RatSWD Working Paper Series

www.ratswd.de

RatSWD
Rat für Sozial- und WirtschaftsDaten

184

Datenmanagement und Data Sharing: Erfahrungen in den Sozial- und Wirtschaftswissenschaften

Denis Huschka, Claudia Oellers, Notburga Ott und Gert G. Wagner

August 2011

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD)

Die *RatSWD Working Papers* Reihe startete Ende 2007. Seit 2009 werden in dieser Publikationsreihe nur noch konzeptionelle und historische Arbeiten, die sich mit der Gestaltung der statistischen Infrastruktur und der Forschungsinfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften beschäftigen, publiziert. Dies sind insbesondere Papiere zur Gestaltung der Amtlichen Statistik, der Ressortforschung und der akademisch getragenen Forschungsinfrastruktur sowie Beiträge, die Arbeit des RatSWD selbst betreffend. Auch Papiere, die sich auf die oben genannten Bereiche außerhalb Deutschlands und auf supranationale Aspekte beziehen, sind besonders willkommen.

RatSWD Working Papers sind nicht-exklusiv, d. h. einer Veröffentlichung an anderen Orten steht nichts im Wege. Alle Arbeiten können und sollen auch in fachlich, institutionell und örtlich spezialisierten Reihen erscheinen. Die *RatSWD Working Papers* können nicht über den Buchhandel, sondern nur online über den RatSWD bezogen werden.

Um nicht deutsch sprechenden Nutzer/innen die Arbeit mit der neuen Reihe zu erleichtern, sind auf den englischen Internetseiten der *RatSWD Working Papers* nur die englischsprachigen Papers zu finden, auf den deutschen Seiten werden alle Nummern der Reihe chronologisch geordnet aufgelistet.

Einige ursprünglich in der *RatSWD Working Papers* Reihe erschienenen empirischen Forschungsarbeiten sind ab 2009 in der RatSWD Research Notes Reihe zu finden.

Die Inhalte der *RatSWD Working Papers* stellen ausdrücklich die Meinung der jeweiligen Autor/innen dar und nicht die des RatSWD.

Kontakt: Rat für Sozial- und Wirtschaftsdaten | Mohrenstraße 58 | 10117 Berlin | office@ratswd.de

Datenmanagement und Data Sharing

Erfahrungen in den Sozial- und Wirtschaftswissenschaften

Denis Huschka[1], Claudia Oellers[2], Notburga Ott[3] und Gert G. Wagner[4]

erscheint in: St. Büttner, H.-C. Hobohm und L. Müller (Hg.), Grundlagen des Forschungsdatenmanagements, Bonn 2011

[1] Geschäftsführer des Rates für Sozial- und Wirtschaftsdaten
[2] Wissenschaftliche Mitarbeiterin des Rates für Sozial- und Wirtschaftsdaten
[3] Professorin für Sozialpolitik und öffentliche Wirtschaft an der Ruhr-Universität Bochum und stellvertretende Vorsitzende des Rates für Sozial- und Wirtschaftsdaten
[4] Vorsitzender des Vorstands des Deutschen Instituts für Wirtschaftsforschung (DIW Berlin); Professor für Empirische Wirtschaftsforschung und Wirtschaftspolitik an der Technischen Universität Berlin und Vorsitzender des Rates für Sozial- und Wirtschaftsdaten

Die Menge der für Forschungszwecke zur Verfügung stehenden Daten vergrößert sich beständig (King, 2011). Jedoch werden unter Daten in den verschiedenen wissenschaftlichen Disziplinen ganz unterschiedliche Dinge gefasst. Aus dem lateinischen kommend bezeichnet ein Datum zunächst einmal etwas „Gegebenes". In den Geowissenschaften können Daten Eisbohrkerne sein, aber auch numerische Geokoordinaten. In den Geschichtswissenschaften können Daten das Format alter Dokumente haben. In der Medizin können es auch biologische Proben oder Laborwerte sein. In den quantitativ empirisch arbeitenden Sozial-, Verhaltens- und Wirtschaftswissenschaften ist das „gängige" Format der einschlägigen Daten das von Zahlen als Teil von Datenmatrizen oder Tabellen.

Die unterschiedlichen Phänotypien von Forschungsdaten erfordern spezifische Datenmanagementstrategien. Oft beschreiben die Daten die Ausprägung einer Eigenschaft eines Individuums oder einer Organisation, wie z. B. einer Firma. In diesen, insbesondere in der Medizin, den Sozial-, Verhaltens- und Wirtschaftswissenschaften vorkommenden, Fällen spricht man von personenbeziehbaren oder firmenbeziehbaren Daten, bei deren Be- und Verarbeitung sich automatisch Fragen des Datenschutzes und der Forschungsethik stellen. Auch dies hat Auswirkungen auf das Forschungsdatenmanagement und die Zugänglichkeit dieser Art von Daten.

