Péter Király und Jan Brase

4.3 Qualitätsmanagement

Abstract: Dieses Kapitel gibt einen Überblick über das Datenqualitätsmanagement. Es listet einige Ansätze zum Thema und seine grundlegenden Definitionen auf. Die Datenqualität hängt immer vom Kontext und Zweck der Daten ab, daher haben verschiedene Bereiche unterschiedliche Metriken zur Messung der Daten geschaffen. Es werden einige relevante Bereiche, wie Forschungsdaten, verknüpfte Daten, Datenjournalismus, untersucht, um ihre Messprinzipien und besten Praktiken hervorzuheben. Schließlich werden einige praktische Beispiele, darunter Europeana (die europäische Kulturerbe-Plattform) und Forschungsdaten-Repositorien gezeigt.

Einleitung

Um sich mit dem Bereich Qualitätsmanagement bei Forschungsdaten zu beschäftigen, ist es sinnvoll, zuerst zu betrachten, wie Qualitätsmanagement in anderen Sammlungen von digitalen Inhalten schon seit geraumer Zeit betrieben wird. Im Bereich Kulturerbe beispielsweise hat die Entwicklung von Katalogen eine lange Tradition. Im Laufe der Jahrhunderte entwickelten Museen, Archive und Bibliotheken verschiedene Systeme zur Erfassung ihrer Bestände. Wie wird nun in diesen digitalen Systemen die Qualität sichergestellt?

Zwar gibt keine einheitliche Definition für Qualität an sich, aber ein Großteil der Literatur[1] ist sich einig, dass die Qualität mit der „Eignung für einen Zweck" übereinstimmen sollte. D. h. für die Qualität eines Objekts sollte gemessen werden, wie sehr das Objekt einen bestimmten Zweck unterstützt. Die Hauptziele der Metadaten zum Kulturerbe sind die Registrierung der Sammlung und die Unterstützung der Nutzenden bei der Entdeckung. Die Funktionsanalyse des MARC 21-Formats[2] (das international am weitesten verbreitete Metadatenschema für bibliographische Datensätze) geht weiter und richtet Funktionsgruppen ein, wie z. B. Suche, Identität, Auswahl, Verwaltung, Verarbeitung und Klassifizierung der zugrunde liegenden Schemaelemente in diesen Kategorien.[3] Durch die Analyse der Felder der einzelnen Datensätze können wir also genauer sagen, welche Aspekte der Qualität gut oder schlecht sind.

1 Vgl. z. B. die „metadata assessment bibliography" bei Zotero: https://www.zotero.org/groups/488224/metadata_assessment. Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

2 Vgl. Desley 2003. MARC steht für Machine-Readable Cataloging.

3 Vgl. IFLA 1998; Desley 2003; Library of Congress 2006.

https://doi.org/10.1515/9783110657807-020

Diese Katalogdaten dienen nicht nur der Registrierung und der Entdeckung der Materialien, sie sind auch die Quelle für zusätzliche Forschungen z. B. in den Geisteswissenschaften. Der Katalog enthält viele Sachinformationen, die in anderen Quellen nicht (oder nicht organisiert) verfügbar sind, und so hätte man vor dem Zeitalter der Digitalisierung die gedruckten Kataloge der wichtigsten Sammlungen (z. B. British Library,[4] Library of Congress[5] etc.) in den Lesesälen verschiedener Forschungseinrichtungen finden können. In den letzten zwei Jahrzehnten haben mehrere Forschungsprojekte bestehende Bibliotheksmetadaten an verschiedene Arten von Volltextdatensätzen (z. B. optische Zeichenerkennung oder XML-kodierte Versionen) angehängt, um zusätzliche Facetten für den Analyseprozess zu liefern, wie persönliche oder institutionelle Namen (Autoren, Verlage), geografische Informationen (Erscheinungsorte), Zeitspanne usw.

Nur ein paar Beispiele: KOLIMO (Corpus of Literary Modernism)[6] verwendet TEI-Headers (Text Encoding Initiative)[7], welche Kataloginformationen sowie andere Metadaten enthalten, um Literatur und Sprachmerkmale zu extrahieren, die für einen bestimmten Zeitraum oder für einen bestimmten Autor spezifisch sind. OmniArt[8] ist ein Forschungsprojekt, basierend auf Metadaten des Rijksmuseum[9] (Amsterdam), des Metropolitan Museum of Arts[10] (New York) und der Web Gallery of Art.[11] Sie sammelten 432 217 digitale Bilder mit kuratierten Metadaten (die größte Sammlung dieser Art), um eine kategorische Analyse durchzuführen. Benjamin Schmidt verwendet die HathiTrust[12] digital library und ihre Metadatensätze um Klassifikationsalgorithmen des maschinellen Lernens zu testen, bei denen er die Ergebnisse mit den in den Metadatensätzen verfügbaren Themenüberschriften der Library of Congress vergleichen kann.[13] Die Gemeinsamkeiten dieser Projekte bestehen darin, dass sie die Katalogdaten der Einrichtungen des kulturellen Erbes als primäre Quellen für ihre eigene Forschung verwenden. Es ist offensichtlich, dass die Qualität dieser Daten Auswirkungen auf die Schlussfolgerungen der Forschung haben könnte, und andererseits liegt es außerhalb der Verantwortlichkeiten und

4 Über die Kataloge der British Library und ihrer Vorgänger: http://vll-minos.bl.uk/reshelp/findhelprestype/catblhold/printedcatalogues/printedcats.html.

5 Über den National Union Catalog (USA) siehe https://en.wikipedia.org/wiki/National_Union_Catalog. Digitale Ausgaben verfügbar über HathiTrust: https://catalog.hathitrust.org/Record/000140237.

6 S. https://kolimo.uni-goettingen.de/index.html.

7 S. https://tei-c.org/.

8 Vgl. Strezoski und Worring 2017. Das Projekt ist verfügbar über http://www.vistory-omniart.com/.

9 S. https://www.rijksmuseum.nl/.

10 S. https://www.metmuseum.org/.

11 S. https://www.wga.hu.

12 S. https://www.hathitrust.org.

13 Vgl. Smith 2017.

Möglichkeiten einzelner Forschender (oder sogar einer Forschungsgruppe), die Aufzeichnungen jeweils zu validieren und bei Bedarf zu korrigieren.

Dieser Anwendungsfall von Daten zum Kulturerbe ist in letzter Zeit so häufig geworden, dass er vor zwei Jahren zu einem neuen Begriff geführt hat: „Sammlungen als Daten", bzw. „collections as data". Wie das Santa Barbara Statement on Collections as Data zusammenfasst:

> Seit Jahrzehnten bauen Institutionen für das Kulturerbe digitale Sammlungen auf. Gleichzeitig haben die Forscher auf rechnergestützte Mittel zurückgegriffen, um Fragen zu stellen und nach Mustern zu suchen. Diese Arbeit steht unter einer Vielzahl von Namen, einschließlich aber nicht beschränkt auf: Text-Mining, Datenvisualisierung, Mapping, Bildanalyse, Audioanalyse und Netzwerkanalyse. Mit bemerkenswerten Ausnahmen [...] haben Institutionen des Kulturerbes seltene digitale Sammlungen aufgebaut oder den Zugang gestaltet, um die maschinelle Nutzung zu unterstützen. Wenn man über Sammlungen als Daten nachdenkt, signalisiert dieses die Absicht, diese Herangehensweise zu ändern.[14]

Während einerseits Sammlungen als Datenbewegung die Bedeutung der Wiederverwendbarkeit von Daten des Kulturerbes hervorheben, und wir erwarten, dass diese große und wichtige Bewegung Organisationen dabei unterstützen wird, mehr über die wissenschaftliche Nutzung oder ihre Metadaten nachzudenken,[15] konzentrieren sich ihre Prinzipien andererseits auf den Zugang und die Beseitigung aktueller Barrieren, und sie übersehen dabei jedoch die Qualitätsaspekte. Der Aspekt der Qualitätsbewertung, den wir hier betrachten, wäre ein ergänzendes Element neben den anderen Prinzipien.

1 Metadatenqualität

> Wir erkennen es [d. h. die Qualität der Metadaten], wenn wir es sehen, aber die Vermittlung des vollen Bündels von Annahmen und Erfahrungen, die es uns ermöglichen, es zu identifizieren, ist eine andere Sache.[16]

14 Collections as Data project team 2017. The Santa Barbara Statement on Collections as Data. v2. https://collectionsasdata.github.io/statement/.

