Kurzstudie:
Anforderungen an die Archivierung sozial- und wirtschaftswissenschaftlicher Forschungsdaten

„Also ich hab so viel Zeit in die Aufbereitung der Daten gesteckt, da weiß ich nicht, ob ich persönlich so schnell bereit wäre, diese Daten einfach anderen zugänglich zu machen.“ (H33)

„Ja, also bei Forschungsvorhaben die mit öffentlichen Geldern finanziert werden, da sehe ich das quasi geradezu als Pflicht, die Daten auch anderen Wissenschaftlern zur Verfügung zu stellen.“ (A33)

„Die Frage ist natürlich mit welchem Aufwand das verbunden ist. Also ich wäre bereit, die Daten alle ins Netz zu stellen, hätte allerdings keine Lust, daran einen Monat zu arbeiten, um die einzugeben, dann sage ich nee, also keinen Bock drauf.“ (B37)

Projekt SowiDataNet

Bearbeitung: Patrick J. Droß

Unter Mitarbeit von: Franziska Kuhnt

Berlin, 21. April 2015

Inhalt

1. Einleitung

Im Rahmen des SowiDataNet (SDN) Projektworkshops, der am 23. Juni 2014 am Deutschen Institut für Wirtschaftsforschung Berlin stattfand, konnte bereits eine Vielzahl konkreter Anforderungen an die Entwicklung des SDN-Infrastrukturangebots identifiziert werden[1]. Diskutiert wurde über Erfahrungen in der Datenerzeugung und im Datenmanagement, Möglichkeiten einer standardisierten Beschreibung der Forschungsdaten durch Metadaten, Zugriffsrechte und Recherchemöglichkeiten sowie mögliche Bedenken bei den Forschenden. Mehrfach wurde hervorgehoben, dass seitens der Forschungsförderer – etwa DFG oder EU – zunehmend ein planvoller Umgang mit Forschungsdaten sowie ihre langfristige Aufbewahrung gefordert werden. Das künftige SDN-Repositorium sollte es den Wissenschaftler/innen daher ermöglichen, diesen Anforderungen gerecht zu werden. Einigkeit bestand zudem hinsichtlich des Bedarfs an einer innovativen, leistungsfähigen und zugleich einfach nutzbaren Forschungsdateninfrastruktur. Letzteres ist v.a. deshalb wichtig, weil eine erhöhte Arbeitsbelastung zu den meistgenannten Befürchtungen auf Seiten der Forschenden zählt. Entsprechend sind auch die überzeugenden Vorteile einer zentralen Archivierung und Bereitstellung von Forschungsdaten noch stärker in die Forschungspraxis zu vermitteln: Neue Rechercheoptionen, die erhöhte Sichtbarkeit der Daten in der Scientific Community, sowie die Möglichkeit der Zitation oder der Verknüpfung der Datensätze mit daraus resultierenden Publikationen.

Um die Ergebnisse des Projektworkshops durch weitere Angaben aus dem Forschungsalltag zu unterfüttern, wurden im Zeitraum von Oktober bis Dezember 2014 zehn leidfadengestützte Experteninterviews mit empirisch arbeitenden Forscher/innen aus den Sozial- und Wirtschaftswissenschaften geführt. Ziel war es mehr über deren alltägliche empirische Arbeit, das Spektrum (quantitativer) empirischer Forschung und die Anforderungen an eine künftige Forschungsdateninfrastruktur in Erfahrung zu bringen. Der verwendete Interviewleitfaden (Anhang A) thematisierte entsprechend den derzeitigen Umgang mit den eigenen Forschungsdaten, die Suche nach Daten sowie das Interesse und die Bedarfe hinsichtlich der möglichen Archivierung und Veröffentlichung von Datensätzen in einem webbasierten Repositorium. Der Leitfaden diente als grobes Gerüst für die Gespräche, den eigenen Beiträgen und Schwerpunktsetzungen der Forscher/innen wurde jedoch reichlich Platz eingeräumt. Die Auswahl der Interviewpartner/innen erfolgte über die „Letter of intent“-Kontakte des SDN-Projekts sowie über persönliche Kontakte mit Forscher/innen am WZB und DIW Berlin. Bei der Zusammenstellung des Samples wurde versucht ein möglichst breites Spektrum unterschiedlicher empirischer Forschungsszenarien abzubilden. Ebenso sollten die Sozial- und die Wirtschaftswissenschaften zu gleichen Anteilen vertreten sein. Alle Gespräche wurden aufgezeichnet und anschließend vollständig transkribiert. Die Transkripte wurden durch die Löschung von Klarnamen, Institutsnamen oder Projekttiteln anonymisiert. Zur Codierung des Interviewmaterials kam die Software MAXQDA zum Einsatz.

[1] www.sowidatanet.de/images/pdfs/SowiDataNet_Workshopbericht_Anforderungsanalyse_22.07.2014.pdf

Das Untersuchungssample besteht aus insgesamt zehn Personen in einer Altersspanne zwischen ca. 30 und 60 Jahren, darunter drei Frauen und sieben Männern[2]. Sechs Interviewpartner sind als wissenschaftliche Mitarbeiter/innen an Leibniz-Instituten, drei Befragte an Universitäten und ein/e Interviewpartner/in an einem privaten Forschungsinstitut beschäftigt. Sieben Forscher/innen üben eine leitende Funktion innerhalb einer Abteilung oder Forschungsgruppe aus und haben in den Interviews daher nicht nur über ihre eigenen, sondern auch über die Erfahrungen in ihren jeweiligen Forschungsteams berichtet. Neun von zehn Befragten sind promoviert und fünf als Professor/innen tätig. Jeweils die Hälfte der Interviewten verfügt über einen sozialwissenschaftlichen bzw. über einen wirtschaftswissenschaftlichen Forschungshintergrund, wobei sich in den aktuellen Forschungsthemen diverse fachliche Überschneidungen ergeben. Allen Befragten ist gemeinsam, dass sie auf langjährige Erfahrungen in der quantitativen empirischen Forschung zurück blicken können, bereits eigene Erhebungen durchgeführt haben sowie über fundierte Methodenkenntnisse verfügen.

Das Vorwissen zur Thematik der Interviews ist hingegen eher moderat ausgeprägt. So wurden bislang durch keine/n Gesprächspartner/in selbständig Forschungsdaten in ein Repositorium oder ein vergleichbares Infrastrukturangebot eingestellt. Immerhin drei von zehn Befragten war das GESIS-Angebot *datorium*[3] bekannt, wobei sie sich z.T. irritiert über die Entwicklung eines weiteren Angebots zur Datenarchivierung zeigten. Eine Befragte gab an, dass das institutsinterne Forschungsdatenmanagement bereits erste Datensätze in *datorium* eingestellt hat. In einem weiteren Fall wurden Daten in der Vergangenheit anderen Forschenden über eine private Webseite zugänglich gemacht. Lediglich in zwei Interviews wurde ein explizites Vorwissen in Richtung professioneller Datenarchivierung erkennbar, indem bspw. der Begriff „Metadaten“ durch die Interviewten selbst aufgebracht wurde oder indem konkrete Nachfragen hinsichtlich der Funktionalitäten des SDN-Repositoriums gestellt wurden. Auch das Interesse an Daten-Archivierung und Data-Sharing fällt insgesamt betrachtet gemischt aus. Interessiert und aufgeschlossen zeigten sich immerhin fünf der zehn Interviewten, wobei diese Einstellung teils dennoch mit Bedenken hinsichtlich der Archivierbarkeit ihrer eigenen Forschungsdaten einherging. Zwei Befragte waren zwar nicht prinzipiell skeptisch eingestellt, hatten jedoch Zweifel am Nutzen und der praktischen Umsetzbarkeit des Vorhabens und drei Interviewpartner äußerten eine grundsätzliche Skepsis gegenüber der Idee des Data-Sharing und der möglichen Nachnutzung von Forschungsdaten.

Für die Auswertung der Gesprächsinhalte wird zunächst eine typisierende Beschreibung verschiedener empirischer Forschungsszenarien unternommen, um unterschiedliche Arbeitsweisen und Anforderungen herauszuarbeiten (Abschnitt 2). Anschließend werden all jene Aussagen der Gesprächspartner systematisiert, die sich keinem spezifischem Forschungsszenario zuordnen ließen. Hierbei wurde zwischen Bedenken (Abschnitt 3.1) und Anforderungen (Abschnitt 3.2) hinsichtlich der

[2] Soziodemographische Merkmal wurden im Rahmen der Interviews nicht explizit abgefragt.
[3] https://datorium.gesis.org

Archivierung und Bereitstellung von Forschungsdaten unterschieden. Schließlich erfolgt eine Zusammenfassung zentraler Ergebnisse mit Blick auf erste Schlussfolgerungen für die Weiterentwicklung der SDN-Infrastruktur (Abschnitt 4).

2. Arbeitsweisen und Anforderungen spezifischer Forschungsszenarien

Die nachfolgende Darstellung der unterschiedlichen Forschungsszenarien folgt dem methodischen Ansatz einer heuristischen Typologie. D.h. die Typen wurden auf Grundlage eines theoretischen Vorwissens über die Methoden der empirischen Sozialforschung gebildet, welches bereits explizit in die Konstruktion des Untersuchungssamples eingebracht wurde. Die den einzelnen Typen zugeordneten Merkmale wurden hingegen allein dem Interviewmaterial entnommen und erheben daher keinen Anspruch auf eine vollständige Charakterisierung der unterschiedlichen Forschungsszenarien. Gegenüber den idealtypischen Konstruktionen finden sich in der Forschungspraxis zudem diverse Überschneidungen, insbesondere sind viele Forscher/innen zeitgleich in unterschiedlichen Projektkontexten aktiv.

2.1 Der klassische Survey – *„50 Fragen, 1000 Befragte“*

Der klassische quantitative Survey ist ohne Frage das empirische Forschungsszenario, welches am besten durch die idealtypischen Schritte des Datenlebenszyklus abgebildet wird: auf Grundlage einer Fragestellung wird durch die Forschenden ein standardisiertes Erhebungsinstrument entwickelt und ein Untersuchungssample ausgewählt. Die Befragung geht ins Feld, Daten werden erfasst, aufbereitet und ausgewertet. Schließlich werden Forschungsberichte und Publikationen erstellt, und die Daten könnten potentiell für die Archivierung aufbereitet und anderen Forscher/innen über ein Repositorium zur Verfügung gestellt werden.

