Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

Wie FAIR sind unsere Metadaten?

Eine Analyse der Metadaten in den Repositorien des TIB-DOI-Services

Marleen Burger[i] Anette Cordts[ii] Ted Habermann[iii]

2021

Zitiervorschlag

Burger, Marleen, Anette Cordts und Ted Habermann. 2021. Wie FAIR sind unsere Metadaten?. Eine Analyse der Metadaten in den Repositorien des TIB-DOI-Services. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 3/2021: S. 1-13. DOI: 10.17192/bfdm.2021.3.8351.

[i]Technische Informationsbibliothek Hannover (TIB). ORCID: 0000-0001-6836-1193
[ii]Technische Informationsbibliothek Hannover (TIB). ORCID: 0000-0002-9649-7829
[iii]Metadata Game Changers, Colorado USA. ORCID: 0000-0003-3585-6733

1 Abstract

Im vorliegenden Erfahrungsbericht stellen wir eine Metadatenanalyse vor, welche die Metadatenqualität von 144 Repositorien des TIB-DOI-Services im Hinblick auf die Erfüllung der FAIR Principles, Konsistenz und Vollständigkeit untersucht. Im Ergebnis zeigt sich, dass der Fokus der untersuchten Repositorien schwerpunktmäßig auf der Auffindbarkeit der mit Metadaten beschriebenen Ressourcen (z. B. Forschungsdaten oder Textpublikationen) liegt und im Gesamtdurchschnitt über die Metadaten-Pflichtfelder hinaus nur wenige weitere Metadaten angegeben werden. Dies wirkt sich nachteilig auf die Nachnutzbarkeit von Repositoriumsinhalten und zugehörigen Metadaten aus. Auch mit Blick auf die im Sinne der FAIR Principles angestrebte Verknüpfung der in einem Repositorium enthaltenen Ressourcen mit anderen in Beziehung stehenden persistenten Identifikatoren wie ORCID oder ROR ID sowie die Angabe von DOI-zu-DOI-Beziehungen mit zitierten oder zitierenden Ressourcen bestehen noch ungenutzte Potenziale, die im Sinne einer offenen, zukunftsweisenden Wissenschaft erschlossen werden sollten. Dahingegen zeigt unsere Analyse auch einzelne Repositorien mit umfangreichen Metadaten als Best-Practice-Beispiele auf, an denen sich andere Repositorien orientieren können.

2 Zielstellung des Papers

Die Technische Informationsbibliothek (TIB) vergibt seit 2010 Digital Object Identifiers (DOIs) im Rahmen ihres DOI-Services für verschiedene Objekttypen wie z. B. Forschungsdaten, graue Literatur und wissenschaftliche Software.[1] Ein wesentlicher Bestandteil eines DOIs ist der damit verbundene verpflichtende Metadatensatz. Bei jeder Registrierung eines DOIs müssen Metadaten angegeben und über die Infrastruktur von DataCite zur freien Nachnutzung bereitgestellt werden. Dabei ist eine hohe Metadatenqualität, und damit auch eine hohe Vollständigkeit der Metadaten, erstrebenswert, um eine gute Auffindbarkeit und eine möglichst effektive Form der Nachnutzung erreichen zu können. Um die Metadatenqualität und insbesondere die Vollständigkeit der für die FAIR Principles relevanten Metadatenfelder der über die TIB registrierenden Einrichtungen bzw. Repositorien einschätzen zu können, haben wir im November 2020 eine Metadatenanalyse durchgeführt.[2]

Die zentrale Frage war: Erfüllen die Metadaten der teilnehmenden Einrichtungen des DOI-Services die Kriterien der FAIR Principles[3]?

[1] Seit dem 01.01.2021 bietet die TIB den DOI-Service im TIB DOI Konsortium an.

[2] Die Analyse wurde von Ted Habermann von Metadata Game Changers (https://metadatagamechangers.com/, letzter Zugriff 01.03.2021) durchgeführt.

[3] Wilkinson, M., et al. “The FAIR Guiding Principles for scientific data management and stewardship.”, Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18 (letzter Zugriff 26.05.2021).

In einer quantitativen Analyse wurden die vorhandenen Metadaten in Bezug auf Vollständigkeit und Konsistenz im Sinne einer gleichbleibenden Nutzung der Metadatenfelder betrachtet. Hierzu wurden 55 von den insgesamt 95 DataCite-Metadatenelementen von uns in FAIR-Kategorien eingeteilt, die sich auf verschiedene Aspekte der Metadaten-FAIRness beziehen, und als Grundlage für die Überprüfung von fast 28.000 Metadatensätzen aus 144 Repositorien herangezogen. Ziel dieser quantitativen Analyse war es herauszufinden, inwieweit die FAIR-Kriterien in der Praxis umgesetzt werden und welche Aspekte im Zuge der aktuellen Kuratierungspraxis optimiert werden müssen, um sämtliche Prinzipien der FAIR-Kriterien zu erfüllen.[4] Zudem liefern die in diesem Beitrag vorgestellten Ergebnisse ein Verständnis über die aktuellen Schwerpunkte, die die Einrichtungen bei der Vergabe der Metadaten setzen, und lassen Rückschlüsse darüber zu, welche Metadatenfelder in welchem Ausmaß im Arbeitsalltag der teilnehmenden Einrichtungen verwendet werden. Weiterhin diente unsere Analyse dazu, neue Ansätze als Basis für weiterführende Untersuchungen (z. B. bezüglich der Objekttypen) zu identifizieren bzw. einen Einblick in die Diversität der Daten zu erhalten. Diese Erkenntnisse können nicht nur zur Verbesserung der aktuellen Metadaten eingesetzt werden, sondern auch der Weiterentwicklung des bestehenden Schemas dienen.