15 Ein Bericht aus dem Jahr 2016, der die Nutzung zweier wichtiger britischer Sammlungen des Kulturerbes analysiert, erwähnt, dass „die verfügbaren Zitationsnachweise eine wachsende Literatur zeigen, die mit EEBO [Early English Books Online] oder HCPP [House of Commons Parliamentary Papers]" arbeitet und dass „Verschiebungen zu geisteswissenschaftlichen Datenwissenschaften und datengetriebener Forschung [...] für Wissenschaftler von wachsendem Interesse" sind, vgl. Meyer und Eccles 2016, 51, 52–53.

16 Bruce und Hillmann 2004, 1.

Die National Information Standards Organization (NISO) stellt eine Definition für Metadaten zur Verfügung, als „ structured information that describes, explains, locates, or otherwise represents something else“.[17] Das Interessante an dieser Definition ist die Liste der Verben: beschreiben, erklären, finden und repräsentieren. Metadaten sind keine statische Einheit, sie haben mehrere verschiedene Funktionen und sollten im Kontext anderer Einheiten stehen. Das steht im Einklang mit dem berühmten Qualitätssicherungsslogan „fitness for purpose“. Es gibt verschiedene Definitionen dieses Slogans, unter anderem kann man ihn aufbrechen in:

- Erfüllung einer Spezifikation oder der angegebenen Ergebnisse;
- gemessen an dem, was als Ziel der Einheit angesehen wird;
- zur Erreichung der institutionellen Mission und der Ziele.

Aus diesen Definitionen können wir zwei wichtige Schlussfolgerungen ziehen:

- Die Qualität eines Objekts ist kein absoluter Wert, sie hängt vom Kontext des Objekts ab, welche Ziele die Benutzenden im aktuellen Kontext mit Hilfe des Objekts erreichen möchten.
- Die Qualität ist ein facettenreicher Wert. Da das Objekt unterschiedliche Funktionen haben kann, sollten wir die Erfüllung von ihnen unabhängig voneinander bewerten.

Die Definition von Metadaten durch die NISO passt gut in diesen Rahmen, da sie die Vielschichtigkeit und den Kontext der Metadaten hervorhebt.

In einer aggregierten Metadatensammlung wie z. B. Europeana[18] besteht der Hauptzweck der Metadaten darin, Zugangspunkte zu den Objekten bereitzustellen, die von diesen Metadaten beschrieben werden (und die in den Institutionen, die das Kulturerbe bereitstellen, gespeichert sind). Wenn die in Europeana gespeicherten Metadaten von geringer Qualität sind oder fehlen, kann der Dienst keine Zugangspunkte bereitstellen und der Benutzer wird das Objekt nicht verwenden.

Wie Bruce und Hillmann erklären, könnte eine Expertin bzw. ein Experte erkennen, ob ein bestimmter Metadatensatz „gut“ oder „schlecht“ ist. Wenn wir dieses Wissen formalisieren wollen, müssten wir zuerst die Dimensionen der Qualität, Metriken und Messmethoden festlegen.

17 National Information Standards Organization 2007. „strukturierte Informationen, die etwas anderes beschreiben, erklären, lokalisieren oder anderweitig darstellen“ (deutsche Übersetzung Kiraly/Brase).

18 S. https://europeana.eu/.

2 Metriken in der Literatur

In der Literatur der Metadatenqualitätsbewertung findet man eine Reihe von metrischen Definitionen. Im Folgenden gehen wir auf einige von ihnen ein, die in diesem Zusammenhang als relevant erachtet wurden.

Während es sich auf den Kontext des Kulturerbes bezieht, definiert die oben bereits zitierte Seminararbeit von Bruce und Hillmann Datenqualität.[19] Palavitsinis fasste sie in seiner Doktorarbeit folgendermaßen zusammen:[20]

- *Vollständigkeit – Completeness:* Anzahl der vom Annotator ausgefüllten Metadatenelemente im Vergleich zur Gesamtzahl der Elemente im Anwendungsprofil.
- *Genauigkeit – Accuracy:* In einem genauen Metadatensatz entsprechen die in den Feldern enthaltenen Daten der zu beschreibenden Ressource.
- *Konsistenz – Consistency:* Konsistenz misst den Grad, in dem die bereitgestellten Metadatenwerte dem entsprechen, was durch das Metadaten-Anwendungsprofil definiert ist.
- *Objektivität – Objectiveness:* Grad, in dem die bereitgestellten Metadatenwerte die Ressource unvoreingenommen beschreiben, ohne zu unter- oder übertreiben.
- *Angemessenheit – Appropriateness:* Grad, in dem die angegebenen Metadatenwerte den Einsatz von Suchmechanismen auf dem Repositorium erleichtern.
- *Korrektheit – Correctness:* Der Grad, in dem die in den Metadaten verwendete Sprache syntaktisch und grammatikalisch korrekt ist.

Derselbe Autor listet in einer Analyse der Metadatenqualitätsliteratur, die sich hauptsächlich auf die Metadaten der Learning Object Repositories[21] konzentriert, die folgenden zusätzlichen Dimensionen auf, die von verschiedenen Autorinnen und Autoren vorgeschlagen werden: Zugänglichkeit, Konformität, Währung, Verständlichkeit, Objektivität, Präsentation, Herkunft, Relevanz und Aktualität. Er wiederholt auch die Kategorisierung von Lee et al.[22] die Qualitätsdimensionen betreffend:

- *Intrinsische Metadatenqualität:* stellt Dimensionen dar, die erkennen, dass Metadaten unabhängig vom Kontext, in dem sie verwendet werden, eine angeborene Korrektheit aufweisen können. Bspw. können Metadaten für ein digitales Objekt mehr oder weniger „genau“ oder „unvoreingenommen“ sein.
- *Kontextuelle Metadatenqualität:* erkennt an, dass die wahrgenommene Qualität je nach der jeweiligen Aufgabe variieren kann und dass die Qualität relevant,

19 Vgl. Bruce und Hillmann 2004, 4–10.
20 Vgl. Palavitsinis 2014, 87–88.
21 S. https://en.wikipedia.org/wiki/Learning_object_metadata.
22 Vgl. Lee et al. 2002, 134.

zeitnah, vollständig und in ihrer Höhe angemessen sein muss, um dem Zweck, für den die Informationen verwendet werden, einen Mehrwert zu verleihen.
- *Repräsentative Metadatenqualität:* bezieht sich auf den Grad, in dem die zu bewertenden Metadaten leicht verständlich sind und in einer klaren, prägnanten und konsistenten Weise dargestellt werden.
- *Zugängliche Metadatenqualität:* verweist auf die Leichtigkeit, mit der die Metadaten zugänglich sind, einschließlich der Verfügbarkeit der Metadaten und der Aktualität ihres Eingangs.

Interessant ist ebenfalls die Arbeit von Zaveri et al. über Linked Data Quality (LD Quality).[23] Sie wurde zum meist zitierten Artikel in Bezug auf die Datenqualität. Zaverli et al. untersuchten, welche Qualitätsdimensionen und -metriken von anderen Autorinnen und Autoren vorgeschlagen wurden, und gruppierten einzelne Metriken in die folgenden Dimensionen:
- *Dimensionen der Verfügbarkeit:* Beinhaltet Bewertungen zur Zugänglichkeit, Lizensierung, Vernetzung, Sicherheit und Performance.
- *Intrinsische Dimensionen:* Beinhalten Bewertungen zur syntaktischen Validität, semantische Genauigkeit, Konsistenz, Prägnanz und Vollständigkeit.
- *Kontextuelle Dimensionen:* Beinhalten Bewertungen zur Relevanz, Vertrauenswürdigkeit, Verständlichkeit und Aktualität.
- *Repräsentative Dimensionen:* Beinhalten Bewertungen zur repräsentativen Prägnanz, Interoperabilität, Interpretierbarkeit und Vielseitigkeit.

Einige dieser Metriken sind nur im Zusammenhang mit LD relevant (so fragt die Bewertung der Zugänglichkeit auch Elemente ab, die LD-technologiespezifisch sind, wie SPARQL-Endpunkt[24] oder RDF-Dump[25]). Auf der anderen Seite gibt es viele Metriken, die auch für nicht verknüpfte Metadaten nützlich sind, wie wir in den nächsten Abschnitten noch sehen werden.

