Heinz Pampel und Kirsten Elger

5.6 Publikation und Zitierung von digitalen Forschungsdaten

Abstract: Der vorliegende Beitrag beschreibt gängige Anforderungen und Praktiken bei der Publikation von digitalen Forschungsdaten. Er gibt einen Überblick über relevante Initiativen, Informationsinfrastrukturen und Standards. Über die Perspektive der Publikation hinaus befasst sich der Beitrag mit der derzeitigen Praxis der Zitation von Forschungsdaten und er gibt einen Ausblick auf zukünftige Herausforderungen rund um die dauerhafte Zugänglichkeit und Nachnutzung von Forschungsdaten im Kontext von Open Science.

1 Anforderungen an die Publikation von Forschungsdaten

Die fortschreitende Digitalisierung bietet Forschenden neue Möglichkeiten im Umgang mit digitalen Forschungsdaten. Bereits 2003 hat ein breites Bündnis von wissenschaftlichen Einrichtungen dieses Potenzial in der „Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen“ betont.[1] Im Kern dieser Erklärung steht das Anliegen, alle Ressourcen der wissenschaftlichen Arbeit offen zugänglich und nachnutzbar zu machen. Über den Open Access zu wissenschaftlichen Textpublikationen hinaus, soll auch der offene Zugang zu Forschungsdaten, Metadaten, Software und anderen Quellen der wissenschaftlichen Arbeit sichergestellt werden.[2]

Diese Forderung ist mittlerweile zu einem zentralen Bestandteil der Wissenschaftspolitik geworden, der in Europa unter dem Motto „as open as possible, as closed as necessary“[3] verfolgt wird. Auch im Kontext von G8 wird das Thema erörtert. So haben die G8-Staaten im Jahr 2013 die folgende Forderung formuliert: „Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.“[4] Dieser Standpunkt macht deutlich, dass es eines definierten Rahmens der Publikation von Forschungsdaten bedarf, der auf technischer, rechtlicher, organisatorischer und auch finanzieller Ebene sicherstellt, dass die Forschungsdaten dauer-

1 Vgl. Max Planck Society 2003.
2 Vgl. Klump et al. 2006.
3 Council of the European Union 2016, 8.
4 G8 Science Ministers 2013.

https://doi.org/10.1515/9783110657807-028

haft zugänglich sind und in qualitativ angemessener, nachnutzbarer Form angeboten werden.

Zentrale Instanzen für die Bereitstellung von Forschungsdaten sind digitale Forschungsdatenrepositorien (FDR), die sicherstellen, dass die Daten anhand von definierten Standards gespeichert, dokumentiert, für Menschen und Maschinen in nachnutzbarer Form zugänglich gemacht werden und über Suchdienste auffindbar sind. Diese Publikationsverfahren und damit verbundene Standards haben sich in den letzten Jahren in vielen Fachgebieten als gute Praxis des wissenschaftlichen Arbeitens manifestiert, wie im Folgenden an zwei Fachdisziplinen skizziert werden soll.

In der biomedizinischen Forschung wurde im Human Genome Project[5] bereits im Jahr 1996 Folgendes beschlossen: annotierte Gensequenzen „should be submitted immediately to public databases".[6] Diese Praxis wurde durch weitere Erklärungen, wie den Fort Lauderdale Principles[7] und dem Toronto Statement,[8] weiterentwickelt. So ist es heute in diesem Forschungszweig ein allgemeiner Standard, dass Gensequenzen in fachlichen Repositorien wie z. B. GenBank[9] gespeichert werden. Diese Praxis wird durch die wissenschaftlichen Fachzeitschriften unterstützt, die die Publikation der Gensequenzen in GenBank und anderen fachlichen Forschungsdatenrepositorien zur Bedingung für die Veröffentlichung von wissenschaftlichen Artikeln machen.[10] Mittlerweile gibt es über 1 600, teils sehr spezialisierte, digitale FDR im Bereich der biomedizinischen Forschung über die wissenschaftliche Daten veröffentlicht werden.[11]

Auch in den Erd- und Umweltwissenschaften etablieren sich solche Praktiken. Im Rahmen der Coalition for Publishing Data in the Earth and Space Sciences (COPDESS)[12] arbeitet seit 2014 ein breites Bündnis aus wissenschaftlichen Fachgesellschaften, FDR, Bibliotheken, Verlagen und Forschungsförderern an der Entwicklung und Förderung von abgestimmten Standards, um die Qualität der geowissenschaftlichen Forschungsdaten sicherzustellen und Forschungsdaten als zitierbare Ergebnisse wissenschaftlicher Arbeit anzuerkennen.[13] So verpflichten sich bspw. die Verlage, die das COPDESS Statement of Commitment unterzeichnet haben, bei der Einreichung eines Artikels in ihren Journalen aktiv nach der Veröffentlichung von

5 S. https://www.genome.gov/human-genome-project. Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.
6 Smith und Carrano 1996.
7 Vgl. Wellcome Trust 2003.
8 Vgl. Birney et al. 2009.
9 S. https://www.ncbi.nlm.nih.gov/genbank.
10 Vgl. z. B. Cell 2019; PLOS 2019; Nature 2019.
11 Vgl. Rigden und Fernández 2019, D1.
12 S. https://copdess.org.
13 Vgl. Hanson et al. 2015.

Forschungsdaten, die der wissenschaftlichen Publikationen zugrunde liegen, zu fragen und deren Zitierung in den Artikeln sicherzustellen. Auch wird die Bedeutung von domänenspezifischen FDR hervorgehoben: „Earth and space science data should, to the greatest extent possible, be stored in appropriate domain repositories that are widely recognized and used by the community".[14] Im November 2019 weist das COPDESS Statement of Commitment 44 Unterschriften von Verlagen, Datenzentren und -repositorien, Fachgesellschaften und anderen Initiativen in der Fachcommunity nach. Im Jahr 2018 wurden von der gleichen Gruppe die Ergebnisse des auf das COPDESS Statement of Commitments aufbauenden Enabling FAIR Data Projektes vorgestellt.[15] Im Rahmen dieses Projektes wurden Standards und Empfehlungen zur Publikation von Forschungsdaten gemäß den FAIR-Prinzipien[16] entwickelt.

