiraly3
[5] https://uni-goettingen.academia.edu/PeterKiraly
[6] http://www.linkedin.com/in/peterkiraly
[7] http://www.slideshare.net/pkiraly
[8] https://twitter.com/#!/kiru

- 2007-2010: Hungarian Academy of Science - Eötvös Loránd University's Online Critical Editing special research group (Hungary, part-time job)
- 2005-2008: Tesuji S.r.l., Tesuji Hungary Ltd. (Italy-Hungary) as Java developer
- 2001-2005: Arcanum Database Ltd. (Hungary) as web publication developer, editor
- 1999-2000: National Széchényi Library (Hungary) as manuscript librarian
- 1998-1999: Budapest City Archives (Hungary) as archivist
- 1996-1998: University of Miskolc Library, Archives and Museum (Hungary) as archivist

Voluntary works

- 2017-: organizer of Göttingen Data Science Meetup team
- 2016-: member of Europeana Data Quality Committee[9]
- 2015-: member of The Code4Lib Journal editorial team[10]
- 2015-2017: board member of Göttingen Dialogue in Digital Humanities
- 2012-2014: member of W3C Schema Bib Extend Community Group
- 2008-2014: active member of Drupal community
- 1995-: volunteer of Hungarian Electronic Library[11]

Conference presentations and publications in the context of this thesis

1. *Multilinguality of Metadata. Measuring the Multilingual Degree of Europeana's Metadata*[12] International Symposium on Information Science, Berlin, 2017. March 13-15. Presentation together with Juliane Stiller. Paper was published as [57]
2. *Measuring metadata quality. A quick overview in the context of Europeana metadata*[13] #dariahTeach, Lausanne, 2017. March 23-24. Poster
3. *Multilinguality of Metadata. Measuring the Multilingual Degree of Europeana's Metadata* SI & IT Workshop, Göttingen, 2017. May 10. Presentation together with Juline Stiller
4. *Measuring Metadata Quality and the Europeana use case*[14] Linked Data Quality workshop 2017, Portorož, 2017. May 29. Invited keynote speech

[9] `https://pro.europeana.eu/project/data-quality-committee`
[10] `http://journal.code4lib.org`
[11] `https://mek.oszk.hu`
[12] slides: `http://doi.org/10.13140/RG.2.2.30771.43047`
[13] `http://doi.org/10.13140/RG.2.2.36780.67207`
[14] slides: `http://bit.ly/metadata-ldq4`

5. *Towards an extensible measurement of metadata quality*[15] DATeCH 2017, Göttingen, 2017 June 1-2. Paper published as [35]
6. *"Nothing is created, nothing is lost, everything changes". Measuring and visualizing data quality in Europeana.*[16] ELAG 2017, Athens, 2017. June 6-9. Presentation together with Valentine Charles
7. *Measuring completeness as metadata quality metric in Europeana*[17] Digital Humanities 2017, Montréal, 2017 August 7-11. Extended abstract was published in the Digital Humanities 2017 Conference Abstracts[18], cited by [33].
8. Linked Data Quality Workshop Semantics 2017, Amsterdam, 2017. September 14 – as organizer and presenter.
9. *Evaluating Data Quality in Europeana: Metrics for Multilinguality* (Meta)-Data Quality Workshop (part of TPDL 2017), Thessaloniki, 2017. September 17-21. Presented by Juliane Stiller, the paper was published as [13].
10. *Measuring Metadata Quality in Europeana*[19] and *Measuring library catalogs*[20] ADOCHS meeting, Brussels, 2017 November 21
11. LDCX at Stanford University, 2018 March. Leading workshop and presentation.
12. *Data Quality Workshop*[21] ELAG 2018, Prague, 2018 June 5-7. Workshop together with Anette Strauch and Patrick Hochstenbach with contribution from Mark Phillips[22].
13. *Evaluating Data Quality in Europeana: Metrics for Multilinguality.* MTSR 2018. 12th International Conference on Metadata and Semantics Research, Limassol, 2018 Oct 23-26. Presentation together with Juliane Stiller. Paper published as [37].
14. *Researching metadata quality*[23]. Open Research Knowledge Graph workshop, Hannover, 2018 November 22.
15. *Metadata quality in cultural heritage institutions*[24] ReIReS (Research Infrastructure on Religious Studies) Workshop on FAIR Principle for Digital Research Data Management, Mainz, 2018 November 28.
16. *Measuring Completeness as Metadata Quality Metric in Europeana*[25] Computational Archival Science workshop (part of IEEE Big Data

