

Holger Alda

► So nah und doch so fern? Open-Access-Kriterien beim Zugang und der Nutzung von Forschungsdaten

1 Publikationen und Forschungsdaten

Publikationen und Forschungsdaten haben auf mehreren Ebenen spezifische Verbindungen. Auf der inhaltlichen Ebene benötigen Forschende geeignete Forschungsdaten, wenn sie ihre im Forschungszusammenhang aufgestellten Hypothesen empirisch testen. Auf der Ebene des wissenschaftlichen Fortschritts spielen insbesondere Reichweite und Gültigkeit der verwendeten Forschungsdaten im Abgleich mit den in wissenschaftlichen Publikationen verfolgten inhaltlichen Fragestellungen und Schlussfolgerungen eine entscheidende Rolle bei der Gesamtbewertung einer jeden neu entstehenden Publikation. Auf der publizistischen Ebene geht es um die Frage der Verbindung von Fachveröffentlichung und den in den einzelnen Artikeln genutzten Forschungsdaten respektive ihre jeweilige Dokumentation, Verfügbarkeit und wissenschaftlichen Seriosität.

Der vorliegende Beitrag betrachtet diese drei Ebenen des Verhältnisses von Forschungsdaten und Publikationen in einer Perspektive, die sich an zwei essenziellen Grundgedanken von Open-Access-Publikationen orientiert, d. h., es geht im Wesentlichen um die Frage, inwiefern die in Fachpublikationen verwendeten Forschungsdaten kostenlos und für jedermann frei zugänglich sind bzw. sein sollten. Gerade in Open-Access-Publikationen wirken die den empirischen Tabellen und Abbildungen zugrunde liegenden Forschungsdaten vergleichsweise „nah“ an diesen beiden Grundgedanken. Während ein kostenfreier Zugang natürlich sehr leicht festzustellen ist, ist es hingegen schwieriger zu beurteilen, inwiefern leichte Zugänglichkeit auch für Forschungsdaten gegeben ist. Wie fern sind gegenüber dem Grundgedanken von Open Access (OA) einer freien Verfügbarkeit der Zugang und die Nutzung von Forschungsdaten für die Erstellung von Fachpublikationen?

Um sich einer Antwort auf diese Frage anzunähern, bestimmt Abschnitt 2, was Forschungsdaten überhaupt sind, welche Formate sie haben und auf welche Objekte sie sich

beziehen (können). Forschungsdaten in den für die Berufsbildungsforschung maßgeblichen sozial- und wirtschaftswissenschaftlichen Teildisziplinen beinhalten sehr oft Informationen über Personen oder ihnen in datenschutzrechtlicher Sicht gleichgestellten Erhebungseinheiten wie etwa Betriebe. Es bedarf demnach gewisser Vorkehrungen zum Schutz von persönlichen Informationen, bevor Forschende mit solchen Datenbeständen arbeiten können und dürfen.

Vor diesem Hintergrund wird in Abschnitt 3 die optimale Ausgestaltung der Zugangs- und Nutzungsbedingungen für Forschungsdaten aus Sicht der *Scientific Community* beschrieben. Da dies je nach Schutzbedürftigkeit der Forschungsdaten in unterschiedlich anspruchsvolle und ressourcenintensive Datenzugangsmodelle mündet, werden in Deutschland sozial- und wirtschaftswissenschaftliche Forschungsdaten üblicherweise von Forschungsdatenzentren (FDZ) oder Datenarchiven (DA) im Rahmen der Nationalen Forschungsdateninfrastruktur bereitgestellt.¹ Mit Kenntnis der Zugangs- und Nutzungsbedingungen von solchen Beständen an Forschungsdaten werden diese in einen Kontext mit Fachpublikationen gestellt und die derzeitige Praxis diskutiert und gewertet.

Die Kenntnis der entsprechenden Bestimmungen sowie die Beschreibung der technischen und inhaltlichen Lösungen, die Wissenschaft und Forschung vor allem in den letzten 20 Jahren für das Spannungsfeld von Datenschutz auf der einen und Forschungsfreundlichkeit auf der anderen Seite vor allem durch die Gründung von FDZ entwickelt haben, bestimmen zunehmend das Verhältnis und den Umgang mit Forschungsdaten in den Sozial- und Wirtschaftswissenschaften – gerade für referierte Beiträge in Fachzeitschriften. Aber auch innerhalb dieser auch für die Berufsbildungsforschung maßgeblichen wissenschaftlichen Teildisziplinen sind die zurzeit existierenden technischen und inhaltlichen Lösungen über FDZ dezentral und somit auch immer in gewisser Weise heterogen. Die meisten der nachfolgenden Ausführungen gelten in identischer oder vergleichbarer Form für viele Forschungsdaten (und FDZ/DA) in den Sozial- und Wirtschaftswissenschaften. Zur Begrenzung der Komplexität des Beitrags werden jedoch die nachfolgend diskutierten technischen und inhaltlichen Lösungen in vollständiger Form nur für Zugangs- und Nutzungsbedingungen für Forschungsdaten zur beruflichen Bildung abgebildet, wie sie vom Forschungsdatenzentrum im Bundesinstitut für Berufsbildung (BIBB-FDZ) angeboten werden.