Obgleich im Bereich der Sozial-, Verhaltens- und Wirtschaftswissenschaften in Deutschland datenschutzrechtliche Notwendigkeiten die gemeinsame Nutzung (sozusagen das Teilen von Daten – „*data sharing*") erschweren, nimmt Deutschland eine Vorreiterrolle hinsichtlich des Auf- und Ausbaus einer sozial- und wirtschaftswissenschaftlichen Forschungsdateninfrastruktur ein (vgl. Solga & Wagner, 2007; Habich et al., 2010; Bender et al., 2008). Der Zugang zu einschlägigen Daten hat sich in den vergangenen Jahren für die Wissenschaft deutlich verbessert. Neben den klassischen Datenarchiven (z. B. dem GESIS Datenarchiv für Sozialwissenschaften – vormals Zentralarchiv für empirische Sozialforschung an der Universität Köln) sind alle vom Rat für Sozial- und Wirtschaftsdaten akkreditierten Forschungsdatenzentren (FDZ) und Datenservicezentren (DSZ) Teil dieser Forschungsinfrastruktur. Die FDZ und DSZ als institutionalisierte Orte des *data sharing*, ermöglichen nicht nur den Zugang zu Daten, sondern bieten darüber hinaus einen Service um die Daten herum an. Ein solcher Service ist wegen der komplexen Strukturen vieler Datensätze, und der jeweils beschränkten Aussage-

kraft der Daten (Reichweite, Validität und Reliabilität), welche durch die Operationalisierungen der Erhebungen bedingt sind, nötig und kann am besten von denen geleistet werden, die die Daten produzieren. In den Verhaltenswissenschaften ist eine solche Tradition des *data sharings* noch wenig ausgeprägt. Dies beginnt sich zu ändern (vgl. Weichselgartner, 2011).

So positiv die Entwicklungen hin zu mehr Datenverfügbarkeit im Bereich der Sozial- und Wirtschaftswissenschaften und zuletzt auch in den Verhaltenswissenschaften zu bewerten sind, so aktuell ist aber auch die Frage, wie man die Daten im Rahmen einer geordneten und transparenten Infrastruktur zur Verfügung stellen kann, und wie man den Zugang selbst transparent und nutzerfreundlich regelt.
Für innovative Forschung wird es zunehmend wichtiger multi- und interdisziplinär zu arbeiten. Georeferenzierte Daten, Biomarker, Transaktionsdaten oder auch Datensätze privater Firmen stellen relativ neue und besonders reizvolle Datenquellen dar, durch deren Verknüpfung mit „herkömmlichen" sozialwissenschaftlichen Daten sich innovative Fragestellungen beantworten lassen. Auch die digitale Verfügbarkeit von Daten sowie die technologischen Möglichkeiten im Umgang mit den digitalen Daten (z. B. durch persistente Identifikatoren und verbesserte Computertechnik und -leistungsfähigkeit) sind aus Sicht der Wissenschaft Chance und Herausforderung an ein systematisches Datenmanagement zugleich. Eine besondere Bedeutung wird in Zukunft deshalb der Organisation der Informationen über die Daten zukommen, also die Beschreibung der Inhalte, Qualität, Analysepotenziale, Aussagekraft und insbesondere über Verknüpfungsmöglichkeiten zwischen Datensätzen. Es reicht also nicht jeden einzelnen Datensatz verfügbar zu machen. Für eine breite Nutzung in der Wissenschaft ist ein „Informationsportal" notwendig, in welchem ein an einem bestimmten Thema interessierter Forscher alle erforderlichen Informationen über alle relevanten zur Verfügung stehenden Datensätze finden kann. Wohlgemerkt: ein solches Portal soll und kann nicht die Daten selbst vorhalten, dies ist wie wir unten ausführen, aus rechtlichen Gründen nicht möglich und aus Servicegründen auch gar nicht wünschenswert. Ein solches Portal sollte lediglich die nicht zu unterschätzende Funktion eines Informationsbrokers übernehmen.

1. Data sharing

In den Sozial- und Wirtschaftswissenschaften hat sich in den vergangenen Jahren eine Kultur des Teilens von Daten (*data sharing*) durchgesetzt. Teilen ist deswegen leicht möglich, da die mehrfache Nutzung der Daten diese nicht zerstört (wie das z.B. bei Biomaterial oder Bohrkernen der Fall ist). Das systematische Argument für *data sharing* ist, dass nur die Möglichkeit von Re-Analysen veröffentlichter Ergebnisse diese zu wissenschaftlichen Erkenntnissen macht. Denn Wissenschaft bedeutet, dass Ergebnisse nachprüfbar sind. Hinzu kommt die praktische Überlegung, dass Daten, welche im Rahmen öffentlicher, beispielsweise durch Forschungsförderung finanzierter Unterfangen entstehen, für die breite Forschung zur Verfügung gestellt werden *sollen* und nicht durch einen einzelnen Forscher monopolisiert werden dürfen (der ggf. nur Re-Analysen zur Prüfung von Ergebnissen erlaubt).

Die Überprüfbarkeit von Forschungsergebnissen durch Re-Analysen gehört zu den formalisierten Kriterien guter wissenschaftlicher Praxis, die von der Deutschen Forschungsgemeinschaft (DFG, 1998) erarbeitet wurden. Inzwischen wird beispielsweise in der Ökonomie vermehrt einer von wissenschaftlichen Zeitschriften gestellten Anforderung entsprochen, neben der eigentlichen Publikation auch die zugrundeliegenden Datensätze zu veröffentlichen bzw. im Falle von datenschutzrechtlich sensiblen Daten in geschützten Bereichen zugänglich zu machen.