Dieses klassische Format kommt im Forschungsalltag der Interviewten auch häufig zum Einsatz, jedoch zeigen sich in der Praxis verschiedene Abweichungen vom idealen Verlauf. Zunächst wurde in den Gesprächen mehrfach erwähnt, dass Erhebungen nicht selten als Aufträge an Umfrageinstitute vergeben werden. Der Ablauf des Forschungsprozesses ändert sich hierdurch zwar nicht grundlegend, das „Outsourcing“ der eigentlichen Erhebung führt jedoch dazu, dass entscheidende Arbeitsschritte nicht von den Forschenden selbst durchgeführt werden: *„Also das Umfrageinstitut hat von uns den fertigen Fragebogen bekommen und hat dann den Versand gemacht, den Versand und die Eingabe und alles andere. Dann haben sie uns den fertigen eingegebenen Datensatz übermittelt in SPSS und wir arbeiten jetzt damit weiter“* (M5). Zwar sind die Umfrageinstitute stets dazu angehalten, ihre Arbeiten umfassend zu dokumentieren, doch fallen diese Dokumentationen, nach Angabe eines Forschers, sowohl was Umfang als auch Qualität anbelangt durchaus unterschiedlich aus. Berichtet wird bspw. von einem Methodenbericht, der zwar ausführlich den Versand und die Bearbeitung des Rücklaufs einer schriftlichen Befragung dokumentierte, die

eigentliche Datenaufbereitung, z.B. der konkrete Umgang mit fehlenden Werten, wird jedoch nicht näher erläutert. In der Konsequenz wissen die Primärforscher selbst nicht genau, wie bei der Eingabe der Daten vorgegangen wurde, waren sie an diesen Arbeiten doch schlichtweg nicht beteiligt. Für die Datenarchivierung bedeutet dies, dass neben den Primärforschern auch Umfrageinstitute im Prozess der Datengenerierung eine z.T. zentrale Rolle spielen. Unter Umständen verfügen die Institute dann allein über wichtige Informationen, die zur vollständigen Nachvollziehbarkeit der Datengenerierung notwendig wären. Darüber hinaus ist zu beachten, dass im Falle einer Auftragsvergabe das Umfrageinstitut für die Datenschutz- bzw. Einwilligungserklärung verantwortlich zeichnet. In einem Beispielfall sicherte das Institut den Befragten zu, dass die „erhobenen Daten nicht an Dritte“ weitergegeben werden. Durch den Verzicht auf den Zusatz „persönliche Daten“ ergeben sich in diesem Fall u.U. Einschränkungen in Bezug auf die Möglichkeit, diese Daten überhaupt in einem Forschungsdatenrepositorium zu veröffentlichen.

Als Spezialfall unter den quantitativen Umfragen wurde in einem weiteren Interview der Cross-Country Survey genannt. Solche international vergleichenden Studien sind erwartungsgemäß häufig im Kontext der EU-geförderten Forschungsprogramme anzutreffen. Wie ein Interviewpartner erläutert, wird hierbei zumeist mit einem Basis-Set an Fragen für alle beteiligten Länder und mit Spezialfragen je Land gearbeitet. Das Resultat sind mehrere länderspezifische Datensätze und ein Cross-Country-Datensatz, in welchem die Länderinformationen zusammengeführt werden. Die Datensätze können zwar einzeln ausgewertet werden, die besondere Attraktivität liegt aber in der vergleichenden Perspektive. In einem Datenarchiv müsste sich daher im Optimalfall die Gesamtheit solcher Erhebungen abbilden lassen. Zudem werden Cross-country Surveys zumeist im Projektverbund mit internationalen Partnern erhoben, so dass – bestenfalls vorab – geklärt werden sollte, welche Partner welche Daten archivieren oder ob man sich für eine zentrale Lösung entscheidet. Zu beachten ist auch, dass in den beteiligten Ländern unterschiedliche rechtliche Rahmenbedingungen in Bezug auf die Datenarchivierung bestehen können. In zwei Interviews wurde sogar von Erhebungen im nicht-europäischen Ausland berichtet. Nach Einschätzung der Forschenden kann dies noch größere Hürden für die Datenarchivierung mit sich bringen.

Zu einer weiteren Abweichung vom idealen Untersuchungsverlauf kommt es, wenn Forschende ihre Erhebung als Einschaltung in bereits bestehende Befragungen konzipieren. Dies wird v.a. dann attraktiv, wenn bestehende Erhebungsstrukturen mitgenutzt werden können, um gesonderte Forschungsfragen zu untersuchen. Die Einschaltungen stellen daher häufig eine Erweiterung eines bestehenden Fragebogens dar, teils werden aber auch zusätzliche Sub-Samples untersucht. Durch die Vermengung der eigenen Forschung mit der Forschung Dritter stellt sich in Bezug auf die Datenarchivierung die Frage, ob im Nachhinein eine Trennung der Daten möglich und v.a. ob sie inhaltlich sinnvoll ist. In einem Interview wird bspw. davon berichtet, dass eine aktuelle Erhebung auf einem europaweiten Survey aufsetzt, um eine spezielle Altersgruppe zu befragen. Eine getrennte Archivierung des Subsamples wäre in diesem Fall sicher möglich, ohne die Vergleichsdaten der Hauptuntersuchung wären die Möglichkeiten der Nachnutzung jedoch sehr begrenzt. Zur

Ausgestaltung der konkreten Nutzungsbedingungen im Projekt konnte der Interviewpartner keine Auskunft geben. Bislang war lediglich vorgesehen, die Daten im engen Kreis der beteiligten Wissenschaftler/innen auszuwerten.

2.2 Experimentelle Designs – *„das sind bei uns riesige Datenmengen“*

Im experimentellen Forschungsdesign wird der Einfluss gezielter Stimuli auf Probanden untersucht. Ein Vorteil dieses Designs ist, dass der Einfluss von Drittvariablen durch die Zufallsverteilung der Probanden auf unterschiedliche Versuchsgruppen kontrolliert werden kann. Experimentelle Designs sind häufig in der Ökonomie, der Sozialpsychologie, aber durchaus auch in Soziologie und Politikwissenschaft anzutreffen. Ein typisches Beispiel sind spieltheoretische Versuchsanordnungen. Die experimentellen Daten werden häufig im Labor erhoben, möglich sind jedoch auch experimentelle Feldstudien. Für gewöhnlich sind die Fallzahlen deutlich geringer als in der Umfrageforschung, Untersuchungen mit 20 bis 50 Probanden sind keine Seltenheit. Am Rande eines Gesprächs wurde allerdings von einem experimentellen Trenddesign mit ca. 5.000 Versuchspersonen berichtet. Die interviewte experimentelle Forscherin aus dem Feld der neuroökonomischen Verhaltensforschung stellt hingegen eher einen Spezialfall in Hinblick auf die Messverfahren und die Größe der Daten dar.

Unter kontrollierten Laborbedingungen werden in der Neuroökonomie Experimente mit 20 bis 40 Probanden durchgeführt. Dabei werden in einem komplexen Methodenmix sowohl Verhaltensdaten (Entscheidungen der Probanden) als auch physiologische Daten mittels Eye-Tracking und MRT erfasst. Zusätzlich werden mithilfe eines schriftlichen Fragebogens Einstellungen und Standarddemographie erhoben. Bei der Verarbeitung der Hirnrohdaten und der anatomischen Bilder kommen proprietäre Softwareformate zum Einsatz. Die Fragebogendaten können hingegen leicht in Excel oder SPSS verarbeitet werden.

In Bezug auf das Datenmanagement stellt die Interviewpartnerin von sich aus fest, dass die Größe der Daten im Vergleich zu anderen Forschungsprojekten in der Regel extrem hoch ausfällt. Allein die Hirnrohdaten können je Experiment ca. 100 bis 150 GB ausmachen. Da jedoch während der Analyse weitere Daten generiert werden, sei es nicht ungewöhnlich, dass ein Ordner mit Hirndaten für ein Experiment am Ende der Auswertungen rund 400 GB umfasst. Hinzu kommen Eye-Tracking-Daten, mit ca. einem GB je Proband. Diese enorme Datenmenge bringt bereits im Forschungsalltag erhebliche Probleme mit sich, z.B. in Bezug auf die dauerhafte Speicherung auf Netzlaufwerken, die benötigte Rechenleistung zur Datenverarbeitung oder auch bei der Zusammenarbeit mit anderen Forschenden. Ob eine Archivierung von Forschungsdaten dieser Größe in einem webbasierten Repositorium überhaupt möglich ist, blieb im Gespräch selbst eine offene Frage.

Auch losgelöst vom konkreten Interviewbeispiel weist die Forscherin darauf hin, dass Daten aus experimentellen Designs teils einer komplexen Hierarchisierung folgen, welche sich vom einzelnen

Probanden über verschiedene Versuchsgruppendaten bis zu verschiedenen Datensätzen für das Gesamtexperiment erstrecken kann. Die Verarbeitung und Auswertung der Daten erfolge oftmals in einem Stufenprozess, von Ebene zu Ebene, wobei teils vorverarbeitete Variablen in die höhere Ebene einfließen, teils auch mehrfach Informationen hin und her gespielt werden. Wenn Nachvollziehbarkeit das Ziel ist, wäre es daher enorm wichtig, so die Forscherin, dass sich die Struktur des Designs in Form des kompletten Datensets im Repositorium abbilden lässt. Zudem müssten sämtliche Auswertungsschritte inkl. aller Zwischenschritte nachvollziehbar beschrieben sein.

Allein aufgrund der vielfach sehr geringen Fallzahlen, kommt schließlich der Frage des Datenschutzes in der experimentellen Forschung eine hohe Bedeutung zu. Die interviewte Neuroökonomin erhebt in ihren Experimenten stets Daten, die zu einem gewissen Grad personenbezogen sind. Sie ist sich allerdings unsicher, ob ihre bisherigen Anonymisierungsschritte im Falle einer Veröffentlichung den datenschutzrechtlichen Anforderungen genügen würden: *„Das halte ich für eine Grauzone, also dass der Datenschutz gewährleistet ist, ist mir extrem wichtig, also ich möchte möglichst viele Daten bereitstellen, aber natürlich keine personenbezogenen Daten unanonymisiert von irgendwelchen Probanden veröffentlichen“* (G55). Sollten datenschutzrechtliche Fragen nicht eindeutig geklärt werden können, so die Forscherin, würde sie im Zweifel lieber auf eine Archivierung der Daten verzichten als sich in besagte rechtliche Grauzone zu begeben. Der Bereich der Neuroökonomie ist hierbei freilich erneut ein Spezialfall, da die Daten immer medizinische Informationen über Einzelpersonen enthalten und daher besonders hohe Anforderungen an den Datenschutz stellen.