2.1 FAIR Metriken

Eine der wichtigsten aktuellen Entwicklungen im Bereich des Forschungsdatenmanagements (FDM) war die Formulierung der FAIR-Grundsätze.[26] „Die FAIR-Grund-

23 Vgl. Zaveri et al. 2015.
24 SPARQL ist eine rekursive Abkürzung für „SPARQL Protocol and RDF Query Language". Sie wird verwendet, um Daten im Resource-Description-Framework-Format (RDF-Format) zu durchsuchen oder zu verändern. S. https://www.w3.org/TR/rdf-sparql-query/ und https://www.w3.org/RDF.
25 RDF-Dump nennt man eine herunterladbare Datei, die RDF statements in einem der RDF Serialisierungsformate enthält.
26 Vgl. Wilkinson et al. 2016.

sätze enthalten Richtlinien für die Veröffentlichung digitaler Ressourcen wie Datensätze, Code, Workflows und Forschungsobjekte in einer Weise die sie auffindbar, zugänglich, interoperabel und wiederverwendbar macht."[27] Es wurde zum Ausgangspunkt vieler verschiedener Projekte, die entweder diese Prinzipien umsetzen oder zusätzliche Erweiterungen untersuchen. Eines davon ist FAIRMetrics.[28] Es konzentriert sich auf die Messaspekte der FAIR-Prinzipien: Wie können wir Metriken aufstellen, auf deren Grundlage wir die „Fairness" von Forschungsdaten validieren können?

Die Autorinnen und Autoren schlugen vor, dass gute Metriken im Allgemeinen die folgenden Eigenschaften haben sollten. Sie sollten:

- klar,
- realistisch,
- unterscheidend,
- messbar und
- universell sein.

Es gibt 15 FAIR-Prinzipien, und für jedes gibt es eine Metrik. Jede Metrik beantwortet Fragen wie: „Was wird gemessen?", „Warum sollen wir es messen?", „Wie messen wir es?", „Was ist ein gültiges Ergebnis?", „Für welche digitalen Ressourcen ist das relevant?" usw.

Die Autorinnen und Autoren haben die einzelnen Metriken als Nanopublikationen veröffentlicht und arbeiten an einer Implementierung. Neben den Metriken definierten sie „Maturity Indicator Tests", die als REST API verfügbar sind, unterstützt durch eine Ruby-basierte Software namens FAIR Evaluator[29]. Reifegradindikatoren sind ein offener Satz von Kennzahlen. Über das Kernset (das von der FAIRMetrics vorgestellt wurde)[30] hinaus luden die Autorinnen und Autoren die Forschungsgemeinschaften ein, ihre eigenen Indikatoren zu entwickeln, denn sie betonen: „Wir betrachten FAIR als ein Kontinuum von ‚Verhaltensweisen', die von einer Datenquelle dargestellt werden, um zunehmend die maschinelle Auffindbarkeit und (Wieder-)Nutzung zu ermöglichen."[31] Die Elemente von FAIRmetrics sind die Folgenden:

- F1: *Identifier Uniqueness – Eindeutigkeit des Identifier*: Ob es ein Schema zur eindeutigen Identifizierung der digitalen Ressource gibt.

27 Vgl. Wilkinson et al. 2018.

28 Vgl. Wilkinson et al. 2018; GO FAIR Metrics Group n.d.

29 S. https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/. Der zugrunde liegende Software-Code ist verfügbar unter: https://github.com/FAIRMetrics/Metrics/tree/master/MetricsEvaluator-Code.

30 S. das Metrik Repository der FAIR Metrics Group: https://github.com/FAIRMetrics/Metrics/.

31 https://github.com/FAIRMetrics/Metrics.

- F1: *Identifier persistence – Persistenz des Identifier*: Ob es eine Richtlinie gibt, die beschreibt, was der Anbieter im Falle einer Vernachlässigung eines Identifizierungsschemas tun wird.
- F2: *Machine-readability of metadata – Maschinenlesbarkeit der Metadaten*: Die Verfügbarkeit von maschinenlesbaren Metadaten, die eine digitale Ressource beschreiben.
- F3: *Resource Identifier in Metadata – Identifier in den Metadaten*: Ob das Metadatendokument den global eindeutigen und persistenten Identifier für die digitale Ressource enthält.
- F4: *Indexed in a searchable resource – Indexierung in suchbaren Ressourcen*: Der Grad, in dem die digitale Ressource über webbasierte Suchmaschinen gefunden werden kann.
- A1.1:[32] *Access Protocol – Zugangsprotokoll*: Die Art und Nutzungsbeschränkungen des Zugriffsprotokolls.
- A1.2: *Access authorization – Zugangsauthorisierung*: Spezifikation eines Protokolls für den Zugriff auf eingeschränkte Inhalte.
- A2: *Metadata Longevity – Langlebigkeit der Metadaten*: Die Existenz von Metadaten auch bei Abwesenheit/Entfernung von Daten.
- I1: *Use a Knowledge Representation Language – Verwendung einer Wissensrepräsentativen Sprache*: Verwendung einer formalen, zugänglichen, gemeinsamen und allgemein anwendbaren Sprache zur Wissensrepräsentation.
- I2: *Use FAIR Vocabularies – Verwendung von FAIRen Vokabularien*: Die Metadatenwerte und qualifizierten Beziehungen sollten selbst FAIR sein, z. B. Begriffe aus offenen, von der Gemeinschaft akzeptierten Vokabularen, die in einem geeigneten Wissensaustauschformat veröffentlicht werden.
- I3: *Use Qualified References – Verwendung von qualifizierten Verweisen*: Beziehungen innerhalb von (Meta-)Daten sowie zwischen lokalen und Fremddaten haben eine explizite und „sinnvolle“ semantische Bedeutung.
- R1.1: *Accessible Usage License – Zugängliche Nutzungslizenz*: Das Vorhandensein einer dokumentierten Lizenz, sowohl für die Daten als auch für die zugehörigen Metadaten. Außerdem die Möglichkeit (unabhängig voneinander), die Dokumente zu den Lizenzen abzurufen.
- R1.2: *Detailed Provenance – Detaillierte Herkunftsinformationen*: Den Daten sind Herkunftsinformationen zugeordnet, die mindestens zwei primäre Arten von Herkunftsinformationen abdecken: Wer/was/wann die Daten produziert hat (z. B. für Zitate); Warum/wie die Daten produziert wurden (d. h., um den Kontext und die Relevanz der Daten zu verstehen).

32 Es gibt auch A1 und R1 Prinzipien in FAIR. Diese fehlen in FAIRmetrics.

– R1.3: *Meets Community Standards – Genügt den Standards der Gemeinschaft*: Zertifizierung der Ressource, die den Gemeinschaftsstandards entspricht, durch eine anerkannte Stelle.

Die meisten dieser Metriken messen eher das Datenrepository als einzelne Forschungsdatensätze. Es ist zu beachten, dass FAIRmetrics keine klassischen Metadatenqualitätsmetriken (wie Vollständigkeit, Genauigkeit usw.) abdeckt, so dass selbst bei einer robusten Implementierung noch Raum für zukünftige Forschungen zur Forschungs(meta)datenqualität bleibt und andererseits einige dieser Metriken für Daten zum Kulturerbe anwendbar und nachnutzbar sind (z. B. würden persistente Identifier den Aufnahmeprozess von Europeana unterstützen, so dass eine Metrik zur Identifier persistence hier ein nützlicher Indikator wäre).

2.2 Vokabulare zur Validierung von Linked Data

Die Domäne von LD (oder Semantic Web) basiert auf der „Open World"-Annahme,[33] die besagt, dass Objekte (Entitäten) und Aussagen über diese Objekte getrennt sind, verschiedene Akteurinnen und Akteure könnten Aussagen über dasselbe Objekt erstellen. Praktisch bedeutet das, dass es kein abgeschlossenes Konzept einer Metadatenbeschreibung gibt, da das Objekt keine klaren Grenzen hat. Die traditionellen dateibasierten Systeme haben Schemata, die beschreiben, welche Art von Aussagen über eine Entität gemacht werden können. So besteht beispielsweise das Dublin Core Metadata Element Set 1.1[34] aus 15 Metadatenelementen.