[15] slides: `http://bit.ly/way2datech`
[16] slides: `http://bit.ly/mq-elag2017`
[17] slides: `http://bit.ly/mq-dh2017`
[18] `https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf`
[19] `http://bit.ly/adochs-europeana`
[20] `http://bit.ly/adochs-marc`
[21] `http://bit.ly/elag2018-data-quality-workshop`
[22] Mark Philipps: "UNT Libraries Metadata Quality Interfaces" (video) `https://www.youtube.com/watch?v=ATM3EwixnW8`
[23] slides: `http://bit.ly/qa-orkg2018`
[24] slides: `http://bit.ly/qa-relres-fair`
[25] slides: `http://bit.ly/qa-cas2018`

2018), Seattle, 2018 December 10-13. Paper published as [36].

17. *Adat a könyvtárban* (Data in the library – paper in Hungarian about the changing status of data in LAM). In Hagyomány és újitás a 21. századi könyvtárban (Erdélyi Évszázadok. A Kolozsvári Magyar Történeti Intézet Évkönyve. III.) eds. Rüsz-Fogarasi Enikő, Monok István. Kolozsvár (Romania), 2018. pp. 49-74.

Appendix D.

Declarations

D.1. About identical copies

I hereby confirm, that the electronic version matches the printed version of the thesis. The paper was created with LaTeX, its source code can be found at `https://github.com/pkiraly/metadata-qa-thesis`.

D.2. About independent research

I declare in lieu of an oath, that I completed this thesis independently and that I did not use any other sources or resources than those stated. Furthermore, I declare that no equivalent doctoral studies have been applied elsewhere, and that the proposed thesis or parts thereof have not been submitted elsewhere.

Bibliography

[1] White Paper on Best Practices for Multilingual Access to Digital Libraries. Tech. rep., Europeana, 2016. http://pro.europeana.eu/files/Europeana_Professional/Publications/BestPracticesForMultilingualAccess_whitepaper.pdf.

[2] Al-Gumaei, K. 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.

[3] Albertoni, R., De Martino, M., and Podesta, P. A linkset quality metric measuring multilingual gain in skos thesauri. In *Proceedings of the 2nd Workshop on Linked Data Quality co-located with 12th Extended Semantic Web Conference (ESWC 2015)* (2015), A. Rula, A. Zaveri, M. Knuth, and D. Kontokostas, Eds., CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1376/LDQ2015_paper_01.pdf.

[4] American Library Association, Canadian Library Association, and Chartered Institute of Library and Information Professionals. *Anglo-American Cataloging Rules. 2nd edition.* American Library Association, 2005. http://www.aacr2.org/.

[5] Apache Software Foundation. *Accumulators (in Apache Spark documentation).* https://spark.apache.org/docs/latest/rdd-programming-guide.html#accumulators.

[6] Apache Spark committers. *Source code of Apache Spark's TaskMetrics class.* https://github.com/apache/spark/blob/master/core/src/main/scala/org/apache/spark/executor/TaskMetrics.scala.

[7] Avram, H. D., and Library of Congress. *MARC; its History and implications.* Library of Congress, 1975. http://catalog.hathitrust.org/Record/002993527.

[8] Bade, D. The perfect bibliographic record: Platonic ideal, rhetorical strategy or nonsense? *Cataloging & Classification Quarterly 46*, 1 (2008), 109–133. http://www.tandfonline.com/doi/abs/10.1080/01639370802183081.

