Persistent Identifier, ihre Metadaten und FAIRness

Eine Analyse zur verbesserten Auffindbarkeit von Forschungsressourcen

Stephanie Hagemann-Wilholt, Sara El-Gebali, Britta Dreyer, Paul Vierkant

Abstract

Persistent Identifiers (PIDs) sind essentiell für die Sichtbarkeit wissenschaftlicher Ressourcen. In dem vorliegenden Artikel wird anhand einer Analyse der DOI-Metadaten von deutschen DataCite-Mitgliedern die aktuelle Nutzung standardisierter Metadaten beleuchtet. Metadaten dienen einer verbesserten Auffindbarkeit (Findability) wissenschaftlicher Ressourcen und sind ein elementarer Bestandteil der FAIR-Prinzipien. Die Ergebnisse zeigen Fortschritte, aber auch Optimierungspotenziale, insbesondere bei der Standardisierung und Vollständigkeit der Metadaten. Der Artikel hebt die Relevanz von PIDs und Metadatenstandards für eine langfristige und globale Sichtbarkeit und Nachnutzbarkeit von Forschung hervor.

Persistent identifiers (PIDs) are essential for the visibility of research outputs and resources. This article analyzes the DOI metadata of German DataCite Members to shed light on the current use of standardized metadata. Metadata improves the findability of scientific resources and is an elementary component of the FAIR principles. The results show progress, and potential for optimization, especially in the standardization and completeness of metadata. The article emphasizes the relevance of PIDs and metadata standards for the research's long-term and global visibility and reusability.

Einleitung

Durch wissenschaftliche Forschung werden in kontinuierlich wachsendem Umfang neue Informationen erzeugt und publiziert. Sollen die gewonnenen Erkenntnisse für weiterführende Forschung oder zur Verwirklichung gesellschaftlicher Ziele jetzt oder in Zukunft nutzbar gemacht werden, müssen relevante Wissensbestände in dieser Menge neuer Informationen für jeweils spezifische Fragen letztlich auffindbar sein.[1] Es ist die Aufgabe wissenschaftlicher Informationsinfrastrukturen, diese Bestände (z. B. Textpublikationen, Forschungsdaten, Software) für Forschung, Lehre und weitere Akteur:innen des Wissenschaftssystems und darüber hinaus vorzuhalten. Metadaten sind Daten über Daten, die den Kontext über die Struktur, die Merkmale und den Zweck der Daten selbst liefern. Es handelt sich nicht um die eigentlichen Daten, sondern um die Informationen, die helfen, sie zu beschreiben oder zu kategorisieren. Mit ihrer Hilfe können sie verlässlich und auffindbar für die Nachnutzung zugänglich gemacht werden. Die Autor:innen der FAIR-Prinzipien beschrieben 2016, wie die Auffindbarkeit (Findability), Zugänglichkeit (Accessibility), Interoperabilität (Interoperability) und Nachnutzbarkeit (Re-usability) von Forschungsdaten erhöht werden könnten, um ein effektives und effizientes Forschungsdatenmanagement zu ermöglichen und damit die Nutzung wie Reproduzierbarkeit wissenschaftlicher Erkenntnisse zu unterstützen.[2] Die Anwendung dieser Prinzipien wird seitdem auch auf weitere Forschungsressourcen und -informationen (z. B. FAIR for Software) ausgeweitet und bietet eine grundsätzliche Orientierung für wissenschaftliche Informationsinfrastrukturen.[3]

Für die Realisierung der FAIR-Prinzipien spielen Persistente Identifikatoren (PIDs) eine zentrale Rolle. PIDs dienen der eindeutigen und langfristigen Identifizierbarkeit von Forschungsressourcen und -akteur:innen. Es existiert eine Vielzahl von PIDs proprietärer und nicht-proprietärer Anbieter:innen, die verschiedene Ressourcen und Akteur:innen referenzieren können. Offene, nicht-proprietäre PID-Systeme unterstützen insbesondere die Prinzipien der Interoperabilität und Zugänglichkeit.[4] PIDs erfassen eine Vielzahl von Informationen über Ressourcen, einschließlich Details über Mitwirkende, ihren Entstehungsprozess und umfangreiche Metadaten. Diese Informationen werden in einem standardisierten Format gespeichert, das die Interoperabilität zwischen verschiedenen Systemen gewährleistet, die Auffindbarkeit verbessert und die Qualitätssicherung unterstützt. Im Rahmen von Forschungsprojekten wird gezielt an der Umsetzung der FAIR-Kriterien in Informationsinfrastrukturen deutscher Hochschulen und Forschungseinrichtungen gearbeitet. Die Verbesserung des Forschungsdatenmanagements innerhalb der deutschen Forschungslandschaft ist eines der grundlegenden Ziele der Nationalen Forschungsdateninfrastruktur (NFDI).[5] PID4NFDI ist ein Projekt innerhalb der

1 Marginson 2017, S. 129-131.

2 Wilkinson et al. 2016.

3 Z. B. Barcelona Declaration on Open Research Information 2024 zu FAIRen Forschungsinformationen und -systemen, Barker et al. 2022 zu FAIRer Forschungssoftware, Burger et al. 2021 zur Anwendung von FAIR auf verschiedene Forschungsressourcen, Hagemann-Wilholt et al. 2020 zu FAIRen Konferenzmetadaten.

4 https://openscholarlyinfrastructure.org [28. August 2024].

5 https://www.nfdi.de/ [28. August 2024].

NFDI-Förderung zur Entwicklung eines Basisdienstes für die Fachkonsortien in der NFDI, der die Nutzung von PIDs in unterschiedlichen Anwendungsfällen ermöglichen, erleichtern oder weiter entwickeln und FAIRes Forschungsdatenmanagement befördern soll. Neben der Verbesserung der Interoperabilität zwischen NFDI-Infrastrukturdiensten und PID-Providern sowie der Harmonisierung von Metadaten zur Registrierung von PIDs werden auch Kommunikations-, Trainings- und Steuerungsmaßnahmen in den Blick genommen, um die Anwendung von PIDs quantitativ und qualitativ zu erweitern.[6] Denn die Nutzbarkeit und Wirksamkeit von PIDs hängt ebenso sehr von technischen Voraussetzungen ab wie von der Akzeptanz gemeinsamer (Metadaten-)Standards innerhalb der Anwender:innen-Communities sowie der Regelung von Verantwortlichkeiten, um Persistenz tatsächlich zu gewährleisten. Auf diese Faktoren zielen auch die Aktivitäten des seit 2023 von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts „PID Network Deutschland. Die Etablierung eines Netzwerks von Akteur:innen aus Wissenschaft und Kultur“[7], um die Nutzung von PIDs weiter voranzutreiben und eine nationale PID-Roadmap zu entwickeln. An beiden Projekten sind DataCite und die Technische Informationsbibliothek (TIB) beteiligt und in beiden Projekten werden gerade komplementäre Landschaftsanalysen vorgenommen, um den Status quo der PID-Nutzung und ihrer Herausforderungen in den NFDI-Konsortien respektive an den Infrastruktureinrichtungen von Forschungs- und Kultureinrichtungen zu erfassen und Verbesserungspotentiale zu identifizieren.