2 Grundlegende Gruppierung von Forschungsdaten

„Forschungsdaten sind Daten, die im Zuge wissenschaftlicher Vorhaben z. B. durch Digitalisierung, Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen. Die Verfügbarkeit und Nachnutzung digitaler Informationen schließt *den möglichst kostenfreien und offenen Zugang zu Forschungsdaten ein ...* Begleitende profes-

1 Vgl. URL: <https://www.nfdi.de/konsortswd-2/> (Stand: 17.06.2022).

sionelle informationswissenschaftliche und informationstechnische Methoden helfen, diese Daten zu sammeln, zu archivieren und für eine Weiternutzung bereitzustellen.“²

Je nach Fachgebiet oder Forschungsvorhaben können Forschungsdaten demnach in Form von Texten, Tabellen, Bildern, Messdaten, Videos, Objekten und vielem mehr vorliegen. Um wissenschaftlichen Qualitätskriterien zu genügen, gibt es an Forschungsdaten grundlegende Anforderungen. Forschungsdaten erfüllen eine replizierbare, objektivierbare und standardisierende Funktion für die *Scientific Community*, wozu neben dem Zugang und der Nutzung der jeweiligen Forschungsdaten in der Regel auch methodische Informationen gehören, die über Reichweite und Gültigkeit der Daten sowie deren methodische Entstehungsgenese Auskunft geben. Je nach Fachdisziplin und den dort jeweils üblicherweise verwendeten Forschungsdaten wird diese Funktionalität auf unterschiedlichen Wegen erreicht.

In der überwiegenden Mehrheit sind Forschungsdaten heutzutage digitale Objekte, d. h., es handelt sich um prinzipiell jederzeit mit geringem Aufwand veränderbare, aber auch vervielfältigbare Objekte. Demnach sind digitale Objekte für die wissenschaftliche Forschung nur geeignet, wenn sie standardisiert sind oder werden, d. h., es muss garantiert sein, dass prinzipiell alle wissenschaftlich tätigen Personen bei Bedarf auf ein exakt identisches digitales Objekt zugreifen können. Der Zugriff und die Nutzung solcher digitalen Objekte durch Wissenschaftler/-innen sollten sich dabei an FAIR-Kriterien orientieren,³ die sich als niedrigschwelliger Zugang zu Forschungsdaten unter Geltung des Gleichheitsgrundsatzes zusammenfassen lassen.

Im einfachsten Fall findet sich dann ein wissenschaftlich anerkannter Hoster für die jeweiligen Forschungsdaten und macht sie über das Internet für die Öffentlichkeit verfügbar. Die *Scientific Community* kann dann ihre Diskurse über die Validierung von Forschungsdaten über verschiedene Arbeiten hinweg gerade im Zeitverlauf vergleichsweise niedrigschwellig leisten. Ein solches Vorgehen zur Bereitstellung und Nutzung von Forschungsdaten ist oft besonders gut in den Naturwissenschaften bzw. allen wissenschaftlichen Fachdisziplinen möglich, in denen im weiteren Sinne Daten zu Dingen bzw. Objekten vorliegen.

Liegen hingegen – wie es auch in der Berufsbildungsforschung bzw. allgemeiner in den Sozial- und Wirtschaftswissenschaften häufig der Fall ist – Forschungsdaten über Personen (oder diesen gleichgestellten Einheiten wie etwa (Ausbildungs-)Betrieben) vor, so wäre eine freie Verfügbarmachung der entsprechenden Daten nur unter sehr engen Voraussetzungen möglich. Denn personenbezogene Daten unterliegen einem gewissen Schutz, d. h., sie dürfen nicht *frei* verfügbar sein bzw. zugänglich gemacht werden.

Für wissenschaftliche Forschungszwecke können solche Datenbestände genutzt werden, wenn vor der Auswertung der entsprechenden Forschungsdatensätze geeignete Anonymisierungsmaßnahmen durchgeführt wurden, mit denen gleichzeitig das wissenschaftliche

2 URL: <https://www.allianzinitiative.de/archiv/forschungsdaten/> (Hervorhebung nicht im Original; Stand: 05.07.2022).

3 Vgl. URL: <https://www.go-fair.org/go-fair-initiative/vision-and-strategy/> (Stand: 17.06.2022).

Auswertungspotenzial der Forschungsdaten nicht unangemessen einschränkt wird. Es gibt Datenbestände, bei denen die Balance von ausreichender Anonymisierung und verbleibendem wissenschaftlichen Auswertungspotenzial der Daten sehr gut gewahrt werden kann, und andere, wo dies nur mit einem wesentlich höheren Aufwand erreicht werden kann. Der folgende Abschnitt beschreibt, welche grundlegenden Anonymisierungsniveaus sich daraus ergeben und welche Zugangs- und Nutzungskonditionen für Forschungsdaten daraus folgen.