2.3 Prozessproduzierte Daten – *„zu klären wäre, was man einstellen darf“*

Das Spektrum prozessproduzierter Daten umfasst im Prinzip alle Daten, die im Verlauf der sozialen Prozesse „von selbst“, ohne Eingriff der Forschenden, produziert und mehr oder weniger systematisch gesammelt werden. Häufig handelt es sich um Daten, die im weitesten Sinne aus staatlichem Verwaltungshandeln resultieren und über amtliche Register erfasst werden. Beispiele wären Daten der Sozialversicherungen, Bevölkerungs- oder Kriminalitätsstatistiken. Als Quelle kommen aber auch private Datenarchive bzw. Daten von Unternehmen in Frage. Teils sind die Daten leicht zugänglich, da sie von Behörden oder Organisationen im Netz bereitgestellt werden, teils werden sie nur restriktiv zugänglich gemacht oder sogar ausschließlich kommerziell angeboten.

Der Arbeitsaufwand, den Forschende in die Aufbereitung prozessproduzierter Daten investieren, ist häufig enorm: *„Es ist grundsätzlich vieles verfügbar, aber sehr mühsam einzusammeln. Und alle, die so was machen, setzen immer jede Menge Studenten und Doktoranden ran, um irgendwie diese Daten zusammenzusammeln“* (D13). In einem Beispiel wurden etwa Handbücher einer internationalen Luftverkehrs-Vereinigung mit Angaben aus 160 Ländern gescannt und anschließend per Hand in sozialwissenschaftlich verwertbare quantitative Daten überführt. Der fertige Datensatz ist verhältnismäßig einfach, zur Interpretation der Daten werden jedoch unterschiedliche Kontextinformationen benötigt, die bei einer Archivierung der Daten mitgeführt werden müssen. Der

Forscher stellt sich zudem die Frage, ob nicht im Sinne einer vollständigen Nachvollziehbarkeit sogar die ursprünglichen Rohdaten, in diesem Fall also die gescannten Dokumente, zur Verfügung gestellt werden müssten, da bei der Verarbeitung des Rohmaterials bereits erste Veränderungen vorgenommen wurden. Aufgrund des hohen Arbeitsaufwands bei der Aufbereitung vermutet der Forscher zudem, dass seine Kollegen eher zögerlich seien, was die Freigabe dieser Daten anbelangt.

In einem zweiten Beispielfall werden auf Grundlage prozessproduzierter Daten Energiemarktmodelle berechnet. Der Forscher verwendet hierzu Daten der Bundesnetzagentur, unterschiedlicher Netzbetreiber, Kraftwerksdatenbanken sowie verschiedene kommerzielle Informationen. Die Quellen sind so vielfältig, das zur Berechnung einzelner Modelle auf 40 bis 50 Datenfiles zugegriffen wird. Der Prozess der Zusammenstellung der Daten aus den heterogenen Quellen wird nach Einschätzung des Forschers derzeit nicht in einer Form dokumentiert, die für Dritte nachvollziehbar wäre. Aufgrund der heterogenen Quellenlage wird zudem wiederholt die Frage der Nutzungsrechte angesprochen: *„Ich finde das sehr gut mit dem Repositorium, aber das wäre dann halt die Frage, so grundsätzlich, was darf man in einem Repositorium ablegen und in welcher Form. Selbst wenn die Daten im Netz verfügbar sind, haben wir letztlich keine Lizenz, die einfach zu reproduzieren“* (D11). Um solche Daten zu veröffentlichen, müssten die Forscher sich also auf freie Quellen beschränken oder, was realistischer sei, die freien von den kommerziellen Daten trennen. Bei gewachsenen Datenbeständen, in die mit der Zeit alles Mögliche eingeflossen ist, *„kann aber niemand mehr so genau sagen, ob da nicht Sachen drinstecken, die man nicht hochladen darf“* (D115). Erschwerend kommt hinzu, dass prozessproduzierte Daten häufig im Zeitverlauf fortgeschrieben werden, da am aktuellen Rand ständig neue Informationen erzeugt werden. Die Archivierung müsste daher die Möglichkeit der laufenden oder periodischen Aktualisierung bieten, und es stellt sich die Frage in welchen Abständen eine Archivierung sinnvoll wäre.

Der Forscher schlägt aufgrund dieser speziellen Bedingungen vor, in seinem Fall nur die Metadaten und den Modellcode (einfache Syntax-Files in proprietären Formaten) zu veröffentlichen. Ohne die Daten sind die Ergebnisse der Modelle jedoch freilich nicht replizierbar: *„Man kann sich dann halt die Modellformulierung ansehen, aber kann es halt nicht nachrechnen, das bringt sozusagen nur einen gewissen Fortschritt in der Transparenz“* (D13).

2.4 Sekundärdatennutzung – *„Hybride Datenbanken“*

Auf den ersten Blick kommt die Arbeit mit Sekundärdaten nicht als Nutzungsszenario für die Archivierung von Forschungsdaten in Frage. In den Gesprächen mit den Wissenschaftler/innen wurde indes an verschiedenen Stellen deutlich, dass Nutzer von Sekundärdaten sich nicht darauf beschränken, neue Auswertungen durchzuführen. Sie verarbeiten die Daten in unterschiedlichster Form weiter, sei es indem sie verschiedene Sekundärdatenquellen zusammenführen und dafür neu aufbereiten müssen, sei es, dass sie durch eigene Berechnungen neue Variablen generieren oder die

Daten mit neuen Informationen anreichern. Diese Aufbereitung und Weiterverarbeitung führt oftmals zu völlig neuen Analyseoptionen.

So wurde in einem Interview von einem Forschungsprojekt berichtet, das die aufwendige Verknüpfung einer europaweiten Erhebung mit Zensusdaten aus den USA zum Ziel hatte. Die Ergebnisse ließen für bestimmte Indikatoren erstmals einen Vergleich zwischen einzelnen europäischen Ländern und den USA zu. In einem zweiten Beispiel wurden volkswirtschaftliche Langzeitreihen der letzten 50 Jahre aufbereitet, ebenfalls mit dem Ziel, ökonomische Kennziffern im OECD-Kontext vergleichbar zu machen. Mehrfach wurde durch die Interviewpartner/innen zudem die Verknüpfung von Mikro- und Makrodaten angesprochen. Ein Forscher spricht diesbezüglich von der Generierung *„hybrider Datenbanken"* (E7). Mikrodaten, zumeist Ergebnisse quantitativer Befragungen, werden dabei mit unterschiedlichsten Makro-Indikatoren, bspw. ökonomischen Kennziffern, Informationen zu politischen Systemen oder Umwelteinflüssen angereichert. Ein Sonderfall ist die Verknüpfung mit Regionaldaten bzw. geographischen Informationssystemen (GIS). Dies ermöglicht das Erstellen kartografischer Produkte, das Aufdecken räumlicher Muster und das Errechnen raumbezogener Prädiktoren, die wiederum in weitere Analysen einfließen können.

An den Beispielen wird schnell deutlich, wie schwierig es sein dürfte, genauer zu definieren, ab wann in diesen Fällen tatsächlich qualitativ neues Datenmaterial generiert wird. Unstrittig ist jedoch sicher, dass die Forschenden in die Aufbereitung der Sekundärdaten zum Teil sehr viel Zeit investieren und die bestehenden Daten durch ihre Arbeit aufwerten. Aus Perspektive des Forschungsdatenmanagements ist es daher sicherlich sinnvoll, auch neu aufbereitete Sekundärdaten zu archivieren und nachnutzbar zu machen. Gleichwohl bestehen bei den Forschenden diesbezüglich erhebliche Bedenken: *„Nun ist aber Folgendes, jetzt berechne ich also Wachstumsraten dieser Indikatoren, also wo liegen dann die Eigentumsrechte, bei dem, der die Originaldaten bereitstellt oder indem ich diese Daten in Excel eingebe und der Computer spuckt mir dann neue Ergebnisse aus, und dann gibt es ja noch Berechnungen mit Filtern, und ich führe Glättungen durch, und dann kriegt man so Langzeittrends, das ist dann schon eine Streitfrage, das ist schwierig, da gibt's sicher unterschiedliche Meinungen"* (B37). Grundsätzlich ist der Forscher, der hier von seiner Arbeit mit Sekundärdaten berichtet, bereit seine Daten in einem Repositorium zu veröffentlichen. Er möchte jedoch Sicherheit dahingehend haben, dass er hierdurch nicht die Rechte Dritter verletzt. Eine Trennung der eigenen Anteile der Datenaufbereitung von den Originaldaten ist allerdings nur schwer vorstellbar, insbesondere wenn der innovative Gehalt im hybriden Zusammenspiel unterschiedlicher Daten besteht. Für die Dokumentation von weiterverarbeiteten Sekundärdatenquellen stellt sich schließlich die grundsätzliche Frage, wie eine Beschreibung durch Metadaten erfolgen kann, wenn der Datensatz eine Verknüpfung verschiedener Datenquellen darstellt.

2.5 Mixed Methods – *„die Zukunft der empirischen Forschung“*

Unter Mixed Methods Forschung wird zumeist die Kombination quantitativer und qualitativer Methoden verstanden. So wurde in einem Interview bspw. von einem Forschungsprojekt zur Nutzung des öffentlichen Raums berichtet, in dem qualitative Interviews den Ausgangspunkt für die Entwicklung eines standardisierten Fragebogens bildeten. Zusätzlich wurde in einem Feldexperiment mit ethnografischen Methoden gearbeitet (Beobachtungen, Kartenexperimente, Fotografien), und schließlich kam in einem speziellen Teilprojekt Eye-Tracking-Technik zum Einsatz. Als Datenmaterial liegen im Projekt somit Transkripte, quantitative Datensätze, Beobachtungsprotokolle, Fotos, Karten sowie Eye-Tracking Daten vor. Die interviewte Projektleiterin ist von den Chancen und Möglichkeiten des Mixed Methods Ansatzes überzeugt: Mit qualitativen Methoden könne *„ein feinerer Strich gezeichnet“* werden, mit quantitativen Methoden sind *„mehr generalisierende Aussagen möglich“* (F10). Die Kombination beider Formate ergänze sich optimal und berge viele Möglichkeiten für die Zukunft. Auch ein weiterer Interviewpartner sieht aktuell einen deutlichen Trend zum Mix im Forschungsdesign, der Brücken über die alten Methodenlager schlagen wird: *„Die Zukunft der empirischen Forschung wird Mixed Methods sein“* (C61).