Wenn wir z. B. die Farbe eines Buches in diesem Schema neu aufnehmen möchten, können wir das nicht direkt tun. Natürlich können wir diese Informationen in ein semantisch generischeres Feld einfügen, wie z. B. „Format", aber dann verlieren wir die Spezifität, und die Farbe wird zusammen mit anderen Merkmalen wie Größe, Abmessungen usw. gespeichert. Im Kontext von LD ist die Situation anders: Wir können leicht eine neue Eigenschaft einführen und eine Anweisung erstellen, aber wir verlieren die Kontrolle über das Schema. Wir können nicht sagen, ob die neue Eigenschaft gültig ist oder nicht.

Um dieses Problem zu lösen, hat die W3C die Arbeitsgruppe RDF Data Shapes[35] eingerichtet, um „eine Sprache zu entwickeln zur Definition struktureller Einschrän-

33 S. https://en.wikipedia.org/wiki/Open-world_assumption.
34 S. https://www.dublincore.org/specifications/dublin-core/dces/1999-07-02/.
35 S. https://www.w3.org/2014/data-shapes/wiki/Main_Page.

kungen für RDF-Grafiken".[36] Eines der Ergebnisse dieses Ansatzes ist die Shapes Constraint Language (SHACL).[37]

SHACL hat ein Vokabular definiert (siehe Tab. 1) auf dem man Validierungsregeln erstellen kann. Es werden keine direkten Metriken festgelegt, aber diese Einschränkungen sind sehr nützliche Bausteine eines Datenqualitätsmesssystems. Die Implementierung von SHACL basiert auf LD, aber die Definitionen sind auch in anderen Kontexten sinnvoll.

Tab. 1: Kernbedingungen in SHACL

Kategorie	Einschränkungen
Kardinalität	minCount, maxCount
Typen von Werten und Klassen	datatype, nodeKind
Formen	node, property, in, hasValue
Wertebereich	minInclusive, maxInclusive, minExclusive, maxExclusive
Stringbasiert	minLength, maxLength, pattern, stem, uniqueLang
Logische Einschränkungen	not, and, or, xone
Abgeschlossene Formen	closed, ignoredProperties
Einschränkungen für Eigenschaftspaare	equals, disjoint, lessThan, lessThanOrEquals
Nicht validierende Einschränkungen	name, value, defaultValue
Qualizierte Formen	qualiedValueShape, qualiedMinCount, qualiedMaxCount

Im Rahmen des Europeana Data Quality Committee[38] planen wir, häufig auftretende Metadatenprobleme (oder „Anti-Patterns") mit SHACL zu definieren.

2.3 Organisation von Themen nach verantwortlichen Akteuren

Christopher Groskopf, der einen Leitfaden zur Erkennung von Datenproblemen für Datenjournalisten geschrieben hat,[39] verfolgt einen anderen Ansatz. Er verfasste einen praktischen Leitfaden, d. h. er organisiert Probleme basierend darauf, wer sie lösen kann. Seine wichtigsten Botschaften sind:

36 S. https://www.w3.org/2014/data-shapes/charter.

37 S. https://www.w3.org/TR/shacl. Wir sollten feststellen, dass es für das gleiche Problem einen anderen Ansatz gibt: Shape Expressions (ShEx), verfügbar unter: http://shex.io.

38 S. https://pro.europeana.eu/project/data-quality-committee.

39 Vgl. Groskopf 2015.

- Sei skeptisch bezüglich der Daten.
- Überprüfe mit einer explorativen Datenanalyse.
- Überprüfe früh, überprüfe oft (check it early, check it often).

Seine Kategorisierung ist die folgende:
Probleme, die die Quelle lösen sollte:
- Werte fehlen.
- Nullen ersetzen fehlende Werte.
- Daten fehlen, die da sein sollten.
- Zeilen oder Werte sind doppelt.
- Die Rechtschreibung ist inkonsistent.
- Die Reihenfolge der Namen ist inkonsistent.
- Datenformate sind inkonsistent.
- Einheiten sind nicht angegeben.
- Die Kategorien sind schlecht gewählt.
- Feldnamen sind nicht eindeutig.
- Die Herkunft ist nicht dokumentiert.
- Verdächtige Zahlen sind vorhanden.
- Die Daten sind zu grob.
- Die Summen weichen von der veröffentlichten Gesamtmenge ab.
- Spreadsheet hat 65 536 Zeilen.[40]
- Spreadsheet hat Daten in 1900 oder 1904.[41]
- Text wurde in Zahlen umgewandelt.

Probleme, die man selber lösen sollte:
- Text ist verstümmelt.
- Daten sind in einem PDF.
- Daten sind zu feinkörnig.
- Daten wurden von Menschen eingegeben.
- Aggregationen wurden auf fehlenden Werten berechnet.
- Die Probe ist nicht zufällig.
- Margin-of-error ist zu groß.
- Margin-of-error ist unbekannt.[42]
- Die Probe ist verzerrt.
- Daten wurden manuell verändert.

40 Die maximale Anzahl von Zeilen in älteren Versionen von MS Excel Tabellen war 65 536.
41 Das Standarddatum, ab dem MS Excel alle anderen Daten berechnet, ist der 1. Januar 1900, 1. Januar 1904 in der Mac-Version.
42 Die Fehlermarge ist ein Maß für die Genauigkeit eines statistischen Ergebnisses. Ist dieser Wert zu groß (Groskopf schlägt 10 Prozent als Grenze vor), ist das Ergebnis ungenau. Fehlt der Wert oder wird er nicht berechnet, kennen wir die Genauigkeit überhaupt nicht.

- Inflation verzerrt die Daten.
- Natürliche/saisonale Schwankungen verzerren die Daten.
- Zeitrahmen wurde manipuliert.
- Bezugsrahmen wurde manipuliert.

Probleme, bei denen eine externe Expertin bzw. ein externer Experte helfen sollte:

- Autorin bzw. Autor ist nicht vertrauenswürdig.
- Der Sammelprozess ist undurchsichtig.
- Daten bestätigen unrealistische Präzision.
- Es gibt unerklärliche Ausreißer.
- Ein Index maskiert die zugrundeliegende Variation.
- Die Ergebnisse wurden p-gehackt.[43]
- Benford's Gesetz scheitert.[44]
- Zu gut, um wahr zu sein.

Probleme, bei denen eine Entwicklerin bzw. ein Entwickler helfen sollte:

- Die Daten werden zu den falschen Kategorien oder Regionen zusammengefasst.
- Daten befinden sich in gescannten Dokumenten.

Groskopfs Liste ist keine Definition allgemeiner Metriken, sondern ein Katalog von „Anti-Patterns".[45] Sie wurde in Reflexion zum Kontext des Datenjournalismus erstellt, und das bedeutet, dass dieser Ansatz im Vergleich zu den Daten des Kulturerbes ein kleinerer Ansatz ist, sowohl in Bezug auf die Anzahl der Beitragenden als auch auf die Anzahl der Datensätze. Andererseits ist der einzige Zweck dieser Daten die Verwendung in der Datenanalyse, so dass der Datenjournalist als Editor während des Datenreinigungsprozesses mehr Freiheit hat als eine Bibliothekarin bzw. ein Bibliothekar, die bzw. der mehrere Szenarien zur Datenwiederverwendung berücksichtigen sollte. Trotz dieser Unterschiede erhalten Projekte des Kulturerbes auch Anregungen von Groskopfs Liste.

43 Der P-Wert misst das Niveau der statistischen Signifikanz. Es gibt bekannte Beispiele für das Hacken des Wertes von p, was zu irreführenden Schlussfolgerungen führt.

44 Das Benford'sche Gesetz besagt, dass Zahlen an der Anfangsposition großer Zahlen nicht gleichmäßig verteilt sind. Es kann als erster Test verwendet werden, um zu überprüfen, ob die Zahlen nicht evtl. gehackt worden sind. S. https://en.wikipedia.org/wiki/Benford's_law.

