KonsortSWD ist das NFDI Konsortium für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften. Für die äußerst vielfältigen Datentypen und Forschungsmethoden bauen die Beteiligten im Rahmen der NFDI eine bereits bestehende Forschungsdateninfrastruktur aus und ergänzen neue integrierende Dienste. Basis sind die heute 41 vom Rat für Sozial- und Wirtschaftsdaten akkreditierten Forschungsdatenzentren (FDZ). FDZ sind Spezialsammlungen zu jeweils spezifischen Forschungsdaten, z. B. aus der qualitativen Sozialforschung, und können so Forschende auf Basis einer ausführlichen Expertise zu diesen Daten beraten. Neben der Unterstützung der FDZ baut KonsortSWD auch neue Dienste in den Bereichen Datenproduktion, Datenzugang und Technische Lösungen auf.

KonsortSWD is the NFDI consortium for the social, behavioural, educational and economic sciences. The stakeholders are expanding an existing research data infrastructure within the NFDI to accommodate these highly diverse types of data and research methods, and adding new integrating services. The 41 research data centres (RDCs) already accredited by the German Data Forum constitute the basis for this. The RDCs are special collections of specific research data, e.g. from qualitative social research, and can thus advise researchers on the basis of their detailed expertise on the relevant data. In addition to supporting the RDCs, KonsortSWD is also establishing new services in the areas of data production, data access and technical solutions.

ANDREAS BLÄTTE, ANNA FRÄSSDORF, JAN-OCKO HEUER, UTE HOFFSTÄTTER, CHRISTOPH LEONHARDT, LAURA MENZE, BERNHARD MILLER, KATI MOZYGEMBA, DAGMAR PATTLOCH, JULIA RAKERS, SILKE REINEKE, THOMAS RUNGE, FRIEDERIKE SCHLÜCKER, THOMAS SCHMIDT, KNUT WENZIG, CHRISTOF WOLF

Eine Dateninfrastruktur für die Gesellschaftswissenschaften

Unterstützung in der Arbeit mit Forschungsdaten durch KonsortSWD

Anforderungen an Daten in den Gesellschaftswissenschaften

Daten, mit denen unsere Gesellschaft beschrieben und analysiert wird, sind so vielfältig wie die Fragen, die sie zu beantworten helfen. Verhalten lässt sich in Experimenten beobachten, Einstellungen werden oft in Umfragen erfasst, amtliche Daten erlauben Einblicke in Einkommensverhältnisse oder Altersvorsorge, Protokolle vermitteln ein Verständnis von Verhandlungsprozessen in Parlamenten, Interviews können der einzige Weg sein, die Motive hinter Entscheidungen zu verstehen, Leistungstests der gewählte Weg, um die Kompetenz von Schüler*innen zu erfassen.

Daten, die Forscher*innen benötigen, um Theorien zu überprüfen oder Muster zu finden, enthalten oft persönliche und schützenswerte Informationen über Menschen oder Organisationen. Nicht selten sind diese Informationen für andere Zwecke, z. B. der amtlichen Statistik, gesammelt worden und können nicht ohne weiteres beforscht werden. Auch wissenschaftliche Umfragedaten enthalten oft personenbezogene – und damit schützenswerte – Informationen. Während dies eine besondere Herausforderung für die Forschung ist, existieren gleichzeitig auch spannende Daten aus öffentlichen Quellen. Die Herausforderung ist hier, Datenschätze zu finden und sicherzustellen, dass sie die erforderliche Qualität und die richtigen Inhalte haben.

Maßstab für eine FAIRe[1] Forschungsdateninfrastruktur können nur die Bedarfe der Forschenden sein. Abgeleitet aus diesen Bedarfen entstehen Dienste, die Unterstützung bei der Arbeit mit Forschungsdaten anbieten. Diese können sich z. B. auf die Erschließung, Dokumentation, Verknüpfung, Auswertung oder langfristige Sicherung von Daten beziehen. Das Angebot an solchen Diensten auf- und auszubauen ist Zielsetzung von KonsortSWD, dem Konsortium für die Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften in der Nationalen Forschungsdateninfrastruktur (NFDI).

Dieser Beitrag gibt einen Überblick über die Vielfalt an Daten und die Forschungsdatenzentren (FDZ) als erprobten und belastbaren Weg zu ihrer Bereitstellung. FDZ sind eine Recherche- und Erschließungsinfrastruktur, an der vielfach aber auch Datenprodukte erstellt werden und Forschende mit diesen Daten arbeiten können. Darüber hinaus beschreibt der Beitrag den Auf- und Ausbau von Diensten und Angeboten, die die Nutzung dieser und anderer Forschungsdaten unterstützen.

Zunächst verortet er aber die Einbindung der Forschenden als grundlegendes Architekturmerkmal der NFDI in die Prozesse von KonsortSWD.[2]

KonsortSWD und seine Communitys

Sicherzustellen, dass Dienste zum Forschungsdatenmanagement (FDM) geeignet sind, die Forschung zu unterstützen, ist ein zentrales Anliegen der NFDI. Bund und Länder beschreiben als Voraussetzung daher: »Die NFDI wird von Nutzern von Forschungsdaten und von Infrastruktureinrichtungen ausgestaltet, die dazu in und zwischen Konsortien zusammenarbeiten.«[3] Um dies zu leisten, müssen die Konsortien in der jeweiligen Fachgemeinschaft verankert sein. KonsortSWD definiert seine Community dabei vorrangig über Disziplinen. Es wird angestrebt, die Forschung mit den dort üblichen, vielfältigen Datentypen zu unterstützen. Disziplinen, für die eine Infrastruktur bereitgestellt wird, sind insbesondere die Sozialwissenschaften (z. B. Anthropologie, Politikwissenschaft, Soziologie), Verhaltenswissenschaften (Psychologie), Bildungswissenschaften (Erziehungswissenschaft und Bildungsforschung) und Wirtschaftswissenschaften (Volks- und Betriebswirtschaftslehre).

Für die Zusammenarbeit mit den Communitys bestehen zwei gut eingespielte Wege: Zum einen der Ständige Ausschuss Forschungsdateninfrastruktur (FDI-Ausschuss, s. u.), vor allem aber der Rat für Sozial- und Wirtschaftsdaten (RatSWD), der seit 2020 vollständig aus Mitteln des Konsortiums finanziert wird. Er besteht seit 2004 als Beratungsgremium der Bundesregierung und ist ein Forum, in dem Wünsche der Forschenden an die Dateninfrastruktur der einschlägigen Fächer formuliert werden. Der Rat besteht dabei zur Hälfte aus gewählten Wissenschaftler*innen, die auf Vorschlag von derzeit 15 Fachgesellschaften nominiert werden.[4] Die andere Hälfte seiner Mitglieder stammt aus Einrichtungen der Forschungsdateninfrastruktur.[5] Dank dieser Konstruktion kann der RatSWD schon seit fast zwei Jahrzehnten Wünsche aus verschiedenen Forschungsdatenkulturen artikulieren und wichtige Impulse für die Infrastruktur setzen.

Die Vielzahl der Fachgesellschaften im RatSWD lässt bereits die Pluralität der Forschungsansätze und Datentypen erahnen. Für Forschungszwecke erhobene Umfragedaten sowie amtliche Daten (z. B. aus Erhebungen der Statistischen Ämter oder der Rentenversicherung) und Daten aus Unternehmen (inkl. Markt- und Preisinformationen) oder der Zivilgesellschaft (bspw. aus Vereinen) sind schon länger Teil der Infrastruktur. Zunehmend spielen auch Audio-, Video- sowie Textdaten eine Rolle. Aus Sicht des Zugangs ist dabei vor allem die Granularität der Daten von großer Wichtigkeit: Textdaten können (z. B. als Interviews) genauso personenbezogen sein wie amtliche oder Umfragedaten und sind damit in besonderem Maße sensibel – also aus forschungsethischen und/oder datenschutzrechtlichen Gründen schützenswert. Aggregierte Informationen, wie zum Beispiel Kennzahlen zur Wirtschaftsleistung, sind es hingegen nicht. Auf diese Unterschiede kann die FDZ-Struktur eine Antwort geben.