Die zentrale Herausforderung bei der Archivierung von Forschungsdaten aus Mixed Methods Projekten ergibt sich entsprechend aus der Kombination der unterschiedlichen Forschungsmethoden. Im Idealfall müssten nach Meinung einer Interviewpartnerin die unterschiedlichen Datensätze in ihrer Gesamtheit archiviert werden, da *„der Reiz ja gerade darin liegt, dass man über die verschiedenen Materialtypen hinweg Vergleiche ziehen kann“* (C5). Bisherige Ansätze werden aufgrund der strikten Trennung von quantitativen und qualitativen Daten als unzureichend eingeschätzt, da sie *„das Material auseinanderreißen, was ja genau das ist, was man als Forscherin nicht möchte“* (C5). Benötigt werden hingegen speziell Archivierungskonzepte für Mixed Methods Daten, die nicht nur die unterschiedlichen Datentypen, sondern das gesamte Studiendesign abbilden. Nur so wäre bspw. nachvollziehbar, wie aus einer offenen Befragung in einem zweiten Schritt ein standardisierter Fragebogen entwickelt wurde. Hilfreich wäre zudem die Möglichkeit, für unterschiedliche Datentypen separate Zugriffsrechte zu vergeben, da die Anforderungen je nach Datentyp stark variieren. So könnten bspw. die quantitativen Daten direkt freigegeben werden, während qualitative Transkripte nur auf Nachfrage bei der Primärforscherin einsehbar wären. Hohe Anforderungen bestehen schließlich in Bezug auf den Datenschutz, was v.a. auf den teils erheblichen Anteil des qualitativen Materials zurückzuführen ist (vgl. hierzu Abschnitt 2.6).

Neuen Konzepten, die den Umgang mit Forschungsdaten und deren dauerhafte Ablage und Beschreibung erleichtern, steht die interviewte Mixed Methods Forscherin sehr offen gegenüber. Insbesondere werden hierin auch Chancen gesehen, die Fülle der Daten strukturierter zu verwalten. Jedoch wird auch eine gewisse Skepsis deutlich, was den Zeit- bzw. Arbeitsaufwand anbelangt, der für eine gute Dokumentation im Mixed Methods Bereich notwendig wäre. Ohne die Bereitstellung

zusätzlicher finanzieller bzw. personeller Mittel für diese Aufgaben sei es im Forschungsalltag perspektivisch kaum leistbar, den neuen Anforderungen gerecht zu werden.

2.6 Qualitative Forschung – *„eine Großbaustelle“*

Auch wenn die qualitative Forschungspraxis nicht ausdrücklich Gegenstand der Interviews war, gab es in den Gesprächen doch eine Vielzahl an Hinweisen und Nachfragen zu den Möglichkeiten der Archivierung qualitativer Daten. Zunächst ist festzuhalten, dass auch der qualitative Bereich ein ganzes Spektrum unterschiedlicher Forschungsmethoden umfasst: Dies reicht von klassischen Experteninterviews oder Fokusgruppen über ethnografische Forschung (teilnehmende Beobachtung, Feldnotizen, Fotografie), der qualitativen Codierung von Sekundärtexten (Zeitungsartikel, Parteiprogramme) bis hin zur Arbeit mit Videoaufzeichnungen. All diese Methoden stellen spezifische Anforderungen an eine Archivierung der Daten, auf die im Rahmen der Interviews jedoch im Detail nicht eingegangen werden konnte.

Die systematische Dokumentation und Archivierung qualitativer Daten wird von einem Interviewpartner als *„Großbaustelle“* (C5) bezeichnet. Es mangele an Infrastrukturangeboten, so dass die Daten aktuell zumeist bei den Forschenden verbleiben. Aufzeichnungen aus älteren Projekten liegen häufig noch als Tonbandmaterial vor, weshalb ein fortlaufender Datenverlust befürchtet wird. Eine grundlegende Frage sei dabei, in welcher Phase der Datenaufbereitung das Material archiviert werden soll. Transkribierte Interviews oder die dazugehörigen Originalaufzeichnungen könnten prinzipiell leicht zur Verfügung gestellt werden. Zur Nachvollziehbarkeit der Interpretationen wären jedoch die codierten Datenbanken aus Softwareprogrammen zur computergestützten qualitativen Daten- und Textanalyse deutlich besser geeignet. Für die Beschreibung qualitativer Daten müssten zudem spezielle Metainformationen erfasst werden, die u.U. deutlich detaillierter ausfallen können als die Metadatenstandards für quantitatives Material. Im Unterschied zu den quantitativen Daten werden zur Interpretation des Materials ferner umfangreiche Kontextinformationen benötigt. Allein die Interviewsituation bzw. die Gesprächsatmosphäre kann hier von Bedeutung sein. Gerade hieraus speisen sich jedoch auch grundlegende Zweifel, ob die qualitativen Daten durch Forschende, die selbst nicht an deren Erhebung beteiligt waren, überhaupt sinnvoll nachgenutzt werden können.

Als eine wesentliche Herausforderung bei der Archivierung qualitativer Daten wurde darüber hinaus mehrfach die Frage der Anonymisierung angesprochen, da das Material häufig eine Vielzahl personenbezogener Informationen enthält. Unter qualitativ arbeitenden Forscher/innen sind daher Befürchtungen bzgl. der Weitergabe der Daten weit verbreitet. Auch können die Forschungsfelder deutlich kleiner ausfallen als in der quantitativen Forschung üblich (wenn bspw. innerhalb einer Organisation oder in einem bestimmten Wissenschaftsfeld geforscht wird). In diesen Fällen sind Personen, trotz paraphrasierter Texte und geänderter Namen, leicht anhand von längeren Interviewpassagen identifizierbar. Dies hätte jedoch schwerwiegende Auswirkungen auf das Vertrauensverhältnis zwischen Forschern und Interviewpartnern zur Folge: *„Es gibt die Angst, dass*

der Vertrauensschutz der Forschungssubjekte, die einen ins Feld reingelassen haben, missbraucht wird, aber auch, dass man selber bloßgestellt wird, da die Feldnotizen natürlich viel mehr von der Person des Forschenden offen legen als z.B. eine standardisierte Befragung" (C13). Die Lösung dieses Problems sieht der Gesprächspartner allein in einer deutlichen Verlängerung der Karenzzeiten, die Rede ist hier von Zeiträumen zwischen zehn und zwanzig Jahren.

Grundsätzlich besteht nach Ansicht des interviewten Forschers jedoch durchaus ein Bedarf, qualitatives Material zu archivieren und verfügbar zu machen, zum einen mit Blick auf die Möglichkeiten der Nachnutzung, aber auch in Hinblick auf die Transparenz innerhalb des qualitativen Forschungsfeldes. Hierfür werden jedoch gute Archivierungskonzepte benötigt, die den Aufwand für die Forschenden möglichst gering halten und den spezifischen Eigenschaften des Materials gerecht werden[4].

2.7 Auftragsforschung und private Institute – *„Ping-Pong-Effekte"*

Durch eine Interviewpartnerin wurden schließlich die Rahmenbedingungen der Forschung und hierbei insbesondere der mögliche Einfluss der Forschungsförderer thematisiert. So gelten Objektivität und Unabhängigkeit nach wie vor als fundamentale Grundlagen wissenschaftlicher Arbeitsethik. Gleichwohl sind Forschungsaufträge von Unternehmen oder Verbänden in der Praxis durchaus keine Seltenheit. Handelt es sich um reine Auftragsforschung, kann dies erhebliche Konsequenzen für die Wissenschaftler/innen mit sich bringen. Die Rechte an den Ergebnissen ihrer Arbeit liegen in diesen Fällen zumeist klar beim jeweiligen Auftraggeber. Dies betrifft Projektberichte ebenso wie die erhobenen Forschungsdaten. Betreffen die Ergebnisse bspw. Interna eines Unternehmens, ist stets damit zu rechnen, dass die Auftraggeber kein Interesse daran haben, die Resultate in vollem Umfang einer breiteren Öffentlichkeit zugänglich zu machen. Etwas unübersichtlicher ist die Situation, wenn der Einfluss des Auftraggebers weniger eindeutig erkennbar ist. Unter Umständen handelt es sich zwar nicht um formale Forschungsaufträge, gleichwohl können sich Einschränkungen durch informelle Regelungen und Verpflichtungen gegenüber den Geldgebern ergeben. Ein typisches Bsp. ist eine Förderung durch Ministerien, welche Forschungsergebnisse und Daten unter Umständen zunächst nur intern oder mit einer bestimmten (politischen) Zielstellung nutzen wollen. Bei der Datenarchivierung ist somit zu berücksichtigen, dass Forschungsförderer unter bestimmten Bedingungen einen einschränkenden Einfluss auf die Verfügungsmöglichkeit der Primärforscher über ihre Forschungsdaten geltend machen können.

