

Sozialwissenschaft

NATASCHA SCHUMANN, OLIVER WATTELER

Forschungsdaten in den Sozialwissenschaften

Die Sozialwissenschaften bestehen aus verschiedenen Fachdisziplinen, die sich mit gesellschaftlichen Fragestellungen befassen. Zur Untersuchung von sozialer Wirklichkeit werden Daten benötigt. Diese werden in unterschiedlicher Art und Weise erhoben und in vielen Fällen für weitere Forschungsvorhaben zur Verfügung gestellt. Damit die Daten für andere von Nutzen sind, müssen sie entsprechend dokumentiert, aufbereitet und archiviert werden. Dies geschieht in Forschungsdatenzentren, Datenservicezentren ebenso wie in Datenarchiven. Bei sozialwissenschaftlichen Erhebungen werden regelmäßig Daten von Individuen erhoben, das bedeutet, dass der Datenschutz eine große Rolle spielt. Fragen bezüglich Datenmanagement und Langzeitarchivierung gewinnen auch in den Sozialwissenschaften zunehmend an Bedeutung.

Social science is comprised of various specialized disciplines which deal with inquiries into societal issues. Investigations of social reality require data. These are acquired in a variety of ways and means and frequently made available to other research projects. In order for data to be made accessible to others, they must be appropriately documented, processed and archived. This is done in research data centers, data service centers, and also in data archives. Since social research surveys often collect data from individuals, data protection is a significant concern. Issues related to data management and digital preservation are hence becoming more important in social science.

DATEN IN DEN SOZIALWISSENSCHAFTEN

Als Sozialwissenschaften kann man jene Wissenschaften sehen, »die die menschliche Gesellschaft, gesellschaftliche Gruppen, einzelne Individuen in ihrer Beziehung zu anderen oder Einrichtungen und Institutionen von Gesellschaften sowie materielle und kulturelle ›Güter‹ als Ausdruck des Zusammenlebens von Menschen zum Gegenstand haben.«¹ Zu ihnen zählen die Soziologie, die Politikwissenschaften, die Wirtschaftswissenschaften, sowie die Sozial- und Wirtschaftspsychologie. Allen gemeinsam ist die Untersuchung von Ausschnitten gesellschaftlicher Wirklichkeit.

Ein wichtiges Mittel zur Untersuchung sozialer Wirklichkeit sind Daten. Prinzipiell ist der Begriff der Daten eng mit dem Konzept der Messung verbunden, also der Zuordnung von Zahlen zu Objekten oder Ereignissen.² Obwohl es keine einheitliche Definition des Begriffes in den Sozialwissenschaften gibt, können Daten jedoch als »eine Abbildung des untersuchten Ausschnittes der gesellschaftlichen Wirklichkeit«³ bestimmt werden.

Ein allgemein bekanntes Beispiel für Daten ist die volkswirtschaftliche Gesamtrechnung mit dem allgemein bekannten Indikator Bruttoinlandsprodukt (BIP). Dieses betrug 2012 für die Bundesrepublik rund 2,6 Billionen Euro (Angabe: Statistisches Bundesamt) und besteht aus Größen, die im Rahmen der Gesamt-

rechnung eine »umfassende, systematische quantitative Beschreibung [...] einer Volkswirtschaft« ermöglichen. In diese Gesamtrechnung gehen Angaben über Vermögen, Arbeitnehmer, Arbeitsstunden, Zahlungsbilanz etc. ein.⁴ Kurz gesagt, ist das Bruttoinlandsprodukt ein Maß für die Wirtschaftsleistung einer Volkswirtschaft. Erhoben werden diese Daten von der größten Einrichtung der so genannten Amtlichen Statistik, dem Statistischen Bundesamt.

An dieser Stelle interessieren uns nun zwei Dinge, die eng mit der Vorstellung von sozialwissenschaftlichen Daten zusammenhängen:

