Bibliothekartag 2011

Warum Bibliotheken – im Prinzip – ideale Datenarchive sind*

Gert G. Wagner
DIW Berlin (Deutsches Institut für Wirtschaftsforschung)
Rat für Sozial- und Wirtschaftsdaten (RatSWD)

Neben klassischen Veröffentlichungen (Aufsätze in Fachzeitschriften und Bücher) spielt der Zugang zu "Forschungsdaten" in nahezu allen wissenschaftlichen Disziplinen eine zunehmende Rolle. Dabei versteht man unter Daten in den verschiedenen wissenschaftlichen Disziplinen ganz unterschiedliche Dinge. 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.

Sogar für die Geisteswissenschaften werden „Daten", über die traditionelle Literatur hinaus, bedeutsam. Auf Basis einer bereits immens großen Datenbasis von digitalisierten Texten können statistische Analyseverfahren sehr gut eingesetzt werden. Erste Analysen von über fünf Millionen Büchern, die etwa 4 Prozent aller jemals weltweit gedruckten Bücher reprä-

*Große Teile dieses Artikels sind dem Beitrag „Datenmanagement und Data Sharing: Erfahrungen in den Sozial- und Wirtschaftswissenschaften" entnommen (von Denis Huschka, Claudia Oellers, Notburga Ott und Gert G. Wagner, erscheint in: Handbuch Forschungsdatenmanagement, hrsg. von Stephan Büttner, Hans-Christoph Hobohm und Lars Müller, Bad Honnef 2011: Bock+Herchen).

sentieren, wurden sogar bereits in einer Spitzen-Fachzeitschrift wie *Science* abgedruckt (Jean-Baptiste Michel et al., Quantitative Analysis of Culture Using Millions of Digitized Books, Science, Vol. 331, 2011, S. 176-182).

Und für innovative Forschung wird es in vielen Bereichen zunehmend wichtiger, multi- und interdisziplinär zu arbeiten. Georeferenzierte Daten, Biomarker (einschießlich Gensequenzen), Transaktionsdaten (z. B. von Telefongesellschaften) oder auch Datensätze privater Firmen stellen relativ neue und besonders reizvolle Datenquellen dar, durch deren Verknüpfung mit „herkömmlichen“ Daten sich zum Beispiel in den Sozial-, Verhaltens- und Wirtschaftswissenschaften innovative Fragestellungen beantworten lassen (vgl. Rainer Schnell, Biological Variables in Social Surveys; Julia Lane, Administrative Transaction Data; und Bernhard Engel, Transaction Data: Commercial Transaction Surveys and Test Market Data. Alle in: Rat für Sozial- und Wirtschaftsdaten (Hg.), Building on Progress - Expanding the Research Infrastructure for the Social, Economic and Behavioral Sciences, Budrich UniPress: Opladen & Farmington Hills, MI, 2011, S. 367-412).

Da die Nutzung verschiedenster Datenquellen eine immer größere Rolle spielt, wachsen Bibliotheken nahezu automatisch neue Aufgaben im Bereich des Datenmanagementns, der Datendokumentation und des „Ausleihens“ von Forschungsdaten zu. Ob die Bibliotheken diese Aufgaben erfolgreich bewältigen können, hängt von den Bibliotheken selbst ab, aber natürlich auch von den „Produzenten“ von Forschungsdaten. Dieser Beitrag versucht über das Aufgabenfeld und mögliche Lösungsansätze einen kurzen Überblick zu geben. Der Beitrag ist von einem „Daten-Pratiker“ aus Sicht der Sozial- und Wirtschaftswissenschaften geschrieben (vgl. auch Roland Habich, Ralf K. Himmelreicher und Denis Huschka, Zur Entwicklung der Dateninfrastruktur in Deutschland, RatSWD Working Paper Nr. 157, Berlin 2010, http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_157.pdf). Der Artikel nimmt jedoch in Anspruch auch für andere Disziplinen et-

was Sinnvolles sagen zu können. Die im Vortragstitel getroffene Feststellung wird am Ende begründet.

Data sharing

Die Überprüfbarkeit von Forschungsergebnissen durch Re-Analysen gehört zu den formalisierten Kriterien guter wissenschaftlicher Praxis, die von der Deutschen Forschungsgemeinschaft 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 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 und im Rahmen der Verwaltung als sog. prozessproduzierte Datensätze wie beispielsweise die Daten der Bundesagentur für Arbeit oder der Sozialversicherungen. 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.

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 Da-

ten 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.

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 Unsicherheiten die eigenen Daten zu teilen auf der anderen Seite. Aus Befragungen weiß man, dass die Gründe, warum Daten – und dies trifft v.a. auf Daten aus kleineren wissenschaftlichen Erhebungen zu (Erich Weichselgartner, Disziplinenspezifische Aspekte des Archivierens von Forschungsdaten am Beispiel der Psychologie, RatSWD Working Paper Nr. 179, Berlin 2011: http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_179.pdf) – 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. z. B. Martin Feijen**,** What Researchers Want, SURF Foundation , February 2011: http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf).