Die Ermöglichung einer Nachnutzung der Daten durch deren Übermittlung an geeignete Datenarchive oder andere Orte ist seit langem Bestandteil der Förderrichtlinien der Deutschen Forschungsgemeinschaft (DFG, 2010) und der entsprechenden Förderprogramme des Bundesministeriums für Bildung und Forschung (BMBF). Die konsequente Umsetzung dieser Verpflichtung ist freilich in den verschiedenen wissenschaftlichen Disziplinen unterschiedlich.

Öffentlich finanziert entstehen Daten auch im Rahmen der Politiksteuerung und durch die amtliche Statistik (vgl. Hahlen, 2009) und im Rahmen der Verwaltung als sog. prozessproduzierte Datensätze wie beispielsweise die Daten der Bundesagentur für Arbeit oder der Sozialversicherungen. Auch in diesen Bereichen hat sich inzwischen eine Kultur des *data sharing* durchgesetzt. Viele Ressortforschungseinrichtungen und die Statistischen Ämter verfügen heute über

Forschungsdatenzentren, welche den Zugang zu den jeweiligen Daten ermöglichen. Diese Entwicklungen wurden maßgeblich durch den Rat für Sozial- und Wirtschaftsdaten (RatSWD) angestoßen, dessen Arbeit inzwischen als Modell für weitere Wissenschaftsbereiche dient (vgl. Kommission Zukunft der Informationsinfrastruktur, 2011; Wissenschaftsrat, 2011).

Ein weiteres Argument für *data sharing* basiert auf der Erkenntnis der Datenproduzenten, dass eine Sekundärnutzung von Daten wissenschaftliche Vorteile bringt. *Data sharing* ermöglicht wissenschaftlich wertvolle Rückkopplungsprozesse, so dass die Datenproduzenten die Qualität ihrer Daten und die Effektivität ihrer Datenerhebungen und –analysen erhöhen können, wenn sie in intensivem Austausch mit der Forschung stehen. Aber auch die Forschungsergebnisse der Datenproduzenten werden durch eine intensive externe Auswertung bekannter und damit auch deren Reputation.

Damit Forschungsdaten im Rahmen einer Sekundärnutzung richtig verwendet werden können, ist eine gute Dokumentation der Daten Voraussetzung. Diese Arbeit am Datensatz erfolgt bislang in der Regel ohne entsprechende Würdigung durch die *Scientific Community*, also die Gemeinschaft aller Forschenden. Dadurch ist es gerade für Spitzenforscher relativ unattraktiv, Zeit und Energie in die Erhebung von qualitativ hochwertigen Daten, deren Dokumentation und Nachnutzung zu investieren. Datensätze werden i. d. R. nicht im Literaturverzeichnis von Veröffentlichungen zitiert und entsprechend erntet der Datenproduzent keine Zitate. Aber Zitate sind die Währung, mit der Wissenschaftlerinnen und Wissenschaftler innerhalb der *Scientific Community* entlohnt werden. Eine Verbesserung der „Belohnungsstrukturen“ für diese Arbeiten trüge somit zu einer Verbesserung der Datenverfügbarkeit bei. Durch die Kennung eines Datensatzes mit einem persistenten Identifikator (zum Beispiel in Form eines *Digital Object Identifiers* (DOI)) in Verbindung mit einer Autorenidentifikation könnte die wissenschaftliche Arbeit an der Produktion eines Datensatzes kenntlich und zitierfähig gemacht werden (vgl. GESIS, 2011).

Trotz aller Fortschritte im Bereich des *data sharings* besteht weiterhin eine deutliche Diskrepanz zwischen der Forderung nach einem freien Zugang insbesondere zu öffentlich finanzierten Daten auf der einen Seite, sowie Vorbehalten und Unsi-

cherheiten die eigenen Daten zu teilen auf der anderen Seite. Aus Studien weiß man, dass die Gründe, warum Daten – und dies trifft v.a. auf Daten aus kleineren wissenschaftlichen Erhebungen zu – nicht zur Weiternutzung bereitgestellt werden, vielfältig sind: Sie reichen von banaler Ressourcenknappheit – eine ordentliche Dokumentation der Daten erfordert zeitliche und personelle Ressourcen – bis hin zu Unsicherheiten über die Frage, wem die Daten eigentlich als Eigentümer gehören und der daraus resultierenden nicht geklärten Verantwortlichkeit (vgl.: PARSE insight, 2010; Feijen, 2011).

Es sind also neben rechtlichen Fragen vor allem Bemühungen nötig, um das Weitergeben von Daten inklusive einer notwendigen Dokumentation der Daten so einfach und ressourcensparend wie möglich zu gestalten. Auf der technischen Ebene gibt es hier seit langem entsprechende Entwicklungen: die Data *Documentation Alliance* bemüht sich um einen internationalen Standard bei der Beschreibung (Dokumentation) von Daten der Sozial-, Verhaltens- und Wirtschaftsforschung (vgl.: DDI Alliance, 2009). Inzwischen sehen sich Datenarchive wie die GESIS zunehmend als Dienstleister und bieten umfangreiche Serviceleistungen und Hilfestellungen.

Neben ressourcenökonomischen Überlegungen können aber auch forschungsökonomische Überlegungen ausschlaggebend für die zu beobachtende Zurückhaltung mancher Forscher und mancher Disziplinen beim *data sharing* sein, beispielsweise die Befürchtung, dass sich eine Veröffentlichung des Datensatzes nachteilig auf die eigene wissenschaftliche Karriere auswirken kann. Piwowar et al. (2007) konnten jedoch unlängst in einer Studie nachweisen, dass das Teilen von Daten mit höheren Zitationsraten verbunden ist.