Ergänzend zu den laufenden Aktivitäten in PID4NFDI und PID Network wird in der vorliegenden Analyse der Fokus auf die Auffindbarkeit wissenschaftlicher Ressourcen durch Digital Object Identifiers (DOIs) gelegt. Ziel ist es, zu zeigen, wie die Möglichkeiten zur standardisierten Erschließung und Verbreitung von Forschungsressourcen sowie ihrer Metadaten über die Registrierung von DOIs aktuell genutzt werden und wo Optimierungspotentiale liegen, um die Auffindbarkeit dieser Ressourcen und damit ihren langfristigen Impact noch zu verbessern.

Methode

PIDs schaffen wichtige Voraussetzungen für die grundsätzliche Auffindbarkeit (Findability) von Forschungsressourcen. Um zu zeigen, inwieweit deutsche Hochschul- und Forschungseinrichtungen bei der Bereitstellung von Forschungsressourcen bereits die Anforderungen an das erste der FAIR-Prinzipien erfüllen, liegt der inhaltliche Fokus der vorliegenden Analyse auf der detaillierten Betrachtung der vier Findable-Prinzipien:[8]

- F1. Forschungsressourcen und -akteur:innen erhalten einen weltweit einzigartigen und dauerhaft gültigen Identifier.
- F2. Forschungsressourcen und -akteur:innen werden mit umfangreichen Metadaten beschrieben.
- F3. Metadaten beinhalten die Identifier der Forschungsressourcen und -akteur:innen.
- F4. Metadaten werden in einer durchsuchbaren Datenbank registriert oder indiziert.

Ziel der vorliegenden Erhebung ist es, die Landschaft der DOI-Registrierung von DataCite-Mitgliedern in Deutschland abzubilden. Bei den untersuchten Mitgliedern handelt es sich um Non-Profit-Organisationen mit Hauptsitz in Deutschland: Dies umfasst im Wesentlichen Hochschulen, außeruniversitäre Forschungseinrichtungen, Ressortforschungseinrichtungen und weitere Organisationen aus dem wissenschaftlichen Kontext. Insgesamt wurden 241 Konsortialmitglieder innerhalb von vier Konsortien sowie 16 direkte Mitglieder, die zusammen 691 Repositorien repräsentieren, erfasst. Mit Stand vom 8. August 2024 haben diese Einrichtungen insgesamt 5.244.999 auffindbare DOIs registriert, deren Metadaten analysiert wurden. Dieses Datenkorpus bietet in Größe und Reichweite einen umfassenden Überblick über das deutsche wissenschaftliche Ökosystem.[9]

Kern der inhaltlichen Untersuchung ist eine quantitative Analyse der Vollständigkeit der in DataCite hinterlegten Metadaten anhand des DataCite-Metadatenschemas.[10] Das DataCite-Metadatenschema bietet derzeit 20 Metadaten-Properties zur Beschreibung von Forschungsoutputs an: Sechs verpflichtende zur Abdeckung notwendiger bibliographischer Angaben wie z. B. Titel, Urheber:in oder Veröffentlichungsjahr; sechs empfohlene (z. B. Mitwirkende, Beschreibungen) und acht optionale (z. B. Referenzen auf die zugrunde liegende Förderung durch Dritte). Es ermöglicht Anpassungen und Erweiterungen, um spezifische Metadatenanforderungen der Nutzer:innen-Community zu erfüllen. Die Datenanalyse erfolgte mittels der browserbasierten und offenen Analyseplattform OpenSearch, die auf der Suchmaschine Elasticsearch aufbaut.

6 http://pid.services.base4nfdi.de/ [28. August 2024].

7 Bertelmann et al. 2023.

8 Zwischen den auf der Webseite von FORCE11 veröffentlichten FAIR Data Principles und den in “FAIR Guiding Principles for scientific data management and stewardship” von Wilkinson et al. aufgeführten Kriterien gibt es eine Inkongruenz hinsichtlich der Reihenfolge der Kriterien F3 und F4, siehe https://force11.org/info/the-fair-data-principles/ [28. August 2024] und Wilkinson et al. 2016. Die vorliegende Analyse verwendet die Reihenfolge der Findable-Kriterien aus der Publikation von Wilkinson et al. 2016.

9 Im Rahmen dieser Erhebung wurden keine deutschen Beiträge bei Zenodo oder anderen nicht in Deutschland ansässigen DataCite-Mitgliedern berücksichtigt.

10 https://schema.datacite.org [28. August 2024].

- Jahresabonnement PRINT (10 Ausgaben) ab 99,00 Euro
- Jahresabonnement Elektronische Ausgabe (10 Ausgaben) ab 74,00 Euro
- Jahresabonnement Print und Elektronische Ausgabe ab 121,00 Euro

(Abonnementpreise jeweils inkl. MwSt. zzgl. Versandkosten)