Durch die Verankerung von Anforderungen zur Publikation der Forschungsdaten in den Data Policies von Förderorganisationen, wissenschaftlichen Einrichtungen und Zeitschriften nimmt das Themenfeld im Bereich des wissenschaftlichen Publikationswesen eine zunehmend wichtigere Position ein,[17] die durch die wissenschaftspolitische Verankerung der FAIR-Prinzipien[18] in der europäischen Forschungsförderung weiter an Bedeutung gewinnt.[19] Diese Prinzipien wirken auf die Praxis, wie digitale Forschungsdaten veröffentlicht werden. Im Kern tangieren vier FAIR-Prinzipien die Veröffentlichungspraxis der Forschungsdaten, ihre Metadaten und der Repositorien, auf denen die Daten gespeichert werden: So müssen Forschende, unterstützt durch Einrichtungen der Informationsinfrastruktur, sicherstellen, dass Forschungsdaten auffindbar (Findable), zugänglich (Accessibe), interoperabel (Interoperable) und wiederverwendbar (Reusable) sind.

2 Datenpublikation auf Repositorien

Datenpublikationen sind eigenständige, zitierbare und dauerhafte Veröffentlichungen von digitalen Forschungsdaten in einem FDR. FDR sind digitale Informationsinfrastrukturen für Forschungsdaten. Sie stellen die dauerhafte Zugänglichkeit von Forschungsdaten sicher, indem sie die Forschungsdaten speichern, mit persistenten Identifikatoren, z. B. dem Digital Object Identifier (DOI), eindeutig adressieren, deren Auffindbarkeit sicherstellen und sie so einer definierten Gruppe an Nutzerinnen und Nutzern zur Verfügung stellen. Die FDR und ihre Services sind durch Form und

14 CODPESS 2015.
15 Vgl. Stall et al. 2018.
16 Vgl. Wilkinson et al. 2016.
17 Vgl. Pampel und Bertelmann 2011; Bloom et al. 2014.
18 Vgl. Wilkinson et al. 2016.
19 Vgl. European Commission 2016.

Formate der Forschungsdaten geprägt, die sie speichern und zugänglich machen. Ein FDR ist als technisches und organisatorisches System zur Sicherung und dauerhaften Zugänglichkeit der Forschungsdaten zu verstehen.

Eine Datenpublikation fördert die Transparenz der Forschung, in dem Fachkolleginnen und -kollegen sowie weitere interessierte Personen die erzeugten Ergebnisse nachprüfen können. Sie ermöglicht die Nachnutzung der Daten in neuen Kontexten und stellt darüber hinaus die Anerkennung der Forschenden, die die Daten erhoben haben, sicher.[20]

Mit Blick auf ihre Nutzerinnen und Nutzer können vier Typen von FDR unterschieden werden: institutionelle, disziplinspezifische, multidisziplinäre und projektspezifische.[21] Im Folgenden werden einige Beispiele gegeben (zur Recherchemöglichkeit für FDR sei auf das Ende dieses Abschnittes verwiesen).

Ein *institutionelles FDR* steht den Angehörigen einer wissenschaftlichen Einrichtung zur Speicherung ihrer Daten zur Verfügung. Beispiele sind Edinburgh DataShare,[22] das an der University of Edinburgh betrieben wird, und Open Data LMU[23] an der Ludwig-Maximilians-Universität München.

Ein *disziplinspezifisches FDR* ist beispielsweise GFZ Data Services, welches am Deutschen GeoForschungsZentrum GFZ betrieben wird.[24] Während die Daten auf Edinburgh DataShare und Open Data LMU die Vielfalt der Disziplinen der jeweiligen Universität widerspiegeln, ist GFZ Data Services auf geowissenschaftliche Daten und fachspezifische Software spezialisiert. Als Teil des Fachinformationsdienstes Geowissenschaften (FID GEO)[25] steht es der breiten geowissenschaftlichen Fachcommunity als Infrastruktur zur Verfügung.[26] Als fachlicher Service unterstützt das Repositorium neben dem DataCite Metadaten Standard[27] diverse in den Geowissenschaften genutzten Fachstandards für Metadadaten, wie bspw. die ISO 19115 Geographic Information – Metadata der Internationalen Organisation für Normung (ISO)[28] oder das Directory Interchange Format (DIF)[29] der US-Raumfahrtbehörde NASA. Diese Fachstandards bieten die Möglichkeit, die Beschreibung der Daten durch fachspezifische Begriffe aus kontrollierten Vokabularien und Ontologien zu

20 Vgl. Kaden 2016.
21 Vgl. Pampel et al. 2013.
22 S. https://datashare.is.ed.ac.uk.
23 S. https://data.ub.uni-muenchen.de.
24 S. http://dataservices.gfz-potsdam.de.
25 S. http://www.fidgeo.de/.
26 Vgl. Achterberg et al. 2018.
27 S. https://schema.datacite.org.
28 S. https://www.iso.org/standard/53798.html, https://www.iso.org/standard/67039.html.
29 S. https://earthdata.nasa.gov/esdis/eso/standards-and-references/directory-interchange-format-dif-standard.

ergänzen und somit einen wichtigen Beitrag zur Verbesserung der Auffindbarkeit und Inhaltserschließung bzw. -dokumentation von Forschungsdaten zu leisten.