3 Datensatzproduktlinien und Datenzugriffsmöglichkeiten

Üblicherweise werden Forschungsdaten, die Informationen zu einzelnen Personen bereitstellen, in die nachfolgenden Datenproduktlinien überführt. Datenproduktlinien repräsentieren verschiedene Anonymisierungsniveaus von Forschungsdaten und legen damit gleichzeitig Zugangs- und Nutzungsbedingungen für das jeweilige Forschungsdatenprodukt fest.

Public-Use-File (PUF)

Die Forschungsdaten enthalten keinerlei Informationen mehr, mit denen sich einzelne Erhebungseinheiten re-identifizieren lassen (z. B. das Sozialprestige eines Berufs). Solche Forschungsdaten können frei ins Netz gestellt und von jedem/jeder eingesehen und genutzt werden. Üblicherweise bilden dann nicht mehr einzelne Personen, sondern eine Aggregatenebene wie etwa Berufe, Wirtschaftszweige oder Regionen eine einzelne Datenzeile. Der Personenbezug wird in PUF demnach meist durch eine Aufsummierung der Einzelwerte von Personen zu einer Gruppenebene entfernt. Bei PUF gibt es keinerlei (Vorab-)Kontrolle der Verwendungszwecke, d. h., sie sind zumindest theoretisch auch weitestgehend uneingeschränkt für jedwede kommerzielle und private Nutzung verfügbar.

Scientific-Use-File (SUF)

Die Forschungsdaten enthalten mit höchster Wahrscheinlichkeit keinerlei Informationen, mit denen sich Erhebungseinheiten re-identifizieren lassen. Eine Re-Identifikation von Erhebungseinheiten ist nicht anzunehmen. Für wissenschaftliche Forschungszwecke kann eine zeitlich befristete Überlassung von SUF förmlich beantragt werden. Es gibt zwar in diesen Fällen eine Kontrolle der Verwendungszwecke, aber nach Überlassung des Datenbestandes haben Forschende (zeitlich befristet) die volle Verfügungsgewalt über den jeweiligen Forschungsdatenbestand.

Campus-Files (CF)

CF sind eine Unterart von SUF, die hauptsächlich von Studierenden zum Schreiben von Seminar-, Bachelor- und Masterarbeiten sowie für universitäre Lehrzwecke eingesetzt bzw. verwendet werden. Re-Identifikationsrisiken werden vollständig und mit absoluter Sicherheit ausgeschlossen, indem – beispielsweise – nur eine Teilstichprobe mit weniger detaillierten

Klassifikationen (in der Regel Einsteller, etwa für Berufskennziffern, Wirtschaftszweige u. Ä.) aus einem SUF erzeugt wird.

Sondervariablen (SV), Datenfernverarbeitung (DFV), Gastaufenthalt (GWA)

In diesen Fällen können die Datenbestände nicht mit vertretbarem Aufwand anonymisiert werden. Eine Re-Identifikation von Erhebungseinheiten ist mit den verfügbaren Informationen in dem Datensatz unter Umständen möglich. Solche Datenbestände können demnach Forschenden nicht überlassen werden. Stattdessen werden Auswertungsmöglichkeiten für solche Datenbestände geschaffen, bei denen die Forschungsdaten die Heimatinstitution bzw. ihren Entstehungsort niemals verlassen. Entweder erfolgt die Übermittlung der Auswertungen der Forschenden bzw. aller ihre Ergebnisdateien aus der Arbeit mit den entsprechenden Forschungsdaten über teil- bis voll automatisierte Datenfernverarbeitungssysteme (DFV) oder die Forschenden fertigen die Auswertungen an in einer kontrollierten IT-Umgebung vor Ort in der Institution, die die entsprechenden Forschungsdatenbestände führt (Gastwissenschaftler/-innen-Aufenthalt, GWA). In beiden Fällen werden die aus den Auswertungen hervorgehenden Forschungsergebnisse einer automatisierten bzw. zusätzlichen menschlichen Kontrolle auf etwaige Re-Identifikationsrisiken unterzogen, *bevor* die Ergebnisse an die Forschenden – etwa zur freien Verwendung im Rahmen von Fachpublikationen – übermittelt werden.

Mit Ausnahme der PUFs gibt es demnach keinen vollständig freien Forschungsdatenzugang für jedermann, sondern einen je nach Anonymisierungsgrad abgestuften Datenzugang exklusiv und ausschließlich für wissenschaftliche Forschungszwecke. Dies wiederum setzt einen Hoster voraus, der die entsprechenden Forschungsdaten vorhält und bei berechtigten wissenschaftlichen Auswertungsinteressen einen möglichst niedrigschwelligen Datenzugang ermöglicht. Was vor diesem Hintergrund ein niedrigschwelliger Zugang bezüglich der Bereitstellung von Datenprodukten für wissenschaftliche Forschungszwecke leisten muss, wird deutlich, wenn man sich den Geneseprozess von Forschungsdaten genauer ansieht.