www.libess.de

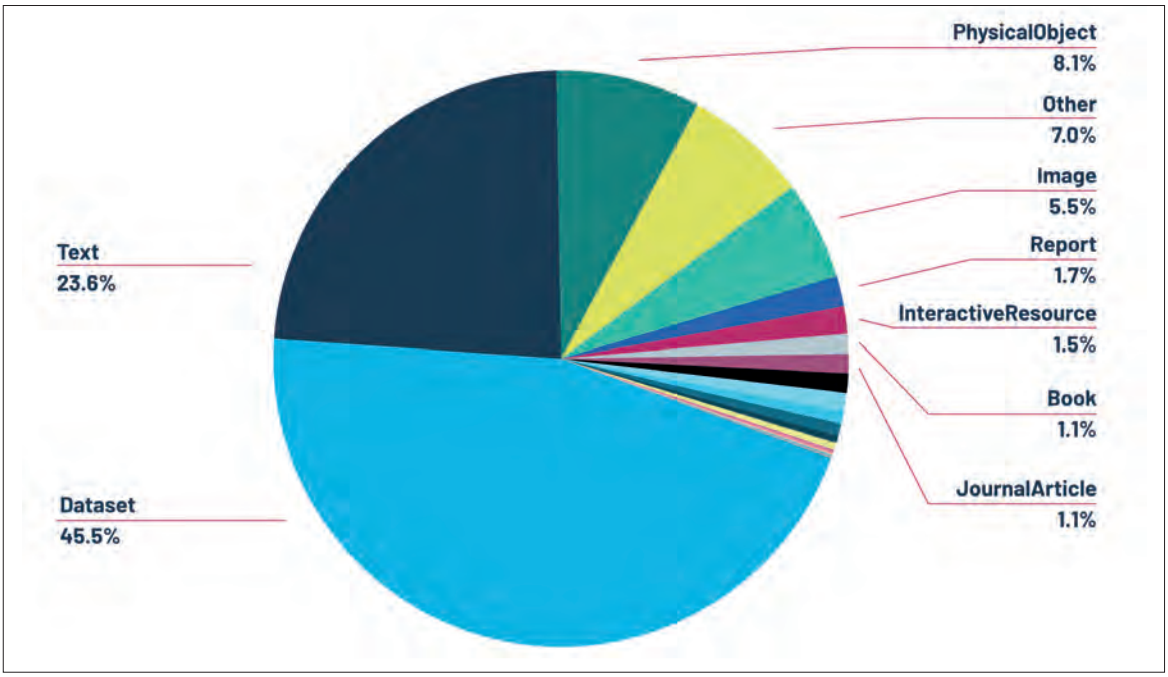

Abbildung 1: Relative Verteilung der ResourceTypes auf die Anzahl registrierter DataCite-DOIs in Deutschland (Stand: 2024-09-06)

Analyse und Diskussion

F1. Forschungsressourcen und -akteur:innen erhalten einen weltweit einzigartigen und dauerhaft gültigen Identifier

PIDs machen es möglich, Forschungsressourcen und ihre Metadaten leicht zu finden und korrekt zu zitieren. Das Prinzip F1 ist besonders wichtig, weil ohne diese dauerhaften Kennungen durch Identifier die anderen FAIR-Prinzipien nur schwer umsetzbar sind. Da auch die Publikation von Forschungsressourcen jenseits von Textpublikationen und Forschungsdaten immer wichtiger und diverser geworden ist, um deren Entstehungsprozess transparenter abbilden zu können und eine sinnvolle Nachnutzung zu befördern, hat sich die Beschreibung von Forschungsressourcen über PIDs weiter ausdifferenziert (siehe Abb. 1). Nachfolgend werden einige der verbreitetsten nicht-proprietären PID-Systeme skizziert.

Der am weitesten verbreitete PID im wissenschaftlichen Bereich ist der Digital Object Identifier (DOI) zur eindeutigen Referenzierung wissenschaftlicher Ressourcen. DOIs erlauben eine langfristige Identifikation der jeweiligen Objekte und werden seit Ende der 1990er Jahre eingesetzt, um insbesondere digitale Ressourcen nachhaltig verfügbar zu machen. DataCite gehört zu den weltweit führenden Registrierungsagenturen für DOIs und bildet einen Großteil des Korpus von Forschungsressourcen ab, die über DOIs verfügbar gemacht werden.

Basierend auf der Zunahme an in DataCite Fabrica registrierten DOIs schneiden wissenschaftliche Einrichtungen in Deutschland im internationalen Vergleich insgesamt sehr gut bei der Registrierung von DOIs ab und erfüllen damit die Anforderung F1. Seit 2009 wurden in Deutschland 7,9% der weltweit über DataCite registrierten DOIs für 30 verschiedene Ressourcetypen wie Textpublikationen, Forschungsdaten, audiovisuelle Medien, Software, Instrumente etc. von den 257 direkten und Konsortialmitgliedern erfasst. Seit Jahren ist ein kontinuierlicher Anstieg der DOI-Registrierung in Deutschland zu beobachten. 2018 waren es insgesamt 1,62 Millionen DOIs, die über DataCite registriert wurden; 2020 bereits 2,77 Millionen DOIs und 2024 bis zum Stichtag der vorliegenden Analyse 5,24 Millionen DOIs. Deutschland liegt damit in den Aktivitäten der DOI-Registrierung derzeit auf dem zweiten Platz hinter den führenden USA, wo 384 wissenschaftliche Einrichtungen bisher 17,8 Millionen DOIs registriert haben.

Ein weiterer PID ist der Uniform Resource Name (URN), der über nationale Knotenpunkte wie zum Beispiel die Deutsche Nationalbibliothek (DNB) Netzpublikationen zugänglich macht. Insgesamt wurden bis Ende 2023 56,8 Millionen URNs über die DNB vergeben und 14,9 Millionen Netzpublikationen erschlossen.[11]

Die Open Researcher and Contributor iD (ORCID iD) identifiziert und verknüpft Autor:innen und andere Mitwirkende im Forschungsprozess, wie z. B. Datenkurator:innen, Datenanalyst:innen etc., eindeutig und dauerhaft mit Forschungsressourcen. Derzeit gehören 91 Forschungseinrichtungen dem ORCID Deutschland Konsortium an. Es ist somit weltweit das zweitgrößte ORCID-Konsortium, nach der ORCID US-Community mit 207 Mitgliedern. Insgesamt wurden weltweit 8,65 Millionen ORCID iDs registriert. In Deutschland werden aktuell 224.198 ORCID-Records ak-

11 DNB 2024, S. 4 und 10.

tiv verwaltet und stehen damit an sechster Stelle hinter Großbritannien mit 363.505 ORCID iDs.[12]

Organisationen haben dynamische Strukturen, die sich ständig verändern: Sie ändern ihre Namen, fusionieren, spalten sich, werden aufgelöst oder neu gegründet. Diese Veränderungen erschweren es, Forschungsorganisationen eindeutig mit Forschungsressourcen und Forschenden zu verknüpfen. Die Research Organization Registry (ROR), ein globales Register, das von der Forschungscommunity betrieben wird und offene, dauerhafte Identifikatoren für Forschungseinrichtungen bereitstellt, schließt diese Lücke.[13] Von insgesamt 108.557 Organisationen weltweit haben bisher 1.920 Organisationen in Deutschland aktive ROR-Einträge.