3 Begriffsdefinitionen und Einordnung in den Kontext

Als Deutsche Zentrale Fachbibliothek für Technik sowie Architektur, Chemie, Informatik, Mathematik und Physik unterstützt die TIB die DOI-Registrierung in ihren Schwerpunktfächern seit über 15 Jahren. Im Jahr 2009 gründete sie mit sechs weiteren Mitgliedern den gemeinnützigen Verein DataCite.[5] Maßgebliche Motivation für die Gründung von DataCite und des DOI-Services an der TIB war es, die Zitierbarkeit und Nachnutzbarkeit von Forschungsdaten zu ermöglichen. Dies wurde u. a. erreicht, indem ein auf Forschungsdaten spezialisiertes Metadatenschema entwickelt und bei der DOI-Registrierung ein verpflichtendes Set an Metadaten angegeben werden musste, welches die Zitierfähigkeit der Ressource gewährleistete. Darüber hinaus gibt es weitere Metadatenfelder, welche Kontextinformationen bereitstellen. Mit allen Einrichtungen, die über die TIB DOIs registrierten, wurden Verträge abgeschlossen, die die Bereitstellung der Metadaten verbindlich machten. So wurde sichergestellt, dass Metadatensätze mit Hilfe eines DOIs zitiert und aufgefunden werden konnten. Die aktuelle Fassung des Metadatenschemas ist die Version 4.4, die durchgeführte Analyse erfolgte auf Basis der Version 4.3[6].

[4] Ein Teil der hier vorgestellten Ergebnisse wurde im Rahmen der PIDapalooza 2021 präsentiert: Taller, Nelli, Cordts, Anette, & Burger, Marleen. “Ein PID-Festessen für die Forschung.”, Januar 2021. http://doi.org/10.5281/zenodo.4478638, (letzter Zugriff 01.03.2021)

[5] DataCite. “DataCite’s Value.”, https://datacite.org/value.html, (letzter Zugriff 01.03.2021)

[6] DataCite Metadata Working Group. “DataCite Metadata Schema Documentation for the Publication and Citation of Research Data.”, Version 4.3. 2019. https://doi.org/10.14454/7xq3-zf69https://schema.datacite.org/, (letzter Zugriff 01.03.2021)

Darüber hinaus arbeitet DataCite mit anderen Organisationen und Einrichtungen zusammen, um die im Zuge der DOI-Registrierung gewonnenen Metadaten für Mehrwertdienste bereitzustellen. Die Qualität der verfügbaren Metadaten ist dabei maßgeblich für die Effektivität der Mehrwertdienste, welche die DOI-Metadaten nachnutzen. Beispielhaft kann hier der Service Event Data[7] genannt werden, welcher Verbindungen zwischen verschiedenen Veröffentlichungen aufzeigt.[8] Vor diesem Hintergrund ist eine möglichst umfassende Nutzung der Metadatenfelder in hoher Qualität erwünscht.

Was bedeutet das konkret? Der Qualitätsbegriff wird hier anwendungsbezogen verstanden und muss daher im Kontext des zu prüfenden Sachverhalts bewertet werden. Als Indikator für Metadatenqualität kann der aus der Wirtschaftswissenschaft stammende und inzwischen vielfach verwendete Term “fitness for use”[9] herangezogen werden. Das bedeutet im Hinblick auf die untersuchten Objekte, dass diese zunächst auffindbar und schließlich ausreichend beschrieben sein müssen, um eine sinnvolle Nachnutzung zu ermöglichen. Zur Beurteilung von Metadaten bezüglich ihrer Auffindbarkeit und Nachnutzbarkeit sowie weiterer damit verknüpfter Aspekte eignen sich die FAIR Principles.[10] Hierbei handelt es sich um Richtlinien, welche die datengestützte Wissenschaft befördern, indem sie die Auffindbarkeit (“Findability”), Zugänglichkeit (“Accessiblity”), Interoperabilität (“Interoperability”) sowie die Nachnutzbarkeit (“Reusability”) von Daten propagieren. Vor dem Hintergrund der wachsenden Datenmengen und verschiedener Quellsysteme unterstützen die Prinzipien maßgeblich die Auffindbarkeit und Nutzung für vielfältige Anwendungsszenarien. Die FAIR Principles wurden erstmals 2016 publiziert und adressieren Bedarfe der Menschen- und Maschinenlesbarkeit.