4 Genese und Auswertung von Forschungsdaten: von Projektdaten zu Forschungsdaten

Bei einer (neuen) Befragung von Personen oder Betrieben entstehen zunächst sogenannte Primärdaten. Diese Roh- bzw. Projektdaten müssen in aller Regel noch aufbereitet werden, um mit ihnen nach wissenschaftlichen Kriterien arbeiten zu können. Primärdaten entstehen dabei in der Regel im Rahmen von Forschungsprojekten, in denen die jeweilige Datenerhebung für die Beantwortung primärer Forschungsfragen finanziert wird.

Oftmals ist eine Projektauflage, dass spätestens mit Abschluss der Projektarbeiten die den Projektergebnissen zugrunde liegenden und neu generierten Forschungsdaten für die Forschung allgemein zugänglich gemacht werden. Primärforschende veröffentlichen ihre Forschungsergebnisse oft in referierten Zeitschriften. Diese Journals erwarten in der Regel,

dass die der eingereichten Arbeit zugrunde liegenden Forschungsdaten prinzipiell für die *Scientific Community* zugänglich sind. Daher entwickeln Primärdaten generierende Projekte oftmals ein Eigeninteresse an der Veröffentlichung der Projektdaten als Forschungsdaten.

Um Forschungsdaten, die auf Personen oder Betrieben basieren, für die *Scientific Community* zugänglich machen zu können, müssen zuallererst die primären Forschungsdaten in die im vorherigen Abschnitt genannten Datenprodukte (PUF, SUF, SV, DFV und GWA) überführt werden. Dies geschieht über Zuordnungen der einzelnen Merkmale oder Gegebenheiten in Forschungsdatensätzen zu ihrem Re-Identifikationspotenzial und ist in der Regel eine anspruchsvolle, detaillierte und zeitauswendige Aufgabe, bei der Erfahrungswissen sicherlich nicht schädlich ist. Die Erstellung der entsprechenden Datenprodukte ist immer der erste Schritt der Veröffentlichung von Forschungsdaten, weil mit der Anonymisierung zunächst ganz schlicht gesetzlichen Vorgaben gefolgt wird. Im zweiten Schritt werden die in die verschiedenen Anonymisierungsstufen zerlegten Forschungsdaten dann in möglichst niedrigschwellige Datenzugangs- und -nutzungsmodelle überführt.

Für PUF ist es lediglich erforderlich, den PUF über das Internet frei zugänglich zu machen. Das ist über das Einstellen auf einer Website leicht erreichbar.

Schon bei SUF bzw. CF ist das nicht mehr möglich. Hier muss etwa vorher sichergestellt werden, dass der Datenzugang in Form eines SUF/CF ausschließlich für wissenschaftliche Forschungszwecke erfolgt. Üblicherweise geschieht dies durch ein Antragsverfahren, in dem der wissenschaftliche Forschungszweck für den Datenzugang dargelegt wird. Für die sich anschließende, zeitlich befristete Überlassung der SUF/CF an die Forschenden wird in der Regel ein Nutzungsvertrag abgeschlossen, in dem sich die Forschenden insbesondere verpflichten, die überlassenen Forschungsdaten ausschließlich für die beantragten wissenschaftlichen Forschungszwecke zu nutzen, den SUF/CF nicht an Dritte weiterzugeben und die Forschungsdaten mit Ablauf der vereinbarten Nutzungsdauer unwiderruflich zu löschen.

Für die anderen drei Sorten von Forschungsdaten (SV, DFV, GWA) ist selbst die zeitlich befristete Überlassung von Forschungsdaten wie bei SUF/CF aufgrund von Re-Identifikationsrisiken für einzelne Erhebungseinheiten nicht möglich. Die entsprechenden Forschungsdaten dürfen aus Datenschutzgründen die Institution, in der sie erhoben wurden bzw. vorliegen, nicht verlassen. Die einzigen beiden Möglichkeiten, potenziell der gesamten *Scientific Community* einen Zugang zu solchen oftmals besonders interessanten Forschungsdatenbeständen zu schaffen, sind, sie entweder über kontrollierte DFV oder einen GWA zugänglich zu machen. Bei der DFV geben externe Personen Auswertungen der Forschungsdaten bei der Institution, in der die Daten entstanden sind, in Auftrag und erhalten vollständig anonymisierte Ausgabedateien zurück. Bei GWA wird Forschenden in der Institution, in der die Forschungsdaten entstanden sind, ein abgeschotteter Arbeitsplatz eingerichtet, mit dem sie die Daten vor Ort auswerten können. Die Mitnahme und freie Verwendung von Forschungsergebnissen sind erst nach einer datenschutzrechtlichen Prüfung auf vollständige Anonymität möglich.