Weitere fachliche FDR sind z. B. die Infrastrukturen GenBank[30] und PANGAEA.[31] GenBank wird seit 1982[32] von der National Library of Medicine (NLM) in den USA betrieben und wird von Forschenden aus aller Welt zur Speicherung von DNA-Sequenzen genutzt. Das Repositorium weist gemeinfreie Gendaten von fast 420 000 Spezies nach. Die einzelnen Datensätze werden durch eine von Genbank vergebene Accession Number adressiert.[33]

PANGAEA ist Mitglied des World Data Systems des International Science Councils (ISC) und definiert sich als „Data Publisher for Earth & Environmental Science". Das Repositorium wurde von 1995 bis 1997 aufgebaut und wird an der Universität Bremen sowie am Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung betrieben.[34] PANGAEA steht Forschenden aus vielen verschiedenen Bereichen der Erd- und Umweltwissenschaften zur Verfügung, hat jedoch ein besonders ausgeprägtes Sammlungsprofil im Bereich der marinen Geowissenschaften und der Paläoklimaforschung. Alle Datensätze werden mit einem DOI adressiert. Die Forschungsdaten sind meist unter der Creative-Commons-Lizenz „Namensnennung" publiziert.[35]

Im Bereich der *multidisziplinären FDR* sind die Dienste Zenodo und Figshare populär. Zenodo[36] wird am CERN betrieben und wurde im Rahmen des EU-Projektes OpenAIRE[37] entwickelt.[38] Figshare[39] wird von der Firma Digital Science betrieben und versteht sich als „The All In One Repository".[40] Das generische Profil beider Dienste hat den Nachteil, dass disziplinäre Standards nicht unterstützt werden. So sind Forschungsdaten in beiden Repositorien nur ein Publikationstyp unter vielen und damit die Auffindbarkeit der Daten über digitale Kataloge herausfordernd.

Ein weiterer Typ sind *projektspezifische FDR*. Ein Beispiel ist Digital Pantheon[41], auf dem digitale Modelle und zugehörige Daten des antiken Pantheon in Rom gespeichert und offen zugänglich gemacht werden. Auch dieses Repositorium adressiert jedes seiner Datensätze mit einem DOI. Die Daten werden unter der Creative-

30 S. https://www.ncbi.nlm.nih.gov/genbank.
31 S. https://pangaea.de.
32 Vgl. Cravedi 2008.
33 Vgl. Sayers et al. 2019, D94.
34 Vgl. Diepenbroek et al. 1999, 717.
35 Vgl. PANGAEA 2019.
36 S. https://zenodo.org.
37 S. https://www.openaire.eu.
38 Vgl. Zenodo n.d.
39 S. https://figshare.com.
40 Vgl. Hyndman 2018.
41 S. http://repository.edition-topoi.org/collection/BDPP.

Commons-Lizenz „Namensnennung – Nicht-kommerziell – Weitergabe unter gleichen Bedingungen“ lizenziert.[42]

Über diese Repositorientypen hinaus gibt es auch Portale, die Daten aus verschiedenen eigenständigen Quellen zusammenführen.[43]

Zur Identifikation von FDR[44] empfiehlt sich der internationale Dienst re3data – Registry of Research Data Repositories.[45] Durch sein umfassendes Metadatenschema,[46] welches sowohl technische als auch inhaltliche Informationen umfasst, hilft dieser Service einer breiten Nutzergruppe (von Wissenschaftlerinnen und Wissenschaftlern über Citizen Scientists bis zu Forschungsförderern) bei der Identifikation von geeigneten Repositorien. Im November 2019 weist das Verzeichnis über 2400 Repositorien nach.[47] Eine Analyse von Kindling et al., basierend auf re3data, zeigt, dass die Landschaft der FDR sehr heterogen und wenig standardisiert ist.[48]

Mit steigender Anzahl an Repositorien erwachsen auch Anforderungen an deren Vergleichbarkeit im Hinblick auf Vertrauenswürdigkeit und Standardisierung.[49] Hierfür haben sich in der Vergangenheit verschiedene Zertifikate entwickelt (u. a. Data Seal of Approval, ICS World Data System, DIN-Norm 31644 „Kriterien für vertrauenswürdige digitale Langzeitarchive“ oder auch ISO 16363 „Audit and certification of trustworthy digital repositories“), die von einigen Repositorien erlangt wurden. Die im Rahmen der Research Data Alliance (RDA)[50] entwickelte CoreTrustSeal-Zertifizierung[51] ist die gemeinsam von Data Seal of Approval und ICS World Data System entwickelte Zertifizierung, die sich als erster Schritt eines globalen Zertifizierungsnetzwerks betrachtet, welches auch die „extended level certification“ der DIN-Norm 31644 und die „formal level certification“ von ISO 16363 mit einschließt.[52]

3 Praktiken der Publikation von Forschungsdaten

Die Publikation von Forschungsdaten auf Repositorien kann durch verschiedene Veröffentlichungsstrategien praktiziert werden. Angelehnt an Dallmeier-Tiessen

42 Vgl. Topoi n.d.
43 S. a. Abschnitt 3.
44 Vgl. Pampel et al. 2013.
45 S. https://www.re3data.org.
46 Vgl. Rücknagel et al. 2015.
47 S. Beitrag von Scholze, Goebelbecker und Ulrich, Kap. 2.2 in diesem Praxishandbuch.
48 Vgl. Kindling et al. 2017.
49 Vgl. Klump et al. 2011.
50 S. https://www.rd-alliance.org.
51 S. https://www.coretrustseal.org.
52 Vgl. CoreTrustSeal n.d.

(2011) und Pampel et al. (2012) können folgende Publikationsstrategien unterschieden werden.[53]

Veröffentlichung der Forschungsdaten als eigenständiges Informationsobjekt in einem Datenrepositorium: Diese Strategie zielt darauf ab, die Daten ohne begleitenden Artikel in einer Fachzeitschrift zu veröffentlichen. Die Autorinnen und Autoren des Datensatzes gehen davon aus, dass die durch das Repositorium erfassten Metadaten und ggfs. bereitgestellte README-Dateien ausreichen, um den Datensatz nachnutzen zu können.