[9] Bellini, E., and Nesi, P. Metadata quality assessment tool for Open Access cultural heritage institutional repositories. In *Information Technologies for Performing Arts, Media Access, and Entertainment* (2013), pp. 90–103. https://link.springer.com/chapter/10.1007%2F978-3-642-40050-6_9.

[10] Brown, N. M., Mendenhall, R., Black, M. L., Moer, M. V., Zerai, A., and Flynn, K. Mechanized margin to digitized center: Black feminism's contributions to combatting erasure within the digital humanities. *International Journal of Humanities and Arts Computing 10*, 1 (2016), 110–125. https://doi.org/10.3366/ijhac.2016.0163.

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

[12] Canali, L. *Apache Spark Performance Troubleshooting at Scale, Challenges, Tools, and Methods*, Oct. 2017. https://www.youtube.com/watch?v=JoQ8m-kM_ZY.

[13] Charles, V., Stiller, J., Bailer, W., Freire, N., and Király, P. Evaluating data quality in europeana: Metrics for multilinguality. In *Joint Proceedings of the 1st Workshop on Temporal Dynamics in Digital Libraries (TDDL 2017), the (Meta)-Data Quality Workshop (MDQual 2017) and the Workshop on Modeling Societal Future (Futurity 2017) (TDDL_MDQual_Futurity 2017) co-located with 21st International Conference on Theory and Practice of Digital Libraries (TPLD 2017), Grand Hotel Palace, Thessaloniki, Greece, 21 September 2017* (2017), A. Caputo, N. Kanhabua, P. Basile, S. Lawless, D. Gavrilis, C. Papatheodorou, Gîfu, and D. Trandabat, Eds., CEUR. http://ceur-ws.org/Vol-2038/paper6.pdf.

[14] Collections as Data project team. The Santa Barbara statement on collections as data. v2, 2017. https://collectionsasdata.github.io/statement/.

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

[16] Delsey, T. Functional analysis of the marc 21 bibliographic and holdings formats. Tech. rep., Library of Congress, 2002. Prepared for the Network Development and MARC Standards Office Library of Congress. Second Revision: September 17, 2003. https://www.loc.gov/marc/marc-functional-analysis/original_source/analysis.pdf.

[17] Ellefi, M. B., Bellahsene, Z., Breslin, J., Demidova, E., Dietze, S., Szymanski, J., and Todorov, K. RDF dataset profiling. a survey of features, methods, vocabularies and applications. *Semantic Web* (2017). http://www.semantic-web-journal.net/content/rdf-dataset-profiling-survey-features-methods-vocabularies-and-applications.

[18] Fowler, D., Barratt, J., and Walsh, P. Frictionless Data: Making Research Data Quality Visible. *International Journal of Digital Curation 12*, 2 (May 2018), 274–285. http://www.ijdc.net/article/view/577.

[19] Gavrilis, D., Makri, D.-N., Papachristopoulos, L., Angelis, S., Kravvaritis, K., Papatheodorou, C., and Constantopoulos, P. Measuring quality in metadata repositories. In *Proceedings from the 19th International Conference on Theory and Practice of Digital Libraries (TPDL)*, S. Kapidakis, C. Mazurek, and M. Werla, Eds. Springer International Publishing, 2015, pp. 56–67. https://link.springer.com/chapter/10.1007%2F978-3-319-24592-8_5.

[20] GO FAIR Metrics Group. FAIR metrics. http://fairmetrics.org/.

[21] Groskopf, C. The quartz guide to bad data. *Quartz* (2015). https://qz.com/572338/the-quartz-guide-to-bad-data/.

[22] Gueguen, G. Metadata quality at scale: Metadata quality control at the Digital Public Library of America. *Journal of Digital Media Management 7*, 2 (2019), 115–126. https://www.ingentaconnect.com/content/hsp/jdmm/2019/00000007/00000002/art00003.

