Tanja Friedrich und Jonas Recker

5.1 Auffindbarkeit und Nutzbarkeit von Daten

Abstract: In diesem Beitrag wird thematisiert, wie und wo Daten für die Forschung auffindbar sind und welche Faktoren darüber entscheiden, ob und wie Forschende Daten im Rahmen ihrer eigenen Arbeit oder zu Replikationszwecken nutzen können. Hierbei stehen verschiedene Agierende mit ihren jeweiligen Perspektiven im Fokus der Betrachtung – neben den Datennutzenden auch die Datenproduzierenden und die Informationsinfrastrukturen, die Forschungsdaten speichern und zugänglich machen. Aufbauend auf einer Analyse des Status quo gibt das Kapitel praktische Empfehlungen, wie die Auffindbarkeit und Nutzbarkeit von Forschungsdaten verbessert werden können.

Einleitung

Die Auffindbarkeit und (Nach-)Nutzbarkeit von Forschungsdaten zu verbessern ist erklärtes Ziel diverser Initiativen zum Aufbau und zur Förderung von Forschungsinfrastrukturen.[1] Findability und reusability gehören außerdem zu den 2014 entwickelten FAIR Guiding Principles for scientific data management and stewardship.[2] Diese Prinzipien sind aktuell Gegenstand verschiedener Initiativen zur Verbesserung von Forschungsdateninfrastrukturen für die Wissenschaft.[3]

Angesichts weiter wachsender Datenmengen und steigender Diversifizierung von Datenarten wird das Auffinden geeigneter Daten zur Nachnutzung zunehmend zur Herausforderung.[4] Das Problem, keine geeigneten Forschungsdaten zu finden, ist einer der wichtigsten Hinderungsgründe für die Nachnutzung von Daten.[5] Die Nutzung aufgefundener Daten in der eigenen Forschung oder zum Zweck der Replikation von publizierten Forschungsergebnissen ist ebenfalls mit Herausforderungen verbunden. Diese ergeben sich aus technischen, rechtlichen und ethischen Rahmenbedingungen sowie aus der intellektuellen Zugänglichkeit und Verstehbarkeit der Daten.

Im ersten Teil dieses Kapitels wird der Status quo der Auffindbarkeit von Forschungsdaten aus den Perspektiven der Datenproduzierenden, der Nutzenden und

1 Vgl. Kommission Zukunft der Informationsinfrastruktur 2011; Wissenschaftsrat 2012; European Commission 2016; Deutsche Initiative für Netzwerkinformation 2018.
2 Vgl. Wilkinson et al. 2016.
3 S. Beitrag von Linne et al., Kap. 3.2 in diesem Praxishandbuch.
4 Vgl. Gregory et al. 2018, 1.
5 Vgl. Shen 2015, 172.

https://doi.org/10.1515/9783110657807-023

der Infrastruktur beschrieben.[6] Dabei werden aktuelle Probleme dargestellt und es werden Empfehlungen zur Verbesserung der Auffindbarkeit gemacht. Der zweite Abschnitt behandelt die Nutzbarkeit von Forschungsdaten unter Berücksichtigung der genannten Dimensionen (Technik, Recht, Ethik, Zugänglichkeit, Verstehbarkeit), die die Nutzbarkeit bedingen. Auch hier werden Empfehlungen für eine Verbesserung und Weiterentwicklung auf Grundlage des Status quo gemacht.

1 Auffindbarkeit von Daten

Die Ursachen für Probleme beim Auffinden von Daten sind vielfältig. Eine Herausforderung ist z. B. die breit verteilte Datenhaltung. Daten entstehen in den unterschiedlichsten Kontexten: in Unternehmen und sozialen Medien, in Behörden und in der Wissenschaft.[7] Selbst bei isolierter Betrachtung des wissenschaftlichen Bereichs wird die breite Dispersion von Forschungsdaten deutlich: Sie sind verteilt auf zahlreiche disziplinäre oder interdisziplinäre, institutionelle und institutionsübergreifende Datenrepositorien.[8]

Der weitaus größte Teil theoretisch nachnutzbarer Daten ist schon allein deshalb nicht auffindbar, weil die Datenproduzierenden diese nicht verfügbar machen. Die Motivation, die in einem Forschungsprojekt erhobenen Daten mit der Forschungscommunity zu teilen, ist nach wie vor zu gering.[9] Vorgaben seitens Forschungsförderungseinrichtungen und Anreizsysteme sollen Forschende zum Data Sharing motivieren, bisher jedoch mit unzureichendem Erfolg.[10]

Data Sharing spielt auch für Unternehmen eine große Rolle. Allerdings haben diese naturgemäß kein Interesse daran, ihre Daten für die breite Öffentlichkeit zugänglich zu machen, sondern beschränken ihre Aktivitäten auf das Teilen ihrer Daten mit Kunden und Geschäftspartnern.[11] Offene Verwaltungsdaten (Open Government Data) wiederum stehen gemäß Gesetzgebung[12] zunehmend der Öffentlichkeit zur Verfügung. Die Auffindbarkeit dieser Daten ist allerdings bislang durch vielfach unzureichende Beschreibung mit Metadaten erschwert.[13]

6 S. a. Beitrag von Henrich, Gradl und Jegan, Kap. 5.2 in diesem Praxishandbuch für eine technischere Perspektive zum Data Retrieval.
7 S. a. Beiträge zu den entsprechenden Datenmärkten in Kap. 3 in diesem Praxishandbuch.
8 Vgl. Rat für Informationsinfrastrukturen 2019, 44 f.
9 Vgl. Fecher und Puschmann 2015, 146.
10 Vgl. Tenopir et al. 2015, 4.
11 Vgl. Fedkenhauer et al. 2017, 11.
12 Vgl. E-Government-Gesetz vom 25. Juli 2013 (BGBl. I S. 2749), das zuletzt durch Artikel 15 des Gesetzes vom 20. November 2019 (BGBl. I S. 1626) geändert worden ist.
13 Vgl. Chapman et al. 2019, 258.

Generell lässt sich festhalten, dass es unabhängig von Sektor oder Disziplin der Datenentstehung häufig an einer einheitlichen Dokumentation und einer nachnutzungsorientierten Inhaltsbeschreibung mangelt.[14] Ursächlich sind hier nicht nur die unterschiedlichen Kontexte in denen die Daten entstehen, sondern auch das Fehlen genauer Kenntnisse der Bedürfnisse und Praktiken bei der Datensuche und daraus ableitbarer Dokumentationsstandards.

Systemperspektive: Wo Daten aufzufinden sind

Daten entstehen im Kontext und als Bezugspunkte zur Wirklichkeit.[15] Viele Daten werden durch geplante Beobachtung erhoben, andere entstehen als Beiprodukt menschlichen (Online-)Verhaltens oder automatisierter Prozesse. Ob und wie Daten auffindbar sind, hängt zunächst mit ihrer Verfügbarkeit zusammen. Bei weitem nicht alle Daten sind überhaupt zur Nachnutzung bestimmt, weshalb viele Daten überhaupt nicht auffindbar gemacht werden oder sind. In der Wissenschaft produzierte Daten werden immer häufiger verfügbar gemacht, nicht zuletzt weil dies zunehmend Voraussetzung für Forschungsförderung ist (siehe z. B. die Leitlinien zur Sicherung guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft[16]). Doch Richtlinien allein führen nicht zwangsweise dazu, dass Forschungsdaten auch tatsächlich veröffentlicht werden.[17] In manchen Fällen stehen der Datenveröffentlichung sogar rechtliche, ethische oder zuweilen auch praktische Gründe entgegen.[18]

Eine verteilte, informelle Datenhaltung steht der professionellen Datenarchivierung zum Zweck der Nachnutzung gegenüber.[19] Entstehen in der Forschung Daten, werden sie in der Regel zwar lokal oder in kleinen Netzwerken gespeichert oder sogar mit anderen geteilt. Eine langfristige Archivierung in einem der breiteren Forschungsgemeinschaft zugänglichen Repositorium findet jedoch nur zum Teil statt. Selbst wenn lokal gehaltene Daten auf persönlichen oder institutionellen Webservern zur Verfügung gestellt werden, entziehen sie sich der Auffindbarkeit (ganz abgesehen von der Problematik der langfristigen Sicherung der Daten). Repositorien reichern archivierte Daten mit Metadaten an, die die Daten mit für die Nachnutzung relevanten Informationen mehr oder weniger ausführlich beschreiben. Dadurch werden Daten im Katalog des jeweiligen Repositoriums, aber auch in Metasuchpor-

14 Vgl. Fecher und Puschmann 2015, 149.
15 Vgl. Borgman 2015, 17 f.
16 Vgl. Deutsche Forschungsgemeinschaft 2019 und hierzu den Beitrag von Putnings, Kap. 1.3 in diesem Praxishandbuch.
17 Vgl. Borgman 2015, 206.
18 S. Abschnitt 2.2 in diesem Beitrag.
19 Vgl. Kitchin 2014, 29 f.

talen (z. B. DataCite[20]) und in Datensuchmaschinen (z. B. Google dataset search[21]) auffindbar.