1. Das BIP ist eine aus einzelnen Messungen zusammengesetzte Größe, die als Indikator für volkswirtschaftliche Leistung genutzt wird. Diese volkswirtschaftliche Leistung kann jedoch selbst nicht direkt beobachtet werden. Man spricht daher von einem Konzept.⁵ Andere, stärker soziologische Konzepte sind z. B. Umweltbewusstsein, Integration oder Identität.
2. In das BIP geht u. a. die gesamte Leistung deutscher Arbeitnehmerinnen und Arbeitnehmer ein. Die dem BIP zugrunde liegende volkswirtschaftliche Gesamtrechnung besagt jedoch nichts über die Gerechtigkeit der Einkommensverteilung, die Arbeitsbedingungen oder den Anteil prekärer Beschäftigungsverhältnisse. Diese sind jedoch für die Sozialwissenschaften interessante Phänomene, da sie etwas über die Entstehung der Wirtschaftsleistung aussagen. So wuchs in den vergangenen Jahren bei guter Wirtschaftsleistung die Zahl sogenannter atypischer Beschäftigungsverhältnisse, etwa Beschäftigungen in Teilzeit oder mit Befristungen. Diese Beschäftigungsverhältnisse können allerdings soziale Risiken wie Verarmung bei gleichzeitiger Beschäftigung mit sich bringen, was wiederum eine Herausforderung für den Sozialstaat darstellt.⁶

Gute wirtschaftliche Leistung und ein Anstieg atypischer Beschäftigungsverhältnisse scheinen also zusammenzuhängen. Hier könnte man nun sozialwissenschaftliche Fragestellungen anschließen: Warum steigt die Zahl atypischer Beschäftigungsverhältnisse? Hat die steigende Zahl atypischer Beschäftigter einen positiven Einfluss auf die Wirtschaftsleistung? Welche sozialstaatlichen Maßnahmen flankieren diese Veränderung des Arbeitsmarktes? Welche Auswirkungen



Natascha Schumann

Foto: privat



Oliver Watteler

Foto: privat

**Datenerhebung zur
Analyse sozialer
Wirklichkeit**

haben die sozialen Risiken auf Gesundheit und Wohlbefinden der betroffenen Beschäftigten? Derartige Fragen bilden die Basis sozialwissenschaftlicher Forschung und Analysen auch mit Hilfe von Daten. Nicht alle Fragen können jedoch mit Daten der Amtlichen Statistik beantwortet werden. Daher wird im Folgenden erörtert, mit welchen Daten in den Sozialwissenschaften gearbeitet wird und welche Einrichtungen sich um ihre Erhebung und Bereitstellung kümmern.

DATENARTEN UND -FORMATE

Die in den Sozialwissenschaften verwendeten Daten hängen eng mit der Methode zusammen, die zur Untersuchung sozialer Phänomene verwendet wird. Diekmann spricht von einem »Werkzeugkasten«, zu dem er standardisierte Umfragen, Beobachtungen, Experimente und Inhaltsanalysen zählt.⁷ Die Anwendung dieser Methoden hängt u. a. mit dem Ziel der Untersuchung zusammen. Ist ein Bereich noch eher wenig erforscht, wird man explorativ vorgehen, um zu Hypothesen, also widerspruchsfreien Annahmen über den zu untersuchenden Ausschnitt gesellschaftlicher Wirklichkeit zu gelangen. Diese können etwa über Beobachtungen gewonnen und mit einer größer angelegten Befragung überprüft werden. Oft werden in den Sozialwissenschaften nach dem sogenannten Forschungsdesign, also der Anlage des Forschungsvorhabens, qualitative und quantitative Daten unterschieden.⁸

Die qualitativ arbeitende Sozialforschung »hat den Anspruch, Lebenswelten »von innen heraus« aus der Sicht der handelnden Menschen zu beschreiben. Damit will sie zu einem besseren Verständnis sozialer Wirklichkeit(en) beitragen und auf Abläufe, Deutungsmuster und Strukturmerkmale aufmerksam machen«⁹. Die qualitative Sozialforschung arbeitet in der Regel mit Transkripten, also Verschriftlichungen offener Interviews. Untersuchungspersonen wird in dieser Art des Interviews Raum zu freien Antworten geboten, und das Gespräch wird nur durch Leitfragen strukturiert. Die Texte können anschließend interpretiert oder mithilfe entsprechender Programme statistisch ausgewertet werden (z. B. MaxQDA). Basis der Transkripte sind oft Audio- und Videoaufnahmen, die auch digital vorliegen können. Sie stellen neben den Textdateien der Transkripte weitere Datenarten dar, mit denen sich u. a. Datenarchive befassen.¹⁰

Der Begriff der quantitativen Daten bezieht sich regelmäßig auf Forschungen, in denen Messungen und statistische Analyseverfahren zum Einsatz kommen. Anders als bei eher qualitativen Ansätzen werden hier Theorien in Anlehnung an naturwissenschaftliche Methoden getestet. D. h. man würde z. B.