Schließlich wurde ein Interview mit einem Gesprächspartner geführt, der in einem privaten Forschungsinstitut tätig ist. Auch in diesem Kontext werden teils umfangreiche Forschungsdaten produziert, die potentiell für breitere wissenschaftliche Forschungszwecke bereitgestellt werden

[4] Hier wird bspw. auf das Archiv für Lebenslaufforschung (ALLF) in Bremen verwiesen, dessen Vorarbeiten und archivierten Daten inzwischen im Angebot Qualiservice aufgegangen sind (http://www.qualiservice.org/).

könnten. Nach Einschätzung des Interviewpartners ist die systematische Dokumentation und Archivierung empirischer Daten jedoch, allein aufgrund des erhöhten Arbeitsaufwands, außerhalb von Universitäten und größeren Forschungseinrichtungen kaum leistbar. Im konkreten Beispiel werden Daten daher interessierten Nutzern nur auf direktem Wege zur Verfügung gestellt. Hierfür wird ein Datennutzungsvertrag geschlossen, der u.a. die Verpflichtung beinhaltet die Forschungsdaten nicht an Dritte weiterzugeben. Die Forschenden müssen zudem vorab eine Skizze einreichen, die verdeutlicht, zu welchem Zweck sie die Daten verwenden werden. Ein wichtiges Ziel dieser Vorgehensweise ist es, die bestmögliche Kontrolle über die Nachnutzung zu behalten. Man entschied sich daher auch bewusst gegen eine Archivierung der Daten in einem Forschungsdatenrepositorium. Der Forscher ist zudem der Ansicht, dass mögliche Interessenten ohnehin überwiegend durch Publikationen auf die Daten aufmerksam werden und sich dann direkt bei den Primärforschern melden. Wären die Daten in einem Repositorium hinterlegt, müsste man diese Anfragen dann stets erst an den Infrastrukturanbieter weiterleiten. Im ungünstigsten Fall würde dieser sich dann erneut bei den Forschenden melden, bspw. wenn Daten nur mit Zustimmung der Primärforscher freigegeben werden sollen. Daher wurden „Ping-Pong-Effekte“ (M35) und ein zusätzlicher Arbeitsaufwand befürchtet. Dennoch wird die Weitergabe von Forschungsdaten insgesamt positiv gesehen, insbesondere da man sich hiervon einen Legitimationsgewinn gegenüber den Geldgebern verspricht: Werden Daten in anderen Forschungsprojekten oder von Doktoranden genutzt, sei dies stets auch ein Beleg für die Relevanz der eigenen Forschungstätigkeit und ein gutes Argument für künftige Forschungsanträge, so der Forscher (M35).

2.8 Synopse typischer Forschungsszenarien

Insgesamt wird deutlich, dass sich das Feld empirischer Forschungspraxis in den Sozial- und Wirtschaftswissenschaften als äußerst heterogen erweist. Unterschiedliche Forschungsansätze bringen unterschiedliche Herausforderungen in Hinblick auf Datenmanagement und Archivierung der Forschungsdaten mit sich. Es sind dabei erwartungsgemäß vor allem die Unterschiede in den Forschungsmethoden bzw. -designs, die zu speziellen Bedarfen führen. Tabelle 1 stellt die Forschungsszenarien und ihre besonderen Merkmale im Überblick dar.

Tabelle 1: Übersicht besonderer Merkmale unterschiedlicher Forschungsszenarien

Szenario	Besondere Merkmale
Survey	• Beteiligung von Umfrageinstituten • Cross-Country-Surveys • Einschaltungen
Experimentelle Designs	• Im Einzelfall große Datenmengen • Teils komplexe Forschungsdesigns/Datenstrukturen • Sensible Daten – hohe Datenschutzanforderungen
Prozessproduzierte Daten	• Teils komplexe Datenstrukturen • Nutzungsrechte an Daten unklar • Dynamische Daten
Sekundärdatennutzung	• Verknüpfung unterschiedlicher Datenquellen • Komplexe Datenstrukturen • Nutzungsrechte an Daten unklar
Mixed Methods	• Teils komplexe Forschungsdesigns/Datenstrukturen • Verschiedene Datentypen (qualitativ/quantitativ) • Hohe Datenschutzanforderungen
Qualitative Forschung	• Hohe Datenschutzanforderungen • Spezielle Dokumentationsinhalte (Kontextinformationen)
Auftragsforschung und private Forschungsinstitute	• Einfluss der Forschungsförderer • Spezielle Rahmenbedingungen, Interessen u. Kapazitäten

Aus der Gesamtschau der einzelnen Merkmale lassen sich sechs zentrale Punkte ableiten. Diese können zudem in entsprechende Leitfragen übersetzen werden, die Forschende an das Datenmanagement stellen, wenn sie mit der Anforderungen konfrontiert werden, dass ihre Forschungsdaten zukünftig umfassend dokumentiert, archiviert und anderen Forschenden zur Verfügung gestellte werden sollen:

1. Geringe Fallzahlen (z.B. in Experimenten), nicht standardisierte Methoden (z.B. qualitative) oder besonders sensible Inhalte einer Erhebung (z.B. medizinische Daten) führen zu erhöhten Anforderungen an den ***Datenschutz***.
   - *Frage: Wann & Wie muss ich meine Daten anonymisieren, um bei der Weitergabe über ein Repositorium den rechtlichen Datenschutzanforderungen zu genügen?*

2. Neben den Forschenden selbst können ***weitere Akteure*** in den Prozess der Datengenerierung involviert sein (z.B. Umfrageinstitute, Projektpartner). Dies kann Nutzungsrechte tangieren, ist jedoch insbesondere relevant für die Bereitstellung der beschreibenden Metainformationen.
   - *Frage: Welche Informationen muss ich für die Dokumentation der Forschungsdaten bereitstellen?*

3. In den Projekten der Forschenden kommen teils ***komplexe Forschungsdesigns*** zum Einsatz. Das Ergebnis sind ebenso ***komplexe Datenstrukturen***. Die Erwartung ist, dass sich diese Komplexität der Forschung in einem Repositorium abbilden und dokumentieren lässt.
   - *Frage: (Wie) lassen sich die teils komplexen Designs/Datenstrukturen meiner Untersuchungen im Repositorium abbilden?*

4. Daten werden häufig zu einem festen Messzeitpunkt erhoben. Es gibt jedoch auch ***dynamische Datenbestände***, die fortlaufend mit neuen Informationen angereichert werden.
   - *Frage: Wie kann ich Datensätze im Repositorium archivieren, denen ich fortlaufend neue Informationen zuspiele?*

5. In Sonderfällen können auch innerhalb der Sozial- und Wirtschaftswissenschaften ***größere Datenmengen*** anfallen, v.a. wenn technische Messinstrumentarien zum Einsatz kommen. (z.B. MRT oder Eye-Tracking).
   - *Frage: (Wie) lassen sich große Datenmengen in ein webbasiertes Repositorium hochladen?*

6. Die Frage der ***Nutzungsrechte*** muss für die Forschenden hinreichend geklärt sein. Forschende wollen diesbezüglich absolute Sicherheiten und keine rechtlichen Grauzonen!
   - *Frage: Welche Daten darf ich in das Archiv einstellen und welche nicht?*

3. Bedenken und Anforderungen der Forschenden

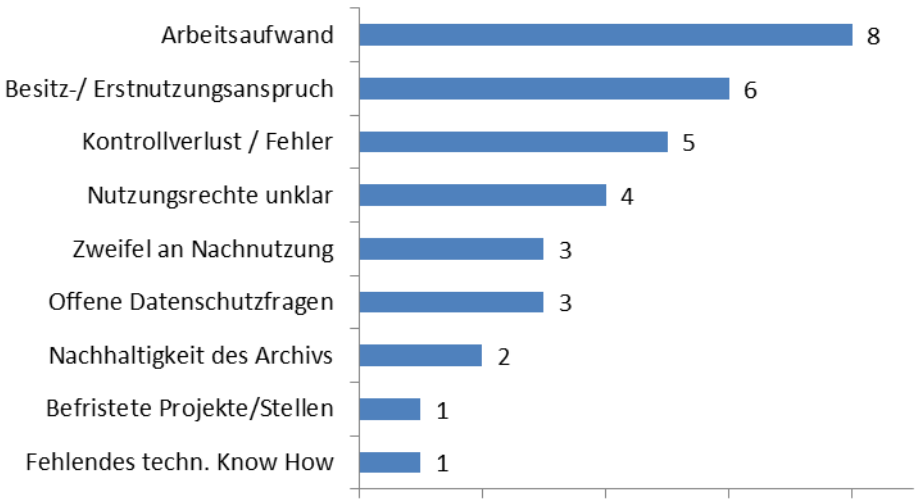
Bereits in den ersten Interviews wurde deutlich, dass es den Gesprächspartner/innen schwerfallen würde, Auskunft über konkrete Anforderungen an ein webbasiertes Forschungsdatenrepositorium zu geben. Dies ist v.a. dadurch begründet, dass aus der Gruppe der Interviewten bislang niemand auf umfangreiche praktische Erfahrungen in der Datenarchivierung zurückblicken kann. Dennoch wurden in den Gesprächen vielfältige Hinweise formuliert, welche im Folgenden systematisch aufbereitet sind. Dabei wird zwischen Bedenken der Forschenden hinsichtlich der Archivierung und

Veröffentlichung ihrer Daten (3.1) und den Anforderungen an die potentielle Nutzung eines Infrastrukturangebots (3.2) unterschieden.

3.1 Bedenken hinsichtlich der Archivierung und Bereitstellung von Forschungsdaten

Im SDN-Workshop zur Anforderungsanalyse wurden bereits zahlreiche Bedenken der Forschenden thematisiert. Dabei wurde deutlich, dass es sich zumeist weniger um technische, sondern v.a. um nicht-technische Hemmnisse bzw. Barrieren handelt. Ein Ergebnis, dass sich auch mit den Resultaten vorliegender Studien zum Thema Data-Sharing deckt[5]. Die qualitativen Interviews bestätigen diese Erkenntnisse auf breiter Linie. Zugleich wurden jedoch auch einige zusätzliche Aspekte angesprochen. Abbildung 1 stellt die Bedenken der Forschenden sortiert nach der Anzahl ihrer Nennung dar. Mehrfachnennungen in einem einzelnen Interviews wurden dabei als einfache Angabe gewertet.

Abbildung 2: Bedenken der Forschenden sortiert nach Anzahl der Interviews mit entsprechender Nennung



Von beinahe allen Gesprächspartner/innen wird die Befürchtung geteilt, dass mit der Archivierung und insbesondere mit der hierfür im Vorfeld notwendigen Dokumentationen und Aufbereitung der Daten ein zusätzlicher Arbeitsaufwand einhergeht: *„Die Frage ist natürlich, sagen wir mal, mit welchem Aufwand das verbunden ist. Also ich wäre bereit, die Daten alle ins Netz zu stellen, ich hätte allerdings keine Lust, daran irgendwie einen Monat zu arbeiten, um die irgendwie einzugeben, dann sage ich nee, also keinen Bock drauf“* (B37). Als Hintergrund werden wiederholt geringe Spielräume in Bezug auf die finanziellen bzw. zeitlichen Ressourcen angeführt. Die Personalkapazitäten in

[5] Vgl. Feijen, Martin (2011): What researchers want. Utrecht: SURF-Foundation. Zugriff am 18.03.2015 unter http://www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/What_researchers_want.pdf

Forschungsprojekten werden, nach Einschätzung der Wissenschaftler/innen, immer knapper kalkuliert. Die Folge ist bereits heute vielfach ein starke Aus- bzw. teils sogar Überlastung im Forschungsalltag. Der Gedanke an zusätzliche und vermeintlich unbezahlte Arbeitsaufgaben stößt daher auf wenig Begeisterung.