Der Research Activity Identifier (RAiD) für Forschungsaktivitäten und -projekte ist ein weiterer PID, der aus den Bedürfnissen der Forschungsgemeinschaft heraus vom Australian Research Data Commons (ARDC) entwickelt wurde.[14] Dieser PID nutzt die DOI-Registrierungsagentur DataCite und stellt für Forscher:innen, Forschungseinrichtungen, Forschungsförderorganisationen, Infrastrukturanbieter:innen und andere Akteur:innen konsistente und aktuelle Informationen über Forschungsprojekte während des gesamten Forschungszyklus bereit. In einer global vernetzten Wissenschaftsgemeinschaft ist die Zusammenarbeit über institutionelle und geografische Grenzen hinweg üblich. PIDs erleichtern diesen Austausch, da sie die Interoperabilität zwischen verschiedenen Datenbanken, wissenschaftlichen Zeitschriften und Repositories sicherstellen, Mehrfacheingaben und Fehler verhindern und auf diese Weise erhebliche Kosten in der Erfassung und Verwaltung von Metadaten in Publikationsdatenbanken, Forschungsinformationssystemen etc. einsparen.[15] Dieses Verständnis hat sich in den wissenschaftlichen Einrichtungen in Deutschland bereits weitgehend etabliert, und die Registrierung von PIDs für Forschungsressourcen und -akteur:innen nimmt kontinuierlich zu.

F2. Forschungsressourcen und -akteur:innen werden mit umfangreichen Metadaten beschrieben

Metadaten beschreiben, wie Forschungsoutput generiert wurde, wer den Forschungsoutput erhoben, bearbeitet bzw. veröffentlicht hat und unter welchen Lizenzbedingungen er wiederverwendet werden darf. Metadaten liefern somit die notwendigen Kontextinformationen für die korrekte Interpretation und Einordnung von Forschungsressourcen und -akteur:innen. Diese Informationen sollten auch maschinenlesbar sein. Je umfangreicher die Metadaten sind, desto besser kann der Forschungsoutput gefunden und nachgenutzt werden. Eine Analyse der Vollständigkeit von Metadaten kann daher Aufschluss geben über deren Mehrwerte.

Das DataCite-Metadatenschema klassifiziert sechs Metadatenproperties (siehe Abb. 2) als empfohlen, die nach Möglichkeit aufgenommen werden sollten. Die Properties „Sachgebiet" (Subject) und „Beschreibung" (Description) sind zentral für die Verbesserung der Auffindbarkeit. Durch die Aufnahme umfangreicher deskriptiver Metadaten können Nutzer:innen die gefundenen Daten und ihren Entstehungskontext besser nachvollziehen. Diese deskriptiven Metadaten können inhaltliche Erschließungen, z. B. detaillierte Beschreibungen durch Abstracts,

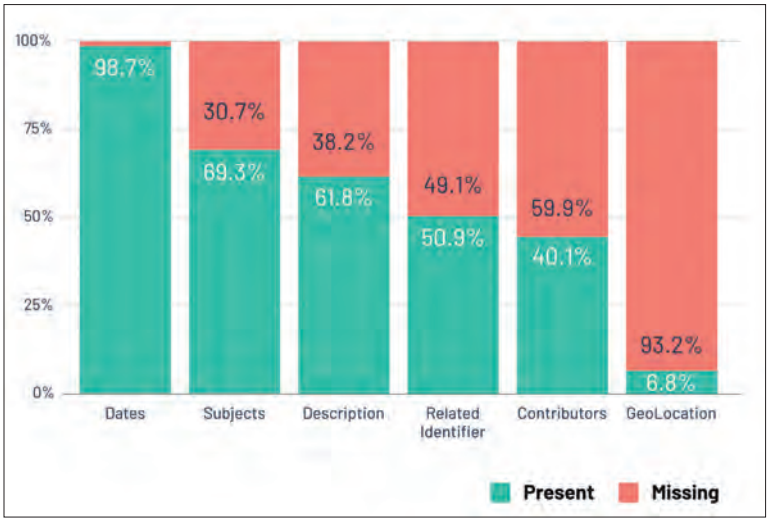

Abbildung 2: Analyse der Vollständigkeit empfohlener Metadatenproperties von DataCite-DOIs (Stand: 2024-09-06)

Zusammenfassungen der angewandten Methode(n) oder Inhaltsverzeichnisse ebenso umfassen wie Sachgebietserschließungen und Verschlagwortungen.

Während bei zeitlichen Angaben (Dates) eine hohe Abdeckung festzustellen ist, fehlen inhaltsbezogene Geoinformationen (Geolocation) – z. B. Angaben zum Ort, an dem Daten erfasst wurden oder der durch sie beschrieben wird – fast vollständig. Darüber hinaus enthalten derzeit nur 44 Prozent der DataCite-DOIs Angaben zu Mitwirkenden (Contributors). Diese Metadaten-Property ist für eine umfassende Darstellung aller Personen und Einrichtungen, die an der Erstellung und Verbreitung von Forschungsergebnissen beteiligt sind, unerlässlich. Durch die Anerkennung verschiedener Rollen, die über die des primären Urhebers hinausgehen, erhöhen diese Metadaten die Ge-

12 https://info.orcid.org/resources/orcid-statistics/ [24. August 2024].

13 https://ror.org [26. August 2024].

14 https://raid.org [26. August 2024].

15 Brown et al. 2022.

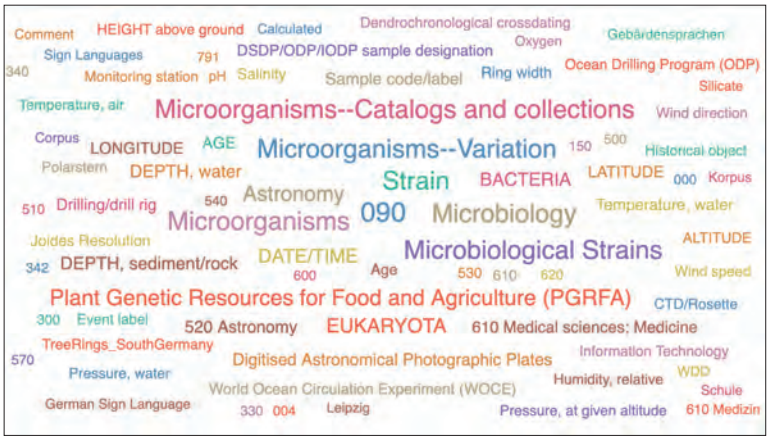

Abbildung 3: Wortwolke mit den 500 am häufigsten verwendeten Schlagwörtern und Fachklassifikationen in DataCite-DOI-Metadaten (Stand: 2024-09-06)

nauigkeit des Datensatzes und unterstützen das Qualitätsmanagement von Forschungsressourcen.