das oben erwähnte Konzept »atypisches Beschäftigungsverhältnis« genauer beschreiben und operationalisieren, sprich ihm beobachtbare Indikatoren zuordnen (Dauer des Verhältnisses, Arbeitszeit, Anzahl der Beschäftigungsverhältnisse pro Arbeitnehmer, etc.). Solche Indikatoren können etwa mithilfe einer Umfrage erfasst werden. Quantitative Daten liegen regelmäßig als Datentabellen für statistische Analyseprogramme wie SPSS, SAS oder Stata vor.¹¹ Grundlegende Daten, die als Fakten betrachtet werden können, liefert häufig die Amtliche Statistik, also neben dem Statistischen Bundesamt die Statistischen Landesämter oder das Institut für Arbeitsmarkt- und Berufsforschung (IAB). Zusammenhänge zwischen solchen Fakten und Präferenzen oder Einstellungen der Befragten werden dann in sozialwissenschaftlichen Forschungsprojekten untersucht. Watteler nennt als Beispiele für das Zusammenspiel von amtlichen Daten und sozialwissenschaftlichen Forschungsdaten u. a. die Bekanntgabe des amtlichen Endergebnisses einer Bundestagswahl durch den Bundeswahlleiter und die Bearbeitung z. B. von Fragen der Nichtwahl durch die empirische Wahlforschung oder die Darstellung der Häufigkeit von Scheidungen durch die Scheidungsstatistik und die Analyse der Gründe für solche Trennungen durch die Familiensoziologie.¹²

INFRASTRUKTUR

Im Vergleich zu manch anderen Disziplinen besteht in den Sozialwissenschaften eine relativ lange Tradition der Datenaufbereitung und Datenarchivierung. Bereits kurz nach dem Zweiten Weltkrieg entstanden die ersten Einrichtungen, die mit der Sammlung sozialwissenschaftlicher Daten begannen. In den USA wurde 1947 das Roper Center for Public Opinion Research¹³ gegründet, das Daten ab den 1930er Jahren archiviert. Ab den 1960er Jahren wurden weitere Einrichtungen für sozialwissenschaftliche Daten gegründet. In Deutschland entstand 1960 das Kölner Zentralarchiv für Empirische Sozialforschung (heute das GESIS Datenarchiv für Sozialwissenschaften¹⁴), in den USA im Jahr 1962 das Interuniversity Consortium for Political and Social Research (ICPSR)¹⁵ in Ann Arbor, das Steinmetz Archiv in den Niederlanden (heute aufgegangen in Data Archiving and Network Service – DANS)¹⁶ im Jahr 1964 sowie das UK Data Archive (UKDA)¹⁷ in Colchester 1967. Inzwischen bestehen allein in Europa 23 sozialwissenschaftliche Datenarchive, die sich im Council of European Social Science Data Archives (CESSDA)¹⁸ zusammengeschlossen haben. Auf internationaler Ebene agiert die International Federation of Data Organizations for Social Science (IFDO)¹⁹ als Zusammenschluss der Archive.

Untersuchungsmethoden

qualitative und
quantitative Daten

Einrichtungen für sozial-
wissenschaftliche Daten

Auch in der Bundesrepublik Deutschland sind in den vergangenen Jahren im Bereich sozialwissenschaftlicher Forschungsdaten einige Initiativen ergriffen worden. Die vom Bundesministerium für Bildung und Forschung (BMBF) 2001 eingerichtete »Kommission zur Verbesserung der informationellen Infrastruktur« hat als eine von mehreren Empfehlungen²⁰ die Einrichtung eines Rates für Sozial- und Wirtschaftsdaten (RatSWD)²¹ formuliert. Dieser wurde 2004 eingerichtet und wird vom BMBF finanziert. Die Zielsetzung besteht in der nachhaltigen Verbesserung der Forschungsdateninfrastruktur für die empirische Forschung und damit einhergehend auch der Steigerung der internationalen Wettbewerbsfähigkeit.

Der RatSWD versteht sich als Schnittstelle zwischen Datenproduzenten und Datennutzern und hat u. a. den Auftrag, Aktivitäten zu koordinieren und Begutachtungs- und Akkreditierungsverfahren durchzuführen. In ihm sind Wissenschaftlerinnen und Wissenschaftler aus unterschiedlichen Einrichtungen vertreten. Im »Ständigen Ausschuss Forschungsdateninfrastruktur«, in dem alle Datenzentren vertreten sind, werden Themen der Datenweitergabe und des Datenmanagements diskutiert und der fachliche Austausch gefördert.