Ein oft vorgebrachtes Argument gegen *data sharing* ist das des Datenschutzes. Personenbeziehbare Daten (aber auch Daten der Wirtschaftsforschung, welche Branchen- oder Firmengeheimnisse beinhalten), die im Rahmen von wissenschaftlichen Erhebungen und Interviews oder auch klinischen Studien erhoben werden, sind in den meisten Fällen datenrechtlich sensitiv. Hier gilt es die Daten selbst und deren Weitergabe (technisch) so zu organisieren, das allen Datenschutz- und Persönlichkeitsschutzaspekten in perfekter Weise Rechnung getragen wird. Datenschutz ist jedoch niemals ein grundsätzliches Argument gegen das *data sharing*.

Um den in Anfängen bereits begonnenen Paradigmenwandel im Bereich *data sharing* erfolgreich weiterzubefördern, ist ein Dialog zwischen Wissenschaft, Wissenschaftsförderern, Datenschützern und wissenschaftlichen Verlagen notwendig. Die Aufgabe der Forschungsförderer wird es dabei sein, mehr als bisher auf die Erstellung und Umsetzung von Datenmanagement- und Datenverwertungsplänen als Bestandteil ihrer Förderpolitik zu achten (vgl. Winkler-Nees, 2011). Ein solcher Dialog sollte in geeigneter Weise durch Gremien wie den RatSWD koordiniert werden, welche sich auch der besonderen Aufgabe der Bündelung der Interessen der Wissenschaft gegenüber Datenproduzenten und Politik widmen sollten. Weitere Herausforderungen bestehen in der Etablierung und Weiterentwicklung einer Kultur des *data sharing,* beispielsweise durch die Schaffung von Anreizsystemen zur Würdigung der Arbeit an Datensätzen. Neue Arten von Daten (beispielsweise. Biomarker oder Geomarker) und deren Verknüpfbarkeit mit herkömmlichen Surveydaten stellen den Datenschutz vor immer neue Herausforderungen.

2. Data Access

Sozial-, verhaltens- und wirtschaftswissenschaftliche Daten weisen oft Charakteristika auf, die rechtliche und forschungsethische Überlegungen notwendig machen. Weiterhin sind sie aufgrund ihrer in vielen Fällen komplexen Strukturen schwierig zu handhaben. Beide Aspekte erfordern eine besondere Organisation des Datenzugangs, d. h. der Forschungsdateninfrastruktur.

Rechtliche Aspekte

Die bereits angedeutete Komplexität der rechtlichen und forschungsethischen Fragen, welche – mit gutem Grund – den Zugang zu sensiblen Daten, insbesondere im Bereich der Wirtschafts- und Sozialwissenschaften, einschränken, macht Überlegungen darüber notwendig, wie der Zugang zu Daten in gleichzeitig effizienter, aber rechtlich und forschungsethisch einwandfreier Form organisiert werden kann. Im Prinzip gilt: je gehaltvoller die Daten, desto interessanter sind sie für die Wissenschaft, aber desto sensibler sind sie auch. Hinlänglich anonymisierte – d. h. zusammengefasste und vergröberte Daten bieten einen umfangreichen

Datenschutz – jedoch zunehmend begrenzte Auswertbarkeit. Für viele Fragestellungen sind aggregierte Daten oder Individualdaten in anonymisierter Form völlig ausreichend. Solche Daten werden bereits heute als *Public Use Files* oder für die universitäre Ausbildung als sogenannte CAMPUS[1] Files durch viele öffentliche Datenproduzenten angeboten. Andere Fragestellungen verlangen jedoch nach Individualdaten, die zusätzlich mit weiteren Merkmalen, beispielsweise über das Wohnumfeld der Befragten oder Daten aus biologischen Proben der Befragten verknüpft werden. Hierdurch steigt das Deanonymisierungsrisiko und ethische Erwägungen müssen angestellt werden.

Wenngleich es hier keine generelle Lösung geben kann, bietet sich ein kontinuierlicher Austausch der Datenproduzenten über jeweilige technische Neuerungen und rechtliche Entwicklungen an. Generell gilt, dass der Daten- und Persönlichkeitsschutz durch die Anwendung entsprechender Vorkehrungen strikt und umfassend entlang der Gesetze eingehalten werden muss, dies jedoch niemals ein Argument dafür sein kann Daten nicht zugänglich zu machen. Allerdings erschweren diese Besonderheiten die Umsetzung eines einfachen Zugangs zu den Daten, die durch angepasste technische und infrastrukturelle Lösungen, d. h. durch ein intelligentes Datenmanagement, überwunden werden können. Viele Produzenten sensibler Daten, besonders jene der amtlichen Statistik und der Ressortforschung, können ihre Daten nicht in herkömmliche Archive geben und so einen Zugang für die Forschung ermöglichen. Die praktikable Lösung ist das Angebot eigener Zugangswege, deren Konformität mit den jeweiligen Gesetzen kontinuierlich geprüft und gewährleistet werden kann.