Einschlägige Datenbestände für die Communitys: Forschungsdatenzentren (FDZ)

FDZ als Spezialsammlungen

FDZ sind Infrastruktureinrichtungen mit dem Ziel, forschungsnahen und rechtskonformen Zugang zu Forschungsdaten zu gewährleisten. Dies ist wichtig, wenn schützenswerte Daten nicht als Open Data bereitgestellt werden können. FDZ schaffen damit einen Interessenausgleich zwischen datennutzenden Wissenschaftler*innen, datenhaltenden Institutionen und/oder den datengenerierenden Studienteilnehmer*innen. FDZ sind damit auch Spezialsammlungen für Forschungsdaten, die ausführliche Expertise zu ihren Daten besitzen. In dieser Funktion unterscheiden sich FDZ von Repositorien, in denen frei zugängliche Daten zur Verfügung gestellt werden.

Mit der Bereitstellung eines rechtskonformen Datenzugangs ermöglichen FDZ die Nachnutzung von (sensiblen) Forschungsdaten. Diese Nachnutzung sichert einerseits die Reproduzierbarkeit von Forschungsergebnissen im Sinne guter wissenschaftlicher Praxis und erlaubt andererseits, neue Fragestellungen mit vorhandenem Datenmaterial zu untersuchen.[6] Auch werden durch FDZ Daten für die wissenschaftliche Nutzung zugänglich, die ursprünglich gar nicht für Forschungszwecke erhoben wurden (z. B. Daten aus der amtlichen Statistik, den Sozialversicherungen oder des Ausländerzentralregisters).[7]

FDM ist notwendig, um Forschungsdaten von einem »Rohformat« in eine Form zu bringen, die eine Nachnutzung (auch) durch nicht unmittelbar an der Datenerhebung beteiligte Forschende ermöglicht. Dementsprechend gehören Datenaufbereitung, Datendokumentation und Langzeitarchivierung zu den Kernaufgaben der FDZ.[8] Hierfür ist ein hohes Maß an Datenkenntnis nötig. Indem zahlreiche datenanbietende Institutionen ein eigenes FDZ betreiben, ermöglichen sie mit ihrer Expertise eine »höchstmöglichste Qualität bei der Erschließung und Aufbereitung«.[9] Gleichzeitig verbleiben bei diesem dezentralen Ansatz die Daten überwiegend bei den Datenanbieter*innen. Dies erlaubt auch die Berücksichtigung von sensiblen Merkmalen, die beispielsweise von staatlichen Stellen erhoben werden.[10] Darüber hinaus bieten FDZ mit einem fachlichen oder methodischen Schwerpunkt (z. B. für qualitative Daten) ihre spezifische Expertise auch Primärforschenden an, die ihre Daten an ein FDZ übergeben und für die Sekundärnutzung zur Verfügung stellen möchten.[11]

Ein besonderes Merkmal vieler Forschungsdaten in den KonsortSWD Disziplinen stellt deren Personenbezug dar. Dies beinhaltet »alle Informationen, die sich auf eine identifizierte oder identifizierbare Person beziehen«.[12] In Bezug auf natürliche Personen sind Merkmale, wie z. B. die politische Meinung, ethnische Herkunft, Gesundheitsinformationen oder sexuelle Orientierung Kategorien, die aufgrund ihrer Sensibilität zusätzlich Schutz erfordern.[13] Wie können solche Daten zur wissenschaftlichen Nachnutzung zur Verfügung gestellt und gleichzeitig die Interessen der betroffenen Personen gewahrt werden? Dies erfolgt in erster Linie auf Grundlage der Einwilligung der Betroffenen in die Datenerhebung, -verarbeitung und -veröffentlichung (diese muss durch die Forschenden eingeholt werden). Darüber hinaus kommen Maßnahmen zum Einsatz, die die Identifizierung von betroffenen Personen verhindern (Anonymisierung oder Pseudonymisierung der Daten sowie ein vertraglich festgelegtes Verbot von De-Anonymisierungsversuchen in Datennutzungsverträgen).[14] Je stärker Daten dabei durch Löschung oder Reduzierung des Informationsgehalts anonymisiert werden, desto geringer ist allerdings deren Forschungspotenzial und damit der Wert für eine Nachnutzung.

FDZ bieten daher unterschiedliche Zugangswege und -formate an, die sich im Grad der Anonymisierung der Daten unterscheiden. An Gastwissenschaftsarbeitsplätzen in einem gesicherten Raum können Forschende auf sensible Daten zugreifen (On-Site-Nutzung). Eine abgesicherte Off-Site-Nutzung der Daten erfolgt mittels Remoteverbindung im Sinne einer (kontrollierten) Datenfernverarbeitung, bei der sämtliche Analysen auf einem Server des datenhaltenden Instituts erfolgen und Ergebnisse nur nach Kontrolle durch das FDZ exportiert werden können. Schließlich besteht die Option, die Daten per Download, E-Mail oder Postversand bereitzustellen. Der Grad der Anonymisierung reicht von Scientific Use Files (SUF, mindestens faktisch anonymisiert, für wissenschaftliche Forschung) über Campus Use Files (CUF, für universitäre Lehre bestimmt, stärker anonymisiert als SUF) bis hin zu vollständig anonymen Public Use Files (PUF, auch für Weitergabe außerhalb wissenschaftlicher Forschung geeignet).[15] Für die Nutzung zugangsbeschränkter Daten muss in aller Regel ein entsprechender Vertrag abgeschlossen werden, der die Wahrung gesetzlicher Bestimmungen regelt. Die Komplexität rechtlicher Vorgaben macht die Expertise hierzu in den FDZ umso bedeutsamer (s. Abschnitt Harmonisierung).

Seit Gründung der FDZ der Statistischen Ämter des Bundes und der Länder 2001 hat sich ein stetig wachsendes Netzwerk etabliert, das seit 2009 den »Ständigen Ausschuss Forschungsdateninfrastruktur« (FDI-Ausschuss) für den Austausch nutzt.[16] Dieser verfolgt das Ziel, die Qualität und Quantität des Datenangebots sowie den Datenzugang für Forschende zu verbessern. Mitglied im FDI-Ausschuss werden FDZ nach erfolgreicher Akkreditierung durch den RatSWD.[17] Mit Stand Dezember 2021 sind 41 FDZ (vorläufig) akkreditiert.[18] Das Datenangebot der FDZ reicht von Sozial-, Wirtschafts-, Bildungs-, und Gesundheitswissenschaften über Psychologie und Korpuslinguistik bis hin zu Humangeographie. Die Datensuche in den FDZ kann für die meisten FDZ zentral hier erfolgen: https://www.konsortswd.de/datenzentren/datensuche-in-den-fdz/

Bis Ende 2020 boten FDZ insgesamt 4.971 Datensätze an, wobei ein Datensatz aus mehreren Einzelstudien bestehen kann (und die angebotene Zahl an Studien somit weit höher liegt).[19] Als ein Indikator für die Bedeutung der FDZ in der Forschungslandschaft gilt der Umfang der Datennutzung. Bis Ende 2020 gab es an 37 FDZ 38.219 bestehende Datennutzungsverträge. An 17 FDZ können Datensätze ohne Nutzungsvereinbarung bezogen werden. 2020 wurden diese frei verfügbaren Datensätze mehr als 68.000-mal heruntergeladen. Die Anzahl der Datennutzenden, außerhalb der Trägereinrichtung des FDZ lag 2020 bei insgesamt 43.703 Personen. Auf Basis der bereitgestellten Forschungsdaten entstanden zudem insgesamt 2.906 Publikationen, wobei referierte Zeitschriften den größten Anteil ausmachen. Die Zahl steigt kontinuierlich, ist aber vermutlich zu niedrig, da Forschende oft die Rückmeldung an FDZ versäumen.[20]

Qualitätssicherung

Angesichts heterogener Datentypen und unterschiedlicher Qualitätsstandards in den Disziplinen sind die Anforderungen an ein System zur Qualitätssicherung groß. Dabei wird die Qualitätssicherung der Forschungsdaten jeweils in den FDZ vorgenommen. Interne Qualitätssicherungsmaßnahmen von Daten sowie von Prozessen für deren Erhebung, Dokumentation und Bereitstellung schließen z. B. Beiräte oder regelmäßige Evaluationen von Trägerorganisationen ein. Daneben profitiert das FDZ-Netzwerk von der Verbindung mit dem RatSWD.