Eine wesentliche Aufgabe des RatSWD besteht in der Akkreditierung von Forschungsdaten- und Datenservicezentren (FDZ und DSZ). Ziel dieser Zentren, die in der Regel auf einen thematischen, bzw. daten-

typischen Schwerpunkt spezialisiert sind, ist ein vereinfachter und unkomplizierter Zugang zu Daten bei gleichzeitig hoher fachlicher Qualität in Bezug auf die Aufbereitung der Daten und Beratung der Nutzerinnen und Nutzer. Die Akkreditierung erfolgt nach definierten Kriterien und Mindeststandards. Die Bereitstellung der Daten erfolgt als faktisch anonymisierte »Scientific Use Files« oder als absolut anonymisierte »Public Use Files« (Definition s. u.).

Neben den FDZs bestehen weitere datenproduzierende wissenschaftliche Einrichtungen: Zu nennen sind an Universitäten verankerte Projekte und Sonderforschungsbereiche,²² die sozialwissenschaftlichen Max-Planck-Institute, z. B. für Gesellschaftsforschung in Köln, für Bildungsforschung in Berlin oder für Demographie in Rostock, sowie die sozial- und wirtschaftswissenschaftlichen Institute der Leibniz-Gemeinschaft, wie das Mannheimer Zentrum für Europäische Sozialforschung (MZES), das Deutsche Jugendinstitut (DJI) in München, das Deutsche Institut für Wirtschaftsforschung (DIW) und das Wissenschaftszentrum Berlin (WZB), beide Berlin, oder das Leibniz-Institut für Sozialwissenschaften GESIS.

Bund und Länder finanzieren neben den genannten außeruniversitären Einrichtungen auch die sogenannte Ressortforschung. Einrichtungen der Ressortforschung sind in der Regel Bundesministerien angeschlossen oder nachgeordnet und dienen als Ratgeber für politische Entscheidungen. Bekannte Bundesämter

Rat für Sozial- und Wirtschaftsdaten

Forschungsdaten- und Datenservicezentren

- Forschungsdatenzentrum des Statistischen Bundesamtes (FDZ-Bund)
- Forschungsdatenzentrum der Statistischen Ämter der Länder (FDZ-Länder)
- Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB)
- Forschungsdatenzentrum der Rentenversicherung (FDZ-RV)
- Forschungsdatenzentrum im Bundesinstitut für Berufsbildung (BIBB-FDZ)
- Forschungsdatenzentrum am Institut zur Qualitätsentwicklung im Bildungswesen (IQB)
- Forschungsdatenzentrum des Sozio-ökonomischen Panels (FDZ-SOEP)
- Forschungsdatenzentrum ALLBUS bei GESIS
- Forschungsdatenzentrum Internationale Umfrageprogramme bei GESIS
- Forschungsdatenzentrum Wahlen bei GESIS
- Forschungsdatenzentrum des Survey of Health, Ageing and Retirement in Europe (SHARE)
- Forschungsdatenzentrum des Deutschen Zentrums für Altersfragen (FDZ-DZA)
- Forschungsdatenzentrum PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID)
- Forschungsdatenzentrum des Beziehungs- und Familienpanels (FDZ pairfam)
- Forschungsdatenzentrum Ruhr am RWI (FDZ-Ruhr)
- LMU-ifo Economics & Business Data Center (EBDC)
- Forschungsdatenzentrum "Gesundheitsmonitoring" am Robert Koch-Institut (RKI)
- Forschungsdatenzentrum German Microdata Lab (GML)
- Internationales Datenservicezentrum des Forschungsinstituts zur Zukunft der Arbeit (IZA)
- Datenservicezentrum Betriebs- und Organisations-Daten (DSZ-BO)
- Forschungsdatenzentrum (FDZ) Bildung am Deutschen Institut für Internationale Pädagogische Forschung (DIPF)
- Forschungsdatenzentrum der Bundeszentrale für gesundheitliche Aufklärung (BZgA)
- Forschungsdatenzentrum Wissenschaftsstatistik
- ZEW-Forschungsdatenzentrum (ZEW-FDZ)
- National Education Panel Study (NEPS)

Abb. 1: Liste der vom RatSWD akkreditierten Forschungsdatenzentren

Vernetzung

sind etwa das Statistische Bundesamt (Destatis), das Bundesamt für Bauwesen und Raumordnung (BBR), das Bundesamt für Migration und Flüchtlinge (BAMF) oder das Umweltbundesamt (UBA). Im Auftrag des Bundes sind ferner u. a. tätig:

- das Bundesinstitut für Berufsbildung (BIBB),
- das Bundesinstitut für Bevölkerungsforschung (BiB) – heute ein Teil des Statistischen Bundesamtes,
- das Institut für Arbeitsmarkt- und Berufsforschung (IAB) als Forschungseinrichtung der Bundesagentur für Arbeit (BA),
- das Robert-Koch-Institut (RKI), Berlin.