Die organisatorischen und technischen Voraussetzungen für die Auffindbarkeit von Daten leisten verschiedene Infrastrukturangebote. Für die akademische Forschung sind besonders Repositorien für das Auffinden von Daten relevant, denn diese sind gleichzeitig Such- und Speicherort archivierter Daten. Der Forschung stehen zahlreiche institutionengebundene (zum Beispiel für Mitglieder einer Universität) oder institutionsübergreifende Repositorien (zum Beispiel für Forschende einer bestimmten Disziplin) zur Verfügung. Forschungsdatenrepositorien (FDR) können über entsprechende Suchdienste, zum Beispiel das Registry of Research Data Repositories (re3data)[22] aufgefunden werden. Fachspezifische Repositorien werden häufig von außeruniversitären Forschungseinrichtungen bereitgestellt, z. B. das geo- und umweltwissenschaftliche Repositorium Pangaea,[23] das vom Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) und dem Zentrum für Marine Umweltwissenschaften (MARUM) betrieben wird. An Universitäten entstehen immer häufiger fächerübergreifende Repositorien, die nach Daten der Universitätsangehörigen durchsucht werden können. Insbesondere im Bereich der fächerübergreifenden Repositorien gibt es auch Initiativen von Non-Profit-Organisationen (z. B. Dryad,[24] Open Science Framework[25]) oder von kommerziellen Anbietern (z. B. Figshare[26]), deren Bestände wesentlich größer sind als beispielsweise die der einzelnen universitären Anbieter.

Die bestehende, breit verteilte Archivierung von Forschungsdaten bedeutet, dass die Suche nach Forschungsdaten häufig sehr aufwendig ist. Es sind daher einige Dienste entstanden, die Forschungsdatenbestände aggregiert nachweisen, z. B. das Angebot GFBio,[27] das vom Konsortium German Federation for Biological Data betrieben wird und die übergreifende Suche in den Beständen von neun Datenzentren aus dem Bereich der Biologie ermöglicht. Das fächerübergreifende Dataverse-Projekt der Harvard University[28] hat eine Software entwickelt, mit der Einrichtungen weltweit Repositorien einrichten können, deren Bestände im Harvard Dataverse gemeinsam durchsucht werden können. Eine fächerübergreifend föderierte Suche bietet das internationale Konsortium DataCite an, in dem alle Datensätze auffindbar

20 S. https://datacite.org. Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.
21 S. https://toolbox.google.com/datasetsearch.
22 S. https://re3data.org.
23 S. https://www.pangaea.de.
24 S. https://datadryad.org.
25 S. https://osf.io.
26 S. https://figshare.com.
27 S. https://www.gfbio.org.
28 S. https://dataverse.org.

sind, die durch eine DataCite-Mitgliedsinstitution einen registrierten Digital Object Identifier (DOI) erhalten haben.

Auch wissenschaftliche Zeitschriftenverlage haben ein Interesse am Auf- und Ausbau von Forschungsdateninfrastrukturen. Sie fordern in der Regel, dass die einer Veröffentlichung zugrundeliegenden Daten mindestens zum Zweck des Peer-Reviews, idealerweise aber auch zugänglich für die Wissenschaft zur Verfügung gestellt werden. Einige Verlage bieten sogar eigene Repositorien an (z. B. das *Journal of Cell Biology*[29]), während andere mit existierenden Repositorien kooperieren, in denen die zu den Publikationen gehörenden Daten veröffentlicht werden (z. B. kooperiert die *Zeitschrift für Soziologie* mit dem Replikationsserver SowiDataNet|datorium[30]). Die meisten Verlage geben generelle Auswahlkriterien für ein Repositorium oder direkt eine ganze Liste von Repositorien an, bei denen Daten archiviert werden können. Ein in der Verlagswelt noch relativ neues Phänomen sind dezidierte Datenzeitschriften (Data Journals), in denen nicht die Beschreibung von Forschungsergebnissen, sondern von Forschungsdaten im Vordergrund steht (z. B. das *Research Data Journal for the Humanities and Social Sciences*[31]). Diese Zeitschriften ermöglichen die Veröffentlichung von Daten durch ein Peer-Review-Verfahren. Welche Relevanz Datenzeitschriften für das Auffinden von Forschungsdaten haben, bleibt noch abzuwarten. Einige Datenzeitschriften wurden bereits wieder eingestellt. Verlage bieten aber inzwischen auch generische Möglichkeiten der Datensuche an, zum Beispiel in Form des von Elsevier betriebenen Mendeley Data.[32] Auch hier werden Metadaten aus verschiedenen Repositorien (z. B. Dryad, Pangaea) aggregiert und durchsuchbar gemacht. Der Datenbankanbieter Web of Science bietet mit dem Clarivate Data Citation Index[33] eine ähnliche, allerdings kostenpflichtige Datensuche an, die die einzelnen Datensatznachweise um in anderen Datenbanken des Web of Science enthaltene Literaturnachweise ergänzt, die diese Datensätze zitieren. Seit 2018 bietet auch Google unter dem Namen Dataset Search einen Dienst an, der darauf ausgerichtet ist, möglichst alle im World Wide Web verfügbaren Forschungsdaten als solche zu identifizieren, zu indexieren und so durchsuchbar zu machen.

Alle aktuellen Bemühungen, die in Richtung einer umfassenden Durchsuchbarkeit aller Forschungsdatenbestände gehen, müssen grundlegende Herausforderungen adressieren. Diese reichen von ganz grundlegenden Problemen wie der Frage, welche Objekte überhaupt als Forschungsdaten zu identifizieren sind, bis hin zum Problem mangelnder Standardisierung der Metadaten, die von den datenhaltenden Stellen generiert werden.

29 S. https://rupress.org/jcb.
30 S. https://www.gesis.org/replikationsserver/home.
31 S. https://brill.com/view/journals/rdj/rdj-overview.xml.
32 S. https://data.mendeley.com/research-data/.
33 S. https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index.

Neben den an Universitäten und außeruniversitären Einrichtungen entstehenden Forschungsdaten spielen offene Verwaltungsdaten (Open Government Data) eine wichtige Rolle für Forschende aus allen Sektoren, aber auch für Privatpersonen.[34] Organisationen wie die Open Knowledge Foundation (OKF)[35] fordern öffentliche Verwaltungen daher auf, ihre Daten, soweit rechtlich möglich, einer breiten Öffentlichkeit zur Nutzung zugänglich zu machen. Die OKF hat unter anderem die Entwicklung der Software CKAN[36] vorangetrieben, die weltweit von Institutionen der öffentlichen Verwaltung genutzt wird, um Verwaltungsdaten zugänglich zu machen, in Deutschland zum Beispiel vom Datenportal Govdata.[37] Aktuell stehen offene Verwaltungsdaten noch nicht in ausreichendem Umfang zur Verfügung und sind vielfach nicht in einer Art und Weise beschrieben, dass sie über Portale wie CKAN oder in anderen Kontexten leicht auffindbar sind.[38]

Unabhängig von den genannten Initiativen existieren für und in Unternehmen ganz andere Infrastrukturen für das Datenmanagement. Die dort zum Einsatz kommenden Data Warehouses[39] müssen abweichende Voraussetzungen erfüllen als FDR in der Wissenschaft. Für Unternehmen ist zwar auch wichtig, dass die Infrastruktur effizientes Data Sharing ermöglicht, sei es intern, mit Kunden oder anderen Unternehmen,[40] Data Warehouses müssen aber gleichzeitig sicherstellen, dass zu keiner Zeit ein unautorisierter Zugriff auf die Daten möglich ist. Sie müssen sowohl höchst interoperabel und effizient arbeiten, als auch Kunden- und Unternehmensdaten nach außen sichern und schützen.