2010 haben RatSWD und FDZ gemeinsam ein Konzept entwickelt, wie die Qualität der Forschungsdateninfrastruktur nachhaltig sichergestellt und kontinuierlich weiterentwickelt werden kann. Dazu haben die Beteiligten Mindeststandards und Kriterien für die Akkreditierung von FDZ ausgearbeitet. So wird eine wissenschaftsinterne, nutzerorientierte Zertifizierung von Datentreuhandmodellen realisiert, die sicherstellt, dass Daten im Bereich der Sozial-, Verhaltens-, Bildungs- und Wirtschaftsdaten – im Sinne der FAIR-Prinzipien – zugänglich, dokumentiert und langfristig verfügbar sind.

Seit 2016 begleitet und überwacht eine Monitoringkommission gemeinsam mit dem RatSWD die Entwicklung der FDZ-Landschaft.[21] Voraussetzung für die Akkreditierung sind unter anderem Informationen zu Datenangebot, Struktur der jeweiligen Einrichtung, in-

ternen Archivierungs- und Qualitätssicherungskonzepten oder Dienstleistungen für Nutzende. Über eine Akkreditierung entscheidet der RatSWD. Voraussetzung ist, dass ein FDZ seit mindestens sechs Monaten Daten anbietet, die nachweislich bereits für wissenschaftliche Zwecke genutzt wurden.

Ein regelmäßiges, öffentlich einsehbares Berichtswesen sichert darüber hinaus die Einhaltung der Akkreditierungskriterien. Erfüllt ein FDZ die Qualitätskriterien nicht mehr in erforderlichem Maße, kann eine Akkreditierung widerrufen werden. »Langfristig hat die Akkreditierung durch den RatSWD zu vergleichbaren Qualitätsstandards und dadurch erleichtertem Transfer zwischen genuin in der Wissenschaft und außerwissenschaftlich erhobenen Daten und Datensätzen geführt.«[22] Um darüber hinaus Nutzenden die Möglichkeit zu geben, Mängel im Datenangebot eines akkreditierten FDZ zu melden, kann eine Meldung an die Beschwerdestelle des RatSWD erfolgen.

Angebote und Erfahrungen der FDZ

Der folgende Abschnitt gibt einen beispielhaften Einblick in die Arbeit von zwei der derzeit 41 akkreditierten FDZ, die sich in dem von KonsortSWD getragenen FDI-Ausschuss (s. o.) vernetzen und für deren Nutzende das Konsortium neue Dienste entwickelt.

Das Archiv für Gesprochenes Deutsch (AGD)

Das Archiv für Gesprochenes Deutsch (AGD) ist ein Forschungsdatenzentrum für Korpora des gesprochenen Deutsch. Es richtet sich in erster Linie an Forschende der Sprachwissenschaft, die in der Interaktionsforschung, der Variationslinguistik, der Korpuslinguistik oder weiteren linguistischen Teildisziplinen mit audiovisuellen Sprachdaten des Deutschen arbeiten. Das AGD ist als Teil des Programmbereichs »Mündliche Korpora« des Leibniz-Instituts für Deutsche Sprache seit 2014 akkreditiert und *Participant* bei KonsortSWD.[23]

Der Bestand des AGD umfasst digitalisierte Bestände von etwa 100 Korpora des mündlichen Sprachgebrauchs mit insgesamt ca. 11.000 Stunden Audio- und Videoaufnahmen sowie 14.000 Transkripten und zugehörigen Metadaten. Etwa die Hälfte dieser Korpora sind erst rudimentär erschlossen, die wichtigsten und vom Umfang her größeren Teile des Datenbestandes haben aber eine umfassende Aufbereitung erfahren und werden der wissenschaftlichen Öffentlichkeit über die Datenbank für Gesprochenes Deutsch (DGD) und einen persönlichen Archivservice zur Nachnutzung angeboten. Diese Ressourcen fallen im Wesentlichen in drei Klassen: Gesprächskorpora, die die sprachliche Interaktion in verschiedensten Gesellschaftsbereichen abbilden, z. B. das Forschungs- und Lehrkorpus Gesprochenes Deutsch oder das Korpus Gesprochene Wissenschaftssprache Kontrastiv, Variationskorpora, die die regionale Variation des Sprachgebrauchs im Deutschen abbilden, z. B. das Korpus Deutsche Mundarten (Zwirner-Korpus) oder das Korpus Deutsch in Namibia sowie Interviewkorpora, die über biografische/narrative Interviews einerseits auf den Sprachgebrauch spezifischer Sprechertypen (z. B. deutschsprachige Emigranten in Israel) fokussieren, andererseits aber auch einen zeithistorischen Aspekt aufweisen, der sie über die Linguistik hinaus für sozialwissenschaftliche oder Oral History-Studien interessant macht, z. B. die Bonner Längsschnittstudie des Alterns (BOLSA) oder das Berliner Wendekorpus.

Das AGD wirbt zudem fortlaufend neue Korpora für seine Bestände ein. Es übernimmt Korpora aus abgeschlossenen Projekten oder Nachlässen, bietet aber auch ein Kooperationsmodell an, in dem Forschungsprojekte bereits während der Planung zu Fragen des FDM und digitaler Methodik beraten und begleitet werden und im Gegenzug die entstehenden Daten dann zum Abschluss ins Archiv überführen.

Die Angebote des AGD werden von der Forschungscommunity sehr gut angenommen, wie mehr als 14.000 Registrierungen bei der DGD (seit 2012) und die steigende Anzahl an Datenübernahmen, Beratungs- und Kooperationsanfragen zeigen. Besonders in den letzten Jahren zeichnen sich über die (germanistische) Linguistik hinaus auch zunehmend interdisziplinäre Perspektiven ab, sowohl durch Übernahme von Daten, die ursprünglich in anderen disziplinären Kontexten entstanden sind (z. B. BOLSA, s. o.), als auch durch Nachfragen aus anderen Disziplinen zur Nachnutzung von originär sprachwissenschaftlichen Daten (etwa die Analyse der Mundartkorpora aus den 1950er-Jahren durch ein kulturhistorisch forschendes Projekt). Nicht zuletzt deutet sich an, dass die ursprünglich für die Linguistik entwickelten Tools und Standards, die im AGD zum Einsatz kommen, auch für das FDM in anderen Disziplinen von Interesse sind. So können etwa Sammlungen qualitativer Interviews aus der Sozialforschung, die in KonsortSWD vor allem über *QualidataNet* (s. u.) repräsentiert sind, oder videographierte Unterrichtskommunikation in der Bildungsforschung von (semi-)automatischen Annotationsverfahren profitieren, die für die Linguistik entwickelt wurden.

Gleichwohl bringt die steigende Nachfrage das FDZ-AGD trotz stetiger Bemühung um Effektivierung und Optimierung der Arbeitsabläufe an Kapazitätsgrenzen. Das Modell der »frühen Kooperation« (s. o.), das gegenüber einer Datenübernahme nach Projektabschluss den Aufwand für die Datenaufbereitung deutlich reduziert, wird daher in Zukunft präferiert werden.

Das FDZ der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin

Als Ressortforschungseinrichtung des Bundes führt die Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA) vielfältige Studien zum Thema Arbeit und Gesundheit durch. Sie betreibt ein eigenes FDZ, das 2021

durch den RatSWD akkreditiert wurde und *Participant* bei KonsortSWD ist.