Um die Metadatenqualität wie in der hier vorliegenden Studie auf Ebene der Repositorien beurteilen zu können, muss zunächst für alle aus einem Repositorium stammenden, untersuchten Metadatensätze bewertet werden, inwieweit die FAIR Principles erfüllt werden.

Auf dieser Grundlage lassen sich dann

1. die Vollständigkeit der benötigten bzw. gewünschten Informationen hinsichtlich der FAIR Principles sowie
2. die Konsistenz bezüglich der Verwendung eines bestimmten Metadatenfeldes für die in einem Repositorium enthaltenen Metadatensätze ermitteln.

Die Vollständigkeit der Metadaten ist die wesentliche Basis für die Evaluierung ihrer Qualität und gleichzeitig ein relativ einfach und automatisiert zu erhebendes Kriteri-

[7] Crossref. “Event Data.”, https://www.crossref.org/services/event-data/, (letzter Zugriff 01.03.2021)
[8] DataCite. “Event Data.”, https://datacite.org/eventdata.html (letzter Zugriff 01.03.2021)
[9] Neely, M. Pamela. “The product approach to data quality and fitness for use: a framework for analysis.”, 2005. http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/TheProductApproach2DQFitness4Use.pdf (letzter Zugriff 01.03.2021)
[10] Wilkinson, et al. “The FAIR Guiding Principles for scientific data management and stewardship.”, Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18 (letzter Zugriff 26.05.2021)

um.[11] Erst ein Metadatensatz, der möglichst der möglichst umfassende Metadaten beinhaltet, bietet eine hohe Qualität, kann Mehrwerte schaffen und die effektive Nutzung durch weitere Services ermöglichen. Die konsistente Befüllung von Metadatenfeldern kann als Indikator für qualitätsgesicherte implementierte Workflows seitens der Repositorien verstanden werden. Denn die Erstellung vollständiger Metadatensätze ist ressourcenintensiv und erfordert den Einsatz fachkundigen Personals sowie bewährter Kuratierungsprozesse. Wenn die Metadatenfelder konsistent verwendet werden, dann liegen gut strukturierte Metadaten vor, welche eine automatische Prozessierung und Nachnutzung erleichtern. In dieser Form erfüllen sie in stärkerem Maße die im Rahmen der FAIR-Kriterien postulierte Maschinenlesbarkeit.

Das von den untersuchten Repositorien verwendete DataCite-Metadatenschema eignet sich vor allem zur primären Beschreibung des Objektes sowie seiner Verbindungen zu anderen Objekten, setzt also einen klaren Fokus auf die Auffindbarkeit. Es handelt sich um ein generisches Schema, d. h. es dient der Beschreibung unterschiedlicher Ressourcen- bzw. Objekttypen, die alle mit demselben Schema beschrieben werden, ohne dass eine Profilierung bezüglich der Art der Ressource erfolgt. Die auf Basis dieses generischen Metadatenschemas erfassten Metadaten wurden daher in dieser Analyse in ihrer Gesamtheit auf Ebene der untersuchten Repositorien betrachtet. Eine weitergehende Differenzierung bezüglich der Ressourcentypen wäre nicht möglich gewesen, da das Feld zur Beschreibung des Objekttyps erst seit 2016 verpflichtend istund entsprechend viele DOIs nicht oder lediglich manuell ausgewertet hätten können bzw. müssen.

Im Kontext der Bewertung von Metadatenqualität und Metadaten-FAIRness sollten immer auch Rahmenbedingungen und Arbeitsweise der Praxis berücksichtigt werden. Hier spielt das Konzept der minimalen Metadaten eine Rolle. Letztere sollen den Aufwand der Veröffentlichung eines Datensatzes möglichst gering halten. In der Regel werden die Elemente für minimale Metadatensätze daher so ausgewählt, dass die Auffindbarkeit (“Findability”) des Objektes sichergestellt ist. Um eine potentielle Nachnutzung von Datensätzen optimal zu ermöglichen, müssen jedoch wesentlich mehr Informationen hinterlegt werden. Andernfalls gehen wichtige Kontextinformationen wie zum Beispiel die Verknüpfung eines Datensatz zu einer Publikation verloren. Minimale Metadaten müssen immer abhängig vom jeweiligen Anwendungsfall zusammengestellt werden, um die ideale Balance zwischen möglichst geringem Aufwand und dem Mehrwert der Datenbeschreibung zu finden. Im generischen DataCite-Metadatenschema existieren sechs Pflichtfelder, die damit den minimalen Metadatensatz bilden: “Identifier”, “Creator”, “Title”, “Publisher”, “PublicationYear” sowie “ResourceType”. Hinzu kommt die “ResourceURL”, welche zusätzlich angegeben werden muss.