Komplexitätsaspekte

Ein Charakteristikum sozial-, verhaltens- und wirtschaftswissenschaftlicher Daten ist deren Vielfältigkeit und deren oft hypothesenbezogene Entstehung. Die Spannweite reicht von einfachen Tabellen, in denen Makrodaten als Zahlenkolonnen dargestellt werden, über Interviewtranskripte und daraus gewonnenen qualitativen Daten, bis hin zu komplizierten Längsschnittdatensätzen, die aus sich fortlaufend verändernden und erweiternden Datenbanken bestehen, in denen mehrere Tausend Einzelitems für mehrere Tausend Personen über die Zeit verknüpf-

1 http://www.forschungsdatenzentrum.de/campus-file.asp [Zugriff am 10.08.2011].

bar gespeichert sind. Voraussetzung für die Nutzung verschiedener Datensätze sind nicht nur Investitionen in eine adäquate Statistik- und Methodenausbildung und ein „Erlernen“ des Umgangs mit den Besonderheiten (insbesondere der Messkonzepte) eines bestimmten Datensatzes auf Seiten der Nutzer, sondern vor allem auch ein geeignetes Serviceangebot von Seiten der Datenproduzenten. Dieser Service kann nur sehr begrenzt durch die „herkömmlichen“ Datenarchive geleistet werden, auch hier sind alternative Lösungen gefragt, da Forschungsdaten oft nur mit Hilfe von Zusatzwissen (Metadaten) sinnvoll interpretierbar sind.

Beispielsweise werden Messverfahren und Skalen auf der Basis von Annahmen entwickelt, in der Hoffnung, sie mögen messen, was beabsichtigt ist. Selbst scheinbar eindeutige Daten, wie die des Haushaltseinkommens sind komplexe Konstrukte: So macht es einen Unterschied, ob man neben den Gehältern der Haushaltsmitglieder auch Einkünfte durch Mieten oder Kapitalerträge zum Haushaltseinkommen hinzuzählt. Auch den zur Schätzung fehlender Angaben verwendeten Imputationsverfahren liegen komplexe Annahmen zu Grunde. Neben einer zu liefernden möglichst standardisierten, aber die Daten vollständig beschreibenden Dokumentation besteht oftmals ein Bedarf an intensiver fachlicher Beratung der Sekundärnutzer. Diese Beratungsleistung kann jedoch in der Regel nur durch die Datenproduzenten selbst, und nicht von Archiven oder Bibliotheken geleistet werden.

Vor diesem Hintergrund einer sehr komplexen und mit unterschiedlichen Anforderungen an Datenschutz und Service zu charakterisierenden Datenlandschaft haben sich in den Sozial-, Verhaltens- und Wirtschaftswissenschaften verschiedene Akteure und Modelle etabliert, welche den Zugang zu Daten ermöglichen und ein den jeweiligen Bedürfnissen entsprechendes Niveau an Service bieten.

3. Modell I: Datenzugang über disziplinenspezifische oder themenspezifische zentrale Datenarchive

Archive, in denen in der Regel disziplinen- oder themenspezifische Datensätze gesammelt werden, stellen für Wissenschaftler oftmals eine erste Anlaufstelle bei der Suche nach geeigneten Daten für ihr jeweiliges Forschungsvorhaben dar. Hier können sie Unterstützung bei Recherche und Datenzugang sowie gelegentlich auch bei der Analyse der Daten (Methodenfragen) erhalten.

Auf der anderen Seite stellen Datenarchive für die Datenproduzenten eine komfortable Möglichkeit dar, ihre Daten sichtbar, auffindbar und somit für die wissenschaftliche Nachnutzung verfügbar zu machen. Hierzu gehört der Zugang zu den eigentlichen Forschungsdaten wie zu den dazugehörenden Dokumentationen, den sog. Metadaten (Informationen über Daten). Durch entsprechende Nutzerverträge können darüber hinaus basale datenschutzrechtliche Aspekte bei der Weitergabe Berücksichtigung finden.

Aufgabe von Archiven ist es, eine technologisch adäquate und nutzerorientierte Bereitstellung und Archivierung der Daten zu ermöglichen. Da die Archive aber nicht die Produzenten der Daten sind, ist eine diesbezügliche Zusammenarbeit mit den Datenproduzenten notwendig, welche für die Qualität der Daten verantwortlich zeichnen. Archive sollten durch fachliche Beratung und Unterstützungsleistungen bei der teilweise sehr anspruchsvollen und zeitintensiven Dokumentation und Aufbereitung der Daten, bei der oftmals auch Fragen der Anonymisierung eine zentrale Rolle spielen, aktive Partner der Datenproduzenten sein. Eine weitere Serviceleistung der Archive sollte in der Organisation und Sicherstellung der eineindeutigen Zitierfähigkeit inklusive der Verknüpfung mit den „Autoren“ der Daten bestehen.

Neben (informations-)fachlichen Expertisen und Serviceangeboten verfügen Archive über die technologischen Möglichkeiten der (Langzeit-)Archivierung von Datensätzen, d. h. der Sicherstellung der physischen Existenz und Verfügbarkeit der Daten über lange Zeiträume (vgl. Kap. 3.1). So komfortabel und leistungsfähig die elektronische Datenverarbeitung ist, so unhinterfragt und gefährlich ist sie auch: CDs, DVDs und Festplatten sind sehr anfällig für Fehler und Zerstörung. Während historisch genutzte Hollerithsysteme mit Lochkarten teilweise auch heute noch rekonstruierbar sind, reicht ein Kratzer, ein Computercrash oder ein Computervirus um Datenbestände u. U. unwiederbringlich zu vernichten. Die Langzeitarchivierung ist eine in seiner Wichtigkeit unterschätzte Aufgabe, die von Archiven am besten erbracht werden kann.