Nutzendenperspektive: Wie nach Daten gesucht wird

Nachnutzbare Forschungsdaten stehen trotz der bestehenden Lücken und trotz vorhandener Qualitätsunterschiede in großer Fülle zur Verfügung. Auffindbar sind diese Daten, wie beschrieben, über eine Vielzahl digitaler Dienste, von Repositorien über digitale Zeitschriften bis hin zu Datensuchmaschinen. Dennoch werden potentiell passende Daten häufig nicht gefunden.[41]

Jegliche Vorhaben, Infrastrukturdienste im Sinne der Auffindbarkeit der Daten zu optimieren, sollten berücksichtigen, wie und wo Nutzende tatsächlich nach Daten suchen. Wie die Forschung zu dieser Frage zeigt, erfahren sie von geeigneten

34 S. Beitrag von Schieferdecker, Kap. 2.3 in diesem Praxishandbuch.
35 S. https://okfn.de/.
36 S. https://ckan.org.
37 S. https://www.govdata.de.
38 Vgl. Chapman et al. 2019, 258.
39 Vgl. Bauer et al. 2013, 5.
40 Vgl. Fedkenhauer et al. 2017, 17.
41 Vgl. Rat für Informationsinfrastrukturen 2019, 44.

Daten üblicherweise von anderen Forschenden. Zum Beispiel besuchen sie Konferenzen oder andere Veranstaltungen, wo sie auf Datenquellen stoßen, die ihnen vorher unbekannt waren. Bei diesen und anderen Gelegenheiten führen sie Gespräche über Daten („data talk“[42]), um auf dem Laufenden zu bleiben. Grundlegendes Wissen über relevante Studien und Datenquellen eignen sich Forschende bereits in der akademischen Ausbildung an. Für die empirische Sozial- und Wirtschaftsforschung werden beispielsweise kontinuierlich Mikrodaten aus seit Jahrzehnten laufenden Umfrageprogrammen wie der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften[43] oder des Sozio-ökonomischen Panels[44] bereitgestellt, mit denen Studierende dieser Fachgebiete ihre Ausbildung durchlaufen. Daten wie diese werden gezielt für eine breite Nachnutzung erhoben, aufbereitet und zur Verfügung gestellt. Forschende, die mit diesen Daten arbeiten, müssen nicht erst danach suchen, sondern wissen um die Bestände und kennen die Zugangsmöglichkeiten. Mit wachsender Erfahrung und Einbindung in Forschungscommunities lernen sie die Datenlandschaft immer besser kennen, was bei der Suche nach Daten ebenfalls hilfreich ist. Idealerweise sollten sich Forschende ein solides Wissen über Datenbestände in ihrem Fachgebiet aneignen, um schon bei der Entwicklung von Forschungsfragen einschätzen zu können, ob die zur Beantwortung dieser Fragen notwendigen Daten überhaupt vorhanden und zugänglich sind.[45]

Neue Studien heben die besondere Bedeutung sozialer Kontakte bei der Suche nach Daten hervor.[46] Der Einfluss dieser Kontakte ist dabei nicht auf das Auffinden von Daten beschränkt, sondern zeigt sich besonders im Hinblick auf die Problemlösungspraxis bei der Nachnutzung der Daten.[47] Persönliche Kommunikation ist gängige Praxis in Suchprozessen, Lernprozessen und Problemlösungsprozessen für Forschende unterschiedlicher Disziplinen.[48]

Insbesondere fortgeschrittene Forschende durchsuchen außerdem gezielt und regelmäßig die Literatur ihres Fachgebiets, in der Regel bestimmte Zeitschriften, nach Hinweisen auf zur Nachnutzung geeignete Datensätze.[49] Die aktuelle Forschung legt nahe, dass Fachliteratur die wichtigste Quelle bei der Datensuche ist, unabhängig von der Disziplin.[50] Für diese Suche nach Daten über Literatur ist es hilfreich, wenn die Daten in den Texten unter Verwendung von persistenten Identi-

42 Yoon 2017, 465.
43 S. https://www.gesis.org/allbus/allbus.
44 S. https://www.diw.de/soep.
45 Vgl. Zimmerman 2007, 6.
46 Vgl. Yoon 2017, 469; Gregory et al. 2019b, 429.
47 S. Abschnitt 2.2 in diesem Beitrag.
48 Vgl. Yoon 2017, 463.
49 Vgl. Zimmerman 2007, 13; Gregory et al. 2018, 1.
50 Vgl. Gregory et al. 2019b, 421.

fiern (z. B. DOI) zitiert werden und dadurch leicht aufgefunden werden können.[51] Generell profitieren Forschende bei der Suche nach Daten von einer allgemeinen Vertrautheit mit Forschungstrends und entsprechender Literatur.[52]

Die Websuche spielt daneben eine immer größere Rolle für Personen, die auf der Suche nach Daten sind.[53] Vielen sind dagegen relevante Repositorien unbekannt.[54] Erfahrene Forschende nutzen zwar die vorhandenen Repositorien für die Suche nach Forschungsdaten, vor allem diejenigen, die in ihrer Disziplin als besonders wichtig gelten.[55] Angesichts der verteilten Datenlandschaft wünschen sie sich aber eine zentrale Suchmöglichkeit nach Daten.[56]

Die Bedeutung von Datenkatalogen und Repositorien für die Suche nach Forschungsdaten liegt weniger im Angebot eines Sucheinstiegs. Sie besteht vielmehr darin, dass dort die Forschungsdaten näher beschrieben werden, idealerweise im Kontext der Forschungsprojekte, in denen sie entstanden sind und mit Hinweisen zu relevanter Literatur oder weiteren Daten. Die Qualität der Datendokumentation spielt für die Datensuche eine entscheidende Rolle. Die Dokumentation muss interessierten Nutzenden ermöglichen, die im Datensatz enthaltenen Informationen zu verstehen.[57] Wichtig für die Relevanzbeurteilung ist außerdem, dass Kontextinformationen mitgegeben werden.[58] Für Forschende aus den Sozialwissenschaften konnte gezeigt werden, dass das Lesen der Dokumentation und die Relevanzbeurteilung wesentliche Schritte bei der Datensuche sind, auf die viel Zeit verwendet wird.[59]

Die Interaktion mit Datenbanken und Suchmaschinen bei der Datensuche ist noch unzureichend erforscht. Bei der Entwicklung dieser Dienste wird von einer Schlagwort- oder Stichwortsuche ausgegangen, die auch durch Einbindung entsprechender Terminologie unterstützt werden kann. Für spezifische Dienste gibt es darüber hinaus nicht-textuelle Sucheingabemöglichkeiten, zum Beispiel das Zeichnen chemischer Strukturformeln über eine Eingabemaske.[60] Häufig werden auch Filtermöglichkeiten zur Eingrenzung der Suchergebnisse angeboten.[61] Die eingesetzten Terminologien und Technologien beruhen größtenteils auf Kenntnissen zum Informationsverhalten bei der Literatursuche. Das Wissen darüber, wie inhaltliche Be-

51 S. Beitrag von Pampel und Elger, Kap. 5.6 in diesem Praxishandbuch.
52 Vgl. Zimmerman 2007, 6.
53 Vgl. Gregory et al. 2018, 1.
54 Vgl. Fecher und Puschmann 2015, 149.
55 Vgl. Gregory et al. 2019a, 10.
56 Vgl. Gregory et al. 2019b, 428.
57 S. Beitrag von Dierkes, Kap. 4.1 in diesem Praxishandbuch.
58 S. Abschnitt 2.1 in diesem Beitrag.
59 Vgl. Kern und Mathiak 2015, 203.
60 S. z. B. die chemische Struktursuche von Fisher Scientific, https://www.fishersci.de/de/de/search/chemical/substructure.html.
61 Vgl. Chapman et al. 2019, 252.

schreibungen von Forschungsdaten aussehen müssen, damit Nutzende sie finden können, ist noch unzureichend.[62]