Veröffentlichung der Forschungsdaten in einem Datenrepositorium und Dokumentation im Rahmen eines begutachteten Artikels in einem Data Journal: Diese Strategie ermöglicht, dass der Datensatz umfassend beschrieben wird und im Rahmen eines Peer-Review-Verfahrens nicht nur der Artikel, sondern auch die Originalität und Qualität der Daten sowie deren Zugänglichkeit und Nachnutzbarkeit gesichert werden. Ein Beispiel für einen solches Data Journal ist Earth System Science Data (ESSD) im Verlag Copernicus Publications.[54] Als erstes Data Journal weltweit, veröffentlicht diese Open-Access-Zeitschrift seit 2012 Data Description Articles, welche die (technische) Beschreibung von Datensätzen enthalten, bei gleichzeitiger Veröffentlichung der Daten über FDR. Durch die explizite Open Access Policy von ESSD, die nicht nur die Artikel, sondern auch Daten und Software miteinschließt, leistet ESSD einen wichtigen Beitrag für die Nachnutzung qualitätsgeprüfter Forschungsdaten. Ein weiteres Beispiel[55] ist das Journal Scientific Data von Springer Nature, das sich als „peer-reviewed, open-access journal for descriptions of scientifically valuable datasets, and research that advances the sharing and reuse of scientific data“ bezeichnet.[56]

Veröffentlichung der Forschungsdaten in einem Datenrepositorium und Dokumentation im Rahmen eines Data Reports: Diese Praxis wird z. B. am Deutschen GeoForschungsZentrum GFZ umgesetzt. Die Daten selbst werden über das Repositorium GFZ Data Services veröffentlicht und darüber hinaus in einem intern begutachteten Datenreport ausführlich beschrieben.[57] Durch die Nutzung der „related identifier“ aus dem Metadaten-Schema der DOI-Registrierungsagentur DataCite[58] wird die Verknüpfung der beiden Ressourcen sichergestellt. Somit wird vom Report auf den Datensatz verwiesen und umgekehrt. Der Report hat deskriptiven Charakter und liefert Informationen zu allen Parametern, die für die Nachnutzung der Daten von Bedeutung sind. Die Daten selbst werden nicht von externen Wissenschaftlerinnen und

53 Vgl. Dallmeier-Tiessen 2011, 5–10; Pampel et al. 2012, 63.
54 S. https://www.earth-system-science-data.net.
55 Eine Liste von Data Journals findet sich unter: https://www.forschungsdaten.org/index.php/Data_Journals.
56 Springer Nature 2019.
57 Report: Voigt et al. 2016; Daten: Wziontek et al. 2017.
58 S. https://datacite.org.

Wissenschaftlern begutachtet. Jedoch prüfen die Kuratorinnen und Kuratoren von GFZ Data Services Inhalte und Vollständigkeit der Metadaten und die technische Präsentation der Daten. Der Report wird GFZ-intern begutachtet. Damit ist die Nachvollziehbarkeit der beiden Informationsressourcen gewährleistet.

Veröffentlichung der Forschungsdaten in einem Repositorium als Ergänzung zu einem begutachteten wissenschaftlichen Artikel (Data Supplement oder Enhanced Publication): Eine steigende Zahl von Journalen fordert, dass in sogenannten Data Availability Statements Aussagen zu den Daten, die Grundlage des entsprechenden Artikels sind, gemacht werden (insbesondere zur Zugänglichkeit und den entsprechenden Zugangskonditionen). Um dieser Anforderung nachzukommen wurden seit dem Beginn des 21. Jahrhundert vermehrt Daten als Supplement veröffentlicht. Dies bedeutete lange Zeit, dass Datentabellen oder zusätzliche Illustrationen den wissenschaftlichen Artikeln als Anhang beigelegt wurden, welche selten kuratiert und nur schwer auffindbar waren. Um diese Datenquellen nutzbar zu machen, empfehlen viele Journals heute die Nutzung von FDR anstelle klassischer Datensupplemente.[59]

Einige Journals verlangen inzwischen, dass die Reviewer schon bei der Einreichung eines Aufsatzes Zugang zu den Daten haben, um die Nachvollziehbarkeit der Ergebnisse zu prüfen. Immer mehr Repositorien unterstützen diese Praxis, indem sie bereits vor der Veröffentlichung der Daten Review-Links oder geschützte Zugänge zu den noch unveröffentlichten Datensätzen bereitstellen, die den Gutachterinnen und Gutachtern den Zugang zu den Daten ermöglichen und erlauben, dass Änderungswünsche an oder Ergänzungen zu den Daten im Rahmen der wissenschaftlichen Begutachtung vor der Registrierung der DOI möglich sind. In den meisten Fällen erfolgt die Publikation der Daten gleichzeitig mit der Publikation des Artikels.

4 Auffindbarkeit von Forschungsdatenpublikationen

Um die Auffindbarkeit der Daten zu ermöglichen, werden verschiedene Verfahren verfolgt. Zum einen gibt es wissenschaftliche Suchmaschinen, die neben anderen Publikationstypen auch Forschungsdaten aggregieren und somit auffindbar machen. Beispiele hierfür sind: BASE – Bielefeld Academic Search Engine[60] und OpenAIRE Explore.[61] Beide Suchdienste aggregieren über das Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Metadaten und erlauben so den Zugriff auf Forschungsdaten, die auf verteilten Repositorien gespeichert sind. Ein ähnli-

59 Vgl. Stall et al. 2018.
60 S. https://www.base-search.net.
61 S. https://explore.openaire.eu.

cher, aber auf Forschungsdaten fokussierter Suchdienst ist das im Rahmen von EUDAT[62] entwickelte B2FIND.[63]

Zum anderen gibt es spezielle Suchmaschinen für Daten wie z. B. DataCite Metadata Search[64] und Google Dataset Search.[65] Diese beiden Dienste erlauben den Zugang zu Forschungsdaten aus allen Disziplinen und werden im Folgenden näher beschrieben.