45 Wir verwenden hier Anti-Muster als das Gegenteil von Best Practice: häufig auftretende falsche Metadatenmuster.

2.4 Fazit zu Metriken

Im vorherigen Abschnitt haben wir einige der Metriken und Ansätze vorgestellt. Dies ist kein umfassender Überblick.[46] Was wir zeigen wollten, ist, dass es in verschiedenen Forschungsbereichen oder Tätigkeitsbereichen ganz unterschiedliche Ansätze zur Messung der Metadatenqualität und zur Erkennung einzelner Fragestellungen gibt. Es gibt allgemeine Metriken wie Vollständigkeit, formatspezifische Metriken, wie z. B. diejenigen für verknüpfte Daten, die von Amrapali gesammelt wurden. Einige Metriken messen Daten, aber es gibt Metriken, die sich auf Dienste konzentrieren, die Benutzenden den Zugriff auf Daten erleichtern (z. B. das Vorhandensein verschiedener API-Endpunkte oder herunterladbare Datenspeicher – wir könnten die meisten FAIRmetriken in diese Kategorie eintragen). In einem der frühen Artikel zur Metadatenqualität betonen Stvilia et al[47] dass das von ihnen erstellte Informationsqualitäts-Framework[48] (IQ-Framework) auf eine Datenquelle angewendet werden sollte, indem relevante IQ-Dimensionen ausgewählt werden. Mit anderen Worten, nicht alle Metriken sind in jeder Situation nützlich, wir sollten für jeden Anwendungsfall die Richtige auswählen.

3 Fazit zu Messbarkeit: Europeana

Einer der Autoren dieses Beitrags arbeitete an der Messung der Metadatenqualität von Europeana. Was er nützlich fand – auf Anregung von Stvilia et al.[49] –ist die Mischung aus verschiedenen Qualitätsdimensionen, Kennzahlen und Ansätzen. Die wichtigsten Arten der Datenqualitätsmessung in der Dissertation[50] waren die Folgenden:

1. *Allgemeine strukturelle und semantische Metriken.* Diese Messungen sind die bekanntesten in der Literatur. Basierend auf dem bekanntesten Artikel dieses Forschungsgebietes[51] sind sie:
   - *Vollständigkeit – completeness*: die Existenz von definierten Felder in den Datensätzen,

46 Für diejenigen, die einen allgemeinen Überblick über die Metadaten-Qualitätsmetriken lesen möchten, empfehlen wir die bereits zitierte Doktorarbeit von Palavitsinis 2014.
47 Vgl. Stvilia et al. 2007, 1726.
48 Das Framework enthält Typologien der IQ-Varianz, die betroffenen Aktivitäten, eine umfassende Taxonomie der IQ-Dimensionen sowie allgemeine metrische Funktionen und Methoden der Rahmenoperationalisierung.
49 Vgl. Stvilia et al. 2007.
50 Vgl. Király 2019.
51 Vgl. Bruce und Hillmann 2004; Ochoa und Duval 2009.

- *Übereinstimmung mit den Erwartungen – confromance to expectations*: Schema-Regelprüfung und Informationswert,
- *Zugänglichkeit – accessibility*: wie einfach es ist, den Text des Datensatzes zu verstehen,
- *Logische Konsistenz und Kohärenz – logical consistency and coherence*: Die „Stimmigkeit" der Daten
- *Herkunft – provenance*: die Beziehung zwischen anderen Metriken und dem Ersteller der Daten.

Die *Genauigkeitsdimension* (Vergleich eines vollständigen Datenobjekts und seiner Metadaten) wurde nicht untersucht, da sie den Vergleich von Metadaten und deren Gegenstand – z. B. den Volltext von Büchern – erfordert, die nicht verfügbar waren.

2. *Unterstützung der funktionalen Anforderungen*. Diese Dimension ist eine Variation der Vollständigkeit. Jedes Datenschema wird zur Unterstützung einer Reihe von Funktionen erstellt, wie z. B. Suchen, Identifizieren oder Beschreiben von Objekten. Die Datenelemente unterstützen eine oder mehrere dieser Funktionen und ihre Existenz sowie ihr Inhalt haben Auswirkungen auf diese Funktionen. Ein Beispiel: Ein Timeline Widget erwartet ein bestimmtes Datumsformat; wenn der Feldwert in einem anderen Format ist, ignoriert das Widget es. Diese Familie von Metriken gibt Messungen den Umfang für die Unterstützung der funktionalen Anforderung. Um diese Metriken anzuwenden, sollten wir eine funktionale Anforderungsanalyse des Datenschemas durchführen und die einzelnen Datenelemente (Klassen und Eigenschaften) auf die Funktionen abbilden. Das Ergebnis ist ein Bericht, der sagt, wie die Daten die vorgesehenen Funktionen unterstützen. In Anlehnung an die bei Stvilia festgelegte Terminologie[52] nennen wir diese Aspekte „Sub-dimensions". Das Europeana Data Quality Committee definierte eine Reihe von solchen Sub-dimensions (wie Suchbarkeit, Beschreibbarkeit, Identifizierung, Kontextualisierung, Browsing usw.), die in anderen Metadatenbereichen wiederverwendet werden können. In Bezug auf das MARC 21-Schema hat die Library of Congress zwölf Aufgaben definiert und eine Zuordnung zwischen ihnen und den Datenelementen des Schemas erstellt.[53] Es stellte sich heraus, dass der Ansatz zur Messung der funktionalen Unterstützung eng an die Vollständigkeit gebunden ist, und da die Gesamtzahl der Datenelemente in MARC viel höher ist als die tatsächlich verfügbaren Felder in den Datensätzen, ist nicht nur die Vollständigkeit, sondern auch die funktionale Unterstützung gering.
3. *Existenz bekannter Datenmuster*. Dies sind schema- und domänenspezifische Muster, die in den Datensätzen häufig vorkommen. Es gibt gute Muster, die

52 Vgl. Stvilia 2006, 20.

53 Vgl. Desley 2002; Library of Congress 2006.

gute Datenerstellungspraktiken erkennen lassen, und Anti-Muster, die vermieden werden sollten (wie Datenwiederholung, bedeutungslose Daten usw.). Für einige Bereiche gibt es bereits Musterkataloge, z. B. arbeitet das Europeana Data Quality Committee an einem Europeana-spezifischen Musterkatalog, während Suominen und Hyvönen drei SKOS-Validierungskriterienkataloge untersucht haben.[54] Király zeigte auch einige der Anti-Muster in MARC 21-Aufzeichnungen.[55] Diese Messungen können unter „conformance to expectations" kategorisiert werden.

4. *Multilingualität*. Das Resource Description Framework (RDF) bietet eine leicht anpassbare Technik, um literalen Werten ein Sprachkennzeichen hinzuzufügen, was die Mehrsprachigkeit zu einem wichtigen Aspekt in der vernetzten offenen Datenwelt macht. In Kulturerbe-Datenbanken kann die Übersetzung der beschreibenden Felder (wie Titel, Beschreibung) eine sehr personalintensive Aufgabe sein. Andererseits ist die Wiederverwendung bestehender mehrsprachiger Thesauri für Schlagworte ein relativ einfacher und kostengünstiger Prozess. Für das Messen der Qualität ist das Schöne daran, dass die mehrsprachige Ebene in Metadatenschemata (auch in solchen, die nicht auf RDF-basieren) im Allgemeinen ähnlich ist, so dass die Implementierung abstrahiert werden kann. Das große Problem ist, wie man mit den Verzerrungen umgeht, die durch die unterschiedliche Bedeutung der Datenelemente in den einzelnen Sprachen entstehen. Ein anderes Problem ist die unterschiedliche Kardinalität bei einigen Begriffen: Europeana hat zum Beispiel „Dokument" als Betreffzeile, die in mehr als siebzig Sprachen zugänglich ist, aber es ist an einen großen Teil der Datensätze angehängt (mehr als 20 Prozent), so dass sein Informationswert oder seine Unterscheidungskraft gering ist – wenn der Benutzer nach Dokumenten sucht, erhält er Millionen von Datensätzen. Diese Messung könnte unter „conformance to expectations" und „accessibility" kategorisiert werden.

Der gemeinsame Punkt dieser Metriken ist, dass sie als generische Funktionen implementiert werden können, bei denen Eingabeparameter spezifische Elemente eines Datenschemas sind. Die Funktionen selbst sollten die Details des Schemas nicht kennen, d. h. sie sollten schemaunabhängig sein. Mit anderen Worten: Das Einzige, was wir auf Schemabasis erstellen sollten, ist eine Methode, die sich um die Abbildung der Schema-Elemente und Messfunktionen kümmert und diese generischen Funktionen mit den entsprechenden Metadatenelementen versorgt.

Der Messprozess besteht aus den folgenden Phasen:

1. Datenaufnahme,
2. Messung von Einzelsätzen,

54 Vgl. Suominen und Hyvönen 2012.

55 Vgl. Király 2019b. 164–165.

3. Analyse der Messergebnisse, um eine Gesamtansicht für die gesamte oder eine Teilmenge der Sammlung zu erhalten,
4. Berichterstattung über die Ergebnisse,
5. Diskussion der Ergebnisse innerhalb einer Expertengemeinschaft.