Kathleen Gregory et al. (2018) geben hilfreiche Hinweise nicht nur für die Suche nach Datenquellen, sondern auch in den Datenbanken und Datenrepositorien selbst. Unter anderem weisen sie darauf hin, dass Repositorien hilfreiche erweiterte Suchfunktionen bieten, mit denen sich Suchende näher auseinandersetzen sollten („Make the repository work for you“[63]). Um erfolgreich zu suchen, sei ein strategisches Vorgehen hilfreich, zum Beispiel indem eine bewusste Entscheidung für ein disziplinspezifisches oder generisches Repositorium getroffen werde.[64] Funktionalitäten, wie thematisches Browsing und Filtermöglichkeiten zu nutzen, wird ebenfalls als erfolgversprechend empfohlen.[65] Darüber hinaus gelte es bei der Suche nach Daten nicht nur die inhaltliche Passung zu beurteilen, sondern auch Kriterien wie räumliche und zeitliche Relevanz und Daten- sowie Metadatenqualität.[66]

Infrastrukturperspektive: Dokumentation und Standardisierung

Das Kernproblem der Auffindbarkeit von nachnutzbaren Daten besteht in der Inkongruenz zwischen verfügbaren Daten einerseits und den benötigten und auffindbaren Daten andererseits.[67] Die Digitalisierung hat die vielzitierte Datenflut („data deluge“[68]) ausgelöst, die unendlich viele Nutzungsmöglichkeiten verspricht. In Gestalt von Big Data entstehen aus digitalen Prozessen und digitalem Verhalten permanent neue Daten, auf deren Erhebung Forschende keinen Einfluss nehmen können. Auch die durch die Open-Government-Bewegung immer umfangreicher zur Verfügung stehenden offenen Verwaltungsdaten tragen zum aktuellen Datenreichtum bei.

Die alleinige Masse an vorhandenen Daten bedeutet allerdings nicht, dass genügend Daten zu allen Fragestellungen zur Verfügung stehen. Gerade in Bezug auf Big Data und offene Verwaltungsdaten, die in der Regel ohne forschungstheoretische Einordnung entstehen, bewegen sich Datensuchende in einem Bereich „dunkler Materie“,[69] wenn die Daten nicht mit geeigneten Mitteln auffindbar gemacht werden. Die Notwendigkeit spezifischen Datenmanagements besteht aber auch für die Auffindbarkeit der im Rahmen geplanter Forschungsprojekte erhobener Daten

62 Vgl. Chapman et al. 2019, 261.
63 Vgl. Gregory et al. 2018, 3.
64 Vgl. Gregory et al. 2018, 3.
65 Vgl. Gregory et al. 2018, 3 f.
66 Vgl. Gregory et al. 2018, 4.
67 Vgl. Chapman et al. 2019, 252.
68 Gray 2007, xxx.
69 Borgman 2015, 241.

(„small data“[70]). Diese weisen theoretische und ontologische Bezüge auf, die bereits in der Erhebung der Daten grundlegend angelegt sind. Dadurch sind diese Daten in ihren Analysemöglichkeiten aufgrund ihres Erhebungskontexts beschränkt. Wenn diese Daten auffindbar gemacht werden sollen, stellt sich die grundlegende Herausforderung, dass sie aus ihren Entstehungskontexten isoliert werden (das sog. Mobilitätsproblem[71]). Der Kontext der Datenentstehung ist für die Interpretierbarkeit und Nachnutzbarkeit aber von elementarer Bedeutung.[72]

Gut auffindbar sind Datensätze, wenn sie entsprechend dokumentiert sind, persistente Identifier nutzen, in verschiedenen Formaten vorliegen, in offen zugänglichen Repositorien zur Verfügung stehen und nicht nur dort recherchiert werden können.[73] Beim Vorhaben, Daten auffindbar zu machen, müssen daher andere Vorgehensweisen, Methoden und Instrumente zur Anwendung kommen, als etwa in Bezug auf Zeitschriftenartikel.[74] Dabei kommt der standardisierten Dokumentation durch Datenzentren, -archive und -repositorien besondere Bedeutung zu.[75] Metadatenstandards zur Dokumentation von Forschungsdaten müssen die besonderen Eigenschaften des Informationsträgers berücksichtigen, die sich idealerweise aus den Informationsbedürfnissen der Nutzenden ableiten lassen. Während eine schlagwort- oder volltextbasierte Indexierung bei Textdokumenten zufriedenstellende Ergebnisse liefert, müssen für die Datensuche Relevanzkriterien jenseits thematischer Passung berücksichtigt werden: Aktualität, Zugangsmöglichkeiten, Versionierung, Datenqualität, Erhebungsmethoden, Provenienz und Untersuchungsbereich gehören zu den disziplinunabhängig relevanten Kriterien bei der Datenauswahl.[76]

Über Disziplingrenzen hinweg unterscheiden sich Standards zur Datendokumentation teilweise stark. Zu unterschiedlich sind die Datenarten und Datenformate, die in den jeweiligen Disziplinen verwendet werden. Es ist daher bei der Dokumentation und Archivierung von Daten gleichermaßen notwendig, die disziplinspezifischen Standards im eigenen Bereich zu beachten und auf übergeordneter Ebene Metastandards zu bedienen oder wo nötig zu entwickeln. Die disziplinspezifischen Standards sind vor allem für die Dokumentation und Bereitstellung der Daten in Fachrepositorien notwendig. Beispielsweise sollten sozialwissenschaftliche Umfragedaten unter Anwendung gängiger Standards der Data Documentation Initiative (DDI)[77] repräsentiert werden. Textdaten in Infrastrukturumgebungen für die Geisteswissenschaften benötigen z. B. eine Repräsentation gemäß Text Encoding Initiative

70 Kitchin 2014, 27.
71 Vgl. Borgman 2015, 219.
72 S. Abschnitt 2.2 in diesem Beitrag.
73 Vgl. Mannheimer et al. 2016, 6.
74 Vgl. Chapman et al. 2019, 259.
75 Vgl. Chapman et al. 2019, 260.
76 Vgl. Chapman et al. 2019, 260.
77 S. https://ddialliance.org.

(TEI).[78] Auch für die Annotation zum Zweck der Inhaltsbeschreibung sollten Terminologien aus den jeweiligen Fachgebieten verwendet werden, für Experimentaldaten aus der Biologie z. B. Gene Ontology (GO)[79] oder für diverse Daten aus den Geowissenschaften z. B. der GeoRef Thesaurus.[80] Im Hinblick auf Auffindbarkeit in anderen disziplinären Kontexten ist die Verwendung von Vokabularen aus der Linked Open Data Cloud (LOD cloud)[81] sinnvoll, denn hier bestehen Verknüpfungen zwischen Terminologien unterschiedlicher Fachgebiete.

Da die Websuche nach Daten an Bedeutung gewinnt, genügt es nicht, die Daten nur für die Auffindbarkeit in Repositorien aufzubereiten. Damit sie von Suchmaschinen wie Google als Forschungsdaten indexiert werden können, ist der Einsatz von disziplinübergreifenden Standards wie schema.org-Vokabularen,[82] W3C Semantic-Web-Standards[83] und Sitemaps[84] notwendig.[85] Für offene Verwaltungsdaten ist insbesondere der W3C-Standard DCAT[86] von Bedeutung (z. B. auch Grundlage des Metadatenmodells OGD[87] von GovData).

Da aktuelle Erkenntnisse darauf hinweisen, dass neben Dokumentationsqualität auch forschungsdatenbezogene Literatur und Forschungsdatencommunities für Nutzende eine wichtige Rolle beim Auffinden von Forschungsdaten spielen, sollte die Auffindbarkeit auch in diesen Kontexten unterstützt werden. Zum einen sollten Initiativen zur Verknüpfung von Literatur mit Datensätzen (z. B. Scholix[88]) weiter vorangetrieben werden. Auch Dienste wie der Clarivate Data Citation Index kommen diesem spezifischen Suchverhalten der Datennutzenden entgegen. Zum anderen sollte der Austausch über Daten innerhalb der Forschungsdatencommunities wo möglich unterstützt werden, z. B. durch datenorientierte Workshops bei relevanten Konferenzen und perspektivisch über die fachlichen NFDI-Konsortien.[89]

78 S. https://tei-c.org.
79 S. http://geneontology.org.
80 S. https://www.americangeosciences.org/information/georef/thesaurus.
81 S. https://lod-cloud.net.
82 S. https://schema.org.
83 S. https://www.w3.org/standards/semanticweb.
84 S. https://www.sitemaps.org.
85 Vgl. Wu et al. 2019, 9 f.
86 S. https://www.w3.org/ns/dcat.
87 S. https://www.govdata.de/standardisierung.
88 S. http://www.scholix.org.
89 S. https://www.dfg.de/nfdi.