Zusammenfassung: Modell Datenarchiv

Systematik: Der Datenproduzent gibt seine Daten und deren Dokumentation in standardisierter Form an ein Archiv weiter.

Vorteile: Das Archiv kümmert sich um Zugang, Distribution, Vertragsangelegenheiten, Langzeitverfügbarkeit der Daten und bietet den Sekundärforschern bei der Auswertung der Daten einen basalen Service um die Daten herum. Dieses Modell ist insbesondere geeignet für im Rahmen von Forschungsprojekten entstandene Datensätze, in denen Wissenschaftler zeitlich begrenzt als Datenproduzenten fungieren, und durch die Archivierung deren dauerhafte Verfügbarkeit sichergestellt wird.

Nachteile: Ein Archiv kann den Service um die Daten herum nur im begrenzten Maße leisten – in der Regel können inhaltliche Fragen nicht beantwortet werden. Es erfolgt bislang faktisch keine systematische Sammlung und Verknüpfung von bereits mit denselben Daten gefertigten Analysen und Papieren. In dieser Frage sollte die Zusammenarbeit mit den Forschungsbibliotheken und Verlagen angeregt und intensiviert werden. Archive können hier koordinierend fungieren. Ein weiterer Nachteil besteht darin, dass datenrechtlich hoch sensible Daten nicht ohne weiteres in allgemeinen Archiven gespeichert und verarbeitet werden dürfen. Herausforderungen: Durch die Entwicklung und Verbesserung der Standards bei der Weitergabe von Daten und deren Beschreibung durch Metadaten verbessert sich die Zugänglichkeit und die Benutzerfreundlichkeit der Daten. Die dauerhafte Herstellung eines Links zwischen Datenproduzenten, Bibliotheken und Verlagen schafft die Voraussetzungen für eine adäquate Würdigung der Arbeit an den Daten und die umfangreiche Bereitstellung von Analysen mit den Daten.

4. Modell II: Zugang zu den Daten und Serviceleistungen durch Forschungsdatenzentren[2]

Eine zweite – in jüngerer Vergangenheit erfolgreich implementierte Variante des Datenzugangs – besteht im Angebot der Forschungsdatenzentren. Dieses Modell scheint sich insbesondere für potente Datenproduzenten zu bewähren und etabliert zu haben, die dauerhaft Daten zur Verfügung stellen (z. B. statistische Ämter) und/oder besonders komplizierte Datensatzstrukturen anbieten (z. B. prospektive Längsschnitterhebungen) und deshalb eine enge Verbindung zwischen Datenpro-

2 Die Aufgabenfelder von Forschungsdatenzentren und Datenservicezentren lassen sich heute, auf der Basis der gemachten Erfahrungen nicht mehr eindeutig trennen. Im Folgenden beziehen wir uns Forschungsdatenzentren und Datenservicezentren gleichermaßen, ohne letztere immer zu nennen. Der Begriff Forschungsdatenzentrum scheint sich auch international durchzusetzen.

duzent und Datennutzer wünschenswert ist. Auch die Einhaltung des Datenschutzes kann in „eigenen“ FDZ durch die Datenproduzenten oft einfacher gewährleistet werden.

In den Sozial-, Verhaltens- und Wirtschaftswissenschaften haben sich ausgehend von einer Empfehlung der Kommission zur Verbesserung der Informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) aus dem Jahr 2001 in den Folgejahren die ersten vier Forschungsdatenzentren und zwei Datenservicezentren gegründet (das Forschungsdatenzentrum des Statistischen Bundesamtes, das Forschungsdatenzentrum der Statistischen Ämter der Länder, das Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung und das Forschungsdatenzentrum der Rentenversicherung, das Servicezentrum für Mikrodaten des Leibniz-Instituts für Sozialwissenschaften (GESIS/MISSY), das Internationale Datenservicezentrum des Forschungsinstituts zur Zukunft der Arbeit (IZA)). Ziel dieser Datenzentren war und ist es, die jeweiligen amtlichen Daten einer wissenschaftlichen Verwendung zur Verfügung zu stellen. Dies war bis dahin aufgrund der Vorgaben des Bundesdatenschutzgesetzes, des Statistikgesetzes und Sozialgesetzbuches bezüglich der zum großen Teil personenbeziehbaren Daten nicht ohne weiteres möglich. In der Zwischenzeit sind neben den genannten sechs Datenzentren eine ganze Reihe weiterer Forschungsdatenzentren hinzugekommen, die über den Rat für Sozial- und Wirtschaftsdaten akkreditiert und organisiert werden (http://www.ratswd.de/dat/fdz.php). Derzeit (Stand Sommer 2011) gibt es 19 vom RatSWD akkreditierte Datenzentren. Auch Daten, die für eine wissenschaftliche Nachnutzung anfänglich nur schwer zugänglich waren, wie es zum Beispiel im Bereich der Bildungsdaten der Fall war, konnten auf diese Weise erschlossen werden.