Für die Subject-Property ist quantitativ eine relativ gute Abdeckung zu erkennen, allerdings ist auch eine große Diversität in den Metadaten zu beobachten. Eine Wortwolke (Abb. 3) verdeutlicht die Vielfalt der verwendeten Schlagworte und Fachklassifikationen. Das DataCite-Metadatenschema erlaubt die systematische Nutzung von Fachklassifikationen – standardmäßig der OECD (Organisation for Economic Co-operation and Development) FOS (Fields of Science); die Angabe anderer Schemata wie beispielsweise der DFG-Fachsystematik oder der Library of Congress Subject Headings (LCSH) ist ebenso möglich.[16] Das DataCite-Metadatenschema lässt aber auch freie Verschlagwortungen und weitere Annotationen zu, was eine der möglichen Ursachen für die Diversität von Metadaten ist. Das analysierte Datenset umfasst 3,6 Mio. DOIs mit freier Annotation, 2,1 Mio. DOIs, bei denen eine Fachklassifikation angegeben wurde, jedoch nur 614.268 DOIs mit Angabe der Schema-URI und lediglich 442.561 mit Angabe des Klassifikationscodes. Bei der Nutzung von Schemata und Klassifizierungscodes zur Sacherschließung bestehen daher Optimierungspotenziale in der Standardisierung, um Forschungsressourcen noch effektiver für die Nachnutzung auffindbar zu machen.

Unter den acht optionalen Eigenschaften des DataCite-Metadatenschemas ist die Erfassung der Sprache (Language), in der die Ressource vorgehalten wird, stark vertreten. Bei der Erfassung der Lizenzinformationen (Licenses) hingegen bestehen Verbesserungsbedarfe, da diese besonders wichtig für die Wiederverwendbarkeit der Daten sind. Es ist jedoch ein positiver Trend (Abb. 4) zur Registrierung von mehr Forschungsressourcen mit offenen Lizenzen zu beobachten: Es ist sowohl eine allgemeine Zunahme bei der Registrierung von Lizenzinformationen

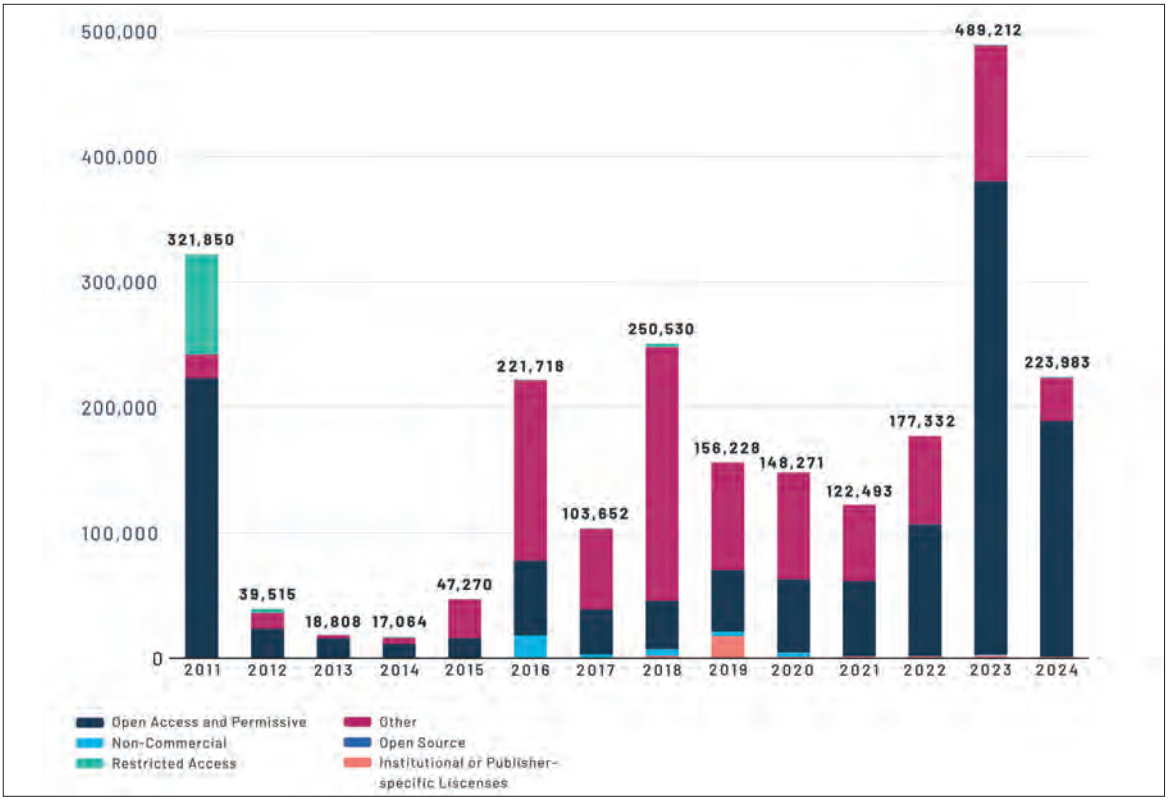

Abbildung 4: Trend in der Erfassung von Lizenzinformationen über die Zeit, stratifiziert nach Lizenztypen in DataCite DOI-Metadaten (Stand: 2024-09-06), für die Lizenzkategorisierung siehe Tabelle 1

16 OECD 2007; DFG 2024; LCSH: https://www.loc.gov/aba/cataloging/subject/ [28. August 2024].

Tabelle 1: Lizenz-Kategorisierung

Open Access and Permissive Creative Commons Attribution (CC BY) Lizenzen, CC0, Public Domain Mark	Non-Commerial-Lizenzen, die die Nutzung auf nicht-kommerzielle Zwecke beschränken.	No Derivatives-Lizenzen, die die Nutzung erlauben, aber Änderungen oder Ableitungen verbieten.	Restricted Access-Lizenzen, die den Datenzugriff einschränken oder auf bestimmte Zwecke beschränken, einschließlich institutioneller Lizenzen.	Institutionelle oder verlagsspezifische Lizenzen Diese Lizenzen werden oft von bestimmten Institutionen vergeben, und ihre Bedingungen können von Open Access bis zu eingeschränktem Zugang variieren.	Open-Source Lizenzen für Software oder Code, die in der Regel die Nutzung, Änderung und Weitergabe unter Open-Source-Bedingungen erlauben.
Creative Commons Attribution 4.0 International	Creative Commons Attribution Non Commercial 3.0 Unported	Creative Commons Attribution-NoDerivs 3.0 Unported	Der Datenzugang ist eingeschränkt (Moratorium, sensible Daten, Lizenzbeschränkungen)	Baseline Surface Radiation Network License 1.0	GNU General Public License v3
Creative Commons Zero v1.0 Universal	Creative Commons Attribution Non Commercial Share Alike 3.0 Unported	Creative Commons Attribution Non Commercial No Derivatives 4.0 International	Lizenzierung unbekannt: Bitte kontaktieren Sie den/die Hauptautor(en)	Alfred-Wegener-Institut für Polar- und Meeresforschung	GNU Lesser Public License v2.1
CC BY 4.0	CC-BY-NC-SA 3.0 DE	CC BY-NC-ND 3.0	Institut für Meeresforschung in Bremerhaven	KITopen License	
Public Domain Mark					

als auch eine Verschiebung hin zu offenen Lizenzen wie Creative Commons (CC) BY 4.0 zu beobachten.[17] Dieser Trend spiegelt ein wachsendes Bestreben zu mehr Transparenz und Zugänglichkeit von Forschungsressourcen im Sinne eines offenen Wissenschaftssystems (Open Science) wider.