Das FDZ-BAuA stellt Daten aus hauseigenen Studien zur Nachnutzung zur Verfügung. Das Datenangebot konzentriert sich zunächst auf Daten aus quantitativen Befragungen und wird derzeit über zwei Zugangswege bereitgestellt: Scientific Use Files (SUFs) zur Nutzung außer Haus machen den Kern des Angebotes aus. Sie stehen ausschließlich für wissenschaftliche Zwecke zur Verfügung, und ihre Nutzung erfordert einen Vertragsabschluss zwischen der BAuA und den Forschenden. Aktuell werden SUFs zu zwei repräsentativen Panelstudien von Erwerbstätigen in Deutschland – der BAuA-Arbeitszeitbefragung und der Studie zur mentalen Gesundheit bei der Arbeit – angeboten. Mit Public Use Files (PUFs) wurde darüber hinaus ein niedrigschwelliges Angebot für die breite Fachöffentlichkeit geschaffen. Diese stärker anonymisierten Daten stehen zum freien Download zur Verfügung, und es gelten einfachere Nutzungsbedingungen. Der erste verfügbare PUF enthält eine sog. Job-Exposure-Matrix auf Basis der BIBB/BAuA-Erwerbstätigenbefragung 2018. Datenangebot und Zugangswege des FDZ-BAuA sollen zukünftig weiter ausgebaut werden.

Ein Blick zurück: In seiner Evaluation im Jahr 2017 unterstützte der Wissenschaftsrat die Gründung eines FDZ an der BAuA nachdrücklich. Die zwei wichtigsten Voraussetzungen hierfür – wertvolle Daten und der Wille zum Teilen – lagen vor. Der Aufbau des FDZ-BAuA erfolgte daraufhin in mehreren Schritten: Zunächst wurde ab 2018 in einem Aufbauteam die Expertise aus unterschiedlichen Bereichen des Hauses – datenerhebende Fachgruppen, Forschungs- und Entwicklungsmanagement, Justiziariat, Datenschutz, IT, strategische Kommunikation – gebündelt, um die Voraussetzungen für die FDZ-Gründung zu erarbeiten. Mitte 2019 konnte das FDZ dann mit einem ersten Datenangebot in den Pilotbetrieb starten und mit der Rekrutierung des Kernteams Anfang 2021 schließlich in den Regelbetrieb übergehen. Interesse am Datenangebot des FDZ-BAuA ist seit seiner Gründung vor allem bei Forschenden aus der Arbeits- und Organisationspsychologie, den Arbeitswissenschaften und der Arbeitsmedizin, aber auch aus den breiteren Sozial- und Wirtschaftswissenschaften zu beobachten.

Im Aufbauprozess galt es, vielfältige Fragen zu klären, wie beispielsweise: Welche Daten können angeboten werden? Wie soll mit teilweise konfligierenden Anforderungen in Bezug auf den Zeitpunkt der Verfügbarmachung von Daten umgegangen werden? Wer darf zu welchen Nutzungsbedingungen Datenzugang erhalten?

Mit den ersten eigenen praktischen Erfahrungen wurden die ursprünglichen Verfahren weiterentwickelt und auch Anregungen des RatSWD aus dem Akkreditierungsverfahren aufgegriffen. Für die Weiterentwicklung des FDZ-BAuA ergibt sich eine besondere Herausforderung aus der großen, in der BAuA vertretenen interdisziplinären Bandbreite, die es zusammenzubringen gilt. Das FDZ-BAuA wurde für seinen weiteren Ausbau durch die Akkreditierung des RatSWD gestärkt und möchte – im Austausch mit den anderen akkreditierten FDZ – kontinuierlich hinzulernen, beispielsweise im Hinblick auf Datenzugangswege, die Bekanntmachung der Daten oder das Nutzendenmanagement. Das ermutigt zum einen die Forschenden des Hauses, ihre »Datenschatztruhen« zu öffnen, und zum anderen die Fachcommunitys, die angebotenen Daten für eigene Forschungszwecke zu nutzen.

Übergreifende Bedarfe: Vielfalt an Daten und Ausbau der Infrastruktur

Mit dem RatSWD und den FDZ als Basis baut KonsortSWD eine bestehende Infrastruktur aus. Hierfür hat das Konsortium die Bedarfe der Communitys gesammelt und bearbeitet sie in drei inhaltlichen Bereichen (Task-Areas): Datenproduktion, Datenzugang und Technische Lösungen. Die Task-Area zur Datenproduktion folgt der Einsicht, dass schon bei der Erzeugung

1 Dienste von KonsortSWD

von Forschungsdaten die Grundlagen für die FAIRness dieser Daten gelegt und damit die Nachnutzbarkeit verbessert wird (s. u. die Aktivitäten zu Textdaten). Im Bereich Datenzugang gilt die Aufmerksamkeit einer Reihe von Vereinfachungen für Forschende. Durch ihre dezentrale Natur und die verschiedenen rechtlichen Rahmenbedingungen in den datenanbietenden Institutionen haben sich in der Entwicklungsgeschichte der Infrastruktur in unseren Disziplinen teils unterschiedliche Standards (z. B. bei Datennutzungsverträgen) gebildet. Auf der technischen Seite stehen im Sinne von FAIR die Auffindbarkeit der dezentral vorhandenen Daten sowie die Versorgung mit persistenten Identifikatoren (Nachnutzbarkeit) im Zentrum. Abbildung 1 zeigt die Dienste, die derzeit auf- und ausgebaut werden, ebenso wie zentrale Mechanismen zum Ausbau der Infrastruktur (rechte Spalte) und den RatSWD als einer wesentlichen Rückbindung an die Communitys.

Einige ausgewählte Aktivitäten sollen im Folgenden vorgestellt werden.

Auf dem Weg zu mehr Harmonisierung im FDZ-System

FDZ entstehen in der Regel an Daten generierenden Institutionen mit dem Ziel, diese Daten zur Verfügung zu stellen. Ursprünglich für die Erschließung amtlicher Daten gedacht, entstehen inzwischen immer mehr wissenschaftliche Datenzentren.[24] Dementsprechend weisen FDZ diverse institutionelle Kontexte, Rechtsgrundlagen, Datentypen und Ressourcen auf. So bestehen unterschiedliche Arbeitsabläufe, es existiert noch wenig standardisiertes Material (z. B. Verträge), und nicht alle Prozesse sind systematisch beschrieben. KonsortSWD erarbeitet daher Unterstützungsangebote[25] für zentrale Bereiche und Aufgaben der FDZ. So sollen Referenzmodelle und -dokumente entstehen. Diese dienen auch als Vorbereitung für die Akkreditierung von FDZ durch den RatSWD sowie die Zertifizierung nach dem CoreTrustSeal.[26] Beides sind wichtige Belege für die Vertrauenswürdigkeit eines FDZ.

Konkret sollen allgemeine Referenzmodelle für die zentralen Funktionsbereiche Datenzugriff und Datenaufnahme erstellt werden. Diese bauen auf zentralen Konzepten wie dem OAIS (Open Archival Information System)[27] und den FAIR-Prinzipien[28] auf. Die allgemeinen Referenzmodelle sollen FDZ helfen, diese zentralen Funktionsbereiche zu harmonisieren, auch wenn Daten oder Prozesse aufgrund der Entstehungsgeschichte heterogen sein mögen. Zentrale Referenzdokumente in diesen Funktionsbereichen sind Datennutzungsverträge im Bereich der Datennutzung bzw. Datenaufnahmeverträge bei der Aufnahme externer Daten. Auch diese werden bislang meist von der jeweiligen Institution erstellt und weisen trotz einer sehr ähnlichen Zielsetzung Unterschiede in Gliederung, Umfang und inhaltlicher Ausgestaltung auf. Mithilfe juristischer Expertise entstehen derzeit harmonisierte Mustervertragsvorlagen für die Datennutzung und Datenaufnahme, die dann von den FDZ eingesetzt werden können.[29]

Neben der Akkreditierung durch den RatSWD (s. Qualitätssicherung) kann das international anerkannte CoreTrustSeal als peer-review-gestützte Selbstevaluierung zur Qualitätssicherung erworben werden. Die Selbsteinschätzung auf Grundlage eines Anforderungskatalogs mit 16 Kriterien wird anschließend durch zwei unabhängige Expert*innen mit abschließender Diskussion im CoreTrustSeal Board begutachtet. Sowohl bei der Akkreditierung als auch bei dem Zertifizierungsverfahren besteht ein Bedarf an Unterstützung. Hier sollen Unterstützungsmaterialien und -workshops erarbeitet werden, um FDZ auf dem Weg zu formalen Qualitätskriterien zu unterstützen und damit die Qualität der gesamten Infrastruktur zu verbessern.