[11] Margaritopoulos, M., T. Margaritopoulos, I. Mavridis und A. Manitsaris (2012), Quantifying and measuring metadata completeness. J. Am. Soc. Inf. Sci., 63: 724-737. https://doi.org/10.1002/asi.21706 (letzter Zugriff 29.06.2021)

4 Methodik

Für die Auswahl der zu untersuchenden Datensätze wurde Mitte November 2020 eine Liste zusammengestellt, die alle in Deutschland ansässigen und zu diesem Zeitpunkt den DOI-Service der TIB nutzenden Einrichtungen enthielt. Diese Einrichtungen verfügten insgesamt über 144 Verwaltungseinheiten im DataCite-Registrierungssystem, sogenannte Repositorien. Alle Repositorien, die zu diesem Zeitpunkt exakt 300 oder weniger als 300 Metadatensätze bzw. vergebene DOIs verfügten, wurden vollständig erfasst, d. h. alle Metadatensätze wurden in die Analyse einbezogen. Für die anderen 72 Repositorien mit mehr als 300 Datensätzen wurden Zufallsstichproben von jeweils 300 Datensätzen ausgewählt. Dabei reicht die Spannweite der Repositorien von 1 bis zu 200.000 Metadatensätzen. Die gezogene Grenze bei 300 registrierten Datensätzen ist ein pragmatischer Ansatz, der das zu bewältigende Datenvolumen begrenzt, und gleichzeitig eine hohe Repräsentativität gewährleistet. Insgesamt wurden nahezu 28.000 Datensätze untersucht und über 50 % aller Repositorien vollständig erfasst. Bei den wissenschaftlichen Ressourcen, zu denen die Datensätze gehören, handelt es sich überwiegend um Textpublikationen. Weitere Ressourcentypen sind u. a. Forschungsdaten, Bildmaterialien, AV-Materialien und Datenkollektionen. Die Metadaten, die in dieser Analyse einbezogen worden sind, sind allesamt auf der Grundlage des DataCite-Metadatenschemas erstellt worden, welches in mehreren Versionen vorliegt (zum Zeitpunkt der Analyse bis Version 4.3). Es wurde bei der Erhebung nicht berücksichtigt, welche Version des Schemas jeweils verwendet wurde. Bei der Auswahl der Metadatensätze erfolgte keine Begrenzung auf einen bestimmten Zeitraum.

Im Rahmen unserer Metadatenanalyse wurden die vorliegenden Metadaten daran bemessen

1. inwieweit alle von uns als relevant eingestuften Informationen, die für die Umsetzung der FAIR Principles notwendig sind, bereitgestellt werden und die Metadaten somit vollständig sind und
2. inwieweit bei den Metadaten eines Repositoriums für alle enthaltenen wissenschaftlichen Ressourcen jeweils die gleichen Metadatenfelder ausgefüllt worden sind, die bereitgestellten Metadaten also konsistent sind.

Für die Bewertung der Vollständigkeit der TIB-DataCite-Repositoriums-Metadaten und der Relevanz einzelner Metadatenfelder wurden die vier FAIR-Anwendungsfälle (Auffindbarkeit, Zugänglichkeit, Interoperabilität und Nachnutzbarkeit) im Hinblick auf Empfehlungen für wesentliche und unterstützende Felder kategorisiert. Da es sich bei den FAIR Principles um Empfehlungen auf der konzeptuellen Ebene handelt, wurden die zu analysierenden Metadatenfelder zunächst auf Documentation Concepts[12] gemappt[13], welche anschließend auf Vorhandensein hin analysiert wurden.

[12] Habermann, Ted. „MetaDIG recommendations for FAIR DataCite metadata"(2019). https://doi.org/10.5438/2chg-b074 (letzter Zugriff 29.06.2021)
[13] Habermann, Ted. „DataCite Metadata Schema-FAIR Principles-Mapping"(2021).

Bei der Kategorisierung der Metadatenfelder wurde ein besonderes Augenmerk auf die Auffindbarkeit gelegt, da diese eine Voraussetzung für die Erfüllung der weiteren Kategorien der FAIR-Prinzipien, etwa die Nachnutzbarkeit, darstellt. Neben der Auswahl von Metadatenfeldern, die von substanzieller Bedeutung für eine gute Auffindbarkeit sind ("Findable Essential"), wurden ergänzende Metadatenfelder zur Kategorie "Findable Support" zusammengefasst. Die restlichen drei Bestandteile der FAIR Principles wurden jeweils zu "AIR Essential" und "AIR Support" zusammengefasst.

Die untersuchten Metadatensätze wurden zunächst in der Gesamtschau, über alle erfassten Repositorien hinweg, hinsichtlich des Erfüllungsgrades der FAIR Principles ausgewertet, um Aussagen über die Metadatenqualität im Durchschnitt aller Repositorien treffen zu können. Darüber hinaus erfolgten Detailauswertungen für diejenigen Repositorien, die bei bestimmten FAIR-Kategorien besonders umfassende und konsistente Metadaten aufweisen, um Best-Practice-Beispiele aufzeigen zu können. Inwieweit eine konsistente Verwendung einzelner Metadatenfelder vorliegt, lässt sich ermitteln, indem der Durchschnittswert der Häufigkeit bei der Befüllung des entsprechenden Metadatenfeldes betrachtet wird. Hierbei wurde für die einzelnen Repositorien untersucht, wie oft ein entsprechendes Metadatenfeld von einem Repositorium verwendet wurde, nicht jedoch wie homogen die Inhalte angegeben werden.