Was den konkreten Arbeitsaufwand anbelangt, haben die Forschenden realistischerweise nicht nur die eigentliche Erfassung ihrer Daten, sondern insbesondere auch die notwendigen Vorarbeiten im Blick. So ist die Datendokumentation derzeit häufig nur für den internen Gebrauch konzipiert, bspw. um Informationen für die eigene Weiterverwendung oder für Kolleg/innen festzuhalten. Hierbei wird bislang eher unsystematisch vorgegangen: *„Ich weiß nicht, wie es woanders ist, aber bei uns gibt's das im eigentlichen Sinne nicht, so eine strukturierte Dokumentation, jeder macht das irgendwie auf seine Art. […] Wenn man es sozusagen mal für die Allgemeinheit vernünftig aufbereiten wollte, würde man das wahrscheinlich ganz anders konzipieren.*" (D33/59). Die Daten sind also nicht in einer Form dokumentiert, die man direkt einer breiteren Community zugänglich machen könnte. Gleiches trifft auf die Dokumentation konkreter Auswertungsschritte zu. Syntaxfiles werden im Moment vor allem so angelegt, „dass sie laufen". Bestünde das Ziel darin, solche Syntaxfiles anderen Forschenden zugänglich zu machen, wäre dies ebenfalls mit einem zusätzlichen Aufwand verbunden: *„Oft gibt es ja auch so gewachsene Prozesse, man macht irgendwas mal in einer Form und erweitert es dann mit der Zeit. Die Syntaxfiles sind eben gewachsen und wenn es keinen Druck gibt zu investieren in Sachen, die eigentlich laufen, nur damit sie schöner und strukturierter sind, macht man es halt im Zweifel immer nicht. So wie das immer ist. Also ja das wären halt Dinge, wo man wahrscheinlich sagen müsste, wenn man was teilt, wird es wahrscheinlich noch mal mehr Arbeit machen."* (D59). Durchweg auffällig ist, dass die Forschenden beim Stichwort „Dokumentation" überwiegend an die Dokumentation ihrer Datenanalysen denken und weniger an eine grundlegende Beschreibung ihrer Datensätze durch Metainformationen.

Als zweithäufigster Punkt wurde der ebenfalls bereits aus dem Workshop bekannte Anspruch geäußert, die eigenen Daten zunächst selbst voll auszuschöpfen: *„Also speziell die aufbereiteten Sekundärdaten, da haben wir dann manchmal Anfragen, da sind wir aber sehr zurückhaltend, einfach weil es wahnsinnig viel Arbeit war und weil man jetzt erstmal nichts davon hat, wenn man den Datensatz rausgibt. Also eigentlich müssten wir selber noch erstmal mehr davon veröffentlicht haben, wenn ich das mal so frank und frei sagen kann"* (H31). Die Befürchtung ist, dass man sich durch eine zu frühzeitige Veröffentlichung der Daten selbst der Möglichkeit beraubt, weitere Ergebnisse zu publizieren. *„Also wir haben uns auch nicht so sehr bemüht [die Daten zugänglich zu machen], ich würde auch immer sehen, dass es ja auch ein bisschen gefährlich ist, gerade solche originären Daten, die wenig vorhanden sind, breit zu streuen, weil das natürlich in unserer Hand auch ein Pfund war. Wenn ich das gleich auf den Markt schmeiße, dann ist das natürlich weg, und wir wussten ja auch nicht, inwieweit wir selber noch weitere Analysen damit durchführen"* (K23). Die Grenze zwischen Erstnutzungsanspruch und einem ausschließenden Besitzanspruch ist hierbei fließend. Letzterer wird z.B. dadurch begründet, dass Daten durchaus auch in eigenen Folgeprojekten nachgenutzt werden

können: *„Na ja, wenn ein Projekt abgeschlossen ist, also oft ist es so, dass man die Datensätze auch so ein bisschen von Projekt zu Projekt mitschleppt, also gerade z.B. diese Sekundärdaten, die wir aufbereitet haben, die wurden dann sozusagen in unserer Projektgruppe weitergeführt, einfach weil die Aufbereitungskosten von solchen Daten so hoch sind, dass es eigentlich schade ist, sie nicht in eigenen Projekten weiter zu benutzen. Also von daher ist das so ein fließender Übergang"* (H31). In diesem Zusammenhang ist die exklusive Verfügbarkeit von Forschungsdaten also auch ein „Pfund" für neue Projektanträge, in denen die Möglichkeit, Daten exklusiv für eine neue Fragestellung auszuwerten, durchaus ein zentrales Argument darstellen kann, dass zur Bewilligung der finanziellen Mittel führt.

An dieser Stelle wird deutlich, dass sich die Sichtweise der Forschenden auf ihre Daten deutlich vom Verständnis anderer Wissenschaftsprodukte unterscheidet. So werden Publikationen als abgeschlossenes Werk betrachtet, dessen Nachnutzung durch die formalen Regeln der Zitation klar definiert ist. Im Sinne der Mertonschen Wissenschaftsethik werden sie daher der Community als eine Art Gemeingut zu Verfügung gestellt. Forschungsdaten werden hingegen vielmehr als eine Art Zwischenprodukt der wissenschaftlichen Arbeit verstanden, in das bereits Arbeitszeit sowie eigene Ideen investiert wurden. Sie sind in diesem Sinne Träger eines intellektuellen Kapitals und können wiederholt „verwertet" werden, um in der Forschungsarbeit neu Ergebnisse zu produzieren. Aus Sicht der Primärforscher kommt die Veröffentlichung ihrer Daten daher einem Verzicht auf zukünftige Nutzungsoptionen gleich. Im Konkurrenzbetrieb des wissenschaftlichen Feldes werden entsprechend Wettbewerbsnachteile befürchtet. Die Bereitstellung von Forschungsdaten steht somit exemplarisch für das paradoxe Spannungsverhältnis von Kooperation und Konkurrenz im modernen Wissenschaftsbetrieb.

In engem Zusammenhang mit dem Besitz- bzw. Erstnutzungsanspruch steht die Befürchtung, dass man als Forscher/in durch die Weitergabe der Daten die Kontrolle über die eigene Arbeit verlieren könnte: *„Nein, man hat ja auch keine Kontrolle mehr darüber. Der Do-File ist total schnell verschickt so, ich habe aber unglaublich viele Stunden da reingesteckt, nee das würde ich nicht machen, und ich glaube auch andere würden das nicht wollen"* (H39). Dahinter steht zum einen der Wunsch, die inhaltliche Deutungshoheit über die Daten zu behalten. Man möchte vermeiden, *„dass sich die Zahlen, die Daten oder der Datensatz im Prinzip selbständig macht und man darauf keinen Einfluss mehr hat"* (K35). In zugespitzter Form kann dies sogar zu der Befürchtung führen, andere Forschende könnten die eigenen Auswertungen replizieren, um methodische Fehler oder fehlerhafte Interpretationen nachzuweisen: *„Man hat ja immer die Angst, es guckt sich jemand anders seine Daten an und dann findet er Fehler und man wird dann öffentlich bloßgestellt"* (C13). Unabhängig davon, wie realistisch diese Befürchtungen im Einzelnen auch sein mögen, kann ein faktischer Reputationsverlust im Wissenschaftsbetrieb ohne Frage negative Auswirkungen zur Folge haben. Die öffentlichen Plagiatsdebatten der vergangenen Jahre haben derlei Ängste sicherlich zusätzlich verstärkt.

Die vorhandenen Unsicherheiten bezüglich der Nutzungsrechte sowie offene Fragen hinsichtlich des Datenschutzes wurden bereits in Abschnitt 2 mehrfach angesprochen. Wenn Forschende prozessproduzierte Daten unterschiedlicher Provenienz verarbeiten, Sekundärdaten aufbereiten oder Datenquellen mit unterschiedlichen Nutzungsrechten verknüpfen, bestehen häufig Befürchtungen, dass im Zuge der Archivierung die Rechte Dritter verletzt werden könnten. Spezielle Anforderungen an den Datenschutz sind hingegen vor allem dann zu erwarten, wenn qualitative Daten archiviert werden sollen. Die Identifizierbarkeit einzelner Personen aus diesem Datenmaterial ist häufig nur schwer zu verhindern. Für quantitative Daten bestehen demgegenüber deutlich seltener grundlegende datenschutzrechtliche Bedenken, gibt es doch zumeist einfache Möglichkeiten, die sensible Angaben bzw. personenbezogene Informationen aus den Datensätzen zu entfernen.

Eine Befürchtung, die im Rahmen des Workshops noch nicht thematisiert wurde, betrifft die potentielle Nachnutzung der Inhalte eines Forschungsdatenrepositoriums. Die Nachfrage nach Sekundärdaten aus kleineren Studien, so die Einschätzung mehrerer Interviewpartner, dürfte insgesamt eher gering ausfallen. Hierfür werden unterschiedliche Gründe angeführt: Zum einen würden Forschende eigene Erhebungen vielfach der reinen Sekundäranalyse vorziehen. Umfragen seien heutzutage kostengünstig und schnell zu realisieren, liefern „neue Ergebnisse“ und erhöhen somit den Wert und die Aktualität der eigenen Forschung. Auch für die Begründung neuer Forschungsvorhaben seien eigene Datenerhebungen oftmals besser geeignet. Wenn in Projekten Umfragen durchgeführt werden, bleibe zudem kaum Zeit, ergänzende Sekundärdaten auszuwerten. Der Aufwand, sich einzuarbeiten, sei hoch, die zu erwartenden Ergebnisse hingegen begrenzt. Hinzu kommt, dass man mit den eigenen Fragestellungen in der Sekundärdatenanalyse schnell an Grenzen stoße, da diese stets mit einer spezifischen Zielstellung konzipiert sind. Der Vergleich verschiedener Sekundärdatenquellen untereinander sei zudem aufgrund der unterschiedlichen Designs, Stichproben und Erhebungsinstrumente methodisch fragwürdig. Die Forschenden erwarten daher erhebliche Unterschiede in der Nachnutzung der Daten: Größere und thematisch breiter angelegte Datensätze könnten durchaus häufig nachgefragt werden (K39). Gerade die kleineren und mittleren Befragungen seien jedoch zumeist von zu engem thematischem Zuschnitt und durch ein bis zwei Publikationen der Primärforscher inhaltlich „abgegrast“. Es bestehen somit Bedenken, was das Kosten-Nutzen Verhältnis von Datenarchivierung und Nachnutzung anbelangt. Zumal erwartet wird, dass gerade die größeren und thematischen breiter aufgestellten Datensätze (bspw. das SOEP oder der ALLBUS) auch zukünftig eher über eigene Infrastrukturangebote zugänglich gemacht werden.