Mit den Entwicklungen der letzten Jahre wurden wichtige Schritte getan in Richtung einer nationalen Infrastruktur für sozialwissenschaftliche Daten. Weiterhin besteht allerdings Bedarf, die Rollen und Verantwortlichkeiten zu diskutieren und gegebenenfalls anzupassen, zum Beispiel in Bezug auf die langfristige Sicherung der Daten. Grundsätzlich wäre eine noch stärkere Vernetzung wünschenswert, damit die Nutzerinnen und Nutzer ohne besondere Kenntnisse der Infrastrukturen Zugang zu den gesuchten Daten erhalten. Die Etablierung einer Datenregistrierungsagentur da|ra²³ für sozialwissenschaftliche Forschungsdaten durch GESIS und die Zentralbibliothek für Wirtschaftswissenschaften (ZBW) hat genau dieses Ziel.

Metadaten

Auch muss geklärt werden, wer welche Daten vorhält und in welcher Version. Denn es werden immer wieder Änderungen an Datensätzen vorgenommen, die zu einer neuen Version führen. Hier muss sichergestellt werden, dass z. B. ein Persistent Identifier immer auf die gleiche Version verweist.

Im Bereich der qualitativen Daten gibt es bislang nur wenige Einrichtungen, die Services für Forschungsdaten anbieten. Das Projekt Qualiservice²⁴ als Teil des eScience lab der Universität Bremen kooperiert

mit der Staats- und Universitätsbibliothek Bremen und dem Datenarchiv für Sozialwissenschaften (DAS) des GESIS Leibniz-Instituts für Sozialwissenschaften mit dem Ziel, eine Serviceeinrichtung für Primärdaten der qualitativen empirischen Sozialforschung aufzubauen. Ein weiteres Beispiel ist das Archiv »Deutsches Gedächtnis«²⁵ an der FernUniversität Hagen.

DATENBEREITSTELLUNG

In den Sozialwissenschaften gibt es eine lange Tradition in Bezug auf die Mehrfachnutzung von Daten. Einerseits ist die Produktion von Daten ressourcenintensiv, andererseits lassen sich manche Erhebungen nicht in der gleichen Form wiederholen. Daher ist die Nachnutzung von Forschungsdaten und das »Data-Sharing« ein wichtiges Anliegen von datenerhaltenden Einrichtungen. In den Sozialwissenschaften werden vor allem in der empirischen Sozialforschung Daten erhoben, die explizit für unterschiedliche Fragestellungen erhoben werden. Damit die Forscherinnen und Forscher diese Daten nutzen können, müssen sie sorgfältig aufbereitet sein.

Generell ist es wichtig, so viel Informationen über die Daten zu erhalten wie möglich, damit diese korrekt interpretierbar sind. So sind z. B. zusätzliche Materialien zur Erstellung der Studie hilfreich. Dazu gehören die verwendeten Erhebungs- und Messinstrumente, Methodenberichte, Codepläne u. a. Die darin enthaltenen Informationen bilden die Grundlage für die Erstellung von Metadaten. Metadaten dienen dazu, Nutzerinnen und Nutzern zu helfen, die für sie relevanten Daten zu finden und zu nutzen.

Für quantitative sozialwissenschaftliche Daten hat sich als XML-basierter Metadatenstandard DDI etabliert. Mitte der 1990er Jahre wurde er von der Data Documentation Initiative (DDI)²⁶ entwickelt und seitdem kontinuierlich weiter entwickelt. DDI ermöglicht die Beschreibung, die Verwaltung und Archivierung

F007

I

Manche Leute sagen, dass man den meisten Menschen trauen kann. Andere meinen, dass man nicht vorsichtig genug sein kann im Umgang mit anderen Menschen.
Was ist Ihre Meinung dazu?