[23] Harej, V., and Žumer, M. Analysis of frbr user tasks. *Cataloging & Classification Quarterly 51*, 7 (2013), 741–759. https://doi.org/10.1080/01639374.2013.785461.

[24] Harper, C. Metadata analytics, visualization, and optimization: Experiments in statistical analysis of the digital public library of america (DPLA). *The Code4Lib Journal*, 33 (2016). http://journal.code4lib.org/articles/11752.

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

[26] Hill, T., and Manguinhas, H. Internal dqc problem patterns. Tech. rep., Europeana, 2016. http://bit.ly/2jIXQGU.

[27] IFLA. *Functional requirements for Bibliographic records: final report / IFLA Study Group on the Functional Requirements for Bibliographic Records*. No. vol. 19 in UBCIM publications ; new series. K.G. Saur, München, 1998.

[28] International Federation of Library Associations and Institutions (IFLA). *International Standard Bibliographic Description.*, 2011. http://www.ifla.org/files/assets/cataloguing/isbd/isbd-cons_20110321.pdf.

[29] Isaac, A. Europeana data model primer. Tech. rep., Europeana Foundation, 2013. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf.

[30] ISO. *ISO/IEC 25012. Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model*. International Organization for Standardization, 2000. https://www.iso.org/standard/35736.html.

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

[32] Jockers, M. L. *Macroanalysis: Digital methods and literary history*. Topics in the Digital Humanities. University of Illinois Press, 2013.

[33] Khan, N. A., Shafi, S., and Ahangar, H. Digitization of cultural heritage: Global initiatives, opportunities and challenges. *Journal of Cases on Information Technology (JCIT) 20*, 4 (2018), 1–16.

[34] Király, P. How to run the analysis? a cheat sheet, 2015. http://pkiraly.github.io/cheatsheet/.

[35] Király, P. Towards an extensible measurement of metadata quality. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017* (2017), ACM Press, pp. 111–115. https://doi.org/10.1145/3078081.3078109.

[36] Király, P., and Büchler, M. Measuring completeness as metadata quality metric in europeana. In *2018 IEEE International Conference on Big Data (Big Data)* (2018), IEEE, pp. 2711–2720. https://ieeexplore.ieee.org/abstract/document/8622487.

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

(Cham, 2019), E. Garoufallou, F. Sartori, R. Siatri, and M. Zervas, Eds., vol. 846 of *Communications in Computer and Information Science*, Springer International Publishing, pp. 199–211. http://doi.org/10.1007/978-3-030-14401-2_19.

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

[39] Lahti, L., Marjanen, J., Roivainen, H., and Tolonen, M. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly 57*, 1 (2019), 5–23. https://doi.org/10.1080/01639374.2018.1543747.

[40] Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. AIMQ: A methodology for information quality assessment. *Information & Management 40*, 2 (2002), 133–146.

[41] Library of Congress. *MARC 21 Format for Holdings Data.*, 2000. https://www.loc.gov/marc/holdings/.

[42] Library of Congress. *MARC 21 Format for Bibliographic Data.*, 2018. https://www.loc.gov/marc/bibliographic/.

[43] Metadata Assesment Group in Zotero. *Metadata assessment. A bibliography.* https://www.zotero.org/groups/488224/metadata_assessment.

[44] Meyer, E. T., and Eccles, K. The impacts of digital collections: Early english books online & house of commons parliamentary papers. *SSRN Electronic Journal* (Mar. 2016). http://dx.doi.org/10.2139/ssrn.2740299.

[45] Miksa, S. D. Understanding support of frbr's four user tasks in marc-encoded bibliographic records. *Bulletin of the American Society for Information Science & Technology 33* (2007), 24–26.

[46] Moreux, J.-P. Data mining historical newspaper metadata. In *Proceedings of the IFLA International News Media Conference* (04 2016). https://www.researchgate.net/publication/291833336_Data_Mining_Historical_Newspaper_Metadata.