Diese Phasen bilden eine Schleife; nach Phase 5 endet der Prozess entweder oder geht zurück zu Phase 2, 3 oder 4.

Wie gezeigt wurde, hat die Metadatenqualität mehrere Dimensionen. Für jede Datenquelle sollten wir diejenigen Maßnahmen auswählen, die sowohl theoretisch als auch praktisch zu den Datenquellen passen. Diese Maßnahmen haben jeweils ihren „rechnerischen Fußabdruck“: Die Berechnung erfordert eine bestimmte Menge an Personal- und IT-Ressourcen (und sie sind nicht immer vorhersehbar), wir sollten sie sowohl in Forschungs- als auch in Nicht-Forschungsprojekten berücksichtigen. Ein weiterer wichtiger Aspekt ist die menschliche Komponente: Die Metriken sollten nicht nur aus statistischer Sicht sinnvoll, sondern auch für die Datenpflegenden von Bedeutung sein. Die Metriken sollen einen Entscheidungsprozess über die Änderung der Daten unterstützen. Während der Recherche war dieses der schwierigste Punkt: die Schnittmenge der Interessen der Metadaten-Expertinnen und -Experten zu finden. Es kam immer wieder vor, dass das Ergebnis aus Sicht der Katalogisierer nicht sinnvoll war, so dass es auf Basis der Rückmeldungen verbessert werden musste. Es war eine angenehme Situation, dass die Forschung zusammen mit einer Expertengruppe, dem Europeana Data Quality Committee, durchgeführt wurde, deren Mitglieder ständig Feedback gaben.

4 Forschungsdaten

Welche Metriken außer den bereits besprochenen FAIR-Metriken sind nun im Umgang mit Forschungsdaten anwendbar? CoreTrustSeal[56] ist eine Zertifizierung für Forschungsdatenrepositorien, die auf den DSA-WDS Core Trustworthy Data Repositories Requirements[57] basiert. Die Zertifizierung ist ein Nachfolger des Data Seal of Approval. Ziel ist es nachzuweisen, dass die zertifizierten Repositorien die besten Praktiken des FDM befolgen. Unternehmen sollten ihre Aktivitäten in 15 Bereichen erläutern, wie Datenzugriff, Lizenzen, Workflow, Datenintegrität usw. Es gibt zwei Bereiche, die aus Sicht der Metadatenqualitätsmessung interessant sind: Bewertung

56 S. https://www.coretrustseal.org/.

57 S. https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf.

und Datenqualität. Die Zertifikate enthalten die Antwort des Unternehmens und die Notizen der Zertifizierungsinstitution und sind öffentlich zugänglich.[58]

Zum jetzigen Zeitpunkt gibt es 54 CoreTrustSeal-zertifizierte Repositorien. Die Zertifizierungen sind sehr interessante Dokumente, und zusammen bilden sie eine Art Querschnitt durch den Stand der Technik in den 15 Bereichen der Datenrepositorien. Es scheint, dass sich ihre Aktivitäten zur Daten- und Metadatenqualität auf folgende Themen konzentrieren:

- Einstellen der Liste der empfohlenen und akzeptierten Dateiformate und Prüfen eingehender Dateien daraufhin.
- Dokumentationsaufwand auf verschiedenen Ebenen (allgemein, domänenspezifisch, national) bei der Erstellung von Handbüchern und Leitfäden, sowohl für die Benutzenden als auch für die Betreuenden des Repository.
- Datenkuration durch Expertinnen und Experten – die meisten dieser Repositorien sind nicht vollautomatisch, wenn die hinterlegten Materialien von Expertinnen und Experten sorgfältig überprüft werden. Sie überprüfen sowohl Archivierungsaspekte (Formate, Metadaten) als auch Domänenaspekte (Inhaltsrelevanz).
- Verwaltung sensibler Daten (sichere Datenverwaltung oder Ausschluss nicht anonymisierter Daten).
- Einstellung von Pflichtfeldern, empfohlenen und optionalen Feldern in Bezug auf die Metadatensätze.
- Online-Formularvalidierung für die Metadaten, die über eine Online-Benutzeroberfläche erstellt wurden.
- Anwendung von XML-Validierern in einigen Repositorien, wenn der Metadatensatz voraussichtlich im XML-Format verfügbar ist.

Unter den traditionellen Metadaten-Qualitätsdimensionen wird nur die Vollständigkeit erwähnt und als Synonym für den Fall verwendet, dass alle Pflichtfelder im Metadatensatz verfügbar sind: „Sicherstellen, dass DDI-Felder in den Metadaten ausgefüllt werden, gewährleistet die Qualitätskontrolle der Vollständigkeit“, schreibt das Australian Data Archive[59] zu dem Thema.

Nur ein kleiner Teil der Repositorien erwähnte die Verwendung von kontrolliertem Vokabular und nur ein Repository, nämlich das institutionelle Forschungsdatenrepositorium FDAT der Universität Tübingen erwähnt überhaupt namentlich ein unabhängiges Tool zur Automatisierung der Metadaten-Qualitätsprüfung.[60] Die Worldwide Protein Data Bank[61] erwähnt, dass sie zwei Arten von Darstellungen der

58 S. https://www.coretrustseal.org/why-certification/certified-repositories.

59 S. https://assessment.datasealofapproval.org/assessment_245/seal/html.

60 FDAT, Tübingen verwendet den docuteam packer, s. https://wiki.docuteam.ch/doku.php?id=-docuteam:packer.

61 S. https://assessment.datasealofapproval.org/assessment_281/seal/html.

Datenqualitätsbewertung erstellt hat: eine für Spezialistinnen und Spezialisten sowie eine für Nicht-Spezialistinnen und Nicht-Spezialisten. Die Letztere enthält eine einfache grafische Darstellung, die eine kleine Anzahl von wesentlichen Qualitätskennzahlen hervorhebt. Verschiedene Repositorien erwähnen, dass sie Metadatensätze von guter Qualität als Beispiele in der Dokumentation wiederverwenden.

Es lohnt sich, die Checkliste des Digital Repository of Ireland[62] zu zitieren, in der die empfohlenen Schritte zur Durchführung regelmäßiger Metadatenqualitätssbewertungen beschrieben werden:

- Benennen Sie eine Person oder ein kleines Team von Informationsexpertinnen bzw. -experten, die die Verantwortung für das Audit übernehmen.
- Entscheiden Sie, inwieweit während des Audits festgestellte Fehler in der Live-Datenbank behoben werden.
- Auf vierteljährlicher oder halbjährlicher Basis laden Sie einen Beispielsatz von Datensätzen in die Softwareanwendung OpenRefine hoch.
- Verwenden Sie die Facettier- und Cluster-Tools in OpenRefine, um Fehler wie Rechtschreibfehler, inkonsistente Verwendung der Groß-/Kleinschreibung oder leere Zellen zu identifizieren und zu erfassen.
- Stellen Sie die Dokumentation so zusammen, dass Qualitätsänderungen über einen längeren Zeitraum festgestellt werden können. Dies ist besonders nützlich, wenn das Unternehmen vor Kurzem begonnen hat, neue Katalogisierungsmethoden anzuwenden.

Die am weitesten verbreiteten allgemeinen Metadatenschemata sind die Elemente des Data Documentation Initiative (DDI)[63] Frameworks[64] und The Dublin Core Metadata Initiative's DCMI Metadata Terms.[65] In Bezug auf Metadatenschemata könnte CLARINs Component Metadata[66] als Standard in linguistischen Datenrepositorien angesehen werden.

Eine wichtige Schlussfolgerung aus dieser vorläufigen Analyse ist, dass es eine Art „Marktlücke" sowohl in der Forschung als auch in der Werkzeugentwicklung im Bereich des FDM gibt. Die in den Zertifikaten genannten Elemente der Datenqualität (Vollständigkeit, Formatkonsistenz, Inhaltsrelevanz, Prüfung von Facetten auf Fehler usw.) unterscheiden sich nicht von denen, die man in anderen Metadaten-Domänen finden kann. Es gibt Elemente, die existieren, aber anscheinend nicht die Popularität erreicht haben, die sie verdienen, z. B. die „frictionless data"-Datenbe-

62 S. McCarthy 2014, 4.
63 S. http://www.ddialliance.org.
64 DDI Lifecycle, s. http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation; das DDI Codebook, s. http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation.html.
65 S. http://www.dublincore.org/specifications/dublin-core/dcmi-terms.
66 S. https://www.clarin.eu/content/component-metadata.

schreibung Metadatenformat[67] oder FAIRmetrics.[68] Ganz zu schweigen von den allgemeinen Elementen der Metadatenqualitätsforschung (Dimensionen, Metriken und Werkzeuge), die in diesen Bereich eingeführt werden könnten, zur Zufriedenheit sowohl der Betreibenden der Datenrepositorien, als auch der Metadatenqualitätsforschenden.