2 Nutzbarkeit

Häufig wird in der Literatur die „primäre“ Nutzung von Forschungsdaten durch die erhebenden Wissenschaftlerinnen und Wissenschaftler von einer als „sekundär“ bezeichneten Nachnutzung unterschieden. Bei genauerer Betrachtung erweist sich eine eindeutige definitorische Abgrenzung der beiden Nutzungsarten als durchaus komplex, wenn nicht sogar unmöglich.[90] So stellen van de Sandt et al. fest, dass weder der Charakter der Daten, noch die Nutzenden, der Nutzungszweck oder der Zeitpunkt ein zuverlässiger Indikator für die Unterscheidung von Nutzung und Nachnutzung sind.[91]

Trotz dieser Einschränkungen wird im Folgenden der Fokus primär auf der Nachnutzung von Forschungsdaten liegen, verstanden als eine Nutzung von Daten, die für einen bestimmten Zweck erhoben wurden, konkret zur Beantwortung von Forschungsfragen jenseits des ursprünglichen (Forschungs-)Zwecks.[92] Dies schließt die Nutzung von Daten ein, die nicht primär zum Zweck der Forschung erhoben wurden, wie beispielsweise digitale Verhaltensdaten oder Daten der amtlichen Statistik. Von der Nachnutzung im Sinne der gegebenen Definition ist die Nutzung von Forschungsdaten zum Zweck der (direkten) Replikation zu unterscheiden, die dem „Nachweis der Replizierbarkeit eines bestimmten Forschungsergebnisses unter unabhängigen Bedingungen“ dient.[93] Auch diese Art der Nutzung von Forschungsdaten wird im Folgenden betrachtet werden, wo relevant.

Ob Forschungsdaten außerhalb des originären Projektkontexts genutzt werden können, hängt von einer Reihe von Faktoren ab, die der technischen, der ethisch-rechtlichen sowie der Dimension der intellektuellen Zugänglichkeit zugeordnet werden können. Diese werden im Folgenden näher beleuchtet, bevor Vertrauen der Forschenden in die genutzten Daten als ein weiterer wichtiger, nutzungsentscheidender Faktor betrachtet wird.

2.1 Dimensionen der Nutzbarkeit

In der *technischen Dimension* hängt die Nutzbarkeit von Forschungsdaten wesentlich von der Verfügbarkeit geeigneter Hard- und Softwareumgebungen ab. Einerseits kann die Form der Speicherung und Bereitstellung der Daten ihre Nutzbarkeit beeinflussen. Müssen Daten beispielsweise zunächst von einem Bandspeicher abge-

90 Vgl. Pasquetto, Randles und Borgman 2017, 3–4; van de Sandt et al. 2019, 6–13.
91 Vgl. Sandt et al. 2019, 13.
92 Vgl. Thanos 2017, 1; Zimmerman 2008, 633–34.
93 Erdfelder und Ulrich 2018, 3. Vergleiche in diesem Zusammenhang auch die Unterscheidung von „computational reproducibility“, „replicability“ und „generalizability“ in National Academies of Sciences, Engineering, and Medicine 2019, 1.

rufen werden, oder werden mit mangelnder Bandbreite übermittelt, kann dies insbesondere bei großen Datenmengen die Verfügbarkeit und damit die Nutzbarkeit erheblich einschränken. Andererseits spielen auf der Ebene der Forschungsdateien die Formate, in denen diese vorliegen, eine wesentliche Rolle bei der Ermöglichung der Nachnutzung.

Die Verbreitung eines Dateiformats und der zugehörigen verarbeitenden Software in der Gruppe der (potenziellen) Nutzenden bestimmt wesentlich, ob die Dateien genutzt werden können. Während im primären Projektkontext in der Regel die (fach-)spezifischen Projektbedarfe darüber bestimmen, welche Dateiformate genutzt werden, sollten so früh wie möglich Überlegungen zu geeigneten Formaten für eine spätere Nutzung der Daten angestellt werden – idealerweise schon durch die Primärforschenden im Rahmen der Datenmanagementplanung.[94] Im Zentrum dieser Überlegungen müssen die zukünftigen Nutzungen stehen, die ermöglicht werden sollen. So macht es bei der Wahl eines geeigneten Formats zum Beispiel einen Unterschied, ob der Inhalt einer Datei nur zum Lesen auf einem Bildschirm bestimmt ist, oder ob es möglich sein muss, die in der Datei gespeicherten Informationen zu editieren oder sie anderweitig maschinell weiterzuverarbeiten.[95]

Je größer der Personenkreis ist, für den die Forschungsdaten nutzbar sein sollen, desto wichtiger ist es, auf weit verbreitete und gut zugängliche Dateiformate zu setzen – gerade, wenn eine Nutzung über Disziplingrenzen hinweg möglich sein soll.

Um die Nutzbarkeit langfristig zu erhalten, sollte zudem auf möglichst offene Dateiformate zurückgegriffen werden. Offene Formate sind solche, deren Spezifikationen im Gegensatz zu geschlossenen, proprietären Formaten komplett offen liegen. Hiermit wird es möglich, Software zum Ausführen der Dateien zu entwickeln, ohne dass Einschränkungen durch Eigentumsrechte kommerzieller Hersteller bestehen, die die Nutzbarkeit von Forschungsdaten einschränken oder gar unmöglich machen können, sollte die kommerzielle Software vom Markt genommen werden.[96]

Nicht immer lassen sich die genannten Aspekte – Offenheit und Verbreitung – vereinbaren. Häufig handelt es sich bei weit verbreiteten Formaten um proprietäre. In der Umfrageforschung sind etwa die proprietären Statistikprogramme SPSS[97] und Stata[98] – und damit auch mit ihnen assoziierte Dateiformate – weit verbreitet und stellen somit einen de facto-Standard dar. Dennoch ist die fehlende Offenheit ein wesentliches Risiko für den Erhalt der langfristigen Nutzbarkeit und es sollten frühzeitig – idealerweise schon vor der Erhebung der Daten – Überlegungen dazu ange-

94 S. Beitrag von Dierkes, Kap. 4.1 in diesem Praxishandbuch.
95 Vgl. hierzu beispielhaft DARIAH-DE 2017.
96 Vgl. Dietrich et al. n.d.
97 S. https://www.ibm.com/de-de/analytics/spss-statistics-software.
98 S. https://www.stata.com.

stellt werden, wie dieses Risiko minimiert werden kann. Dies kann beispielsweise durch eine Transformation in ein offenes Archivierungsformat erfolgen, welches bei Bedarf, z. B. mithilfe von entsprechenden Syntaxen oder Code, wieder in ein verbreitetes Nutzungsformat überführt werden kann.[99]

Eine besondere Herausforderung stellt der technische Wandel auch für die Replikation von Forschungsergebnissen dar, die auf einer maschinellen Verarbeitung, insbesondere einer softwaregestützten Analyse, beruhen. So kann es sein, dass verschiedene Versionen ein und derselben Software unterschiedlich „rechnen" und somit unterschiedliche Ergebnisse auf der gleichen Datengrundlage erzielen – etwa, weil Werte anders gerundet werden. Ershova und Schneider (2018) sowie Kim, Poline und Dumas (2018) weisen in diesem Zusammenhang auf die Bedeutung sorgfältiger Dokumentation der in den Analysen verwendeten technischen Systeme hin.

Neben technischen Hürden können auch ethische[100] und rechtliche[101] Anforderungen die Nachnutzung von Forschungsdaten erschweren oder unmöglich machen. Einschränkungen der Nutzbarkeit ergeben sich beispielsweise aufgrund von Eigentumsrechten (Urheberrecht, Patentrecht, etc.) oder sind notwendig, um den Schutz sensitiver Informationen zu gewährleisten. Entsprechend treffen Forschende Nutzungsentscheidungen unter Berücksichtigung der generellen Zugänglichkeit und Lizenzierung der Daten.[102]

Aus urheberrechtlicher Perspektive sind schutzfähige Forschungsdaten[103] in der Regel nur dann nachnutzbar, wenn die Personen, die die Verwertungsrechte innehaben, einer Nutzung durch Dritte zugestimmt haben. Eine solche Zustimmung kann an bestimmte Bedingungen geknüpft sein, die beispielsweise in einem Lizenztext oder einem Nutzungsvertrag festgeschrieben werden. Je restriktiver diese Nutzungsbedingungen im Hinblick auf den Nutzungszweck oder die Veränderung und Weitergabe der Forschungsdaten sind, desto stärker kann die Nutzbarkeit der Daten eingeschränkt sein.