Es sind also neben der Klärung rechtlicher 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 (DDI) bemüht sich um einen internationalen Standard bei der Beschreibung (Dokumentation) von Daten der Sozial-, Verhaltens- und Wirtschaftsforschung.

Neben ressourcenökonomischen Überlegungen dürften aber vor allem

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. Roger S. Day Piwowar und Douglas B Fridsma (Sharing Detailed Research Data Is Associated with Increased Citation Rate, PLoS ONE 2: 3, 2007, e308) 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, dass 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.

Zugang zu Daten und Serviceleistungen durch Forschungsdatenzentren

In einigen Disziplinen werden zentrale Datenarchive aufgebaut, die i.d.R. aber nicht an Bibliotheken angesiedelt sind. Obwohl Bibliotheken Spezialisten in der Langzeitarchivierung, der Dokumentation (einschließlich der

„Versionierung“ ihrer Bestände) und beim Nutzerservice sind. Für die bislang eher marginale Rolle der Bibliotheken dürfte neben der Tradition von Bibliotheken (die auf geschriebene bzw. gedruckte Werke spezialisiert sind) eine Rolle spielen, dass viele Datensätze „leben“, d. h. ständig verändert werden. Etwa weil neue Beobachtungsergebnisse hinzukommen oder weil ältere Daten durch neue Aufbereitungstechniken verbessert werden.

Datenarchive haben sich neben den Bibliotheken entwickelt; nicht zuletzt, weil viele Datensätze sich ständig verändern und diese Veränderungen nur von den Datenproduzenten selbst – nicht jedoch von Bibliothekaren – durchgeführt werden können. Eine in jüngerer Vergangenheit erfolgreich implementierte neue Variante des Datenzugangs besteht im Angebot der sozial-, verhaltens- und wirtschaftswissenschaftlichen Forschungsdatenzentren (http://www.ratswd.de/dat/fdz.php). 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 Datenproduzent und Datennutzer wünschenswert ist. Auch die Einhaltung des Datenschutzes kann im „eigenen“ Forschungs-Daten-Zentrum (FDZ) durch die Datenproduzenten oft einfacher gewährleistet werden. Anfang 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 ist, dass in diesen wissenschaftlich gearbeitet wird und somit der Service für externe Wissenschaftler von Wissenschaftlern ge-

leistet wird.

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.

Schlußfolgerungen

Im Feld der Forschungsdatenzentren – innerhalb der Sozial-, Verhaltens- und Wirtschaftswisenschaften, aber auch darüber hinaus – sollte durch mehr Koordination, Transparenz und Abstimmung eine Verbesserung des Nutzerservices erreicht werden. Auch die Schaffung eines gemeinsamen Portals als „Tor zur gesamten Datenwelt“ einer Disziplin inklusive der Verknüpfungen mit angrenzenden Disziplinen sollte unbedingt geprüpft werden. Da Bibliotheken Spezialisten für Langzeitarchivierung, Dokumentation und Nutzerservice sind, sollten sie hier systematisch einbezogen werden. Und zwar auf der Ebene zentraler Fachbibliotheken wie auch auf der Ebene von Institutsbibliotheken, nämlich der Institute, die Forschungsdaten erzeugen und vorhalten.

Die Etablierung fachspezifischer Portale und darüber hinaus fachübergreifender Meta-Portale, die Nutzern und insbesondere potentiellen Nutzern einen Überblick über und einfache Zugangsmöglichkeiten zu Forschungsdaten (einschließlich der Daten der amtlichen Statistik) anbietet, ist eine naheliegende Aufgabe von zentralen Fachbibliotheken. 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.

Inwieweit „tote“ Datensätze, die sich also nicht mehr verändern, von Bibliotheken archiviert werden sollten oder von spezialisierten Datenarchiven (wie etwa dem der GESIS für sozialwissenschaftliche Datensätze) ist eine offene Frage, die sinnvollerweise auch von Disziplin zu Disziplin unterschiedlich beantwortet werden sollte. Wo es bereits etablierte und funktionierende zentrale Datenarchive gibt, ist eine Migration hin zu Bibliotheken wenig naheliegend. Hier bietet sich schlicht und einfach eine bessere Zusammenarbeit an.

„Lebende“ Datensätze, die sich ständig verändern, sollten bei den Datenproduzenten selbst archiviert, dokumentiert und zugänglich gemacht werden. Die Zusammenarbeit mit Bibliotheken kann auf zwei Ebenen geschehen. Zum ersten sollten die Meta-Informationen über solche dezenteral in „Forschungsdatenzentren“ archivierten Daten über die Daten-Portale von zentralen Fach-Bibliotheken zu finden sein. Zum zweiten sollten die Institutsbibliotheken der Datenproduzenten stärker als bislang mit den dezentralen Forschungsdatenzentren zusammenarbeiten. Wie gesagt: Bibliotheken sind Spezialisten für Langzeitarchivierung, Dokumentation und Nutzerservive. Diese Fähigkeiten sollten dezentral organisierte Forschungsdatenzentren für ihre Zwecke nutzen, indem Institutsbibliotheken und Forschungsdatenzentren systematisch zusammenarbeiten.