Anders als bei Datenarchiven ist zentrales Merkmal der Forschungsdatenzentren der wissenschaftlich unterstützende inhaltliche Service um die Daten herum, der nur erbringbar ist, weil die das FDZ betreibenden Datenproduzenten in der Regel die besten Experten im Umgang mit den eigenen Daten sind. Ein zentraler Aspekt der Akkreditierungsrichtlinien des RatSWD für FDZ und DSZ ist, dass in diesen wissenschaftlich gearbeitet wird und somit der Service für externe Wissenschaftler von Wissenschaftlern geleistet wird.

Obwohl die Forschungsdatenzentren über einen heterogenen Hintergrund verfügen, lässt sich mittlerweile berechtigt von einer gemeinsamen Forschungsda-

teninfrastruktur sprechen, welche unter dem Dach des RatSWD koordiniert wird. Das Akkreditierungsmodell des RatSWD bietet dabei eine Qualitätssicherung der prozeduralen Mechanismen. Die Koordination findet u.a. in der Festlegung gemeinsamer Kriterien und Standards als Antwort auf gemeinsame rechtliche und organisatorische Voraussetzungen, welche das Modell Datenarchiv ausschließen, ihren Ausdruck. Auch die Weiterentwicklung von Verfahren des *on-site* und des gesicherten Fernrechnens, um sensible Daten unter strikter Einhaltung von datenschutzrechtlichen Vorgaben zur Verfügung zu stellen, oder auch die Erstellung von Skalenhandbüchern um Vergleichbarkeit und Verknüpfbarkeit von Daten darzustellen und zu ermöglichen, sind aktuelle Felder der Zusammenarbeit.

Zusammenfassung: Modell Forschungsdatenzentren

Für Datenproduzenten, die aufgrund der Komplexität, der Menge und/oder der Datenschutzsensibilität ihre Daten nicht über Archive zur Verfügung stellen, findet sich im Modell des Forschungsdatenzentrums eine Möglichkeit, ihre Daten systematisch und unter Einhaltung aller rechtlichen Bestimmungen für die Forschung zu öffnen. Die Daten bleiben beim Datenproduzenten, er hat jederzeit die volle Kontrolle und kann so darüber wachen, dass alle Restriktionen jederzeit eingehalten werden. Der Nutzer der Daten hat direkten Kontakt zu Fachkollegen beim Datenproduzenten und erhält konkrete und kompetente Hilfe bei der Auswertung der Daten. Der Datenproduzent bleibt dadurch mit den Entwicklungen der Wissenschaft verbunden und kann durch eine formalisierte Rückkopplung mit außenstehenden Nutzern die Qualität der Daten, die Messmechanismen, Datenerhebungen und Aufbereitungen kontinuierlich verbessern. Auch hat der Datenproduzent in der Regel ein Interesse daran, Publikationen und Analysen zu sammeln, die auf den eigenen Daten beruhen. Somit können themen- und datenzentrierte Wissensdatenbanken entstehen.

Nachteile: Für die Datenproduzenten ist die Einrichtung von Forschungsdatenzentren v.a. in der Einführungsphase ressourcen- und kostenintensiv. Auch ist das Datenangebot in den Datenzentren in der Regel auf die „eigenen“ Datensätze begrenzt, was zu einer dezentralen Verfügbarkeit von Datensätzen – u. U. sogar zum selben Forschungsgegenstand – führt. Es gibt faktisch keinen zentralen Anlaufpunkt oder Ansprechpartner. Derzeit stellen sich deshalb die Zugangswege,

Dokumentationen und Verknüpfungsmöglichkeiten der Daten etwas unübersichtlich dar.

Herausforderungen: Im Feld der FDZ muss durch mehr Koordination, Transparenz und Abstimmung eine Verbesserung des Nutzerservices erreicht werden. Der RatSWD versucht dies durch die Schaffung einer Austauschplattform der Forschungsdateninfrastruktur zu befördern. Auch die Schaffung eines gemeinsamen Portals als „Tor zur gesamten Datenwelt“ einer Disziplin inklusive der Verknüpfungen mit angrenzenden Disziplinen (beispielsweise Sozialdaten mit Biodaten und Geodaten) ist im Gespräch.

5. Zusammenfassung und zukünftige Entwicklungen

In den Sozial- und Wirtschaftswissenschaften hat sich in den vergangenen Jahren eine Kultur des Teilens von Daten (*data sharing*) durchgesetzt. Das heißt, es sind zunehmend interessante Daten für Forschungszwecke verfügbar; die Herausforderung besteht heute in der Organisation dieser Datenwelt. Archive und Datenzentren fungieren als etablierte Orte des Datenzugangs und werden den unterschiedlichen Anforderungen an Datenschutz und der Erbringung von Serviceleistungen um die Daten herum gerecht. Zusammen bilden sie eine funktionierende Forschungsinfrastruktur, die durchaus einen Modellcharakter aufweist.

Die Etablierung eines Portals, das Nutzern und insbesondere potentiellen Nutzern einen Überblick über und einfache Zugangsmöglichkeiten zu sozial-, verhaltens- und wirtschaftswissenschaftlichen Forschungsdaten (einschließlich der Daten der amtlichen Statistik) anbietet, ist ein naheliegender nächster Schritt beim Ausbau der Forschungsinfrastruktur für die die Sozial-, Verhaltens- und Wirtschaftswissenschaften in Deutschland. Zugleich sollte ein solches Portal die Zitation von Datenquellen und ihren Produzenten befördern.