Schließlich wurde mehrfach die Frage gestellt, in welcher Form das SDN-Repositorium nach Ende der Projektlaufzeit weiterbetrieben werden soll. Die Sorge der Forschenden ist hierbei, dass die Nachhaltigkeit neuer Datenportale nicht garantiert werden kann, insbesondere wenn diese durch Projektmittel mit begrenzter Laufzeit finanziert werden. Notwendig sei hingegen die Garantie, dass die archivierten Daten nicht nur für die nächsten fünf Jahre, sondern für deutlich längere Zeiträume aufbewahrt und bereitgestellt werden können: *„Da muss ich mir aber auch sicher sein, dass das also*

nicht nur für die nächsten fünf Jahre ist, sondern für die nächsten 500 Jahre, also das sind ja die Zeiträume, in denen ich denken muss, sonst ist es nicht attraktiv, dann kann ich es auch hier auf den Server liegen lassen" (C5).

Ein letzter Punkt wurde zwar nur in einem Interview direkt angesprochen, betrifft jedoch die Rahmenbedingungen der Forschungsarbeit nahezu aller Gesprächspartner/innen. Diese sind überwiegend in Forschungsgruppen bzw. -projekten tätig. Die datenbezogenen Arbeiten werden zumeist von Hilfskräften, Doktoranden oder wissenschaftlichen Mitarbeiter/innen besorgt. Zentrales datenrelevantes Wissen ist somit an Personen gebunden, die häufig nur für kurze Zeiträume, bestenfalls für die Laufzeit des jeweiligen Projekts, an den Instituten beschäftigt sind. Mit der Personalfluktuation ist somit die Gefahr verbunden, dass dieses Wissen verloren geht, insofern es nicht umfassend dokumentiert wurde.

Insgesamt gab es somit kein Interview, in dem nicht Bedenken der einen oder anderen Art geäußert wurden. Immerhin drei der zehn Interviewpartner/innen vertreten allerdings auch die Ansicht, dass Forschende in der Pflicht stehen, ihr Forschungsdaten der Community bereitzustellen: *„Ja, also bei Forschungsvorhaben die mit öffentlichen Geldern finanziert werden, da sehe ich das quasi geradezu als Pflicht, die Daten auch anderen Wissenschaftlern zur Verfügung zu stellen"* (A33).

3.2 Weitere Anforderungen

Entsprechend der mit Abstand am häufigsten geäußerten Befürchtung, dass mit der Archivierung und Bereitstellung der Forschungsdaten ein zusätzlicher Arbeitsaufwand verbunden sein wird, wurden auch die meisten Anforderungen dahingehend formuliert, die Arbeitsbelastung für Forschende möglichst gering zu halten:

Erstens sollte das Infrastrukturangebot zur Archivierung und Veröffentlichung der Forschungsdaten so einfach wie möglich gestaltet werden. Der Zeitaufwand und die inhaltliche Anforderungen sind möglichst gering zu halten: *„Genau ich glaube das wäre auch für mich eine der Hauptanforderungen, es muss halt niedrigschwellig sein, die Daten loszuwerden"* (M39).

Zweitens müssten für die zusätzlichen Aufgaben Ressourcen bereitgestellt werden. Bislang ist in Forschungsprojekten hierfür kein Geld bzw. Personal vorgesehen: *„Dokumentation, da kriegst du kein Geld für"* (K60). Für die Beantragung entsprechender Mittel wären konkrete Empfehlungen sehr hilfreich: *„Wenn ich einen Projektantrag stelle, müsste es selbstverständlich sein, dass ich [die Datenarchivierung] mit beantragen kann, ich hätte aber ganz gerne realistische Empfehlungen über Zeitpläne, wie viel ich dafür beantragen kann, auf die ich mich berufen kann, dass ich reinschreiben kann, also bei der DFG, ich richte mich nach den Empfehlungen von so und so und für Daten dieses Typs ist so und so viel einzurechnen und dass das so kalkuliert ist, dass das aber auch locker reicht, dann in der bewilligten Zeit auch die Archivierung zu machen. Wenn ich momentan archivieren soll, ohne dass ich Ressourcen dafür bekomme, das ist ja das Problem"* (C33).

Drittens sollten die notwendigen Vorarbeiten (Aufbereitung, Dokumentation, Syntaxfiles etc.) nach einheitlichen Vorgaben strukturiert werden, so dass diese auch schon in Hinblick auf die Archivierung angelegt werden können. *„Es muss auch stärkere Hilfestellung und auch vielleicht ja Handreichungen oder Guidelines geben, wie so eine Dokumentationen zu machen ist, die aber dann hinreichend verständlich sind und auch nicht so kompliziert, dass man da wieder total abgeschreckt ist und das Gefühl hat, man muss da drei Monate jemanden abstellen"* (A37). Solche Guidelines müssten im Optimalfall zu Beginn der Projekte bereitgestellt werden: *„Da wäre es ganz gut, sagen wir mal, wenn man schon in der Anfangsphase irgendwie so was wie eine strukturierte Vorlage hätte, die aber gleichzeitig so ist, dass man möglichst wenig Arbeitsaufwand damit hat, weil das alles immer nebenher läuft"* (C5). Der Worst Case wäre indes, wenn man als Forscher/in am Ende eines Projekts plötzlich mit Anforderungen konfrontiert wird, auf man nicht vorbereitet war, da gerade zum Ende der Projektlaufzeiten der Zeitdruck häufig am höchsten ist (K95).

Viertens wird auch ein Bedarf an personeller Unterstützung bzw. Beratungsangeboten geäußert: *„So, und da könnte jemand das irgendwie ins Netz stellen und es müsste aber auch jemand da sein, der es ins Netz stellt und der auch was davon versteht von dem, was er ins Netz stellt, also der über irgendwelche Sachen z.B. auch stolpert und sagt, also da kann was nicht stimmen"* (B39). Das Spektrum gewünschter Angebote reicht von einer telefonischen Beratung durch den zentralen Anbieter über institutionelle Serviceangebote, rechtliche Beratungen bis zum Wunsch, die Daten einfach abzugeben und sich nicht mehr weiter darum kümmern zu müssen. Einigkeit besteht darin, dass es in jedem Fall Ansprechpartner für Nachfragen und konkrete Unterstützungsangebote geben sollte. Insbesondere Leitungspersonen weisen darauf hin, dass sie die Aufgaben der Datenarchivierung an Mitarbeiter/innen delegieren werden, welche sich dann mit ihren Nachfragen möglichst direkt bei einem Ansprechpartner des Datenarchivs melden sollten.

Neben diesen zentralen Anforderungen, die speziell auf die Minimierung des Arbeitsaufwands abzielen, wurde eine Reihe weiterer Einzelaspekte genannt. Diese sind nachfolgend stichpunkartig dokumentiert:

<u>Technische Anforderungen:</u>

- **Thematische Suche**: Gute thematische Suchoptionen (z.B. nach Forschungsthemen, Fachgebieten) sind ein entscheidendes Merkmal für eine attraktive Nachnutzung. Die archivierten Daten sollten daher mit thematischen Schlagworten versehen werden (K156-158, B10).
- **DOI-Vergabe:** Die Datensätze sollen mit persistenten Identifikatoren versehen werden um das Auffinden von Daten zu erleichtern und deren Zitation zu ermöglichen (C5, C17).
- **Container-Upload**: Aufgrund der u.U. hohen Anzahl der Datensätze in einem Projekt (genannt wurden 40 bis 50) sollte ein Upload der Datensätze in ZIP-Files oder vergleichbaren Formaten ermöglicht werden (D77).

- **Zugriffsrechte**: Der Zugriff auf die Daten sollte durch abgestufte Rechte steuerbar sein. Günstig wäre es, wenn für verschiedene Datensätze aus einem Projekt unterschiedliche Zugriffsrechte vergeben werden könnten: *„… dass z.B. Aufsätze oder Dokumentationen offen sind, bestimmte Daten vielleicht auch und andere dann aber wirklich nur mit Bewilligung des Primärforschers“* (C25).

Nicht-technische Anforderungen:

- **Zielgruppen**: Das Angebot sollte explizit bei der Zielgruppe des wissenschaftlichen Nachwuchses beworben werden (Abschlussarbeiten, Doktoranden, Verwendung in der Lehre). Zum einen wird hier ein hohes Potential für Sekundärdatennutzung gesehen, zum anderen sollte gerade dieser Gruppe die Philosophie der Datenarchivierung und des Data-Sharing vermittelt werden (Transferleistung!) (K169/178).
- **Peer-Review-Verfahren**: Als ein mögliches Nutzungsszenario wurde die Verwendung im Review-Prozess vorgeschlagen (G9). Reviewer könnten über das Repositorium auf Daten zugreifen, die der Publikation zu Grunde liegen und detaillierte Vorschläge in Bezug auf Auswertungsschritte und Interpretation der Ergebnisse unterbreiten.
- **Umfassende Dokumentationsinhalte:** Für die Nachnutzung von Forschungsdaten ist eine umfangreiche und qualitativ hochwertige Dokumentation zwingend erforderlich (K70-72). Die Inhalte sollten idealerweise einem dreistufiger Aufbau folgen: (Roh-)Daten und Datendokumentation, Skripte zur Auswertung sowie eine Übersicht der Forschungsergebnisse. Nur eine derart umfangreiche Archivierung und Dokumentation würde tatsächlich Replikationen ermöglichen (G9).
- **Versions-Management:** Im Moment gibt es diesbezüglich kein standardisiertes Vorgehen: *„Also da gibt es auch so Fragen wie Versionen-Management, weil da so ein Versionen-Salat entstehen kann, aber wir machen das alles so ein bisschen Learning by Doing, es könnte also auch viel besser sein. Also da gibt’s natürlich auch ein Interesse, wenn es da irgendwelche Best Practice gibt […], das wäre natürlich auch nicht schlecht“* (F34).
- **Recherche zu Forschungsstand und Methoden**: Ein Datenarchiv sollte nicht nur den Zweck verfolgen, Sekundärdaten bereitzustellen, sondern auch für Recherchen zu Forschungsstand und verwendeten Forschungsmethoden/-instrumenten nutzbar sein (K193). Damit wäre es auch *„für Leute [interessant], die gar nicht direkt die Daten nutzen wollen, sondern die halt selbst neue Erhebungen machen, aber so ein Datenarchiv auch nutzen können, um sich mal genauer anzugucken, wie haben denn Leute zu einer ähnlichen Fragestellung vor mir Daten erhoben, was gibt’s für Fragen, was für Variablen, welche Ausprägungen, die Frage der Operationalisierung und der Indikatoren, da kann man sich viele Anregungen verschaffen“* (K193-201).