5 Ergebnisse

In der Gesamtschau[14] aller untersuchten Repositorien zeigt sich, dass durchschnittlich 51 % der Metadatenfelder aus der Kategorie "Findable Essential" und 18 % der Metadatenfelder aus der Kategorie "Findable Support" ausgefüllt worden sind. In der Kategorie "Findable Essential" wird in der Regel mindestens eine Vollständigkeit von 40 % erreicht, da in dieser Kategorie sechs von insgesamt 15 Metadatenfeldern Pflichtfelder im DataCite-Metadatenschema sind. Neben den Pflichtfeldern werden in über 50 % der Fälle zusätzlich „Abstract" und/oder "Keywords" angegeben. Dies sind damit auch diejenigen optionalen Metadatenfelder in dieser Kategorie, welche am konsistentesten genutzt werden.

In der Kategorie "Findable Support" sind keine Pflichtfelder enthalten, was sich in den deutlich geringeren Vollständigkeitswerten von im Mittel 18 % widerspiegelt. Bei einer Betrachtung nach der Größe der Repositorien (siehe Abb. 1) ist kein eindeutiger Zusammenhang zwischen Repositoriumsgröße und Vollständigkeit der Metadaten aus den Findable-Kategorien erkennbar, wobei diejenigen Repositorien, die in der Kategorie „Findable Essential" die höchsten Vollständigkeitswerte („Score" von über 70 %) aufweisen, über weniger als 180 Metadatensätze bzw. registrierte DOIs verfügen.

In der Kategorie "AIR Essential" beträgt die Vollständigkeit im Mittel aller untersuchten Repositorien lediglich 14 %. Im Detail bedeutet dies, dass von den 16 dieser Ka-

[14] Habermann, Ted. „Completeness Analysis DataCite Metadata"(2021).

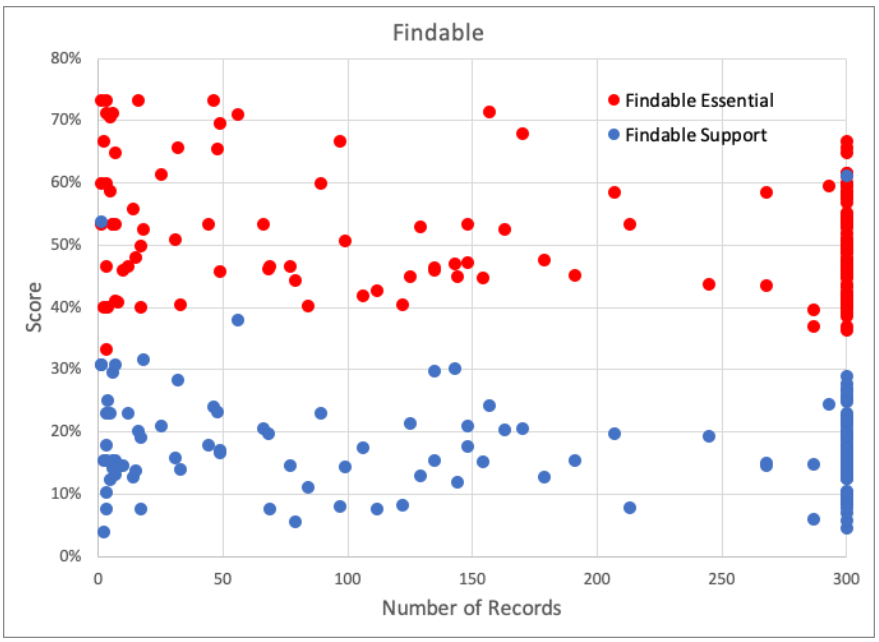

Abbildung 1: Prozentualer Erfüllungsgrad (Score) der FAIR-Kategorien Findable Essential und Findable Support nach Größe der Repositorien (Number of Records). Quelle: Eigene Darstellung.

tegorie zugeordneten Metadatenfeldern stets mindestens ein Feld, nämlich die URL der Landingpage („Resource URL“) als Pflichteingabe, ausgefüllt ist und zusätzlich in einigen Fällen Angaben zu weiteren Metadatenfeldern wie “Resource Size”, “Resource Format” und “Rights” gemacht worden sind. Informationen zu sogenannten “Related Identifiers”, die auf andere persistente Identifikatoren und die zugehörigen wissenschaftlichen Ressourcen (etwa einem Forschungsdatensatz zugehörige Publikationen und umgekehrt, mittels “CitedBy”, “ReferencedBy”,”SupplementTo” etc.) verweisen, sind nur vereinzelt vertreten. Die Vollständigkeit in der Kategorie “AIR Support” erreicht mit nur 4 % den geringsten Durchschnittswert, allerdings handelt es sich bei den hierunter zusammengefassten Metadatenfeldern lediglich um ergänzende Zusatzinformationen; z. B. Informationen zu weiteren Kontaktpersonen neben Autor:in und “Publisher” oder URL der Nutzungslizenzen. Wie in Abbildung 2 erkennbar, werden die höchsten Werte bei „AIR Support“ von sehr kleinen Repositorien mit weniger als 30 DOIs erreicht.