Wie ein solches Portal gestaltet werden sollte, sollte zügig von den verschiedenen Stakeholdern im Bereich der Archivierung diskutiert werden, also Fachbibliotheken, Archiven und Forschungsdatenzentren. Zu klären sind Fragen der langfristigen und sicheren Archivierung sowie des laufenden Services, der für vielgenutzte und noch im Wachsen begriffene Datensätze notwendig ist, um die Nutzung zu unterstützen. Diskutiert werden sollte auch, wie in diesem Zusammenhang die

Anerkennung der „Produktion“ von Forschungsdaten als wissenschaftliche Leistung durch Referenzierbarkeit/ Zitierbarkeit und persistente Identifikatoren für Daten, Datenproduzenten und Forscher verbessert werden kann. Denn nur wenn die Produktion von Forschungsdaten als wissenschaftliche Leistung voll anerkannt wird, wird ihre Qualität und Verfügbarkeit steigen.

Literatur

Bender, S. Himmelreicher, R. Zühlke, S. & Zwick, M., 2009. *Improvement of Access to Data from the Official Statistics.* (RatSWD Working Paper Nr. 118). Online: http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_118.pdf [Zugriff am 09.08.2011].

DFG (Deutsche Forschungsgemeinschaft), 1998. *Sicherung guter wissenschaftlicher Praxis, Denkschrift. Empfehlungen der Kommission „Selbstkontrolle der Wissenschaft“*. Weinheim: Wiley-VCH. Online: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf [Zugriff am 09.08.2011].

DFG (Deutsche Forschungsgemeinschaft), 2010. *Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für Antragstellung und ergänzenden Leitfaden für die Antragstellung von Projekten mit Verwertungspotenzial, für die Antragstellung von Projekten im Rahmen einer Kooperation mit Entwicklungsländern.* (DFG Vordruck 1.02-8/10). Online: http://www.dfg.de/download/programme/emmy_noether_programm/antragstellung/1_02/1_02.pdf [Zugriff am 09.08.2011].

DDA (Data Documentation Alliance), 2009. *What is DDI?* Online: http://www.ddialliance.org/what [Zugriff am 09.08.2011].

Feijen, M., 2011. *What Researchers want*. Utrecht: SURF Foundation (February 2011). Online: http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf [Zugriff am 09.08.2011].

GESIS Leibniz-Institut für Sozialwissenschaften, da|ra Registrierungsagentur für sozialwissenschaftliche Daten, 2011. *Über da|ra*. Online: http://www.gesis.org/dara/home/ueber-dara/ [Zugriff am 09.08.2011].

Habich, R. Himmelreicher, R. K. & Huschka, D., 2010. *Zur Entwicklung der Dateninfrastruktur in Deutschlan*d. (RatSWD Working Paper Nr. 157). Online: http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_157.pdf [Zugriff am 09.08.2011].

Hahlen, J., 2009. Zur Rolle der amtlichen Statistik für eine evidenzbasierte Wirtschaftsforschung und –politik. In: *Wirtschaft und Statistik*. Wiesbaden: Statistisches Bundesamt, S. 1021-1030. Online: http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Querschnittsveroeffentlichungen/WirtschaftStatistik/Gastbeitraege/Wirtschaftsforschung102009,property=file.pdf [Zugriff am 09.08.2011].

King, G., 2011. Ensuring the Data Rich Future of the Social Sciences. *Science*, 331, S. 719-721.

Kommission Zukunft der Informationsinfrastruktur, 2011. *Gesamtkonzept für die Informationsinfrastruktur in Deutschland, Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder*. Online: http://www.leibniz-gemeinschaft.de/?nid=infrastr [Zugriff am 09.08.2011].

PARSE.Insight, 2010. *Insight into digital preservation of research output in Europe*. Online: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf [Zugriff am 09.08.2011].

Piwowar, H. A. Day, R. S. & Fridsma, D. B , 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*, 2(3): e308.

Solga, H. & Wagner, G. G., 2007. *Eine moderne Dateninfrastruktur für eine exzellente Forschung und Politikberatung – Bericht über die Arbeit des Rates für Sozial- und Wirtschaftsdaten in seiner ersten Berufungsperiode (2004-2006).* (RatSWD Working Paper Nr. 1). Online: http://www.ratswd.de/download/RatSWD_WP_2007/RatSWD_WP_01.pdf [Zugriff am 09.08.2011].

Weichselgartner, E., 2011. *Disziplinenspezifische Aspekte des Archivierens von Forschungsdaten am Beispiel der Psychologie*. (RatSWD Working Paper Nr. 179). Online: http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_179.pdf [Zugriff am 12.07.2011].

Winkler-Nees, S., 2011. *Anforderungen an wissenschaftliche Informationsinfrastrukturen.* (RatSWD Working Paper Nr. 180). Online: http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_180.pdf [Zugriff am 09.08.2011].

Wissenschaftsrat, 2011. *Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften.* (Drs. 10465-11). Berlin: Wissenschaftsrat (28.01.2011). Online: http://www.wissenschaftsrat.de/download/archiv/10464-11.pdf [Zugriff am 09.08.2011].