5 Die Rolle von FDZ und Datenarchiven als Publikationsagenten von Forschungsdaten

Der vorherige Abschnitt macht deutlich, dass allein schon die Bereitstellung und – im Falle von SV-, DFV- und GWA-Datensätzen – die Auswertung von Forschungsdaten eine entsprechende IT-Infrastruktur sowie dauerhafte Verlässlichkeit und Kontinuität der durchzuführenden Arbeiten voraussetzen. Insbesondere die nicht freie Veröffentlichung eines Forschungsdatensatzes bringt zwei zentrale Herausforderungen für deren Nutzung mit sich: Es müssen wiederholt für individuelle Forschungsinteressen einzelner Forschender (oder für eine einzelne Forschungsgruppe) nach einem Antrags-, Prüfungs- und Vertragsverfahren exakt die gleichen Forschungsdaten bereitgestellt werden. Hierbei ist für einen freien Wettbewerb der Ideen und damit wissenschaftlichen Fortschritt u. a. sicherzustellen, dass

- ▶ alle mit Forschungsdaten Arbeitenden unabhängig von ihren jeweiligen Forschungsinteressen in Bezug auf Antrags-, Vertrags- und Bereitstellungsverfahren gleichbehandelt werden,
- ▶ alle Gruppen, die individuell den gleichen Forschungsdatensatz beantragen, auch im längeren Zeitverlauf exakt die gleichen Daten bekommen und
- ▶ die Forschungsdaten methodisch-wissenschaftlichen Kriterien genügen, d. h., dass sie nachträglich nicht verändert, verfälscht oder gekürzt werden,
- ▶ der Antrags- und Genehmigungsprozess sowie die Einrichtung des Zugangs zu Forschungsdaten zeitnah erfolgt.

Zur Erzielung von wissenschaftlichen Fortschritten ist es zudem erforderlich, dass sich wissenschaftliche Diskurse auf eine im Zeitverlauf stabile und allgemein akzeptierte Basis – in diesem Fall Forschungsdatensätze – beziehen können. Der hierfür zumindest zurzeit am meisten erfolgversprechende Ansatz bei Forschungsdaten mit ursprünglichem Personenbezug ist dabei die Vergabe von DOI-Nummern (*Digital Object Identifier*).

Sobald Projektdaten in die verschiedenen Datenprodukte (SUF, SV, DFV/GWA) zerlegt sind und sich die entsprechenden (Teil-)Datensätze nicht mehr ändern, erhält man stabile digitale Objekte, die einzeln registriert und somit mit einer einzigartigen Nummer in einem weltweit stabilen DOI-Nummernsystem versehen werden können. Digitale Fingerabdrücke stellen sicher, dass einmal mit einer DOI-Nummer versehene Forschungsdatensätze nicht nachträglich verändert wurden.

Die Vergabe von DOI-Nummern bzw. deren Registrierung erfolgt über nationale Registrierungsagenturen. In Deutschland übernimmt diese Aufgabe für die Sozial- und Wirtschaftswissenschaften da|ra (*data registration agency for the social science and economic data*). da|ra ist über gesis – Leibniz-Institut für Sozialwissenschaften – an die DataCite-Initiative angeschlossen, die als eine der bei der *International Data Foundation* akkreditierten

Agenturen berechtigt ist, doi-Registrierungsprozesse zu initiieren.⁴ Für einen Antrag auf Registrierung von Forschungsdaten bei da|ra, d. h. um einen Forschungs-(Teil-)Datensatz mit einer DOI-Nummer zu versehen, müssen Beantragende diverse Verpflichtungen eingehen:

1. Jede DOI-Nummer für Forschungsdaten beinhaltet einen über die Zeit (> zehn Jahre) stabilen Link, unter dem das digitale Objekt, hier ein neuer Forschungsdatensatz, aufgerufen werden kann. Da ja, wie bereits erläutert, sozial- und wirtschaftswissenschaftliche Forschungsdaten selten im Internet zum Download bereitgestellt werden können, verweisen in diesem Fall DOI-Nummern stabil auf eine *landing page*, auf der die Forschungsdaten inhaltlich beschrieben werden und aus der hervorgeht, wie ein (dauerhaft existierender) Zugang zu dem Forschungsdatensatz für wissenschaftliche Forschungszwecke möglich ist.
2. Für die Forschungsdaten werden sogenannte Metadaten angelegt. Dabei handelt sich um bestimmte Informationen zu dem registrierten Objekt (z. B. Fallzahlen, Erhebungseinheit- und -methode), mit denen u. a. die Reichweite und Gültigkeit der jeweiligen Forschungsdaten beschrieben wird.
3. Es ist für mindestens zehn Jahre sichergestellt, dass die Daten stabil so wie beschrieben und im Zeitverlauf unverändert der *Scientific Community* zur Verfügung gestellt werden.

Für Primärforschende ist es sehr schwer, all diese Voraussetzungen bzw. Zusagen seriös zu erfüllen und ggf. zusätzlich noch die Datenauswertungen anderer Wissenschaftler/-innen über mindestens eine Dekade zu betreuen.