F3. Metadaten beinhalten die Identifier der Forschungsressourcen und -akteur:innen

Zentral für die Auffindbarkeit von Forschungsressourcen ist die Aufnahme von Identifiern in die Metadaten. Das Prinzip F3 besagt zunächst, dass der Identifier, der die Ressource referenziert, auch in den die Ressource beschreibenden Metadatensatz aufgenommen wird.[18] Dies ist technisch unerlässlich, um die Effizienz, Konsistenz, Nachverfolgbarkeit und Automatisierung in der Datenverarbeitung und -verwaltung sicherzustellen. Darüber hinaus können zusätzliche Identifier in die Metadaten aufgenommen werden, die Verknüpfungen zu weiteren Forschungsressourcen und zu Akteur:innen herstellen und somit erhebliche Mehrwerte für die Kontextualisierung der Forschungsressource schaffen. Indem Verbindungen zwischen Textpublikationen und Datensätzen, Software oder anderen Forschungsergebnissen sowie Autor:innen, Datenkurator:innen, Forschungseinrichtungen oder Projektkontexten über Identifier geschaffen werden, können die Entstehungs- und Nachnutzungskontexte besser nachvollzogen werden.[19] Die maschinenlesbaren PID-Metadaten können von Metadatenaggregatoren und Wissensgraphen weiterverarbeitet werden. Die Möglichkeiten zur Verknüpfung von DOIs zueinander sowie zu weiteren PID-Systemen ist im DataCite-Metadatenschema umfassend

17 https://creativecommons.org/licenses/by/4.0/deed.de [26. August 2024].

18 Siehe GO FAIR zu F3: https://www.go-fair.org/fair-principles/f3-metadata-clearly-explicitly-include-identifier-data-describe/ [28.08.2024].

19 El-Gebali 2024.

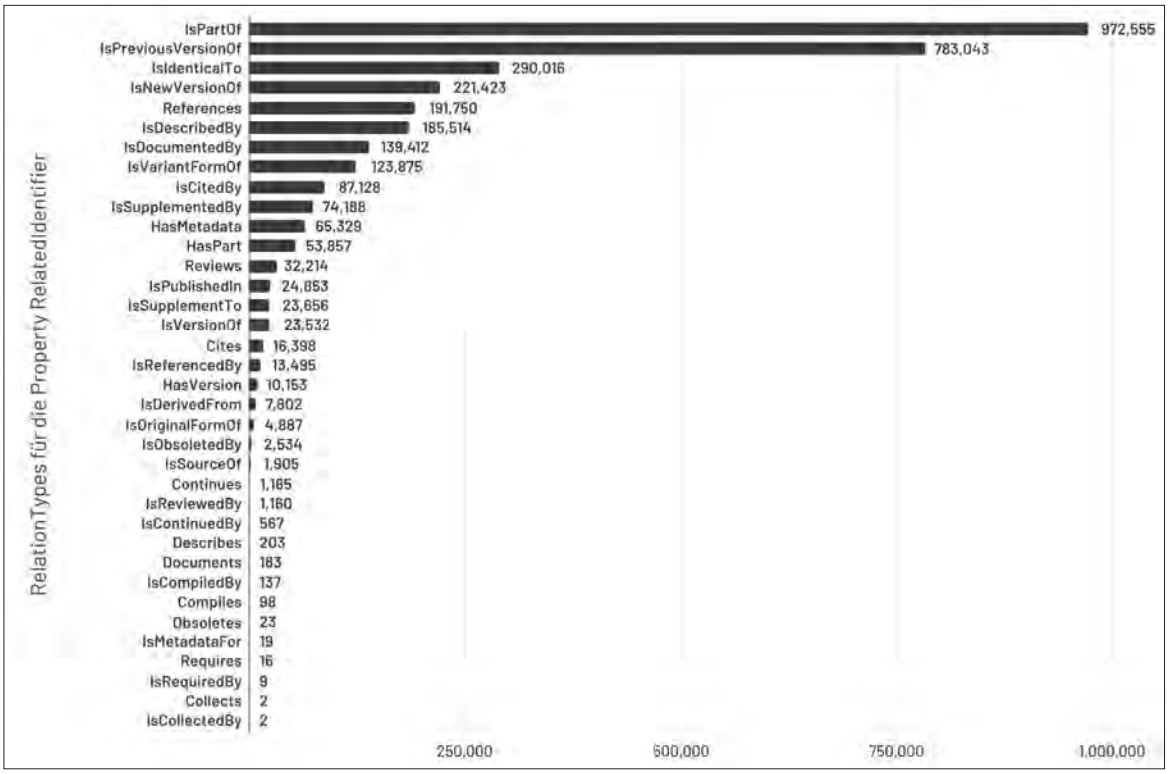

Abbildung 5: Am häufigsten verwendete RelationTypes für die Property RelatedIdentifier in DataCite DOI-Metadaten (Stand: 2024-08-01)

dokumentiert und wird von der DataCite-Registrierungsinfrastruktur unterstützt. Das DataCite-Metadatenschema ermöglicht aktuell die Nutzung von 36 bidirektionalen Relationstypen zur Beschreibung der Beziehung zwischen Forschungsressourcen und -akteur:innen und erlaubt komplexere Beschreibungen von Beziehungen.[20] In den von uns analysierten Metadaten (Abb. 5) wurde der RelationType "IsPartOf" am häufigsten verwendet (972.555 DOIs), gefolgt von der Verknüpfung von Versionen einer Ressource (IsPreviousVersionOf: 783.027 DOIs, IsIdenticalTo: 290.016 DOIs, IsNewVersionOf: 221.423 DOIs). Die Analyse zeigt, dass das Potential zur Abbildung komplexer Beziehungen zwischen Forschungsressourcen und -akteuren noch nicht voll ausgenutzt wird.