DataCite Metadata Search: Dieser Suchdienst erlaubt das Retrieval von Daten- und Softwarepublikationen sowie grauer Literatur, die über die DOI-Registrierungsagentur identifiziert sind, über die entsprechenden Metadaten. Der Dienst kann über eine grafische Benutzeroberfläche und verschiedene maschinenlesbare Schnittstellen adressiert werden. Zu jedem Datensatz finden sich Metadaten und Informationen zu dem Repositorium, das die Daten bereitstellt. DataCite ermöglicht auch die Verknüpfung der Datenpublikationen mit ORCID,[66] dem zentralen Dienst zur Autorinnen- und Autorenidentifikation.[67]

Google Dataset Search: Dieser Suchdienst ging im Jahr 2018 online. Nach eigenen Angaben „nutzt Google schema.org und andere Metadatenstandards, die den Seiten, die Datensätze beschreiben, hinzugefügt werden können".[68] Der Dienst liefert zu jedem Datensatz Informationen zur eindeutigen Kennzeichnung des Datensatzes, ein Veröffentlichungsdatum, Informationen zum Repositorium, die Namen der Autorinnen und Autoren, die verwendete Lizenz sowie, wenn vorhanden, auch Informationen zur Förderorganisation und einen Abstract, der die Daten beschreibt.[69] Darüber hinaus gibt es fachliche Suchdienste, z. B. die Suchmaschine ALBERT.[70] Diese wird von der Bibliothek des Wissenschaftsparks Albert Einstein in Potsdam betrieben. Sie indexiert relevante Informationsressourcen für die geowissenschaftlichen Community in Deutschland und ist ein gutes Beispiel für die gemeinsame Indizierung von Text- und Datenpublikationen.[71]

Ein weiteres Beispiel ist der Data Catalogue des Consortium of European Social Science Data Archives (CESSDA),[72] über den sich Forschungsdaten sozialwissenschaftlicher Repositorien in Europa durchsuchen lassen.[73] Auch fördern vermehrt Verlage über ihre Zeitschriftenportale die Auffindbarkeit der Daten (z. B. Elsevier).

62 S. https://eudat.eu.
63 S. http://b2find.eudat.eu.
64 S. https://search.datacite.org.
65 S. https://toolbox.google.com/datasetsearch.
66 S. https://orcid.org.
67 Vgl. Fenner 2019.
68 Google n.d.
69 Vgl. Burgess und Noy 2018.
70 S. http://bib.telegrafenberg.de.
71 Vgl. Bertelmann et al. 2012.
72 S. https://datacatalogue.cessda.eu.
73 Vgl. Shepherdson und Thiel 2018.

Der Verlag ermöglicht über die Plattform ScienceDirect auch den Zugang zu Daten, die Grundlage eines von Elsevier verlegten Artikels und auf FDR gespeichert sind. Des Weiteren gibt es kostenpflichtige Suchdienste für Forschungsdaten wie z. B. den Data Citation Index von Clarivate, der auch in die Plattform Web of Science integriert ist.

5 Zitation von Forschungsdaten

Forschungsdaten, auf die in einer Publikation Bezug genommen wird, sind entsprechend des Kodexes der guten wissenschaftlichen Praxis zu zitieren:

> „Die Herkunft von im Forschungsprozess verwendeten Daten, Organismen, Materialien und Software wird kenntlich gemacht und die Nachnutzung belegt; die Originalquellen werden zitiert. Art und Umfang von im Forschungsprozess entstehenden Forschungsdaten werden beschrieben."[74]

Dies bedeutet in der Praxis, dass alle Personen, die in die Erhebung und Aufbereitung der Daten involviert sind, als „Creator" eines Datensatzes zu nennen sind.

Die CRediT (Contributor Roles Taxonomy) bietet hierzu Hilfestellungen, indem sie 14 Rollen von Tätigkeiten im wissenschaftlichen Publikationsprozess beschreibt.[75] Bei der Zitierung der Daten sollte der „Joint Declaration of Data Citation Principles"[76] aus dem Jahr 2014 gefolgt werden. Diese beschreiben acht zentrale Aspekte bei der Zitation von Forschungsdaten, inklusive der klaren Empfehlung, dass Forschungsdaten genau wie alle anderen Quellen als Zitate in den Referenzen der wissenschaftlichen Artikel enthalten sein sollen. Das Anliegen der Erklärung ist, dass Forschungsdaten als zitierbares Produkt des wissenschaftlichen Erkenntnisprozesses zu verstehen und den involvierten Personen „credit and attribution" zu garantieren ist.[77]

Bei der praktischen Umsetzung der Zitation von Forschungsdaten empfiehlt sich unbedingt die Nutzung persistenter Identifier, wie z. B. dem DOI. Insbesondere um diesen Identifier und die Registrierungsagentur DataCite sind in den letzten Jahren eine Vielzahl von hilfreichen Services rund um die Zitation und Publikation von

74 Deutsche Forschungsgemeinschaft 2019, 14.
75 S. https://www.casrai.org/credit.html.
76 S. https://www.force11.org/datacitationprinciples.
77 Vgl. Data Citation Synthesis Group 2014.

Forschungsdaten entstanden,[78] so z. B. der Dienst DataCite Event Data, über den u. a. die Zitierung von Forschungsdaten erfasst wird.[79]

Kern der Arbeit von DataCite ist die Entwicklung eines umfangreichen Metadaten-Schemas zur Beschreibung von Forschungsdaten. DataCite empfiehlt in der Version 4.3 des DataCite-Schemas die folgende Zitation für Forschungsdaten:

> Creator (PublicationYear): Title. Version. Publisher. (resourceTypeGeneral). Identifier

Darüber hinaus gibt das Schema Hinweise wie mit der Versionierung von Datensätzen und wie mit dynamischen Daten umzugehen ist.[80] Aktuell werden im Rahmen der Research Data Alliance (RDA) und deren Arbeitsgruppen „Data Citation“[81] und „Data Versioning“[82] wichtige Arbeiten zum Thema verfolgt.