Üblicherweise geben Primärforschende ihre Primärdaten daher an Forschungsdatenzentren bzw. Datenarchive ab. Vom Rat für Sozial- und Wirtschaftsdaten⁵ akkreditierte FDZ sind vor diesem Hintergrund Mitglieder des da|ra-Konsortiums (bzw. können dies jederzeit werden) und demnach eigenständige Publikationsagenten für Forschungsdaten, d. h. einzelne FDZ können Forschungsdaten abhängig von ihrem Anonymisierungsgrad (SUF, SV, DFV/GWA) mit einer DOI-Nummer registrieren. Sie garantieren dann aber natürlich auch über längere Zeiträume alle die in den bisherigen Abschnitten genannten Leistungen, insbesondere den Zugang und die Nutzung von Forschungsdatenbeständen, die aus datenschutzrechtlichen Gründen nicht als SUF Forschenden überlassen werden dürfen. FDZ sind also demnach sowohl Datentreuhänder als auch Datenarchiv, indem sie verbürgen, dass erstmalig mit einer DOI-Nummer registrierte Forschungsdaten für alle Forschenden individuell in exakt der gleichen Art und Weise entweder zeitlich überlassen (SUF, CF) oder für Auswertungen bereitgestellt werden (SV, DFV, GWA).

Vor diesem Hintergrund ist es sinnvoll, wenn die im Original nicht frei und vor einem Antrags- und Vertragsverfahren auch für niemand einsehbaren Forschungsdaten aber zumin-

4 Vgl. URL: <https://www.da-ra.de/about> (Stand: 17.06.2022).

5 Vgl. URL: <https://www.konsortswd.de/datenzentren/akkreditierung/> (Stand: 17.06.2022).

dest öffentlich und frei zugänglich so beschrieben werden, dass man sich auch ohne einen direkten Blick in die Daten sehr weitreichend über den jeweiligen Forschungsdatensatz und seine jeweiligen inhaltlichen und methodischen Charakteristika informieren kann. Das ist für die Beurteilung von empirischen Analyseergebnissen in wissenschaftlichen Publikationen insbesondere im Rahmen von Review-Prozessen bedeutsam, weil es Rückschlüsse auf die Reichweite und Gültigkeit der eingesetzten Forschungsdaten für die Beantwortung der jeweiligen empirischen Forschungsfrage(n) zulässt.

Üblicherweise werden solche Informationen über Forschungsdatensätze als Metadaten bezeichnet. Metadaten geben Auskunft sowohl über den Entstehungskontext der jeweiligen Forschungsdaten als auch deren Charakteristika wie Fallzahlen, Erhebungseinheiten, Erhebungsmethode, Stichprobendesign, Fragebogen und vielem mehr.⁶ Mit Metadaten zu Forschungsdatensätzen lassen sich demnach Reichweite und Gültigkeit der Forschungsdaten bereits gut abschätzen und auf konkrete Anwendungsfälle in wissenschaftlichen Publikationen übertragen. Dies wiederum gestattet – insbesondere im Rahmen von gutachterlichen Prozessen – ansonsten nicht leistbare Rückschlüsse darauf, ob und wie gut die in einem wissenschaftlichen Artikel genutzten Forschungsdaten für die Beantwortung der jeweils entwickelten Fragestellungen bzw. empirische Testung der aufgestellten Hypothesen geeignet sind.

6 Forschungsdaten und Fachpublikationen

Die Nutzung von nicht frei zugänglichen Forschungsdaten für empirische Forschungsergebnisse bringt für referierte Fachbeiträge Besonderheiten mit sich. In aller Regel verlangen referierte Journals, dass die in einzelnen Artikeln den empirischen Auswertungen jeweils zugrunde liegenden Forschungsdaten veröffentlicht sind oder werden. Nahezu alle Verlage (eher selten: einzelne Journals) bieten hierfür ggf. eine eigene Plattform an,⁷ bei der die in einem Artikel ausgewerteten Forschungsdaten hochgeladen werden können. Sie werden dann zusammen mit dem Artikel veröffentlicht, d. h., die Forschungsdaten sind dann in der Regel frei zugänglich. In der Fachpublikation erfolgt meist ein Verweis auf die entsprechend veröffentlichten Forschungsdaten. Die Veröffentlichung von Forschungsdaten über einen Verlag eignet sich insbesondere für Primärforschende, deren Forschungsdaten keinen datenschutzrechtlichen (Zugangs-)Restriktionen unterliegen und die ihre Forschungsdaten aus welchen Gründen auch immer nicht über ein Datenarchiv oder ein FDZ verfügbar machen können oder möchten.

Sind die Forschungsdaten – wie in diesem Artikel ausführlicher beschrieben – hingegen nicht frei zugänglich, sind entsprechende Zugangsinformationen von den Forschenden darzulegen und mit dem Artikel zu veröffentlichen. Hierfür wird häufig die DOI-Nummer

6 Vgl. z. B. BIBB-FDZ, URL: <https://metadaten.bibb.de/> (Stand: 17.06.2022).

7 Zum Beispiel *Mendeley Data*, vgl. URL: <https://data.mendeley.com/> (Stand: 17.06.2022).

herangezogen bzw. auf diese verwiesen. Über die entsprechende *landing page* erschließen sich dann die weiteren für den wissenschaftlichen Begutachtungs- und Verwertungsprozess relevanten und bereits in den vorherigen Abschnitten besprochenen Informationen zu dem Forschungsdatensatz, etwa die Zugangs- und Nutzungskonditionen, aber auch Datendokumentationen wie den Fragebogen sowie methodische Hinweise zur Untersuchung wie etwa das Stichprobendesign oder das Auswahlverfahren.