Unsere Analyse verdeutlicht den starken Fokus auf die Metadatenpflichtfelder, was die grundsätzliche Auffindbarkeit der in den Repositorien befindlichen wissenschaftlichen Ressourcen sicherstellt. Die AIR-Kategorien sind momentan noch schwach vertreten, wodurch wichtiges Potenzial für eine spätere Nachnutzung nicht ausgeschöpft wird. Alles in allem zeigen unsere Ergebnisse, dass das Konzept der Angabe lediglich minimaler Metadaten in der Praxis, zumindest in den hier untersuchten Repositorien, ausgesprochen populär ist.

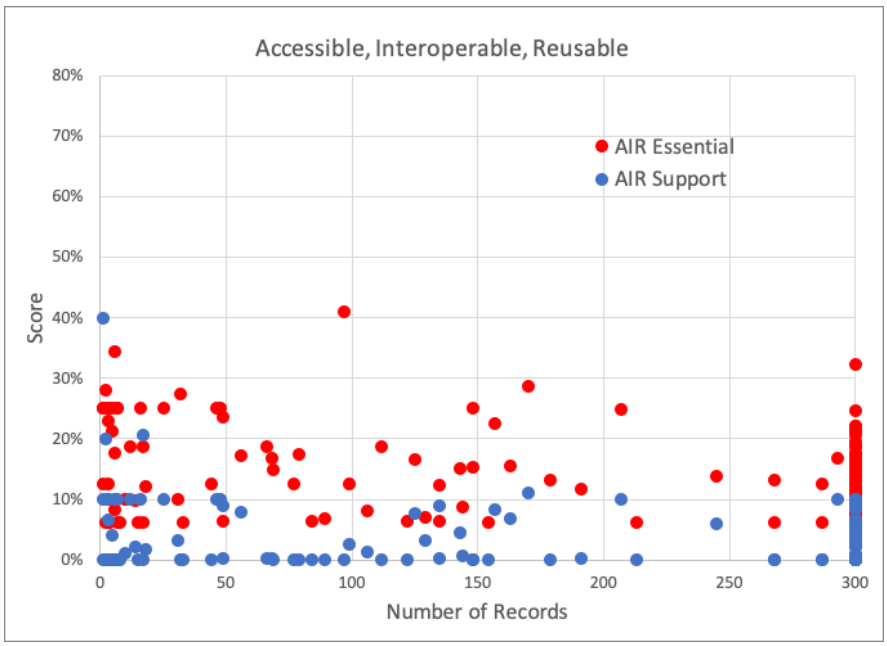

Abbildung 2: Prozentualer Erfüllungsgrad (Score) der FAIR-Kategorien AIR Essential und AIR Support nach Größe der Repositorien (Number of Records). Quelle: Eigene Darstellung.

Auch wenn die Vollständigkeit der für die FAIR Principles relevanten ausgefüllten Metadatenfelder bei Betrachtung der Gesamtstichprobe nur geringe Prozentwerte erreicht, zeigt eine Detailanalyse einzelner Repositorien doch interessante Best-Practice-Beispiele mit Blick auf verschiedene FAIR-Kategorien auf.

Best-Practice-Beispiel 1: Hier handelt es sich um ein großes, interdisziplinär ausgerichtetes Repositorium einer deutschen Universität mit 25.484 wissenschaftlichen Ressourcen, darunter größtenteils Forschungsdaten, das bei “Findable Essential” mit einer Vollständigkeit von 61 % deutlich über dem Durchschnitt liegt und bei “Findable Support” mit 65 % den höchsten Vollständigkeitswert in der Gesamtstichprobe erreicht. In der Kategorie “Findable Essential” ist neben den Pflichtfeldern bei allen Ressourcen auch das Feld “Abstract” ausgefüllt und bei über 80 % der Ressourcen sind zusätzlich Informationen zur Projektförderung und „Keywords“ enthalten. In der Kategorie “Findable Support” werden im Wesentlichen vertiefende Informationen zu den Autor:innen (z. B. ORCID) gegeben und weiterführende Angaben zur Projektförderung (etwa Titel des Forschungsprojektes und “Funder Identifier”) gemacht. Dabei liegt jeweils ein hoher Konsistenzwert von knapp 80 % und mehr vor. In den Kategorien “AIR Essential” und “AIR Support” werden mit Vollständigkeitswerten von 13 % bzw. 1 % jedoch unterdurchschnittliche Werte erreicht.