Disziplinäre FDR bieten im Rahmen ihrer kuratorischen Tätigkeiten vielfältige Beratungsleistungen rund um den Veröffentlichungsprozess an. Diese Kompetenzen sind wichtig, wenn es beispielweise darum geht zu entscheiden, in welcher Granularität die Daten zu veröffentlichen sind. Allgemeine Hinweise sind hier aufgrund der verschiedenen Praktiken in den Fachdisziplinen nur bedingt anwendbar.

Die Anforderungen der gängigen Zitationsstile zur Zitierung von Forschungsdaten variieren. Das Publication Manual der American Psychological Association (APA) sieht z. B. in Version 6 den Publikationstyp „Data set“ vor und darüber hinaus den Publikationstyp „Data file and code book“ für weitere Ressourcen rund um einen Datensatz.[83] The Chicago Manual of Style Online erkennt in seiner Version 17 Forschungsdaten nicht als eigenständigen Publikationstyp an.[84] Auch der Umgang mit Forschungsdaten in Literaturverwaltungsprogrammen variiert. Während EndNote X7 über den Publikationstyp „Data set“ verfügt, ist dieser in Zotero 5.0 nicht vorhanden. Wünschenswert wäre in diesem Feld eine sehr viel stärkere Ausrichtung der Zitationsstile und deren Anwendung in den Literaturverwaltungsprogrammen an dem DataCite-Metadaten-Schema, welches die Anforderungen verschiedener Fächer vereinigt.

Bereits jetzt deuten mehrere Studien darauf hin, dass Zeitschriftenartikel bei denen die zugrundeliegenden Daten offen zugänglich gemacht werden und auf diese

78 Zur Historie der Anwendung des DOI für Forschungsdaten vgl. Paskin 2006 sowie Klump et al. 2016.
79 S. https://datacite.org/eventdata.html.
80 Vgl. DataCite Metadata Working Group 2019.
81 Vgl. Rauber et al. 2015.
82 Vgl. Klump et al. 2020.
83 Vgl. McAdoo 2013.
84 Vgl. The Chicago Manual of Style 2017.

in den Artikeln hingewiesen wird, häufiger zitiert werden als Studien, die ihre Daten nicht veröffentlichen.[85]

6 Zukunft der Publikation von Forschungsdaten

Mit der stetigen Durchdringung der Digitalisierung der Wissenschaft steigt die Notwendigkeit der Maschinenlesbarkeit von digitalen Forschungsdaten mehr und mehr an. Die Schaffung von Interoperabilität, damit Menschen und Maschinen mit digitalen Forschungsdaten arbeiten können, ist ein zentrales Handlungsfeld für die kommenden Jahre. Dazu gehört auch, dass Forschungsdaten, textuelle Publikationen, Software und andere Informationsobjekte stärker vernetzt werden. Die Nutzung des DataCite-Schemas für Metadaten und der DOI zur persistenten Identifizierung der Daten ermöglicht die Anwendung von Frameworks wie „Scholix“,[86] mit denen die Verlinkung der Daten und textuellen Publikationen[87] sichergestellt wird. Damit wird auch entsprechend dem Open-Science-Paradigma ermöglicht, dass Forschungsergebnisse umfassend zugänglich und nachnutzbar gemacht werden. Interoperabilität stellt im Zusammenspiel mit der Publikation der Daten auf nachhaltigen Infrastrukturen sicher, dass Anwendungen des Semantic Webs für die Wissenschaft genutzt werden können.

Die FAIR-Prinzipien formulieren hier zentrale Anforderungen rund um die Publikation von Forschungsdaten, deren Realisierung in den kommenden Jahren eine zentrale Aufgabe für die Wissenschaft, ihre Informationsinfrastrukturen und weitere Dienstleister sein wird. Dabei ist die Wissenschaft gefordert, die Publikation der Daten nach ihren Vorstellungen und Bedingungen zum Wohle der Wissenschaft und der Gesellschaft zu gestalten und die Kommerzialisierung durch Verlage und andere externe Akteure zu verhindern.

Fazit

Die Publikation von Forschungsdaten gewinnt mehr und mehr Aufmerksamkeit. Je nach Disziplin bilden sich verschiedene Publikationspraktiken heraus, in deren Zentren FDR stehen, die die dauerhafte Zugänglichkeit und Nachnutzung der Daten sichern. Die Verankerung der FAIR-Prinzipien in den Leit- und Richtlinien rund um

85 Vgl. Colavizza et al. 2019; Drachen et al. 2016; Dorch et al. 2015; Drachen et al. 2016; Belter 2014; Piwowar und Vision 2013; Piwowar et al. 2007.
86 S. http://www.scholix.org.
87 Vgl. Burton et al. 2017.

das Forschungsdatenmanagement macht deutlich, dass die bisherigen Verfahren der Veröffentlichung von digitalen Forschungsdaten noch am Anfang stehen. Die Realisierung der Vision „FAIRer" Daten stellt Wissenschaft und ihre Informationsinfrastrukturen vor vielfältige Herausforderungen, die es zu diskutieren und zu gestalten gilt.

Da viele digitale Arbeitsmethoden nur angewendet werden können, wenn auch die Daten selber möglichst automatisch und durch Maschinen gefunden, erfasst und analysiert werden können, stellt die Maschinenlesbarkeit der Forschungsdaten eine der zentralen Aufgaben für die Wissenschaft und ihre Serviceeinrichtungen, wie Bibliotheken, Daten- und Rechenzentren dar. Dabei gilt es, die Publikation von Forschungsdaten als technische und organisatorische Aufgabe zu begreifen, bei deren Umsetzung Einrichtungen der Informationsinfrastruktur im Rahmen einer nachhaltigen Finanzierung der Wissenschaft dienen.

Literatur

Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

Achterberg, Inke, Roland Bertelmann, Kirsten Elger, Andreas Hübner, Norbert Pfurr und Mechthild Schüler. 2018. Der Fachinformationsdienst Geowissenschaften der festen Erde (FID GEO). *Bibliotheksdienst* 52 (5/25. April): 391–405. doi:10.1515/bd-2018-0045.