Ein weit verbreitetes Lizenzmodell, das auch für Forschungsdaten häufig Anwendung findet, ist das der Creative Commons-Lizenzen.[104] Ein Vorteil dieser Lizenzen ist, dass sie aufgrund der Verbreitung bei Forschenden einen recht hohen Bekanntheitsgrad haben. Zu beachten ist jedoch, dass Creative-Commons-Lizenzen nicht für alle Datentypen geeignet sind (z. B. Daten mit Personenbezug, siehe unten[105]).

99 Ob dies – insbesondere verlustfrei – möglich ist, hängt selbstverständlich von den spezifischen Formaten ab.

100 S. Beitrag von Rösch, Kap. 1.5 in diesem Praxishandbuch.

101 S. Beitrag von Lauber-Rönsberg, Kap. 1.4 in diesem Praxishandbuch.

102 Vgl. Wu et al. 2019, 5.

103 Zur Frage der Schutzfähigkeit von Forschungsdaten s. Beitrag von Lauber-Rönsberg, Kap. 1.4 in diesem Praxishandbuch.

104 S. https://creativecommons.org.

Bei der Nutzung von Creative Commons (CC) und anderen Lizenzen ist zu beachten, dass durch die Vergabe einer zu restriktiven Lizenz die Nachnutzung erheblich eingeschränkt werden kann. Dies kommt unter anderem dann zum Tragen, wenn Quellen, die unter unterschiedlichen Lizenzen stehen, integriert werden sollen. Hier kann das sogenannte License Stacking dazu führen, dass eine Veröffentlichung des integrierten Produkts überhaupt nicht oder nur unter der restriktivsten Lizenz möglich ist.[106] Ein Datensatz, der unter einer CC Namensnennung-Share-Alike-Lizenz (CC-BY-SA) steht, kann zwar zum Zweck nicht-kommerzieller Forschung mit Daten, die unter der CC-Lizenz-Namensnennung-Nicht kommerziell (CC-BY-NC) integriert werden; der resultierende Datensatz kann aber nicht für die Nutzung durch Dritte lizenziert werden, da die beiden Ausgangs-Lizenzen sich ausschließen. Denn während CC-BY-SA eine kommerzielle Nutzung ausdrücklich erlaubt und untersagt, die Daten oder darauf aufbauende Produkte unter einer restriktiveren Lizenz zu veröffentlichen, erlaubt die CC-BY-NC-Lizenz keinerlei kommerzielle Nutzung der Daten oder auf ihnen aufbauender Produkte.[107]

Eine Einschränkung der Nachnutzbarkeit von Forschungsdaten aus rechtlichen oder ethischen Gründen kann notwendig sein, wenn diese Daten das Recht auf informationelle Selbstbestimmung der teilnehmenden Personen berühren und im Einklang mit dem Datenschutzrecht verarbeitet werden müssen. Auch das Vorkommen bedrohter Tier- oder Pflanzenarten kann eine sensitive Information darstellen, die durch geeignete Maßnahmen geschützt werden muss.[108] Solche Maßnahmen können in einer Veränderung der Forschungsdaten dahingehend bestehen, dass die sensitive Information gelöscht, vergröbert oder anderweitig verfremdet wird (z. B. durch die Verwendung von Pseudonymen). Diese Veränderungen schränken die Nutzbarkeit der Forschungsdaten allerdings ein, da sie das Analysepotenzial teils erheblich mindern können. Alternativ kann zum Schutz von sensitiven Informationen der Zugang zu den Daten restriktiver gestaltet werden, indem spezielle Nutzungsverträge geschlossen werden und/oder die Daten nur in ganz bestimmten und besonders geschützten Umgebungen remote oder vor Ort zugänglich gemacht werden, etwa in einer so genannten Data Enclave oder in einem Secure Data Center, wie es etwa bei GESIS – Leibniz-Institut für Sozialwissenschaften angeboten wird.[109] In solchen Einrichtungen wird der Zugang zu den Daten zugunsten des Erhalts ihrer analytischen Nutzbarkeit bewusst erschwert.

Die dritte Dimension, in welcher sich der Grad der Nutzbarkeit von Forschungsdaten entscheidet, ist die der *intellektuellen Zugänglichkeit* oder *Verstehbarkeit*. Ob

105 Vgl. auch Creative Commons 2019a.
106 Vgl. Mozilla Science Labs n.d.
107 Vgl. Creative Commons 2019b.
108 Vgl. z. B. Chapman und Grafton 2008, 3.
109 S. https://www.gesis.org/angebot/daten-analysieren/secure-data-center-sdc.

Forschungsdaten zur Beantwortung neuer Forschungsfragen nutzbar sind, hängt wesentlich von der Möglichkeit ab, den Inhalt der Daten erfassen zu können.

Forschungsdaten sind nicht „selbsterklärend" und ihre wissenschaftliche Nutzung ist in aller Regel ohne umfassende zusätzliche Informationen darüber, durch wen, warum (Erhebungszweck, Forschungsfrage) und wie (Forschungsdesign, Erhebungsmethode) die Daten erhoben und aufbereitet wurden, nicht möglich. Als einfaches Beispiel können Temperaturwerte dienen, die nur dann verständlich sind, wenn unter anderem bekannt ist wann, wo, mit welchen Instrumenten und unter welchen Bedingungen sie gemessen wurden.[110] Welche Informationen für ein Verständnis notwendig sind, ist einerseits in hohem Maße fach- bzw. datenspezifisch (vgl. Tab. 1 für Beispiele) und hängt andererseits vom jeweiligen Nutzungszweck ab: „a data set is intelligible only when its metadata relates to its intended use."[111] So kann das Fehlen von bestimmten Kontextinformationen einige Nachnutzungen unmöglich machen, während andere Nutzungen hiervon völlig unberührt bleiben.

Angesichts der teils rasanten Herausbildung neuer Forschungsmethoden und -praktiken und der zunehmenden Bedeutung von trans- und interdisziplinärer Forschung ist die Frage, welche Kontextinformationen für zukünftige Nutzungen von Forschungsdaten wohl notwendig sind, nicht mit Sicherheit – möglicherweise noch nicht einmal annähernd – zu beantworten.

So gibt es in der Literatur Hinweise darauf, dass ein umfassendes Verständnis von Forschungsdaten für die Nachnutzung nur möglich ist, wenn die Nutzenden selbst Erfahrungen in der Erhebung und Aufbereitung entsprechender Daten haben. „While standards can be helpful, the results show that knowledge of the local context is critical to ecologists' reuse of data."[112] Auch Pasquetto, Borgman und Wofford weisen darauf hin, dass insbesondere bei der Verwendung von Forschungsdaten zur Beantwortung neuer Forschungsfragen (als „integrative data reuse" bezeichnet) ein Grad von Verständnis der Daten notwendig ist, der kaum über die reine Bereitstellung von Kontextinformationen[113] zu erreichen ist.[114] Dies führt einerseits dazu, dass Forschende im Rahmen einer Nachnutzung mit den Primärforschenden kooperieren. Zum anderen konnte gezeigt werden, dass Nutzende Kontakt zu den Datenproduzierenden (Primärforschenden) und anderen Personen in ihrer Forschungscommunity suchen, wenn sie auf Probleme mit Datensätzen stoßen.[115]

110 Vgl. hierzu Abschnitt 4.3 in Pasquetto, Borgman und Wofford 2019.

111 Thanos 2017, 10.

112 Zimmerman 2008, 631.

113 Hierbei kann es sich beispielsweise um Dokumentation der Datenerhebung und –aufbereitung in Form von Feldtagebüchern, Erhebungsinstrumenten, Methodenreports, Analysecode, oder Beschreibung der Hard-/Software-Umgebung zur Datenerhebung und/oder -verarbeitung handeln.

114 Vgl. Pasquetto, Borgman und Wofford 2019, Abschnitt 4.3.

115 Vgl. Yoon 2017, 466–67; Gregory et al. 2019b, 428–29.

Diese Befunde machen deutlich, dass eine Nachnutzung von Forschungsdaten in neuen Forschungs- und Projektkontexten nur dann überhaupt möglich wird, wenn der originäre Forschungsprozess und die resultierenden Daten möglichst umfassend beschrieben werden. Wie Zimmermann ausführt, kann die Abhängigkeit von implizitem Wissen durch die Verfügbarkeit und Nutzung von fachspezifischen Standards verringert werden.[116] Dies können beispielsweise Standards sein, die bei der Erhebung von Forschungsdaten oder bei der Dokumentation des Erhebungsprozesses Anwendung finden.