Neben den in einer Fachpublikation ausgewiesenen Forschungsergebnissen erfolgt die Indizierung der verwendeten Forschungsdaten meist auf dreierlei Weise:

1. Nennung des Datensatznamens bei Tabellen oder Abbildungen mit empirischen Ergebnissen (FDZ/DA machen hierfür professionelle und praktikable Vorschläge);
2. die in Fachpublikationen verwendeten Forschungsdaten werden in einem eigenen Daten- und Methodenabschnitt vorgestellt und hinsichtlich ihrer Eignung für die jeweiligen Forschungsfragestellungen diskutiert;
3. die Zitation des digitalen Objekts erfolgt in Form einer Literaturangabe. Als Autorinnen bzw. Autoren werden meist die Primärforschenden geführt. Dann folgt der Langname des Forschungsdatensatzes als Titel. Als Herausgeber/-in fungiert in der Regel der/die Publikationsagent/-in der DOI-Nummer, also das FDZ/DA, in dem die jeweiligen Forschungsdaten gehostet werden. Solche Einträge werden in Fachpublikationen in Kurzform im Publikationstext zitiert und die Langform in das Literaturverzeichnis aufgenommen.

Zur Indizierung von Forschungsdaten in Fachpublikationen muss gesagt werden, dass von den hier beschriebenen Möglichkeiten sowohl in einzelnen Journals als auch von einzelnen Autorinnen und Autoren unterschiedlich intensiv Gebrauch gemacht wird. Häufig machen Journals keine Vorgaben zur Zitation von Forschungsdaten, und die in Abschnitt 4 dieses Beitrags beschriebenen Hinweise zur Verfügbarkeit der eingereichten Publikation zugrunde liegenden Forschungsdaten werden dann nur leider allzu oft zu einer Randnotiz. Grundsätzlich steht es Autorinnen und Autoren frei, ob und wie sie ihre Forschungsdaten in ihren Publikationen referenzieren. Dies drückt sich dann in entsprechend heterogenen bzw. wenig systematischen Indizierungen von Forschungsdaten in Fachartikeln aus.

Eine hochstandardisierte, flächendeckend von allen Journals verwendete Indizierung von Forschungsdaten gibt es demnach nicht einmal ansatzweise, und auch in der *Scientific Community* ist man von einer entsprechenden Vereinheitlichung bzw. der Formulierung entsprechender Kriterien weit entfernt. Zwar machen FDZ/DA diesbezügliche Zitations- und Indizierungsangebote, aber diese werden in Fachpublikationen unterschiedlich gut angenommen. Dies erschwert letztendlich auch die systematische Verknüpfung von Fachpublikationen, Forschungsdaten und Autorinnen bzw. Autoren etwa über automatisierte Suchmaschinen bzw. behaftet die entsprechenden Ergebnisse mit großen Unschärfen. Das ist insofern ein Problem, als bedingt durch die (aus datenschutzrechtlichen Gründen zumindest

in Deutschland) nicht anders mögliche Dezentralität von FDZ und DA eine systematische Erschließung von für die Forschung zur Verfügung stehenden Forschungsdatensätzen in Verbindung mit den Fachpublikationen, in denen die einzelnen Forschungsdatensätze zur Produktion von empirischen Forschungsergebnissen verwendet wurden, kaum stattfindet. Dabei sind solche gezielten Recherchemöglichkeiten gerade in ihrer Vollständigkeit immens wichtig, wenn eine neue Forschungsfragestellung verfolgt wird und man sich darüber informieren möchte, wer welche Forschungsfragestellungen mithilfe welcher Forschungsdaten beantwortet hat.

7 Zusammenfassung und Wertung

In den Sozial- und Wirtschaftswissenschaften und damit auch der Berufsforschung scheinen bei einer genaueren Betrachtung der Nähe von Forschungsdaten, wie sie vielleicht gerade bei OA-Fachpublikationen über Tabellen und Abbildungen nahegelegt wird, diese weit hinter einem Dunstschleier von datenschutzrechtlich geschützten Betriebs- oder Personendaten zu verschwinden. FDZ/DA haben aber im Laufe der Zeit Mittel und Wege gefunden, solche geschützten Forschungsdaten in verschiedenen Produktlinien (PUF, SUF, CF, SV, DFV, GWA) für die *Scientific Community* in einer forschungsfreundlichen Form verfügbar zu machen.