Belter, Christopher W. 2014. „Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets." Hg. von Howard I. Browman. *PLoS ONE* 9 (3/26. März): e92590. doi:10.1371/journal.pone.0092590.

Bertelmann, Roland, Sascha Szott und Tobias Höhnow. 2012. „Discovery jenseits von ‚all you can eat' und ‚one size fits all'." *Bibliothek Forschung und Praxis* 36 (3/Januar): 369–376. doi:10.1515/bfp-2012-0050.

Bloom, Theodora, Emma Ganley und Margaret Winker. 2014. „Data Access for the Open Access Literature: PLOS's Data Policy." *PLoS Biology* 12 (2/25. Februar): e1001797. doi:10.1371/journal.pbio.1001797.

Toronto International Data Release Workshop Authors. 2009. „Prepublication data sharing." *Nature* 461 (7261/September): 168–170. doi:10.1038/461168a.

Burgess, Matthew und Natasha Noy. 2018. „Building Google Dataset Search and Fostering an Open Data Ecosystem." Google AI Blog. http://ai.googleblog.com/2018/09/building-google-dataset-search-and.html.

Burton, Adrian, Amir Aryani, Hylke Koers, Paolo Manghi, Sandro La Bruzzo, Markus Stocker, Michael Diepenbroek, Uwe Schindler und Martin Fenner. 2017. „The Scholix Framework for Interoperability in Data-Literature Information Exchange." *D-Lib Magazine* 23 (1/2). doi:10.1045/january2017-burton.

Cell. 2019. Mandatory Data Deposition. https://www.cell.com/cell/authors.

Colavizza, Giovanni, Iain Hrynaszkiewicz, Isla Staden, Kirstie Whitaker und Barbara McGillivray. 2020. „The citation advantage of linking publications to research data." *arXiv* (5. März):1907.02565 [cs]. http://arxiv.org/abs/1907.02565.

COPDESS. 2015. Statement of Commitment from Earth and Space Science Publishers and Data Facilities. http://www.copdess.org/statement-of-commitment/.

Council of the European Union. 2016. The transition towards an Open Science system – Council conclusions (adopted on 27/05/2016). 9526/16. https://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf.

Cravedi, Kathy. 2008. „GenBank Celebrates 25 Years of Service with Two-Day Conference; Leading Scientists Will Discuss the DNA Database at April 7–8 Meeting." National Institutes of Health (NIH). 15. September. https://www.nih.gov/news-events/news-releases/genbank-celebrates-25-years-service-two-day-conference-leading-scientists-will-discuss-dna-database-april-7-8-meeting.

Dallmeier-Tiessen, Sünje. 2011. „Strategien bei der Veröffentlichung von Forschungsdaten. Working Paper." RatSWD Working Paper. http://hdl.handle.net/10419/75349.

DataCite Metadata Working Group. 2019. „DataCite Metadata Schema Documentation for the Publication and Citation of Research Data." Version 4.3. doi:10.14454/7xq3-zf69.

Data Citation Synthesis Group. 2014. Joint Declaration of Data Citation Principles. https://www.force11.org/group/joint-declaration-data-citation-principles-final.

CoreTrustSeal. n. d. About. https://www.coretrustseal.org/about/.

Deutsche Forschungsgemeinschaft. 2019. Leitlinien zur Sicherung guter wissenschaftlicher Praxis. Kodex. https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf.

Diepenbroek, Michael, Hannes Grobe, Manfred Reinke, Reiner Schlitzer und Rainer Sieger. 1999. „Data management of proxy parameters with PANGAEA." In *Use of proxies in paleoceanography: examples from the South Atlantic*, hg. v. G. Fischer und G. Wefer, 715–727. Berlin, Heidelberg: Springer. https://epic.awi.de/id/eprint/637/.

Dorch, S. B. F., T. M. Drachen und O. Ellegaard. 2015. „The data sharing advantage in astrophysics." *arXiv* (8. November): 1511.02512 [astro-ph]. http://arxiv.org/abs/1511.02512.

Drachen, T. M., O. Ellegaard, A. V. Larsen und S. B. Fabricius Dorch. 2016. „Sharing data increases citations." *Liber Quarterly* 26 (2): 67–82. doi:10.18352/lq.10149.

ESSD. n. d. Aims and scope. Abgerufen am 26.11.2019 von: https://www.earth-system-science-data.net/about/aims_and_scope.html.

European Commission. 2016. Guidelines on FAIR Data Management in Horizon 2020. Abgerufen am 26.11.2019 von: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Fenner, Martin. 2019. DataCite's New Search. https://doi.org/10.5438/vyd9-ty64.

G8 Science Ministers. 2013. G8 Science Ministers Statement. https://www.gov.uk/government/news/g8-science-ministers-statement.

Google. n. d. Datensatz. https://developers.google.com/search/docs/data-types/dataset.

Hanson, Brooks, Kerstin Lehnert Kerstin Lehnert und Joel Cutcher-Gershenfeld. 2015. „Committing to Publishing Data in the Earth and Space Sciences." *Eos* 96 (15. Januar). doi:10.1029/2015EO022207.

Hyndman, Alan. 2018. „2018 year in review!" https://figshare.com/blog/2018_year_in_review_/464.

Kaden, Ben. 2016. „Drei Gründe für Forschungsdatenpublikationen." https://www2.hu-berlin.de/edissplus/2016/09/29/gruende-fuer-forschungsdatenpublikationen/.

Kindling, Maxi, Heinz Pampel, Stephanie van de Sandt, Jessika Rücknagel, Paul Vierkant, Gabriele Kloska, Michael Witt, Peter Schirmbacher, Roland Bertelmann und Frank Scholze. 2017. „The Landscape of Research Data Repositories in 2015: A re3data Analysis." *D-Lib Magazine* 23 (3/4). doi:10.1045/march2017-kindling.