- **Datenexpertise:** Als sehr nützlich werden Informationen dazu eingestuft, welche Forscher bereits mit welchen Datensätzen gearbeitet haben. Für mögliche Nachfragen stünden somit konkrete Ansprechpartner zur Verfügung (H33-37).
- **Unabhängige Infrastruktur**: Es wird ausdrücklich begrüßt, dass das SDN-Repositorium als unabhängige und nicht-ökonomische Infrastruktur aufgebaut wird, insbesondere im Vergleich zu kostenpflichtigen Verlagsangeboten (C5).
- **Policy-Ebene**: Auf der Policy-Ebene (Forschungsförderer, Institutspolicies) sollte der Druck auf die Forschenden erhöht werden, spätestens nach bestimmten Karenzzeiten ihre Forschungsdaten anderen Wissenschaftler/innen zugänglich machen (A47). Eine generelle Verpflichtung wird jedoch als unrealistisches Ziel angesehen. Vielmehr sei ein Mix aus konkreten Empfehlungen, kulturellem Wandel und der Bereitstellung der notwendigen Ressourcen und Infrastrukturen empfehlenswert.
- **Konzepte für Mixed Methods Forschung**: Wünschenswert wären Konzepte für die Archivierung von qualitativen und Mixed Methods-Daten (C39). Hier wird zukünftig ein steigender Bedarf erwartet.

4. Ausblick

Als ein zentrales Ergebnis der vorliegenden Kurzstudie kann festgehalten werden, dass die empirische Forschungspraxis in den Sozial- und Wirtschaftswissenschaften durch eine beachtliche Heterogenität gekennzeichnet ist. Zwar können aufgrund der geringen Fallzahl keine repräsentativen Aussagen über die Verbreitung einzelner Forschungsszenarien getroffen werden. Es vermittelt sich in den Interviews jedoch durchweg der Eindruck, dass die Wissenschaftler/innen überwiegend von verbreiteten Forschungsszenarien innerhalb der Sozial- und Wirtschaftswissenschaften berichten. Die unterschiedlichen Bedarfe, die sich hieraus mit Blick auf die Archivierung und Bereitstellung der Forschungsdaten ergeben, sind zum Teil von sehr grundsätzlicher Natur. Insbesondere die qualitative Forschung bringt, bereits was die bloße Archivierung der Daten betrifft, sehr spezielle Anforderungen mit sich, v.a. in Bezug auf die Anonymisierung der Daten. Jedoch konnten auch im Bereich der quantitativen Forschung je nach Erhebungskontext, Design und Methode spezifische Bedarfe identifiziert werden. Diese speisen sich größtenteils aus bestehenden Unklarheiten bzgl. der Nutzungsrechte, des Datenschutzes, der Rolle u. des Einflusses dritter Akteure oder der Komplexität der Datenstrukturen. In Einzelfällen können jedoch durchaus auch die Größe der Daten oder die Notwendigkeit ihrer fortlaufenden Aktualisierung spezielle Fragen aus Sicht der Forschenden aufwerfen.

Für die Weiterentwicklung des SDN-Angebots wäre es notwendig zu klären, inwieweit diesen Anforderungen aus der sozial- u. wirtschaftswissenschaftlichen Datenlandschaft in ihrer Gesamtheit entsprochen werden kann. Da die Entwicklung des Angebots nicht völlig losgelöst von der Aufgabe betrachtet werden sollte, überhaupt eine Praxis der Forschungsdaten-Archivierung und des Data-

Sharing zu etablieren, spricht vieles dafür, sich zunächst auf die einfacheren Anwendungsfälle, bspw. auf die quantitativen Survey-Formate zu fokussieren. Zumal sogar eine Interviewpartnerin anmerkte, dass ein „one-size-fits-all“ Ansatz Gefahr laufen könnte, für die jeweiligen Einzelszenarien nur eine bedingte Passung zu bieten (E35). Gleichwohl sollte ein fachlich ausgerichtetes Repositorium perspektivisch freilich ein möglichst großes Spektrum relevanter Forschungspraxen innerhalb des Feldes bedienen können.

Um diese Frage beantworten zu können, wäre in einem nächsten Schritt näher zu spezifizieren, welche Anforderungen durch das Repositorium selbst bearbeitet werden können und welchen Anforderungen eher durch praktische bzw. organisatorische Unterstützungsleistungen und die Schaffung entsprechender Rahmenbedingungen vor Ort zu begegnen ist. So sind etwa Fragen des Umgangs mit komplexen Designs und Datensets, dynamischen Datenbeständen oder größeren Datenmengen sicherlich klar von Seiten des Repositoriums so beantworten. Auch hinsichtlich möglicher Datenschutz- bzw. Anonymisierungsstandards sowie insbesondere der Unklarheiten bzgl. der Nutzungsrechte wären einheitliche Empfehlungen seitens des SDN-Angebots mit Sicherheit hilfreich. Die konkrete Bearbeitung der letztgenannten Punkte kann jedoch wohl nur am Einzelfall und durch eine Unterstützung an der der jeweiligen Institution erfolgen.

Die Hauptforderung der Forschenden nach einem möglichst geringen Arbeitsaufwand stellt hingegen sowohl eine technische als auch eine organisatorische Anforderung dar. Einerseits ist die einfache Bedienbarkeit, sowohl unter dem Aspekt der Usability als auch in Hinblick auf den zeitlichen Aufwand, seitens des Repositoriums zu garantieren. Ebenso können Orientierungshilfen in Form von Standards oder Guidelines zentral bereitgestellt werden. Andererseits sollten auf lokaler Ebene die notwenigen Ressourcen zur Verfügung gestellt und Hilfestellung bei der Aufbereitung und Einstellung der Daten angeboten werden. Diese Kombination zentraler technischer Infrastrukturleistungen und darauf abgestimmter Informations- und Unterstützungsangebote vor Ort sollte dazu beitragen können, die Bedenken hinsichtlich einer erhöhten Arbeitsbelastung abzubauen.

Mit Blick auf den weiteren Projektverlauf wurde in den geführten Interviews schließlich wiederholt deutlich, dass die Forschenden ihre Anforderungen bislang nur auf einer sehr dünnen Erfahrungsbasis formulieren können. So wurden bisher durch keine/n Interviewpartner/in selbstständig Forschungsdaten in einem Repositorium archiviert oder über ein vergleichbares Angebot bereitgestellt. Es ist somit davon auszugehen, dass weitere Anforderungen artikuliert werden, sobald erste praktische Erfahrungen mit der Nutzung des SDN-Infrastrukturangebots gemacht wurden. Für den Fortgang des Projekts bedeutet dies, dass konzeptionell zu überlegen wäre, wie Anforderungen der Nutzer/innen auch noch zu einem fortgeschrittenen Projektzeitpunkt berücksichtigt werden können.

5. Anhang A

Interviewleitfaden Experteninterviews SowiDataNet

Fragekomplexe: Einstiegsfragen und vertiefende Nachfragen		Fragehintergrund
I. Eigene empirische Arbeit		

Bitte schildern Sie doch kurz, an welchen empir. Forschungsprojekten Sie und/oder Ihre Mitarbeiter bzw. Doktoranden aktuell arbeiten oder zuletzt gearbeitet haben.
 	
• Uns interessiert zunächst, wie die konkrete Forschungsarbeit aussieht:
  • Werden eigene Befragungen durchgeführt? Umfrageinstitut? Experimente durchgeführt? Bestehende Daten neue aufbereitet?
  • Um was für FD handelt es sich? (klassische quantitativ „Rechteckfiles", qualitative Formate? Sonstiges)
  • Werden die FD dokumentiert? Wie genau (Erhebungsmethoden, Aufbereitung, Codebook, Methodenbericht)? Für wen & zu welchem Zweck (intern – extern)? Wie wird das gemacht? Ist eine Person speziell dafür zuständig?
  • Wie groß ist der Kreis der Personen, der mit den Daten arbeitet (eher größeres Team vs. Einzelforscher)?
 	
• Und was passierte mit den FD nach Abschluss eines Projekts?
  • Werden sie aufbewahrt? Wie? Werden sie weiter genutzt? Durch wen?
  • Besteht generell ein Bedarf, die Daten langfristig zu archivieren? (möglich: Verweis auf DFG-Empfehlung und EU 2020)
		Empirische Arbeit
		
		Verwendete Daten
		
		Art der Dokumentation
		
		Verbleib der Daten

II. Suchen und Finden		
Arbeiten Sie selbst auch mit den Daten anderer Forscher/innen? • Wo/Wie haben Sie nach diesen Daten gesucht (Statistische Bundesamt, GESIS Datenbestandskatalog, Institut, Kollegen)? • Wie haben Sie diese Daten erhalten (Konditionen)? • War es einfach an die Daten heranzukommen? Ergaben sich dabei Probleme? • Nutzen Sie auch Fragebögen/Erhebungsinstrumente aus anderen Untersuchungen? Wenn ja, wie waren diese auffindbar?		*Wege zu den Daten* *Fragebögen*
III. Veröffentlichung der eigenen FD		
Stellen Sie eigene Forschungsdaten anderen Forscher/innen zur Verfügung?		*Bedarf/Interesse an FD-Infrastruktur* *Konditionen*
Ja: Wie kann man sich das vorstellen? Wer hat Zugang? Zu welchen Bedingungen?	**Nein:** /	
Können sie sich vorstellen ihre Daten auf einer größeren Online-Plattform für FD aus den Sozial- u. Wirtschaftswissenschaften einzustellen?		
Ja: Zu welchen Konditionen (z.B. frei zugänglich – nur nach Rücksprache)? Gibt es bestimmte Anforderungen von denen Sie sagen würden, die müssten dafür auf jeden Fall erfüllt sein?	**Nein:** Warum nicht? Welche Bedenken bestehen diesbezüglich?	