Brücken zwischen Textdaten und anderen Datentypen

Neben der Unterstützung der FDZ-Landschaft entwickelt KonsortSWD auch datentypbezogene Methoden zur Weiterentwicklung des FDM. So werden z. B. Textdaten, die als *found data*[30] nicht genuin für die Wissenschaft produziert wurden, zunehmend für wissenschaftliche Analysen verfügbar. Beispiele sind Dokumente, die politische und gesellschaftliche Prozesse dokumentieren, wie beispielsweise Plenarprotokolle oder Posts und Kommentare aus Sozialen Medien. Diese *found data* ermöglichen es, neue sozialwissenschaftliche Fragestellungen zu bearbeiten, die mit klassischen Daten[31] nicht vollständig erschlossen werden können. Bei der Analyse dieser »neuen Daten« stellen sich jedoch besondere methodologische, konzeptionelle und technische Herausforderungen. Insbesondere enthalten die Daten meist keine Informationen wie beispielsweise demographische Hintergrunddaten zu den Verfasser*innen, die aber für die sozialwissenschaftliche Analyse entscheidend sein können.[32]

Um *found data* für die Bearbeitung sozialwissenschaftlicher Fragestellungen zu erschließen, ist eine Verknüpfung mit etablierten Daten erforderlich. Beispielgebend ist hier die Kombinationen von Daten aus Sozialen Medien mit klassischen Umfragedaten.[33] Eine spezifische Schwierigkeit bei der Verknüpfung von Textdaten mit anderen Datenarten ist die fehlende Strukturierung von Text. Ist etwa eine Verknüpfung von Social Media Posts und Umfragedaten möglich, sobald beides für dieselbe Person erhoben wird, stellt sich bei der Verknüpfung von Text zunächst die Frage, was verknüpft werden soll. Das ist etwa relevant, weil das Subjekt einer Aussage nicht unbedingt ein oder eine über Metadaten angegebene Autor*in eines Textes ist.

Auch für relativ etablierte Textdaten wie Plenarprotokolle haben sich bislang kaum *best practices* für eine replizierbare Verknüpfung mit anderen Datenarten durchgesetzt. Gleichzeitig stecken in diesen Daten eine Vielzahl potenziell hochrelevanter Bezüge zu anderen

Daten: die diskursive Begleitung der Arbeitslosenstatistik, Diskussionen über Anpassungen der Rentenformel oder aber politikwissenschaftlich relevante Fragen nach dem Ton von Parlamentsreden in Abhängigkeit von Umfragedaten oder Dynamiken des Parteienwettbewerbs.

Ansätze wie der LinkedEP Datensatz, der Plenardebatten des Europäischen Parlamentes und biografische Informationen seiner Abgeordneten umfasst, zeigen hierzu aber mögliche Richtungen auf.[34] Ausgangspunkt des Ansatzes von KonsortSWD ist die Definition eines generischen Prozesses, um Textdaten mit anderen Datenarten verknüpfen zu können. Dafür ist konzeptionell ein mehrstufiger Prozess nötig, der als *entity linking* bezeichnet wird.[35] Am Anfang steht die Erkennung von Entitäten wie Personen, Organisationen, Institutionen oder Orten im fortlaufenden Text. Danach müssen die so erkannten Entitäten (wer spricht oder worüber wird gesprochen) disambiguiert (d. h. eindeutig voneinander unterschieden) und mit eindeutigen Identifikatoren versehen werden. Dieser Schritt ist wichtig, da mehrere Personen oder Orte denselben Namen tragen. Wenn so in einer Rede im Bundestag von »Merkel« gesprochen wird, könnte in der 15., 16. oder 17. Legislaturperiode sowohl die langjährige Bundeskanzlerin Angela Merkel als auch die SPD-Abgeordnete Petra Merkel (MdB zwischen 2002 und 2013) gemeint sein. Solche Verwechslungen gilt es zu vermeiden. Die Zuweisung von eindeutigen Identifikatoren dient der späteren Verknüpfung von Textdaten mit anderen Datensätzen und kann beispielsweise durch den Anschluss an externe Wissensbestände wie Wikidata erfolgen, in denen bereits viele Personen mit Identifikatoren versehen werden.[36] So können verschiedene Entitäten mit demselben (Familien-)Namen oder ähnlicher Bezeichnung eindeutig voneinander unterschieden und mit dem korrekten Eintrag einer externen Ressource verbunden werden. Außerdem ermöglichen solche Identifikatoren, Entitäten mit verschiedenen Bezeichnungen zusammenzufassen, die jedoch dasselbe meinen. Beispielsweise verweisen sowohl die Abkürzung CDU als auch der Ausdruck »Christlich Demokratische Union Deutschlands« auf ein- und dieselbe Partei. Einen in Plenardebatten zum Ausdruck gebrachten Standpunkt einer Partei mit Umfragedaten zu verknüpfen, ist ein Anwendungsszenario für solche Verknüpfungen. Wenn sich beide Entitäten auf denselben Identifier zurückführen lassen, wären neue qualitative wie quantitative Analysen ohne weitere Komplikationen möglich.

Die Verknüpfung von Textdaten und anderen Datenarten birgt also Potentiale zur Bearbeitung neuer und alter Fragestellungen. Gleichzeitig zeigen sich auch konzeptionelle und technische Herausforderungen sowie weiterführende Fragestellungen, die bei KonsortSWD bearbeitet werden. Dabei ist auch die Evaluation von bestehenden und die Entwicklung von neuen Tools zur Verlinkung verschiedener Datenarten, die der wissenschaftlichen Community im Sinne der FAIR-Prinzipien zur Verfügung gestellt werden sollen, ein zentraler Bestandteil.[37]

QualidataNet – Ausbau der Infrastruktur für qualitative Daten

In der qualitativen Sozialforschung reicht die Bandbreite der Datentypen deutlich über die genannten Textdaten hinaus. So werden »vielfach unterschiedliche Methoden (Interviews, Beobachtungen, Dokumentenanalysen, etc.) kombiniert und die Daten in sehr unterschiedlichen Formaten und Medienformen gespeichert […] (z. B. als Textkorpora vorliegende Transkripte von Interviews, Feldnotizen und Beobachtungsprotokollen sowie Bild- , Video- und Tondateien).«[38] Die resultierenden Korpora von komplexen und gering strukturierten Forschungsdaten sind oft nicht erschöpfend ausgewertet und bieten große Potenziale für die Beantwortung weiterer Forschungsfragen. Das Teilen solcher qualitativen Forschungsdaten kann dazu beitragen, eventuellen Belastungen von Beforschten (z. B. bei erkrankten Menschen) und spezifischen Forschungsfeldern entgegenzuwirken – gerade, wenn es sich um kleine Gruppen und dadurch eng begrenzte Forschungsfelder handelt, wie z. B. im Fall von »Superreichen«, oder in Bereichen, in denen Forschung in wichtige Abläufe und Routinen eingreift – wie im Fall der Beforschung von Kindern in der Schule.