F4. Metadaten werden in einer durchsuchbaren Datenbank registriert oder indiziert

Die Sichtbarkeit und folglich auch die Auffindbarkeit von Forschungsressourcen und -akteur:innen hängt maßgeblich davon ab, wo die PIDs und ihre dazugehörigen Metadaten indiziert sind bzw. über welche Wege sie öffentlich zugänglich gemacht werden. Darüber hinaus ist die maschinelle Suche und Abrufbarkeit über standardisierte Schnittstellen (APIs) essentiell für die Kommunikation zwischen verschiedenen Systemen und Informationsinfrastrukturen.

DOIs und ihre dazugehörigen Metadaten können unter Verwendung des DataCite-Metadatenschemas in der Registrierungsumgebung Fabrica oder über Schnittstellen (APIs) registriert werden. Die Metadaten werden über den DataCite-Metadatenspeicher indiziert und langfristig vorgehalten und zur weiteren Nachnutzung über verschiedene Wege öffentlich bereitgestellt. DataCite verfügt über verschiedene, öffentlich zugängliche Schnittstellen (REST API, OAI-PMH) sowie den PID Graph – DataCites Wissensgraph –, über welche die Metadaten maschinell durch Drittanbieter abgefragt und gemäß der CC0-Lizenz weiterverwendet und verarbeitet bzw. angereichert werden können.[21] Über die maschinelle Nachnutzung der Metadaten hinaus basiert das Discovery System DataCite Commons auf den im PID Graphen hinterlegten offenen Metadaten und PIDs und ist somit über eine graphische Benutzeroberfläche frei zugänglich und durchsuchbar.[22] Darüber hinaus stellt DataCite über eine Public Data File sämtliche dort hinterlegten Metadaten über einen Datei-Download der Öffentlichkeit zur Nachnutzung bereit.[23]

Schlussfolgerung

Die vorliegende Analyse der DataCite-DOI-Metadaten zeigt, dass die Forschungslandschaft bereits bedeutende Fortschritte in der Umsetzung der FAIR-Prinzipien gemacht hat, jedoch noch weitere Herausforderungen bestehen.

20 Siehe Dokumentation des DataCite Metadatenschemas zur Property RelatedIdentifier: https://datacite-metadata-schema.readthedocs.io/en/4.5/properties/relatedidentifier/ [26.08.2024].

21 Fenner 2020.

22 https://commons.datacite.org [26.08.2024].

23 Bennett et al. 2024.

Die Anwendung von standardisierten Metadaten hat sich als wirksames Mittel zur Verbesserung der FAIRness von Forschungsressourcen erwiesen, dennoch besteht weiterhin Optimierungsbedarf, insbesondere in Bezug auf die Vollständigkeit und Standardisierung der erfassten Metadaten. Hervorzuheben ist die positive Entwicklung bei der Angabe von Lizenzinformationen in den DOI-Metadaten und der Nutzung offener Lizenzen, die das Engagement für ein transparentes und zugängliches Wissenschaftssystem widerspiegeln. Demgegenüber könnten insbesondere die Standardisierung bzw. Harmonisierung von Fachklassifikationen und die konsistentere Nutzung von Ressourcentypen die Auffindbarkeit und Nachnutzbarkeit erheblich verbessern.

Insgesamt hat die Analyse nach den FAIR-Prinzipien F1-F4 deutlich gemacht, dass PIDs und ihre Metadaten ein grundlegendes Instrument für *Findability* sind, sie aber ihre volle Wirkung nur entfalten können, wenn die zugehörigen Metadaten vollständig und einheitlich erfasst werden. Für eine erfolgreiche Umsetzung braucht es viele Akteur:innen – Forschende, Mitarbeitende in Informationsinfrastrukturen, etc. –, denn die Erstellung und Pflege von Metadaten ist eine kollektive Aufgabe. Die Untersuchung zeigt, dass die Auffindbarkeit von Forschungsressourcen durch die Einbindung umfangreicher Metadaten sowie die Nutzung von Identifiern erheblich gesteigert werden kann. Mithilfe von PIDs lassen sich komplexe Verbindungen zwischen Forschungsressourcen und -akteur:innen darstellen, die Mehrwerte für die Kontextualisierung von Forschungsressourcen und -informationen schaffen.[24] Allerdings sind diese Potenziale noch nicht vollständig ausgeschöpft und könnten auch durch neue technologische Ansätze unter Verwendung von künstlicher Intelligenz erschlossen werden.

Die Ergebnisse unserer Untersuchung unterstreichen die Bedeutung kollektiver Anstrengungen zur Harmonisierung von Metadatenstandards und zur Förderung der Akzeptanz und Implementierung von PIDs in der deutschen Forschungslandschaft. Die fortgesetzte Zusammenarbeit im Rahmen von Projekten wie PID4NFDI und PID Network Deutschland soll diese Herausforderungen angehen und die FAIRness von Metadaten fördern. Langfristig gesehen kann dadurch die Sichtbarkeit, Zugänglichkeit und Nachnutzbarkeit von Forschungsressourcen in Deutschland und darüber hinaus signifikant verbessert werden, was wiederum die Qualität und Reproduzierbarkeit wissenschaftlicher Ergebnisse stärkt.

Bibliographie

Barcelona Declaration on Open Research Information, Kramer, B., Neylon, C., & Waltman, L.: Barcelona Declaration on Open Research Information (1.0), Zenodo, 2024. https://doi.org/10.5281/zenodo.10958522.

Barker, M., Chue Hong, N.P., Katz, D.S. *et al.*: "Introducing the FAIR Principles for research software", in: Sci Data 9, 622 (2022). https://doi.org/10.1038/s41597-022-01710-x.

Bennett, M., Stathis, K., & Gould, M.: Announcing DataCite's First Public Data File, DataCite, 2024. https://doi.org/10.5438/XCNB-0J33.

Bertelmann, R., Buys, M., Kett, J., Pampel, H., Pieper, D., Scholze, F., Sens, I., Burger, F., Dreyer, B., Glagla-Dietz, S., Hagemann-Wilholt, S., Hartmann, S., Schrader, A. C., Schirrwagen, J., Summann, F., Vierkant, P.: PID Network Deutschland. Netzwerk für die Förderung von persistenten Identifikatoren in Wissenschaft und Kultur, Potsdam: Helmholtz Open Science Office 2023. https://doi.org/10.48440/os.helmholtz.059.

Brown, J., Jones, P., Meadows, A., & Murphy, F.: Incentives to invest in identifiers: A cost-benefit analysis of persistent identifiers in Australian research systems, Zenodo, 2022. https://doi.org/10.5281/zenodo.7100578.