Best-Practice-Beispiel 2: Dieses Repositorium enthält 97 wissenschaftliche Ressourcen, darunter überwiegend Forschungsdaten aus dem Gebiet der Meteorologie, und zeichnet sich durch gut dokumentierte und strukturierte Metadaten aus. In der Kate-

gorie “Findable Essential” besteht mit 67 % eine hohe Vollständigkeit und zugleich eine sehr hohe Konsistenz, da alle Ressourcen hier durchgängig über Angaben zu den gleichen Metadatenfeldern verfügen. Passend zur fachlichen Ausrichtung des Repositoriums werden jeweils Informationen zur geographischen Lage der wissenschaftlichen Ressourcen gegeben. Besonders hervorzuheben sind die mit einem Vollständigkeitswert von 41 % vergleichsweise umfassenden Angaben in der Kategorie “AIR Essential”, bei der neben Informationen zu Verwertungsrechten, Dateiformaten und -größe bei allen Ressourcen Angaben zu Methoden und “Related Identifiers” gemacht werden, sodass auch in dieser Kategorie eine hohe Konsistenz erreicht wird. Dagegen hat die Kategorie “Findable Support” nur einen Vollständigkeitswert von 8 % und in der Kategorie “AIR Support” sind überhaupt keine ausgefüllten Metadatenfelder vorhanden.

Beide Best-Practice-Beispiele verdeutlichen somit, dass je nach Schwerpunktsetzung des betrachteten Repositoriums die jeweiligen FAIR-Kategorien unterschiedlich stark repräsentiert sind und die Interpretation der vorhandenen Metadaten eines Repositoriums vor dem Hintergrund seiner Ressourcentypen (z. B. Forschungsdaten, Textpublikationen, Videos etc.), Inhalte und Fachrichtung (interdisziplinär oder fachspezifisch) erfolgen muss.

6 Hypothesen und Schlussfolgerungen zur Vollständigkeit und Konsistenz von Metadaten

Aus den vorgestellten Ergebnissen und Best-Practice-Beispielen lassen sich einige Hypothesen und Schlussfolgerungen bezüglich der Vollständigkeit und Konsistenz von Metadaten ableiten, die in weiterführenden Studien überprüft werden könnten.

1. Eine hohe Konsistenz ist vermutlich im Sinne der oben gegebenen Definition einer gleichförmigen Berücksichtigung der Metadatenfelder bei spezialisierten Repositorien, deren Inhalte ebenfalls homogen vorliegen (etwa nur Forschungsdatensätze aus einem Fachgebiet), leichter zu erreichen (siehe Best-Practice-Beispiel 2).
2. Heterogene Repositorien mit unterschiedlichen Ressourcentypen und Fachgebieten können zwar insgesamt reichhaltige und umfassende Metadaten enthalten, allerdings wird hier eine hohe Konsistenz schwerer zu erreichen sein. Dies liegt darin begründet, dass für unterschiedliche Ressourcentypen und Fachgebiete neben allgemein zutreffenden und grundsätzlich ressourcenunabhängig in gleichem Maße relevanten Metadaten (u.a. die Pflichtfelder) auch verschiedene inhaltsbezogene Metadatenfelder von mehr oder weniger großer Wichtigkeit sein können. So ist beispielsweise die Angabe der geographischen Lage bei Metadaten mit Bezug zu archäologischen Fundstücken von deutlich höherer Relevanz als bei vielen anderen Forschungsthemen. Ein weiteres Beispiel sind Informationen zur Projektförderung (siehe Best-Practice-Beispiel 1), die nur dann

angegeben werden können, wenn die betreffenden Inhalte im Rahmen eines Forschungsprojektes entstanden sind. Eine hohe Konsistenz ist dementsprechend nur für hinreichend ähnlich geartete Inhalte umsetzbar, sofern das Repositorium keine Werte zur Auszeichnung bewusst leer gelassenen Felder verwendet. Im Falle hinreichend homogener Inhalte ist eine hohe Konsistenz aber im Sinne der FAIR Principles (etwa der Auffindbarkeit durch Ermöglichung einer gezielten Suche nach bestimmten Metadatenfeldern über alle Ressourcen eines Repositoriums hinweg) empfehlenswert und vereinfacht darüber hinaus quantitative Analysen von Metadaten. Eine besondere Bedeutung hat Konsistenz auch im Kontext von automatisierter Datenverarbeitung, weil sie maßgeblich dafür ist, inwieweit die Datenquelle verlässliche und brauchbare Informationen liefert, was wiederum die Qualität von darauf aufbauenden Services bestimmt.

3. Vollständige und konsistente Metadaten sind einfacher zu erreichen, wenn entsprechende Workflows, die die Charakteristika der mit Metadaten zu beschreibenden Inhalte berücksichtigen, vorhanden sind.
4. Professionell betriebene Repositorien, für deren Pflege gut qualifiziertes Personal mit ausreichendem Zeitbudget eingesetzt wird, haben bessere Möglichkeiten, entsprechende Workflows umzusetzen und damit eine hohe Metadatenqualität zu erreichen. Größere Einrichtungen (bspw. Universitäten) werden dabei häufiger über finanzielle Mittel verfügen, um ein Repositorium professionell betreiben zu können.