Auf der anderen Seite steht das Teilen qualitativer Forschungsdaten noch ziemlich am Anfang.[39] Dies ist auch deshalb frappierend, weil die qualitative Forschung in vielen Disziplinen von KonsortSWD einen erheblichen Anteil am wissenschaftlichen Erkenntnisprozess hat und u. a. durch die dichte Beschreibung von sozialen Phänomenen zur Theorieentwicklung beiträgt. Die Ursachen für die Zurückhaltung beim Teilen qualitativer Forschungsdaten liegen z. B. darin, dass diese Daten aufgrund ihrer Beschaffenheit – als unstrukturierte und komplexe Daten mit einem oft hohen Anteil an personenbezogenen Informationen – besondere Anforderungen an den Schutz von Persönlichkeitsrechten, an forschungsethische Abwägungen und an Fragen einer guten Forschungsdokumentation stellen. Ein weiterer Grund ist die vergleichsweise schlecht ausgebaute Forschungsdateninfrastruktur für diese Datenart. Zu nennen sind z. B. eine geringere Bekanntheit der vorhandenen FDZ und ihrer Angebote sowie die Fragmentierung des ohnehin noch überschaubaren Datenangebots.

Hier setzt KonsortSWD mit zwei miteinander verzahnten Ansätzen an. Die erste Maßnahme fokussiert auf eine nachhaltige Infrastruktur. Geschaffen wird eine föderierte Archivierungsinfrastruktur für qualitative Forschungsdaten: der Verbund *QualidataNet*. Dieser neugeschaffene Verbund vernetzt Datengebende, Sekundärnutzende und Datenanbietende nachhaltig mit-

einander. Damit wird die bislang fragmentierte in eine integrierte Archivlandschaft für qualitative sozialwissenschaftliche Daten überführt. Dabei fungiert *QualidataNet* als »single point of entry«, der Forschenden Wege zu dem für sie passenden FDZ für die Langzeitarchivierung und Bereitstellung ihrer Forschungsdaten aufzeigt und darüber hinaus bei der Aufbereitung qualitativer Forschungsdaten unterstützt.

Derzeit bilden fünf akkreditierte FDZ, die qualitative Forschungsdaten halten, die Basis für *QualidataNet*. Dies sind das FDZ Qualiservice an der Universität Bremen (Koordination), das FDZ Bildung am DIPF, das Archiv für Gesprochenes Deutsch (AGD) am IDS, das FDZ Betriebs- und Organisationsdaten am DIW und das FDZ für die Hochschul- und Wissenschaftsforschung am DZHW. Gemeinsam entwickeln sie ein Kooperationsmodell für das Netzwerk, ein FAIRes Metadatenset sowie ein kontrolliertes Vokabular, um die internationale Findbarkeit qualitativer Datensätze zu ermöglichen. Der Zugang zu Forschungsdaten wird so vereinfacht, und Daten unterschiedlicher Anbieter können leichter miteinander verknüpft werden. Die Datensätze werden über die Plattform *www.qualidatanet.com* such- und findbar sein.

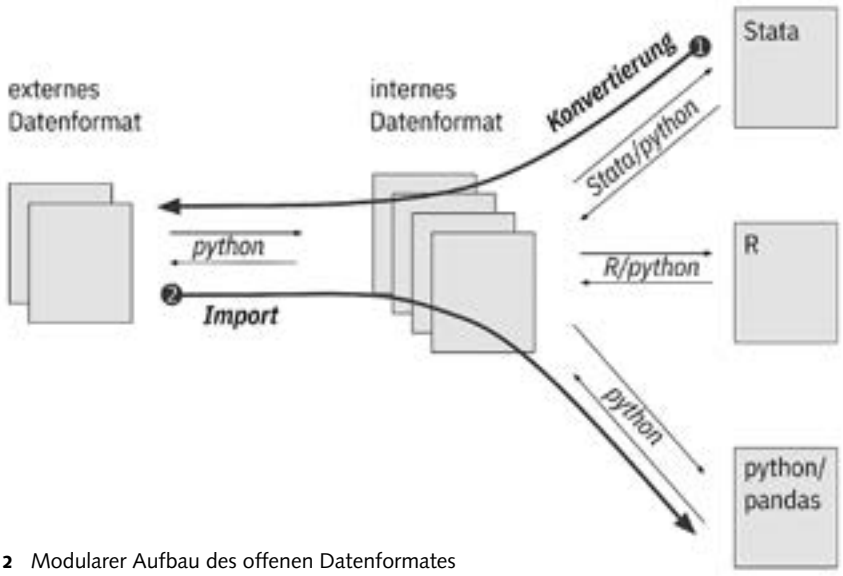
Die zweite Maßnahme zielt auf die Schaffung eines Portfolios an Angeboten für das Management qualitativer Forschungsdaten. Dabei entwickeln die genannten FDZ ein abgestimmtes Angebot von FDM-Instrumenten für qualitative Daten. Dieses Angebot soll Forschende bereits bei der Vorbereitung der Datenerhebung im Forschungsprojekt mit Tools, Vorlagen, Protokollen und Standards unterstützen. Es adressiert die besonderen fachlichen, rechtlichen und ethischen Herausforderungen im Umgang mit qualitativen Daten, damit diese in guter Qualität zur Nachnutzungen bereitgestellt werden können. Der Zugang zu diesen Angeboten wird u. a. über *QualidataNet* ermöglicht.

Sukzessive können auch weitere FDZ, (Fall-)Archive und Repositorien Partner in diesem Netzwerk werden, ihre disziplin- und datenspezifischen Angebote integrieren und damit ihre Datensätze sichtbarer machen. Die Daten selbst verbleiben bei den FDZ mit ihren spezifischen Expertisen. *QualidataNet* ermöglicht Forschenden zudem die Vernetzung mit anderen Datenhaltenden. Der Ansatz, dass Daten dann am besten nachgenutzt werden können, wenn sie beim Datenanbieter mit seiner Expertise und seinem Know-how für die spezifischen Daten verbleiben, ist auch bei *QualidataNet* die Basis für die Qualität der Erschließung neuer wissenschaftlicher Forschungsdaten.

Offener Austausch, offenes Datenformat

Jenseits der besseren Erschließung von Forschungsdaten sollte der Austausch von Forschungsdaten im Sinne von *Interoperability* und *Reusability* (I und R der FAIR-Prinzipien) möglichst nicht auf technische Hürden stoßen. Bei der Weitergabe von digitalen Artefakten für die Sekundäranalyse stellt sich nämlich – selbst bei weitgehend standardisierten, tabellenförmigen Datensätzen in der quantitativen Forschung – die Frage, in welchem Format die Dateien weitergegeben werden. Analysiert werden solche Datensätze heute mit Tabellenkalkulationsprogrammen (z. B. LibreOffice Calc, Excel), statistischer Standardsoftware (z. B. R, Stata) oder auch Spezialanwendungen (z. B. gretl, Latent Gold).

Die verwendeten Programme bringen meist ihre eigenen, oft proprietären Dateiformate für Daten mit, die häufig eine Anreicherung der Daten mit Metadaten,[40] etwa in Form von Variablen- und Wertelabeln, implementieren. Damit andere Programme diese Da-



2 Modularer Aufbau des offenen Datenformates

teien öffnen können, werden Importfilter benötigt. Stehen diese nicht zur Verfügung, kann zwar der Weg über CSV-Dateien gegangen werden; dieser geht aber oft mit einem Informationsverlust einher. Gerade bei älteren Dateiformaten entsteht so oft die Gefahr von Datenverlust, wenn nicht die Informationen im Zuge einer Langzeitarchivierung regelmäßig (und oft manuell) in aktuelle Formate überführt werden.

Die Entscheidung, welches Datenformat Forschende oder FDZ nutzen, ist abhängig von den im Datenmanagement benutzten Werkzeugen, der Software, mit der die Nutzer*innen ihre Analysen durchführen, und der Frage, welche Formate für die Langzeitarchivierung geeignet sind. Oft werden Daten auch in mehreren Formaten erzeugt, um mit dieser Entscheidung einhergehende Zielkonflikte abzumildern. Das kann dazu führen, dass der gleiche Datensatz in unterschiedlichen Dateiformaten bereitgestellt wird, von denen bisweilen auch noch unterschiedliche Sprachversionen für die internationale Nutzung entstehen.