Darüber hinaus werden umfangreiche Metadaten zu diesen Forschungsdatensätzen angeboten, sodass man sich auch ohne einen direkten Zugriff auf die Forschungsdaten ein vergleichsweise genaues Bild von den Daten und ihrer jeweiligen Erhebungsmethodik sowie deren inhaltlicher Reichweite und Gültigkeit machen kann. Insofern sind für Interessierte gut dokumentierte sozial- und wirtschaftswissenschaftliche Forschungsdaten eher nah am OA-Gedanken für Publikationen zu verorten, etwa weil Metadaten zu Forschungsdatensätzen frei im Netz für jedermann verfügbar sind. Der Zugang zu geschützten Forschungsdaten über Personen oder Betriebe ist zwar nur für Forschungszwecke möglich, aber ansonsten so niedrigschwellig wie möglich gehalten und in der Regel kostenfrei für Nutzende.

Wissenschaftlichen Journals genügen diese Kriterien, um die prinzipielle Zugänglichkeit der in Fachpublikationen verwendeten Forschungsdaten nachzuweisen. Als Referenzsystem für Forschungsdaten dienen dabei DOI-Nummern, Klarnamen des Datensatzes als Quellenangaben sowie Zitationsvorschläge für Forschungsdaten und der mit ihnen verbundenen FDZ-Daten- und Methodenberichte. Mit diesen Angaben kann sich die *Scientific Community* die wichtigsten Charakteristika von in Fachpublikationen verwendeten Forschungsdaten und deren Anwendung auf einzelne Forschungsfragestellungen im wissenschaftlichen Diskurs oft auch ohne einen direkten Zugriff auf die Forschungsdaten erschließen.

Dies bedeutet nun keineswegs, dass es einheitliche und allgemein anerkannte Referenz- und Nachweissysteme für Forschungsdaten und Fachpublikationen gibt. Der diesbezügliche Umgang einzelner Journals mit Forschungsdaten ist vergleichsweise heterogen, was insbesondere die Recherche der Verbindung von Fachpublikationen bzw. deren Forschungsfragestellungen mit für eigene Auswertungen nutzbare Forschungsdaten erschwert. Es gibt

zwar – etwa mit ORCID, der *Open Researcher Contributor ID*⁸ – Initiativen, die versuchen, verbesserte Nachweissysteme des Dreiecks Fachpublikationen, Autorinnen/Autoren und Forschungsdaten zu schaffen (vgl. den Beitrag von Kathrin Höhner zu ORCID in diesem Band), aber alle diese Initiativen sind von einem allgemein angewendeten Standard, wie es etwa bei Literaturnachweissystemen der Fall ist, noch weit entfernt. Das (verhältnismäßig junge) Mitdenken von Forschungsdaten mit all den Aspekten ihrer Entstehung und Nutzung innerhalb dieser Prozesse bzw. Systeme sickert erst langsam in eine bereits länger existierende Forschungs- und Publikationskultur ein.

Gleichwohl wird der fundamentalen Bedeutung von Forschungsdaten für den modernen wissenschaftlichen Erkenntnisfortschritt im Zeitverlauf vermehrt Rechnung getragen. In Deutschland drückt sich dies etwa durch eine vermehrte Gründung und Fortschreibung von FDZ/DA aus. In den Sozial- und Wirtschaftswissenschaften und damit auch der Berufsfor-schung führt dies aufgrund der Eigenheit personenbezogener Daten zu einer unausweichlich dezentralen Struktur der FDZ bzw. DA als Herausgeber/-innen bzw. Hoster von Forschungsdaten. Beim Zugang und der Nutzung von Forschungsdaten in der Breite entstehen erhöhte Standardisierungs- und Datendokumentationsbedarfe, weil FDZ-Strukturen notwendigerweise dezentral und damit immer auch in gewisser Art und Weise unvermeidbar heterogen sind.

Summa summarum wird man für den Bereich von Forschungsdaten über Personen oder Betriebe in absehbarer Zeit kaum freiere und offener Zugänge sowie eine möglichst unkomplizierte und die eigentliche Forschungsaktivität wenig beeinflussende Art der Nutzung von solchen Forschungsdaten als über FDZ und DA realisieren können. Für die Forschung erfolgen jedenfalls so zurzeit niedrigschwellige Zugänge und Nutzungen von qualitativ hochwertigen Forschungsdaten jederzeit zeitnah und standardisiert. Ein Zugang zu Forschungsdaten der Sozial- und Wirtschaftswissenschaften über FDZ bzw. DA entspricht somit am ehesten dem OA-Gedanken bei Fachpublikationen. Die Bereitstellung einer entsprechenden Forschungsdateninfrastruktur ist bezüglich der Forschungs- und Metadatenproduktion sowie einer forschungsfreundlichen Nutzung einzelner Forschungsdatensätze kostenintensiv. Die Kosten für ein FDZ entstehen dabei notwendigerweise auf der Ebene der einzelnen Organisationen, in denen Forschungsdaten anfallen und/oder produziert werden, während die *Scientific Community* als Ganzes einen kostenfreien und auch ansonsten vergleichsweise niedrigschwelligen Datenzugang bekommt. Auf diese Art und Weise verzahnen sich universitäre und außeruniversitäre Forschung oftmals besonders eng.

8 Vgl. URL: <https://info.orcid.org/orcid-for-research-resources/> (Stand: 17.06.2022).