Im Jahr 2016 bildeten sich zwei wichtige Gruppen im Bereich des Kulturerbes, die eine eingehende Untersuchung der Datenqualität in bestimmten Segmenten begannen: das Europeana Data Quality Committee (DQC)[69] und die Digital Library Federation Metadata Assessment Working Group (MAWG).[70] Das DQC untersucht die für die Europeana-Sammlung spezifischen Metadatenfragen und ist an der Schaffung des Messrahmens beteiligt. Das MAWG konzentriert sich nicht auf einen bestimmten Dienst und ein bestimmtes Metadatenschema, sondern sammelt relevante Literatur und Anwendungsfälle und versucht, eine Reihe von Empfehlungen zur Bewertung der Metadatenqualität zu formulieren. 2017 startete Auditing Digitalization Outputs in the Cultural Heritage Sector, Belgium,[71] (ADOCHS) mit dem Ziel, den Qualitätskontrollprozess für die digitalisierten Sammlungen der belgischen Nationalbibliothek und des Nationalarchivs zu verbessern. Die Ergebnisse des ADOCHS-Projekts finden sich in den Publikationen von Anne Chardonnens[72] und Ettore Rizza.[73] Ähnliche Aktivitäten der Digital Public Library of America (DPLA) sind bei Gueguen beschrieben.[74]

5 Datenqualitätsprüfung in der Praxis

Es gibt nur wenige Dienste, die eine Datenqualitätsprüfung und Datenkorrekturmechanismen auf der Grundlage der Ergebnisse implementiert haben. Ein sehr schönes Beispiel dafür findet sich bei der University of North Texas Digital Library (UNT DL). Als inhaltliche Drehscheibe für die DPLA kuratiert sie neben den eigenen Materialien zwei externe Sammlungen: das Portal to Texas History und das Gateway to Oklahoma History. Den Workflow zur Qualitätssicherung der Daten haben sie als Teil ihrer Metadaten-Bearbeitungssoftware realisiert. Kuratierende können verschiedene qualitätsbezogene Probleme herausfiltern, die betroffenen Metadatensätze

67 Vgl. Fowler, Barratt und Walsh 2018.
68 Vgl. GO FAIR Metrics Group n.d.
69 S. http://pro.europeana.eu/page/data-quality-committee.
70 S. https://dlfmetadataassessment.github.io.
71 S. http://adochs.be/.
72 S. https://scholar.google.com/citations?hl=en&user=2L_vIJQAAAAJ.
73 S. https://scholar.google.com/citations?hl=en&user=jh_bdOwAAAAJ.
74 Vgl. Gueguen 2019.

auflisten und sie bearbeiten, um die Probleme zu beheben. In einem Screencast für den Metadata Quality Workshop der 2018 ELAG Konferenz[75] zeigten Philipps und Tarver[76] drei unterschiedliche Benutzerinterfaces, um Probleme im Katalog der UNT DL zu entdecken. Das erste listet die Werte auf, die in den einzelnen Feldern gespeichert sind (sie verwenden ein qualifiziertes Dublin Core Schema als Grundlage für ihre Metadatensätze). Als erweiterte Facettenliste kann sie alphabetisch oder nach Häufigkeit sortiert werden. Diese Liste hilft den Kuratierenden, merkwürdige Werte herauszufiltern (z. B. Werte mit unterschiedlicher Interpunktion). Die Zählschnittstelle zeigt an, wie viele Instanzen in einem Datensatz vorhanden sind (z. B. X Datensätze haben eine Instanz, Y hat zwei, während Z keine hat). Philipps erklärte, dass eine Beschreibung entweder ein physischer Typ oder ein Content-Typ sein sollte. Die Schnittstelle zeigt diejenigen Datensätze an, die keinen Typ haben, also Fehler sind. Die letzte (und interessanteste) Schnittstelle zeigt Cluster von Werten an. Dieser Teil der Software verwendet OpenRefine's Clustering-Algorithmen wieder. Beim Clustering wird versucht, verschiedene Werte auf der Basis einer Ähnlichkeit zusammenzuführen. Einer dieser Ähnlichkeitsalgorithmen, der für textuelle Informationen verwendet wird heißt „Fingerprint".

Der Fingerprint-Algorithmus zeigt z. B., dass Schostakowitsch, der russische Komponist, 14 verschiedene Namensformen im Contributor-Feld hat. Die Cluster können nach den extrahierten Schlüsseln, der Anzahl der Variationen, der Anzahl der geclusterten Datensätze u. a. angeordnet werden. Dieser Algorithmus hat zwei spezielle Typen: er kann die Whitespaces oder die Daten, die im Text eines Feldes gefunden werden, ignorieren. Ein anderer Algorithmus könnte für Felder verwendet werden, die hauptsächlich numerische Werte enthalten: „Muster-Maske" ersetzt Zahlen durch Nullen und zeigt so ein Grundmuster, wie z. B. 0000-00-00 oder 0000-0000. Im Falle von Daten erwarten wir keine allzu großen Abweichungen in den sinnvollen Mustern, so dass es relativ einfach ist, nicht interpretierbare Masken, wie z. B. drei Zahlen (die kein gültiges Jahr, Monat oder Tag sein können) herauszufinden.

Ein weiteres Beispiel ist die Qualitätskontrolle von Metadaten in der Nationalbibliothek von Portugal. Ihr System (MANGAS[77] genannt) unterstützt verschiedene Schritte des Qualitätskontrollprozesses wie Validierung, Berichterstattung, Filterung und Korrektur. MANGAS liest die Eingabedaten (das sind UNIMARC-Dateien im XML-Format), erkennt Probleme, kategorisiert sie und erstellt einen Bericht für die Kuratierenden. Wo es möglich ist, gibt es auch Vorschläge für die Korrektur von

75 UNT Libraries Metadata Quality Interfaces – ELAG 2018, s. https://www.youtube.com/watch?v=ATM3EwixnW8.

76 Vgl. Phillips und Tarver 2018.

77 Vgl. Manguinhas und Borbinha 2006.

Fehlern oder, wenn es automatisch durchgeführt werden kann, behebt es diese auch auf der Basis eines von den Kuratierenden vorbereiteten „Korrekturskripts".

In diesen Beispielen haben wir gesehen, dass diese Institutionen eine volle Kontrolle über die Daten haben, d. h. sie haben das Recht, sie zu ändern. Sie haben auch ein gut definiertes Metadatenschema, das ihren Bedürfnissen entspricht, und eine etwas begrenzte Anzahl von Datensätzen, welche keine rechenintensiven Operationen wie z. B. Clustering oder Neu-Indizierung erforderlich macht. Das ist nicht immer der Fall.

Es wurde gezeigt, dass Europeana als Datenaggregator nicht die gleiche Kontrolle über die Daten hat, also kann es Datenqualitätsprobleme nicht auf die gleiche Weise beheben wie z. B. UNT DL und aufgrund der Größe der Daten wären einige der Ansätze in einer ähnlichen Benutzeroberfläche zu langsam. Was Europeana stattdessen tun kann, ist eine Datenqualitätsanalyse durchzuführen, die im Europeana Publishing Framework[78] beschrieben wird, und die Ergebnisse den Datenlieferanten in einem statistischen Dashboard zur Verfügung zu stellen. Die Ergebnisse dieser Analyse stehen auch über die API des Dienstes als zusätzliche Metadatenelemente der einzelnen Europeana-Datensätze zur Verfügung.[79]

Das Swedish National Heritage Board experimentiert mit einem interessanten Projekt namens Wikimedia Commons Data Roundtripping.[80] Roundtripping ist der Name des Arbeitsablaufs, in dem eine Kulturerbe-Institution ihre Daten in Wikimedia Commons veröffentlicht, die Nutzerschaft diese offen verfügbaren Daten anreichern (wie z. B. Übersetzungen von Beschreibungstexten in andere Sprachen hinzufügen, Personen, Namen und Aliasnamen, Orte und Themen identifizieren oder mit Normdaten verlinken und diese zum Abrufen von Beiträgen Dritter von anderen Gedächtnisorganisationen verwenden), dann nehmen die Institutionen diese Daten auf und aktualisieren ihre ursprüngliche Datenbank. Die Daten werden so den bestehenden Qualitätsprüfungsmechanismen von Wikipedia und Verbesserungen von Dritten ausgesetzt durch klassische Crowd-Source-Mechanismen.