Ein offenes Datenformat, das möglichst standardisierte mehrsprachige Metadaten enthält, löst diese Probleme und wird neben entsprechenden Import- und Konvertierungsfiltern als Dienst[41] von KonsortSWD bereitgestellt werden. Durch die integrierten Metadaten entsteht erheblicher Zusatznutzen, weil die Datendokumentation so in das Nutzerinterface des Statistikprogramms integriert werden kann: Neben den Variablen- und Wertelabeln können Informationen über Provenienz (eingesetzte Messmethode, eingeflossene Informationen, Validität, Kodierungsschemata etc.) bereitgestellt werden. Weiterhin können auch Links zu einem Metadatenportal oder PIDs von Variablen[42] enthalten sein. Der Import in das Statistikprogramm sollte dabei so einfach sein wie das Öffnen eines Datensatzes im zugehörigen proprietären Format.

Um Forschenden, aber auch den FDZ die Bereitstellung von Datensätzen in dem neuen Datenformat zu erleichtern, wird es Konvertierungsfilter geben, die eine Umwandlung von aktuell weit verbreiteten proprietären Datenformaten in das neue offene Format ermöglichen.

Die Abbildung 2 zeigt den modularen Ansatz der aktuellen Konzeption. Ein möglicher Importfilter ist mit Pfeil 2 eingezeichnet: Aus dem externen Datenformat erzeugt ein erstes (wiederverwendbares) Programmmodul (programmiert in Python) ein internes Datenformat, das dann mit einem weiteren Programmmodul für den eigentlichen Import sorgt. Der Konvertierungsfilter (Pfeil 1) geht den umgekehrten Weg. Das interne Datenformat stellt ein vollständiges und minimales Datenmodell dar, das als Grundlage für die Diskussion mit der Community dient, welche Features benötigt werden. In der Terminologie des für die KonsortSWD Disziplinen maßgeblichen DDI-Standards[43] würde man etwa von einem DDI-Profil sprechen. Aktuell handelt es sich beim internen Format um vier CSV-Tabellen: eine enthält die Daten, eine die Informationen auf Datensatzebene (z. B. DOI), eine weitere die Informationen auf Variablenebene (z. B. mehrsprachige Variablenlabel und Link in ein Metadatenportal) und die vierte schließlich die Informationen auf Werteebene (mehrsprachige Label).

Um den breiten Einsatz des neuen Formats auch in anderen Konsortien der NFDI und darüber hinaus zu erleichtern, werden die entwickelten Spezifikationen und Codes unter einer freien Lizenz zur Verfügung gestellt.

Herausforderungen

Mit den hier dargestellten exemplarischen Teilprojekten wird KonsortSWD in den kommenden Jahren helfen können, wichtige Lücken in der bestehenden Forschungsdateninfrastruktur in den Sozial-, Verhaltens-, Bildungs- und Wirtschaftswissenschaften zu schließen. Daraus wird eine besser integrierte und FAIRere Infrastrukturlandschaft resultieren, die weiterhin auf ihre dezentrale Expertise aufbaut.

Diese dezentrale Expertise der großen Zahl interessierter Forscher*innen zugänglich zu machen, wird dabei eine wichtige Herausforderung werden. Dieser kann teilweise durch Schulungen und technische Verbesserungen, z. B. bei der Auffindbarkeit von Daten, begegnet werden. KonsortSWD setzt aber auch darauf, seine Verbindungen zu Bibliotheken als den Orten ausbauen zu können, bei denen viele Forschende ihre Suche nach Quellen und Materialien beginnen. Dazu kooperiert KonsortSWD oder die jeweiligen Partner schon heute mit einzelnen Fachinformationsdiensten. Eine stabile und leistungsfähige Infrastruktur für Forschungsdaten setzt in gewisser Weise voraus, dass sich die bestehenden und aufzubauenden Kompetenzen komplementär ergänzen. Somit können mit den Daten, die unsere Gesellschaft beschreiben, Antworten auf noch vielfältigere und interessante Fragen gegeben werden.

Anmerkungen

1 WILKINSON, Mark. D., et al. The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data*, 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18

2 KonsortSWD wird im Rahmen der NFDI durch die Deutsche Forschungsgemeinschaft (DFG) gefördert – Projektnummer: 442494171.

3 https://www.gwk-bonn.de/fileadmin/Redaktion/Dokumente/Papers/NFDI.pdf, S. 1.

4 Dies sind derzeit (alphabetisch): Akademie für Soziologie (AS), Deutsche Gesellschaft für Erziehungswissenschaft (DGfE), Deutsche Gesellschaft für Gesundheitsökonomie (DGGÖ), Deutsche Gesellschaft für Medizinische Psychologie (DGMP), Deutsche Gesellschaft für Politikwissenschaft (DGfP), Deutsche Gesellschaft für Psychologie (DGPs), Deutsche Gesellschaft für Publizistik- und Kommunikationswissen-

schaft (DGPuK), Deutsche Gesellschaft für Soziologie (DGS), Deutsche Gesellschaft für Sozial- und Kulturanthropologie (DGSKA), Deutsche Gesellschaft für Volkskunde (DGV), Deutsche Statistische Gesellschaft (DStatG), Deutsche Vereinigung für Politische Wissenschaft (DVPW), Gesellschaft für empirische Bildungsforschung (GEBF), Verband der Hochschullehrer für Betriebswirtschaft (VHB) sowie der Verein für Socialpolitik (VfS).

5 https://www.konsortswd.de/ratswd/der-ratswd/mandat/ [Zugriff am: 11.11.2021].

6 BUCK, Daniel et al. *Handreichung: Forschungsdatenzentren (FDZ) gründen (in Vorbereitung)*.

7 HOLLSTEIN, Bettina, et al. KonsortSWD: Vom Netzwerk zur integrierten Dateninfrastruktur der Gesellschaftsforschung. In: *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern*, (2). 2021, S. 10–22. https://doi.org/10.17192/bfdm.2021.2.8330

8 BUCK, Daniel et al. *Handreichung: Forschungsdatenzentren (FDZ) gründen* (in Vorbereitung).

9 HOLLSTEIN, Bettina et al. KonsortSWD: Vom Netzwerk zur integrierten Dateninfrastruktur der Gesellschaftsforschung. In: *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern*, (2). 2021, S. 10–22. https://doi.org/10.17192/bfdm.2021.2.8330, S. 11.

10 Ebd., S. 12.

11 Beispielhaft seien die Angebote von Qualiservice für qualitative Daten https://www.qualiservice.org/de/daten-teilen.html, von ZPID für psychologische Daten https://www.psychdata.de/index.php?main=give&sub=uebergabe, vom Verbund Forschungsdaten Bildung für bildungsbezogene Daten https://www.forschungsdaten-bildung.de/daten-teilen und das generische Angebot von GESIS https://www.gesis.org/datenservices/daten-teilen genannt. Im Rahmen von KonsortSWD werden weitere Angebote für die Verbesserung der Datenbereitstellung und des Datenzugangs entwickelt: https://www.konsortswd.de/konsortswd/das-konsortium/services/ [jeweils Zugriff am: 15.11.2021].

12 Art. 4 Nr. 1 DSGVO.

13 RatSWD (Hrsg.). *Handreichung Datenschutz. 2. vollständig überarbeitete Auflage. RatSWD Output 8 (6)*. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD), 2020. https://doi.org/10.17620/02671.50

14 Ebd.

15 Ebd.

16 https://www.konsortswd.de/datenzentren/fdi-ausschuss/ [Zugriff am: 15.11.2021].

17 RatSWD (Hrsg.). *Qualitätssicherung der vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) akkreditierten Forschungsdatenzentren (FDZ). RatSWD Output 8 (5)*. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD), 2017. https://doi.org/10.17620/02671.4

18 https://www.konsortswd.de/datenzentren/alle-datenzentren/ [Zugriff am: 30.11.2021].

19 Die Bereitstellung ist größtenteils kostenfrei. Werden Gebühren erhoben, liegen diese meist im zweistelligen oder niedrigen dreistelligen Bereich.

20 RatSWD (Hrsg.). *Tätigkeitsbericht 2020 der vom RatSWD akkreditierten Forschungsdatenzentren (FDZ).* Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD), in Vorbereitung.