24 Buttrick 2024.

Im Blick: Integration und die kommunale Bevölkerungsentwicklung

Bertelsmann Stiftung (Hrsg.)
Von Arbeit bis Zusammenleben – Wie Flüchtlingsintegration in Kommunen gelingen kann
2024, ca. 120 Seiten, Broschur
Mit separatem Fotoband
18,– € (D)
ISBN 978-3-86793-998-0
Erscheint im Dezember 2024

Fluchtmigration ist zurzeit ein großes Streitthema in der öffentlichen Diskussion. Einerseits wird sie für Missstände jeglicher Art verantwortlich gemacht – sei es knapper Wohnraum, fehlende Kitaplätze oder Kriminalität. Andererseits erfordert der Fach- und Arbeitskräftemangel, dass die schutzsuchenden Menschen schnell in den Arbeitsmarkt integriert werden. Mit diesem Band wird der Gesamtüberblick erweitert.

Bertelsmann Stiftung (ed.)
Wer, wo, wie viele? – Bevölkerungsentwicklung in deutschen Kommunen bis 2040
2024, ca. 180 Seiten, Broschur
20,–€ (D)
ISBN 978-3-86793-976-8
Erscheint im Dezember 2024

Kleinräumige Daten sind für kommunale Planungen wichtig. Besonders interessant ist dabei der Blick in die Zukunft, um vorausschauend die Infrastruktur oder Dienstleistungen für die Menschen in den Städten und Gemeinden entwickeln zu können. Die Zusammensetzung der Bevölkerung verändert sich jedoch durch die Alterung unserer Gesellschaft sowie durch viele vielfältige Wanderungsbewegungen.

www.bertelsmann-stiftung.de/verlag

Verlag BertelsmannStiftung

Burger, M., Cordts, A., Habermann, T.: „Wie FAIR sind unsere Metadaten? Eine Analyse der Metadaten in den Repositorien des TIB-DOI-Services", in: Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern 3 (2021), S. 1-13. https://doi.org/10.17192/bfdm.2021.3.8351.

Buttrick, A.: High-Quality Metadata: A Collective Responsibility and Opportunity, Front Matter, 2024. https://doi.org/10.54900/j2q51-0jj98.

Deutsche Forschungsgemeinschaft (DFG): DFG-Fachsystematik der Wissenschaftsbereiche, Fachgebiete, Fachkollegien und Fächer für die Amtsperiode 2024–2028, Bonn 2024.

Deutsche Nationalbibliothek (DNB): Statistik der Deutschen Nationalbibliothek 2023, Frankfurt a.M. 2024. urn:nbn:de:101-2023092133.

El-Gebali, S.: Connecting the Dots with DataCite DOI Metadata, DataCite, 2024. https://doi.org/10.5438/k81t-zq43.

Fenner, M.: DataCite Commons – Exploiting the Power of PIDs and the PID Graph, DataCite, 2020. https://doi.org/10.5438/F4DF-4817.

Hagemann-Wilholt, S., Hauschke, C., Plank, M.: "ConfIDent – An Open Platform for FAIR Conference Metadata", in: Twentieth-First International Conference on Grey Literature "Open Science Encompasses New Forms of Grey Literature", Hannover, Germany, October 22-23, 2019 (2020), S. 47-51 (GL Conference Series 21). https://doi.org/10.15488/9424.

Marginson, S.: "Global science and national comparisons: beyond bibliometrics and scientometrics" in: Comparative Education, 58 (2), 2021, S. 125-146. https://doi.org/10.1080/03050068.2021.1981725.

Organisation for Economic Co-operation and Development (OECD): Revised Field of Science and Technology (FOS) Classification in the Frascati Manual. Ed. by the Directorate for Science, Technology and Industry – Committee for Scientific and Technological Policy, 2007.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.: "The FAIR Guiding Principles for scientific data management and stewardship", in: Sci Data 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18.

Britta Dreyer
leitet das PID-Kompetenzzentrum an der Technischen Informationsbibliothek (TIB) und ist DataCite's Operations and Finance Director. In beiden Positionen engagiert sie sich in der Strategieentwicklung zu PIDs. Sie hat in den letzten Jahren mit verschiedenen nationalen und internationalen Interessengruppen im Bereich von Open Science zusammengearbeitet, um die Nutzung von PIDs zu fördern und neue PIDs wie die ROR ID zu etablieren. Am PID-Kompetenzzentrum werden das TIB DOI Konsortium sowie das deutsche ORCID-Konsortium geleitet und Hochschul- und Forschungseinrichtungen bei der Implementierung von PIDs nach internationalen Best-Practice-Richtlinien unterstützt.
E-Mail-Adresse: britta.dreyer@datacite.org
ORCID iD: http://orcid.org/0000-0002-0687-5460

Stephanie Hagemann-Wilholt
arbeitet am PID-Kompetenzzentrum der Technischen Informationsbibliothek (TIB) in verschiedenen Projekten wie PID4NFDI, PID Network Deutschland und ConfIDent und widmet sich der Verbreitung von PIDs, der Verbesserung ihrer Metadaten und der Etablierung neuer PID-Typen. Zuvor hat sie in der Geschäftsstelle des Rates für Informationsinfrastrukturen in Göttingen gearbeitet und sich mit den politischen Rahmenbedingungen für ein verbessertes Forschungsdatenmanagement im deutschen Wissenschaftssystem befasst.
E-Mail-Adresse: stephanie.hagemann@tib.eu
ORCID iD: https://orcid.org/0000-0002-0474-2410

Paul Vierkant
ist DataCites Outreach-Manager und trägt zu den DFG-finanzierten Projekten re3data COREF und PID Network Deutschland bei. Vor DataCite arbeitete er für das Helmholtz Open Science Office und verschiedene Universitäten, wo er am Aufbau von Publikations- und Datenrepositorien beteiligt war. Als engagierter Verfechter von Open Science ist er bestrebt, die Idee von Open Science in der Wissenschaftskommunikation zu verbreiten.
E-Mail-Adresse: paul.vierkant@datacite.org
ORCID iD: https://orcid.org/0000-0003-4448-3844

Sara El-Gebali
ist seit 2023 bei DataCite als Metadaten-Spezialistin tätig. Nach dem Abschluss ihrer Promotion in den Biowissenschaften arbeitete sie als wissenschaftliche Kuratorin, Forschungsdatenmanagerin, Projektleiterin und Metadatenspezialistin. Sara ist eine leidenschaftliche Community-Builderin und Verfechterin der Gleichberechtigung und Inklusion von Randgruppen in MINT-Fächern und eine Enthusiastin für alles, was FAIR (Findable, Accessible, Interoperable, and Reusable) ist.
E-Mail-Adresse: sara.elgebali@datacite.org
ORCID iD: https://orcid.org/0000-0003-1378-5495