Klump, Jens, Lesley Wyborn, Robert Downs, Ari Asmi, Mingfang Wu, Gerry Ryder und Julia Martin. 2020. „Principles and best practices in data versioning for all data sets big and small." Version 1.1. *Research Data Alliance*. doi:10.15497/RDA00042.

Klump, Jens. 2011. „Criteria for the Trustworthiness of Data Centres." *D-Lib Magazine* 17 (1/2). doi:10.1045/january2011-klump.

Klump, Jens, Roland Bertelmann, Jan Brase, Michael Diepenbroek, Hannes Grobe, Heinke Höck, Michael Lautenschlager, Uwe Schindler, Irina Sens und Joachim Wächter. 2006. „Data publication in the open access initiative." *Data Science Journal* 5: 79–83. doi:10.2481/dsj.5.79.

Klump, Jens, Robert Huber und Michael Diepenbroek. 2016. „DOI for geoscience data – how early practices shape present perceptions." *Earth Science Informatics* 9 (1): 123–136. doi:10.1007/s12145-015-0231-5.

Max Planck Society. 2003. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://oa.mpg.de.

McAdoo, Timothy. 2013. „How to Cite a Data Set in APA Style." https://blog.apastyle.org/apastyle/2013/12/how-to-cite-a-data-set-in-apa-style.html.

Nature. 2019. Reporting standards and availability of data, materials, code and protocols. https://www.nature.com/nature-research/editorial-policies/reporting-standards

PANGAEA. 2019. License. https://wiki.pangaea.de/wiki/License

Pampel, Heinz und Roland Bertelmann. 2011. „‚Data Policies' im Spannungsfeld zwischen Empfehlung und Verpflichtung." In: *Handbuch Forschungsdatenmanagement*, hg. v. Stephan Büttner, Hans-Christoph Hobohm, und Lars Müller, 49–61. Bad Honnef: Bock + Herchen. http://nbn-resolving.de/urn:nbn:de:kobv:525-opus-2287.

Pampel, Heinz, Hans-Jürgen Goebelbecker und Paul Vierkant. 2012. „re3data.org: Aufbau eines Verzeichnisses von Forschungsdaten-Repositorien. Ein Werkstattbericht." In: *Vernetztes Wissen – Daten, Menschen, Systeme. WissKom 2012*, hg. von Bernhard Mittermaier, 61–73. Jülich: Verlag des Forschungszentrums Jülich. http://hdl.handle.net/2128/4699.

Pampel, Heinz, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher und Uwe Dierolf. 2013. „Making Research Data Repositories Visible: The re3data.org Registry." Hg. von Hussein Suleman. *PLoS ONE* 8 (11/4. November): e78080. doi:10.1371/journal.pone.0078080.

Paskin, Norman. 2005. „Digital Object Identifiers for scientific data." *Data Science Journal* 4: 12–20. doi:10.2481/dsj.4.12.

Piwowar, Heather A. und Todd J. Vision. 2013. „Data reuse and the open data citation advantage." *PeerJ* 1 (1. Oktober): e175. doi:10.7717/peerj.175.

Piwowar, Heather A., Roger S. Day und Douglas B. Fridsma. 2007. „Sharing Detailed Research Data Is Associated with Increased Citation Rate." Hg. von John Ioannidis. *PLoS ONE* 2 (3/21. März): e308. doi:10.1371/journal.pone.0000308.

PLOS. 2019. Data Availability. https://journals.plos.org/plosbiology/s/data-availability

Rauber, Andreas, Asmi, Ari, van Uytvanck, Dieter, and Proell, Stefan. 2015. „Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)." *Zenodo*. doi:10.15497/RDA00016.

Rücknagel, J., P. Vierkant, R. Ulrich, G. Kloska, E. Schnepf, D. Fichtmüller, E. Reuter, u. a. 2015. Metadata Schema for the Description of Research Data Repositories. Version 3.0. doi:10.2312/re3.008.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt und Ilene Karsch-Mizrachi. 2019. „GenBank." *Nucleic Acids Research* 47 (D1/8. Januar): D94–D99. doi:10.1093/nar/gky989.

Shepherdson, John, und Thiel, Carsten. 2018. „The New CESSDA Data Catalogue." *Zenodo*. December 5. doi:10.5281/zenodo.2530106.

Smith, David und Anthony Carrano. 1996. International Large-Scale Sequencing Meeting. Human Genome News 6, Nr. 7. http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n6/19intern.shtml.

Springer Nature. 2019. About. https://www.nature.com/sdata/about

Stall, Shelley, Lynn Yarmey, Reid Boehm, Helena Cousijn, Patricia Cruse, Joel Cutcher-Gershenfeld, Robin Dasler, u. a. 2018. „Advancing FAIR Data in Earth, Space, and Environmental Science." *Eos 99* (5. November). doi:10.1029/2018EO109301.

The University of Chicago Press Editorial Sta. 2017. The Chicago Manual of Style, 17th Edition. University of Chicago Press. doi:10.7208/cmos17.

Topoi. n. d. Digital Pantheon. http://repository.edition-topoi.org/collection/BDPP.

Voigt, C., C. Förste, H. Wziontek, D. Crossley, B. Meurers, V. Pálinkáš et al. 2016. „Report on the Data Base of the International Geodynamics and Earth Tide Service (IGETS)." *Scientific Technical Report STR – Data* 16/08. doi:10.2312/GFZ.B103-16087.

Wellcome Trust. 2003. Sharing data from largescale biological research projects. A system of tripartite responsibility. http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. 2016. „The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3 (1/Dezember): 160018. doi:10.1038/sdata.2016.18.

Wziontek, Hartmut, Peter Wolf, Ilona Nowak, Bernd Richter, Axel Rülke und Herbert Wilmes. 2017. „Superconducting Gravimeter Data from Bad Homburg – Level 1." BKG Federal Agency for Cartography and Geodesy. doi:10.5880/IGETS.BH.L1.001.

Zenodo. n. d. About Zenodo. https://about.zenodo.org.