Tab. 1: Beispiele für benötigte Kontextinformationen nach Domäne

Domäne	Beispiele für benötigte Kontextinformationen
Archäologie	„Site metadata: Site location and background, Excavator, Excavation type and techniques, Cultural sequence, periodization, and affinities, Dating, Recovery metadata, Sampling, Context types.“[117]
Ökologie	„[D]escription of the methods used to obtain an observation or to conduct an experiment, the location of an observation or experiment, and attributes associated with an observed species, such as taxonomic information, physical characteristics, or natural history information.“[118]
Qualitative Sozialforschung	„Methods: Instrument (Abstract), Tool (Abstract), Settings, Data Collection Tool, Analysis Tool, Processing Tool, Data Collection Method, Analysis Method, Processing Method, Data Collection Instrument, Analysis Instrument, Processing Instrument, Data Collection Mode.“[119]

2.2 Vertrauen

Die Forschung zeigt, dass das Vertrauen der Nutzenden in die Forschungsdaten und die datenhaltende Institution wesentlich mit darüber bestimmt, ob Daten nachgenutzt werden. Hierbei wurden verschiedene Faktoren identifiziert, die darüber bestimmen, ob Nutzende einer Information (z. B. einem Datensatz) vertrauen (vgl. Tab. 2).

116 Vgl. Zimmermann 2008, 634–35.
117 Atici et al. 2013, 678.
118 Zimmerman 2008, 633.
119 Hoyle und DDI Qualitative Data Working Group 2012, 5–6.

Tab. 2: Faktoren, die das Vertrauen Forschender in genutzte Daten beeinflussen

Studie	Faktoren (eigene Übersetzung)
Kelton, Fleischmann und Wallace (2008, 367)	Genauigkeit Objektivität Validität Stabilität
Donaldson und Conway (2015, 2440)	Authentizität Verlässlichkeit
Faniel und Yakel (2017, 111)	Identität der Datenproduzierenden Dokumentation, z. B. „completeness or thoroughness of record, evidence of standardized or professional practice“[120] Begutachtete Publikationen über die Daten Hinweise auf frühere Nutzungen Reputation des Repositoriums

Die in Tab. 2 aufgeführten Studien machen deutlich, dass die Entscheidung, ob Forschende Daten ausreichend vertrauen, um diese zu nutzen, wesentlich von (wahrgenommenen) Eigenschaften erstens der Daten und ihrer Dokumentation selbst, zweitens der Datenproduzierenden und drittens der datenhaltenden Institution abhängt.

Bezüglich der Eigenschaften von Daten und deren Dokumentation, wie etwa Genauigkeit, Validität, Verlässlichkeit oder Vollständigkeit, kommt den Primärforschenden, die die Daten erheben und aufbereiten, eine wesentliche Verantwortung zu. Denn viele dieser Eigenschaften leiten sich unmittelbar aus dem Erhebungsprozess und dem nachfolgenden Umgang mit den Daten ab. Werden entsprechende (Kontext-)Informationen nicht bereits während des Forschungsprozesses zum Zweck der Dokumentation festgehalten, können sie nachträglich häufig nicht mehr rekonstruiert werden. Schon hier sollte möglichst eine Orientierung an domänenspezifischen Standards und Best Practices erfolgen, etwa um eine Vollständigkeit der Dokumentation zu gewährleisten.

Eine weitere Anreicherung der Daten mit für die Bewertung ihrer Vertrauenswürdigkeit relevanten Informationen kann Aufgabe der Infrastruktur sein, die die Daten langfristig sichert und zugänglich macht. So können beispielsweise die fortlaufende Verknüpfung mit auf Grundlage der Daten entstandenen Publikationen oder die Bereitstellung von Informationen wie ORCID[121] der Primärforschenden dabei helfen, für die Beurteilung der Vertrauenswürdigkeit der Daten durch die Nutzenden relevante Informationen zugänglich zu machen. Auch eine weitere Aufbereitung der von den Forschenden dokumentierten Kontextinformationen gemäß

120 Faniel und Yakel 2017, 112.
121 S. https://orcid.org.

Domänenstandards (z. B. DDI, MeSH,[122] ADeX[123]) kann Aufgabe der Informationsinfrastruktur sein, die die Daten langfristig sichert und zur Nutzung anbietet.

Wie in Tab. 2 dargestellt, hat nach Faniel und Yakel auch die Reputation des Repositoriums eine Relevanz, wenn Forschende sich für oder gegen die Nutzung von Forschungsdaten entscheiden: „In our DIPIR research we [...] found that data reusers assess trust through repository functions – particularly data processing, metadata application, and data selection – and to a lesser extent repository actions, such as transparency."[124] Vor diesem Hintergrund erscheint es umso wichtiger, dass Informationsinfrastrukturen tief in den jeweiligen Fachcommunities verankert sind und ihre Zielgruppe und ihre Bedürfnisse kennen.[125] Nur so können sie gewährleisten, dass sie relevanter Entwicklungen und Veränderungen gewahr werden und ihre Services und Angebote entsprechend anpassen können.

Fazit

Was die Auffindbarkeit von Daten angeht, besteht eine gewisse Divergenz zwischen Angeboten zur Datensuche und dem tatsächlichen Suchverhalten. Die vorhandenen Repositorien und Portale bieten den Zugang zu den Daten und werden auch genutzt; tatsächlich spielen aber die Datensuche über Literatur, das Kennenlernen von neuen Daten durch Kontakte in der Forschungscommunity und zunehmend auch die Websuche eine größere Rolle für Datennutzende. Diese Praktiken sollten bei der Weiterentwicklung der Forschungsdateninfrastruktur mitgedacht werden. Die Dokumentation spielt eine zentrale Rolle für die Auffindbarkeit von Daten. Die Nutzung der Dokumentation, insbesondere die Verwendung von (fachspezifischen) Terminologien im Suchprozess muss im Sinne einer nutzungsorientierten Weiterentwicklung der Forschungsdatensuchdienste noch weiter erforscht werden.

Wie gut und zu welchen Zwecken Forschungsdaten nutzbar sind, hängt von einer Reihe von Faktoren in unterschiedlichen Dimensionen ab (Technik, Recht und Ethik, intellektuelle Zugänglichkeit). Sowohl die Primärforschenden bzw. Datenproduzierende als auch die Informationsinfrastrukturen, die die Daten archivieren und zugänglich machen, können dazu beitragen, die Nutzbarkeit von Forschungsdaten zu erhöhen. So sollten Forschende schon im Forschungsprozess als Teil des Datenmanagements möglichst genau dokumentieren, wie und warum sie die Forschungsdaten erheben, aufbereiten und analysieren. Hilfestellung bieten Einrichtungen der

122 Medical Subject Headings, s. https://www.nlm.nih.gov/mesh/meshhome.html.

123 Archäologischer DateneXport-Standard, s. https://landesarchaeologen.de/kommissionen/archaeologie-und-informationssysteme/projekte-8.

124 Faniel und Yakel 2017, 110.

125 Vgl. Faniel und Yakel 2017, 118.

Informationsinfrastruktur. Letzteren kommt zudem eine besondere Verantwortung bei der Anreicherung und Standardisierung vorhandener Kontextinformationen zu. Beides kann wesentlich dazu beitragen, die Interpretierbarkeit der Forschungsdaten zu ermöglichen und das Vertrauen der Nutzenden in die Daten zu unterstützen.

Literatur

Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

Atici, Levent, Sarah W. Kansa, Justin Lev-Tov und Eric C. Kansa. 2013. „Other People's Data: A Demonstration of the Imperative of Publishing Primary Data." *J Archaeol Method Theory* 20 (4): 663–681. doi:10.1007/s10816-012-9132-9.

Bauer, Andreas, Holger Günzel und Jens Albrecht. 2013. *Data-Warehouse-Systeme. Architektur, Entwicklung, Anwendung.* Heidelberg: dpunkt-Verlag.

Borgman, Christine L. 2015. *Big data, little data, no data: Scholarship in the networked world*. Cambridge: MIT Press.

Chapman, Adriane, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak und Paul Groth. 2019. „Dataset search: a survey." *The VLDB Journal*. doi:10.1007/s00778-019-00564-x.