Es ist anzunehmen, dass der nach wie vor verbreitete Einsatz von lediglich minimalen Metadaten auf beschränkte Ressourcen für die Kuratierung von Metadaten zurückzuführen ist. Andererseits ist unbekannt, inwieweit standardisierte Workflows von den Einrichtungen erarbeitet wurden, die auch bei der Bearbeitung durch mehrere Personen oder im Falle von Personalfluktuation eine konsistente Verwendung der Metadatenfelder gewährleisten. Um diese Annahmen überprüfen zu können, wäre die Durchführung von darauf aufbauenden Befragungen bei verschiedenen Repositoriumstypen sinnvoll.

7 Fazit

Insgesamt zeigt unsere Analyse, dass die Mehrheit der Einrichtungen, die über den TIB-DOI-Service DOIs registrieren, die FAIR Principles für ihre Metadaten und die damit beschriebenen wissenschaftlichen Ressourcen nicht optimal umsetzt. Durch die geforderten Metadaten-Pflichtfelder, die notwendig sind, um über DataCite einen DOI zu registrieren, werden aber zumindest grundlegende Aspekte für die Auffindbarkeit in allen Repositorien berücksichtigt. Auch hier ist die Angabe weiterführender Metadaten, etwa die Einspeisung eines Abstracts oder von Informationen zum Fachgebiet, jedoch sinnvoll, um die Auffindbarkeit und damit verbunden die Nachnutzung zu fördern. Aufgefallen ist, dass bisher nur äußerst selten weitere persistente Identifikatoren angege-

ben werden, etwa die ORCID der Autor:innen oder die ROR ID[15] als Identifikator für die zugehörige Forschungseinrichtung. Die Verwendung von persistenten Identifikatoren ist im Sinne der Metadaten-FAIRness besonders relevant, um eindeutige Beziehungen zwischen Ressourcen, Personen und Organisationen herstellen zu können. Ähnlich verhält es sich mit der zur Zeit wenig genutzten Möglichkeit, "Related Identifiers" anzugeben, die eine wissenschaftliche Ressource über ihre Metadaten mit anderen zu ihr in Bezug stehenden Forschungsarbeiten, Datensätzen o. ä. verknüpfen können. Die Angabe von miteinander in Beziehung stehenden Ressourcen ermöglicht eine bessere Vernetzung von Forschungsergebnissen mit den zugehörigen Forschungsdaten, Personen, Einrichtungen sowie weiteren zitierten oder zitierenden Arbeiten und ist damit für Services wie Event Data und den PID-Graphen[16] von besonderer Bedeutung. Dies führt zu einer erhöhten Sichtbarkeit und Auffindbarkeit des Forschungs-Outputs und unterstützt daher die Publizierenden direkt, indem der Impact ihrer wissenschaftlichen Arbeiten besser quantifizierbar wird.

Alles in allem zeigt unsere Untersuchung auf, dass insbesondere in Bezug auf eine gute Zugänglichkeit, Interoperabilität und Nachnutzbarkeit in der Breite der Repositorien noch viel Optimierungspotenzial besteht. Die Best-Practice-Beispiele verdeutlichen jedoch auch, dass einzelne Einrichtungen die FAIR Principles bereits gut umsetzen. Diese Beispiele stellen Leuchttürme dar, welche anderen Repositorien mit ähnlicher Ausrichtung eine Orientierung bzgl. der Metadatenkuratierung geben können.

Abschließend lässt sich festhalten, dass die detaillierte Metadatenanalyse u. a. für Repositorienbetreiber:innen wertvolle Informationen über den Status-quo der Vollständigkeit der Metadaten liefert. Es zeigt sich, dass gegenwärtig das Potenzial veröffentlichter Metadatensätze nicht ausgeschöpft wird. Dies trifft folglich auch auf Mehrwertdienste zu, die nur so gut sein können wie die ihnen zugrundeliegende Datenbasis. Auch aus diesem Grund haben die bereitstellenden Repositorien ein Eigeninteresse daran, die Vollständigkeit ihrer Metadaten (insbesondere solcher aus dem Bereich der Findability) zu erhöhen, um die Sichtbarkeit des Forschungsoutputs zu verbessern. Für die TIB als Konsortialführerin leitet sich das Handlungsfeld ab, die Hindernisse im Kuratierungsalltag zu identifizieren und im nächsten Schritt Handlungsempfehlungen für die Implementierung von bedarfsgerechten Workflows praxisnah zu formulieren. Weiterführende Untersuchungen verschiedener Repositoriumstypen und differenzierte Betrachtungen verschiedener Ressourcentypen (z. B. Forschungsdaten, Textpublikationen oder Videos) können darüber hinaus Ansätze für eine Verbesserung der Metadatenqualität und für die Entwicklung passgenauer Services zur Unterstützung der Einrichtungen ans Licht bringen.

[15] ROR. "About.", https://ror.org/, (letzter Zugriff 01.03.2021)
[16] Fenner, M. und A. Aryani "Introducing the PID Graph (Version 1.0).", 2019. https://doi.org/10.5438/JWVF-8A66, letzter Zugriff (01.03.2021)