21 Die Kommission wird für je drei Jahre vom FDI-Ausschuss aus den eigenen Reihen gewählt und genießt so besonderes Vertrauen und Legitimität. Sie besteht aus vier Mitgliedern und zwei Stellvertretenden sowie den Vorsitzenden des RatSWD als Gästen.

22 RfII – Rat für Informationsinfrastrukturen (Hrsg.). *Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel*, Göttingen, 2019, S. 20.

23 Zur Geschichte des AGD und seinem Vorläufer »Deutsches Spracharchiv« (DSAv) siehe Ulf-Michael STIFT und Thomas SCHMIDT. Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In: Institut für Deutsche Sprache (Hrsg.). *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache, 2014, S. 360–375.

24 RfII – Rat für Informationsinfrastrukturen (Hrsg.). *Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel.* Göttingen, 2019.

25 https://www.konsortswd.de/konsortswd/das-konsortium/services/unterstuetzung-fdz-fair-datenzugang/

26 https://www.coretrustseal.org/

27 Consultative Committee for Space Data Systems (Hrsg.). *Reference Model for an Open Archival Information System (OAIS)*. Blue Book: Vol. 1. Washington, 2012. LAVOIE, Brian. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide* (2nd Edition). Digital Preservation Coalition, 2014. https://doi.org/10.7207/twr14-02

28 WILKINSON, Mark. D., et al. The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data*, 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18

29 Dass diese im Einzelnen aufgrund institutioneller oder anderer Verpflichtungen leicht angepasst werden müssen, ist zu erwarten. Dennoch ist davon auszugehen, dass die übergeordnete Struktur allen Beteiligten hilft, die Nutzungsbedingungen leichter zu verstehen.

30 CONNELLY, Roxanne, et al. The role of administrative data in the big data revolution in social science research. In: *Social Science Research*, 59, 2016, S. 1–12.

31 Mit dem Ausdruck »klassische Daten« sind hier Daten nach Connelly et al. (s. o.) gemeint, die von vornherein für den Forschungsprozess produziert werden.

32 STIER, Sebastian, et al. Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. In: *Social Science Computer Review*. 38 (5), 2020, S. 503–516.

33 DE VREESE, Claes H., et al. Linking Survey and Media Content Data: Opportunities, Considerations, and Pitfalls. In: *Communication Methods and Measures*; 11 (4), 2017, S. 221–244. STIER, Sebastian, et. al. Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. In: *Social Science Computer Review*, 38 (5), 2020, S. 503–516.

34 https://linkedpolitics.project.cwi.nl/web/html/home.html

35 AUGENSTEIN, Isabelle, et al. LODifier: Generating Linked Data from Unstructured Text. In: Elena SIMPERL, Philipp CIMIANO, Axel POLLERES, Oscar CORCHO und Valentina PRESUTTI (Hrsg.). *The Semantic Web: Research and Applications*. Berlin/Heidelberg: Springer-Verlag, 2012, S. 210–224.

36 VRANDEČIĆ; Denny und Markus KRÖTZSCH. Wikidata: a free collaborative knowledgebase. In: *Communications of the ACM*, 57 (10), 2014, S. 78–85.

37 Aktuelle Informationen und ein Zeitplan zum Projekt sind unter https://www.konsortswd.de/konsortswd/das-konsortium/services/linking-textual-data/ und https://polmine.github.io abrufbar.

38 HOLLSTEIN, Betina, et al. *KonsortSWD: Vom Netzwerk zur integrierten Dateninfrastruktur der Gesellschaftsforschung. Bausteine Forschungsdatenmanagement* (2), 2021, S. 10–22. DOI: https://doi.org/10.17192/bfdm.2021.2.8330

39 HOLLSTEIN, Betina und Jörg STRÜBING. Archivierung und Zugang zu qualitativen Daten. In: Doris BAMBEY et. al. (Hrsg.). *Archivierung und Zugang zu qualitativen Daten* (Vol.

267). Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD), 2018, S. 1–14.

40 Mit Metadaten sind nicht nur standardisierte Beschreibungen eines Objekts gemeint, wie das etwa bei Dublin Core der Fall ist. Metadaten im hier verwendeten Sinne beschreiben vielmehr die Bestandteile eines Objekts (z. B. Variablen eines Datensatzes) oder bilden sie sogar vollständig ab (z. B. Fragen eines Fragebogens, vgl. ausführlich WENZIG, Knut. Metadaten. In: Nina BAUR und Jörg BLASIUS (Hrsg.). *Handbuch Methoden der empirischen Sozialforschung*. 2. Aufl. Wiesbaden: Springer VS, 2019, S. 1253–1264. https://doi.org/10.1007/978-3-658-21308-4_92).

41 https://www.konsortswd.de/konsortswd/das-konsortium/services/open-data-format/

42 Persistente Identifikatoren für Variablen sind ebenfalls ein Dienst des KonsortSWD: https://www.konsortswd.de/konsortswd/das-konsortium/services/persistent-identifiers/

43 Vgl. https://ddialliance.org/

Verfasser*innen

Andreas Blätte, Professor für Public Policy und Landespolitik, Universität Duisburg-Essen, Duisburg, andreas.blätte@uni-due.de, https://orcid.org/0000-0001-8970-8010

Anna Fräßdorf, wiss. Referentin in der Geschäftsstelle des RatSWD, Wissenschaftszentrum Berlin für Sozialforschung (WZB), anna.fraessdorf@ratswd.de

Jan-Ocko Heuer, wiss. Mitarbeiter, FDZ Qualiservice, Universität Bremen, jheuer@uni-bremen.de, https://orcid.org/0000-0001-8334-0411

Ute Hoffstätter, wiss. Mitarbeiterin im FDZ-DZHW, Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW), Hannover, hoffstaetter@dzhw.eu, https://orcid.org/0000-0001-8692-3428

Christoph Leonhardt, wiss. Mitarbeiter im Projekt Linking Textual Data, Universität Duisburg-Essen, Duisburg, christoph.leonhardt@uni-due.de

Laura Menze, Leiterin FDZ-BAuA, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA), Berlin, menze.laura@baua.bund.de, https://orcid.org/0000-0003-3442-279X

Bernhard Miller, Koordinator KonsortSWD, GESIS – Leibniz-Institut für Sozialwissenschaften, Mannheim, bernhard.miller@gesis.org, https://orcid.org/0000-0002-4385-7245

Kati Mozygemba, wiss. Mitarbeiterin am FDZ Qualiservice, SOCIUM Forschungszentrum Ungleichheit und Sozialpolitik, Universität Bremen, kati.mozygemba@uni-bremen.de, https://orcid.org/0000-0002-0326-1607

Dagmar Pattloch, Mitarbeiterin im FDZ-BAuA, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA), Berlin, pattloch.dagmar@baua.bund.de, https://orcid.org/0000-0002-4793-6805

Julia Rakers, wiss. Mitarbeiterin im Projekt Linking Textual Data, Universität Duisburg-Essen, Duisburg, julia.rakers@uni-due.de

Silke Reineke, Leiterin Archiv für Gesprochenes Deutsch (AGD), Leibniz-Institut für Deutsche Sprache, Mannheim, reineke@ids-mannheim.de

Thomas Runge, komm. Leiter der Geschäftsstelle des RatSWD, Berlin, trunge@ratswd.de

Friederike Schlücker, Koordinatorin Task Area 2 KonsortSWD – Leibniz-Institut für Bildungsverläufe (LIfBi), Bamberg, friederike.schluecker@lifbi.de

Thomas Schmidt, Leiter Research and Infrastructure Support (RISE), Universität Basel, th.schmidt@unibas.ch, https://orcid.org/0000-0003-0026-6450

Knut Wenzig, Mitarbeiter im SOEP-FDZ, DIW Berlin, kwenzig@diw.de, https://orcid.org/0000-0002-2259-0203

Christof Wolf, Sprecher von KonsortSWD, Präsident GESIS, GESIS – Leibniz-Institut für Sozialwissenschaften, Mannheim, christof.wolf@gesis.org, https://orcid.org/0000-0002-9364-9524