Chapman, Arthur D. und Oliver Grafton. 2008. *Guide to best practices for generalising sensitive species occurence data: version 1.0.* Copenhagen: Global Biodiversity Information Facility. https://www.gbif.org/document/80512.

Creative Commons. 2019a. „CC Wiki: Data." https://wiki.creativecommons.org/wiki/data.

Creative Commons. 2019b. „Frequently Asked Questions: Can I combine material under different Creative Commons licenses in my work?" https://creativecommons.org/faq/#can-i-combine-material-under-different-creative-commons-licenses-in-my-work.

DARIAH-DE. 2017. „Empfehlungen für Forschungsdaten, Tools und Metadaten in der DARIAH-DE Infrastruktur." https://wiki.de.dariah.eu/pages/viewpage.action?pageId=38080370.

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

Deutsche Initiative für Netzwerkinformation, Hg. 2018. *Thesen zur Informations- und Kommunikationsinfrastruktur der Zukunft*. doi:10.18452/19126.

Dietrich, Daniel, Jonathan Gray, Tim McNamara, Antti Poikola, Rufus Pollock, Julian Tait und Ton Zijlstra. n. d. „Open Data Handbook: File Formats." http://opendatahandbook.org/guide/en/appendices/file-formats/.

Donaldson, Devan R. und Paul Conway. 2015. „User conceptions of trustworthiness for digital archival documents." *J Assn Inf Sci Tec* 66 (12): 2427–2444. doi:10.1002/asi.23330.

Erdfelder, Edgar und Rolf Ulrich. 2018. „Zur Methodologie von Replikationsstudien." *Psychologische Rundschau* 69 (1): 3–21. doi:10.1026/0033-3042/a000387.

Ershova, Anastasia und Gerald Schneider. 2018. „Software updates: the ‚unknown unknown' of the replication crisis." LSE Impact Blog. http://blogs.lse.ac.uk/impactofsocialsciences/2018/06/07/software-updates-the-unknown-unknown-of-the-replication-crisis/.

European Commission. 2016. H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020. Version 3.0. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Faniel, Ixchel M. und Elizabeth Yakel. 2017. „Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation." In *Curating Research Data Volume 1: Practical Strategies for Your Digital Repository*, hg. v. Lisa R. Johnston, 103–126. Chicago: Association of College and Research Libraries Press.

Fechner, Benedikt und Cornelius Puschmann. 2015. „Über die Grenzen der Offenheit in der Wissenschaft: Anspruch und Wirklichkeit bei der Bereitstellung und Nachnutzung von Forschungsdaten." *Information, Wissenschaft und Praxis* 66 (2–3): 146–150. doi:10.1515/iwp-2015-0026.

Fedkenhauer, Thomas, Yvonne Fritzsche-Sterr, Lars Nagel, Angelika Pauer und Aleksei Resetko. 2017. *Datenaustausch als wesentlicher Bestandteil der Digitalisierung*. Hg. von PricewaterhouseCoopers GmbH, Düsseldorf. https://www.pwc.de/de/digitale-transformation/studie-datenaustausch-digitalisierung.pdf.

Gray, Jim. 2007. „Jim Gray on eScience: A Transformed Scientific Method. Based on the transcript of a talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, on January 11, 2007." In *The Fourth Paradigm. Data-Intensive Scientific Discovery*, hg. v. Tony Hey, Stewart Tansley und Kristin Tolle, xvii–xxxi. Redmond: Microsoft Research.

Gregory, Kathleen, Helena Cousijn, Paul Groth, Andrea Scharnhorst und Sally Wyatt. 2019a. „Understanding Data Search as a Socio-technical Practice." *Journal of Information Science*. doi:10.1177/0165551519837182.

Gregory, Kathleen, Paul Groth, Helena Cousijn, Andrea Scharnhorst und Sally Wyatt. 2019b. „Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines." *Journal of the Association for Information Science and Technology* 70: 419–432. doi:10.1002/asi.24165.

Gregory, Kathleen, Siri Jodha Khalsa, William K. Michener, Fotis E. Psomopoulos, Anita de Waard und Mingfang Wu. 2018. „Eleven quick tips for finding research data." *PLoS Computational Biology 14* (4): e1006038. doi:10.1371/journal.pcbi.1006038.

Hoyle, Larry und DDI Qualitative Data Working Group. 2012. „A Qualitative Data Model for DDI." https://ddialliance.org/sites/default/files/AQualitativeDataModelForDDI.pdf.

Kelton, Kari, Kenneth R. Fleischmann und William A. Wallace. 2008. „Trust in digital information." *J. Am. Soc. Inf. Sci*. 59 (3): 363–374. doi:10.1002/asi.20722.

Kern, Dagmar und Brigitte Mathiak. 2015. „Are there Any Differences in Data Set Retrieval Compared to Well-known Literature Retrieval?" In *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, hg. v. Sarantos Kapidakis, Cezary Mazurek, und Marcin Werla, 197–208. Cham: Springer.

Kim, Yang-Min, Jean-Baptiste Poline und Guillaume Dumas. 2018. „Experimenting with Reproducibility: A Case Study of Robustness in Bioinformatics." *GigaScience* 7 (7): 1–8. doi:10.1093/gigascience/giy077.

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Thousand Oaks: Sage.

Kommission Zukunft der Informationsinfrastruktur. 2011. Gesamtkonzept für die Informationsinfrastruktur in Deutschland: Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder. https://www.hof.uni-halle.de/web/dateien/KII_Gesamtkonzept_2011.pdf.

Mannheimer, Sara, Leila Sterman und Susan Borda. 2016. „Discovery and Reuse of Open Datasets: An Exploratory Study." *Journal of eScience Librarianship* 5 (1): e1091. doi:10.7191/jeslib.2016.1091.

Mozilla Science Labs. n. d. „Open Data Training Primers: Primer 5.3: License Stacking." https://mozillascience.github.io/open-data-primers/5.3-license-stacking.html.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. doi:10.17226/25303.

Pasquetto, Irene V., Bernadette M. Randles und Christine L. Borgman. 2017. „On the Reuse of Scientific Data." *Data Science Journal* 16 (1): 21. doi:10.5334/dsj-2017-008.

Pasquetto, Irene V., Christine L. Borgman und Morgan F. Wofford. 2019. „Uses and Reuses of Scientific Data: The Data Creators' Advantage." *Harvard Data Science Review* 1 (2): 1–35. doi:10.1162/99608f92.fc14bf2d.

Rat für Informationsinfrastrukturen. 2019. Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel. Göttingen. urn:nbn:de:101:1-201911201154165773273 7.

Shen, Yi. 2015. „Research Data Sharing and Reuse Practices of Academic Faculty Researchers: A Study of the Virginia Tech Data Landscape." *International Journal of Digital Curation* 10: 157–175. doi:10.2218/ijdc.v10i2.359.

Tenopir, Carol, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock und Kristina Dorsett. 2015. „Changes in Data Sharing and Data Reuse Practices and Perceptions Among Scientists Worldwide." *PLoS one* 10 (8): e0134826. doi:10.1371/journal.pone.0134826.

Thanos, Costantino. 2017. „Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies." *Publications* 5 (1): 2. doi:10.3390/publications5010002.

van de Sandt, Stephanie, Sünje Dallmeier-Tiessen, Artemis Lavasa und Vivien Petras. 2019. „The Definition of Reuse." *Data Science Journal* 18 (2): 2. doi:10.5334/dsj-2019-022.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. 2016. „The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 1–9. doi:10.1038/sdata.2016.18.

Wissenschaftsrat. 2012. Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. https://www.wissenschaftsrat.de/download/archiv/2359-12.pdf.

Wu, Mingfang, Fotis Psomopoulos, Siri Jodha Khalsa und Anita de Waard. 2019. „Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories." *Data Science Journal* 18 (1): 1–13. doi:10.5334/dsj-2019-003.

Yoon, Ayoung. 2017. „Role of Communication in Data Reuse." *Proceedings of the Association for Information Science and Technology* 54 (1): 463–471. doi:10.1002/pra2.2017.14505401050.

Zimmerman, Ann S. 2008. „New Knowledge from Old Data." *Science, Technology, & Human Values* 33 (5): 631–652. doi:10.1177/0162243907306704.

Zimmerman, Ann. 2007. „Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse." *International Journal on Digital Libraries* 7 (1–2): 5–16. doi:10.1007/s00799-007-0015-8.