Aus dieser Übersicht können wir folgenden Schluss ziehen: Der effiziente Datenqualitätsprozess hat mindestens zwei Hauptphasen: Analyse und Korrektur.[81] Die Auswahl der analytischen Ansätze könnte sich an der Komplexität, der Vielfalt und dem Volumen der Daten orientieren. Die Korrektur könnte nur von den Dateneigentümern durchgeführt werden oder zumindest sollten die Änderungen für und

78 S. https://pro.europeana.eu/post/publishing-framework.

79 Implementierung des Data Quality Vocabulary des World Wide Web Consortiums Vgl. W3C 2016.

80 S. https://outreach.wikimedia.org/wiki/GLAM/Newsletter/February_2019/Contents/Special_story.

81 Es gibt natürlich auch prophylaktische und vorausschauende Qualitätssicherungsmaßnahmen, z. B. Daten Managementpläne (DMP), um sich vorab und während der Projektlaufzeit schon mit möglichen Datenresultaten, -typen etc. zu beschäftigen und wie diese perspektivisch zugänglich gemacht bzw. dokumentiert werden sollen.

durch sie transparent und kontrollierbar sein. Im Falle von Forschungsdatenrepositorien sind die Eigentümer die Forschenden, die ihre Daten hochladen. Gemäß den CoreTrustSeal-Berichten in mehreren Repositorien fungieren Datenkuratierende als Vermittelnde zwischen Forschenden und Daten und/oder als Vermittelnde zwischen Forschenden und Serviceinfrastruktureinrichtungen. Für Self-Service-Repositorien wäre es sinnvoll, ein Data-Quality-Dashboard zu erstellen, in dem die Forschenden das Ergebnis der Qualitätsanalyse sehen und dann über die Korrekturen entscheiden können.

Literatur

Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

Bruce, Thomas R. und Diane I. Hillmann. 2004. „The Continuum of Metadata Quality: Defining, Expressing, Exploiting.“ In *Metadata in Practice*, hg. v. D. Hillman und E. Westbrooks, 238–256: ALA Editions. http://ecommons.cornell.edu/handle/1813/7895.

Delsey, Tom. 2002. „Functional Analysis of the MARC 21 Bibliographic and Holdings Formats.“ https://www.loc.gov/marc/marc-functional-analysis/original_source/analysis.pdf.

Fowler, Dan, Jo Barratt und Paul Walsh. 2018. „Frictionless Data: Making Research Data Quality Visible.“ *IJDC* 12 (2): 274–285. doi:10.2218/ijdc.v12i2.577.

Gavrilis, Dimitris, Dimitra-Nefeli Makri, Leonidas Papachristopoulos, Stavros Angelis, Konstantinos Kravvaritis, Christos Papatheodorou und Panos Constantopoulos. 2015. „Measuring Quality in Metadata Repositories.“ In *Proceedings from the 19th International Conference on Theory and Practice of Digital Libraries (TPDL)*, hg. v. S. Kapidakis, C. Mazurek und M. Werla, 56–67. Cham: Springer International Publishing. doi:10.1007/978-3-319-24592-8_5.

GO FAIR Metrics Group. *FAIR Metrics*. Zugriff: 18. April 2019. http://fairmetrics.org/.

Groskopf, Christopher. 2015. „The Quartz guide to bad data.“ Quartz. https://qz.com/572338/the-quartz-guide-to-bad-data/.

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

Harvey, L. (2004). *Analytic Quality Glossary. Quality Research International*. http://www.qualityresearchinternational.com/glossary. https://www.ingentaconnect.com/content/hsp/jdmm/2019/00000007/00000002/art00003.

IFLA. 1998. *Functional requirements for Bibliographic records: final report/IFLA Study Group on the Functional Requirements for Bibliographic Records. UBCIM publications new series*. München: K. G. Saur.

Király, Péter. 2019. „Measuring Metadata Quality.“ Georg-August-Universität Göttingen. doi:10.13140/RG.2.2.33177.77920.

Király, Péter. 2019b. „Validating 126 million MARC records.“. In *DATeCH2019 Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage Brussels, Belgium – May 08–10, 2019*, hg. v. ACM, 161–168. doi:10.1145/3322905.3322929.

Lee, Yang W., Diane M. Strong, Beverly K. Kahn und Richard Y. Wang. 2002. „AIMQ: A methodology for information quality assessment.“ *Information & Management* 40 (2): 133–146. doi:10.1016/S0378-7206(02)00043-5.

McCarthy, Kate. 2014. „Metadata Quality Control.“ Dublin: Royal Irish Academy. doi:10.3318/DRI.2015.1.

Manguinhas, Hugo und José Borbinha. 2006. „Quality control of metadata: a case with UNIMARC.“ In *International Conference on Theory and Practice of Digital Libraries*, hg. v. Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, Rafael C. Carrasco, 244–255. Berlin, Heidelberg: Springer. doi:10.1007/11863878_21.

Meyer, Eric T. und Kathryn Eccles. 2016. „The Impacts of Digital Collections: Early English Books Online & House of Commons Parliamentary Papers.“ London: Jisc. doi:10.2139/ssrn.2740299.

National Information Standards Organization. 2007. „A Framework of Guidance for Building Good Digital Collections. 3rd ed.“ https://www.niso.org/sites/default/files/2017-08/framework3.pdf.

Network Development und MARC Standards Office. Library of Congress. 2006. „Functional Analysis of the MARC 21 Bibliographic and Holdings Formats.“ https://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html.

Ochoa, Xavier und Erik Duval. 2009. „Automatic evaluation of metadata quality in digital repositories.“ *International Journal on Digital Libraries* 10 (2–3): 67–91. doi:10.1007/s00799-009-0054-4.

Palavitsinis, Nikos. 2014. „Metadata Quality Issues in Learning Repositories.“ Doctoral Thesis at Universidad de Alcalá. https://www.researchgate.net/publication/260424499_Metadata-_Quality_Issues_in_Learning_Repositories.

Phillips, Mark E. und Hannah Tarver. 2018. „Experiments in operationalizing metadata quality interfaces: a case study at the university of North Texas libraries.“ In *Proceedings of the 2018 International Conference on Dublin Core and Metadata Applications*, 15–23. https://www.ingentaconnect.com/content/hsp/jdmm/2019/00000007/00000002/art00003.

Smith, Benjamin. 2017. *A brief visual history of MARC cataloging at the Library of Congress*. http://sappingattention.blogspot.de/2017/05/a-brief-visual-history-of-marc.html.

Strezoski, Gjorgji und Marcel Worring. 2017. „OmniArt: Multi-task Deep Learning for Artistic Data Analysis.“ arXiv preprint https://arxiv.org/abs/1708.00684.

Stvilia, Besiki 2006. „Measuring information quality“. PhD dissertation. https://www.researchgate.net/publication/34172596_Measuring_information_quality.

Stvilia, Besiki, Les Gasser, Michael B. Twidale und Linda C. Smith. 2007. „A framework for information quality assessment.“ *Journal of the American Society for Information Science and Technology* 58 (12): 1720–1733. doi:10.1002/asi.20652.

Suominen, Osma und Eero Hyvönen. 2012. „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.*, hg. v. Annette ten Teije, 383–397. (Lecture Notes in Computer Science Bd. 7603) Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-33876-2_34.

W3C. *Data on the Web Best Practices Data Quality Vocabulary*, https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/.

Wilkinson, Mark et al. 2016. *The FAIR Guiding Principles for scientific data management and stewardship.* https://doi.org/10.1038/sdata.2016.18.

Wilkinson, Mark D., Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Bonino da Silva Santos, Luiz Olavo und Michel Dumontier. 2018. „A design framework and exemplar metrics for FAIRness.“ *Scientific data* 5. doi:10.1038/sdata.2018.118.

Zaveri, Amrapali, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann und Sören Auer. 2015. „Quality Assessment for Linked Data: A Survey.“ *Semantic Web* 7 (1): 63–93. doi:10.3233/SW-150175.