[47] Nanni, F. *The Web as a Historical Corpus: Collecting, Analysing and Selecting Sources on the Recent Past of Academic Institutions.* Dottorato di ricerca in science, cognition and technology, Università di Bologna, 2017. http://amsdottorato.unibo.it/7848/.

[48] National Information Standards Organization (NISO). A framework of guidance for building good digital collections. 3rd ed. Tech. rep., NISO, 2007. https://www.niso.org/sites/default/files/2017-08/framework3.pdf.

[49] Network Development and MARC Standards Office. Library of Congress. Functional analysis of the marc 21 bibliographic and holdings formats. Tech. rep., Library of Congress, 2006. Updated and Revised version. April 6, 2006. https://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html.

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

[51] Ochoa, X., and Duval, E. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries 10*, 2-3 (2009), 67–91. http://doi.org/10.1007/s00799-009-0054-4.

[52] Palavitsinis, N. *Metadata Quality Issues in Learning Repositories*. PhD thesis, Alcala de Henares, Feb. 2014. http://www.slideshare.net/nikospala/metadata-quality-issues-in-learning-repositories.

[53] Park, J.-r. Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly 47*, 3-4 (2009), 213–228. https://www.tandfonline.com/doi/abs/10.1080/01639370902737240.

[54] Radulovic, F., Mihindukulasooriya, N., García-Castro, R., and Gómez-Pérez, A. A comprehensive quality model for linked data. *Semantic Web*, Preprint (2017), 1–22. http://www.semantic-web-journal.net/system/files/swj1247.pdf.

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

[56] Smith, B. A brief visual history of marc cataloging at the library of congress., 2017. http://sappingattention.blogspot.de/2017/05/a-brief-visual-history-of-marc.html.

[57] Stiller, J., and Király, P. 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)* (2017), M. Gäde, V. Trkulja, and V. Petras, Eds., Schriften zur Informationswissenschaft, Verlag Werner Hülsbusch, pp. 164–176.

https://www.researchgate.net/publication/314879735_Multilinguality_of_Metadata_Measuring_the_Multilingual_Degree_of_Europeana's_Metadata.

[58] Strezoski, G., and Worring, M. Omniart: Multi-task deep learning for artistic data analysis. *CoRR abs/1708.00684* (2017). http://arxiv.org/abs/1708.00684.

[59] Stvilia, B., Gasser, L., Twidale, M. B., and Smith, L. C. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology 58*, 12 (2007), 1720–1733. http://onlinelibrary.wiley.com/doi/10.1002/asi.20652/full.

[60] Suominen, O., and Hyvönen, E. Improving the quality of SKOS vocabularies with skosify. In *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012.* (2012), A. ten Teije, Ed., vol. 7603 of *Lecture Notes in Computer Science*, Springer, pp. 383–397. http://dx.doi.org/10.1007/978-3-642-33876-2_34.

[61] Tennant, R. Marc must die. *Library Journal 41*, 4 (2002), 185–194. http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/.

[62] Vogias, K., Hatzakis, I., Manouselis, N., and Szegedi, P. Extraction and visualization of metadata analytics for multimedia learning object repositories: The case of TERENA TF-media network, 2013. https://www.terena.org/mail-archives/tf-media/pdf547CE3lKFt.pdf.

[63] W3C. *Data on the Web Best Practices Data Quality Vocabulary*, dec 2016. https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/.

[64] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. The fair guiding principles for scientific data management and stewardship. *Scientific data 3* (2016). https://www.nature.com/articles/sdata201618.

[65] Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., and Dumontier, M. A design framework and exemplar metrics for fairness. *Scientific data 5* (2018). https://doi.org/10.1038/sdata.2018.118.

[66] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. Quality assessment for linked data: A survey.

Semantic Web 7, 1 (2015), 63–93. http://content.iospress.com/articles/semantic-web/sw175.