☐ Den meisten Menschen kann man trauen

☐ Man kann nicht vorsichtig genug sein

☐ Das kommt darauf an

☐ Sonstiges, und zwar: _____

☐ Weiß nicht

☐ KA

Abb. 2: Beispiel einer Frage aus der ALLBUS (Allgemeine Bevölkerungsumfrage) 2012


```

<var ID="V51" name="V51">
  <labl>VERTRAUEN ZU MITMENSCHEN</labl>
  <qstn>
    <qstnLit><![CDATA[F007<br/>Manche Leute sagen, dass man den meisten Menschen
    trauen kann. Andere meinen, dass man nicht vorsichtig genug sein kann im Umgang mit anderen Menschen.
    <br/>Was ist Ihre Meinung dazu?<br/><br/>1 Den meisten Menschen kann man trauen<br/>2 Man kann
    nicht vorsichtig genug sein<br/>3 Das kommt darauf an<br/>4 Sonstiges, und zwar:<br/>8 Wei&#223;
    nicht<br/>9 Keine Angabe<br/><br/><br/>]]></qstnLit>
  </qstn>
  <catgry missing="N">
    <catValu>1</catValu>
    <labl>MAN KANN TRAUEN</labl>
  </catgry>
  <catgry missing="N">
    <catValu>2</catValu>
    <labl>MUSS VORSICHTIG SEIN</labl>
  </catgry>

```

Abb. 3: Beispiel einer Frage aus der ALLBUS 2012 in DDI-XML

von sozialwissenschaftlichen Datensätzen sowie die Erstellung von Einträgen in elektronischen Katalogen. »DDI hat den Anspruch, den gesamten Forschungsdaten-Lebenszyklus zu unterstützen. DDI Metadaten beziehen sich auf die Studienkonzeption, die Datenerhebung, die Datenbearbeitung und -auswertung sowie auf die Sekundärnutzung und Archivierung.«²⁷ Die Abbildungen 2, 3 und 5 zeigen eine Frage aus der Allgemeinen Bevölkerungsumfrage (ALLBUS), ihre Auszeichnung mit XML-Elementen aus DDI und die Darstellung im Datenportal ZACAT.

Zusätzliche beschreibende und erklärende Informationen werden in sog. Codebüchern erfasst, z.B. Informationen zur Entstehung der Daten, zu den Beteiligten, zum Erhebungszeitraum, den Erhebungsmethoden u.a. Die Beschreibung des Datensatzes beinhaltet die Position der Variablen, Messniveaus, Bedeutung numerischer Codes und die verwendeten Messinstrumente (z.B. Fragebögen).

Wie bereits beschrieben, gibt es nicht nur sehr viele unterschiedliche Arten von Forschungsdaten in den Sozialwissenschaften, sondern auch diverse Einrichtungen, die diese Daten bereitstellen.

Die Forschungsdaten- und Datenservicezentren haben einen thematischen Fokus. In vielen Fällen sind sie selbst Datenproduzenten. Durch die inhaltliche Spezialisierung ist eine tiefe Expertise der Daten selbst und eine umfangreiche Beratung im Hinblick auf die Nutzung der Daten gewährleistet (z.B. in Fragen der Gewichtung oder des Umgangs mit fehlenden Werten). Die Bereitstellung der Daten erfolgt in unterschiedlichen Formen. Besonderes Augenmerk wird auf die Einhaltung datenschutzrechtlicher Bestimmung gelegt, denn gerade in der Sozialforschung werden sehr viele Daten von Personen erhoben. Es muss da-

her gewährleistet sein, dass die Daten keinerlei Rückschlüsse auf die an der Studie beteiligten Personen zulassen. Das bedeutet, dass die Daten anonymisiert werden müssen. Dabei gibt es unterschiedliche Möglichkeiten der Anonymisierung: Absolut anonymisierte Daten werden als sogenannte »Public Use Files« zur Verfügung gestellt und faktisch anonymisierte Daten als sogenannte »Scientific Use Files«.

Bei der absoluten Anonymisierung dürfen keinerlei Rückschlüsse auf die Auskunft gebende Stelle oder Person möglich sein. Dies wird durch das Entfernen einzelner Merkmale oder eine Vergrößerung der Daten erreicht. Die »Public Use Files« stehen allen interessierten Personen zur Verfügung. Anders verhält es sich mit den faktisch anonymisierten »Scientific Use Files«; diese werden so verändert, dass sie nur mit ei-

Anonymisierung der Daten

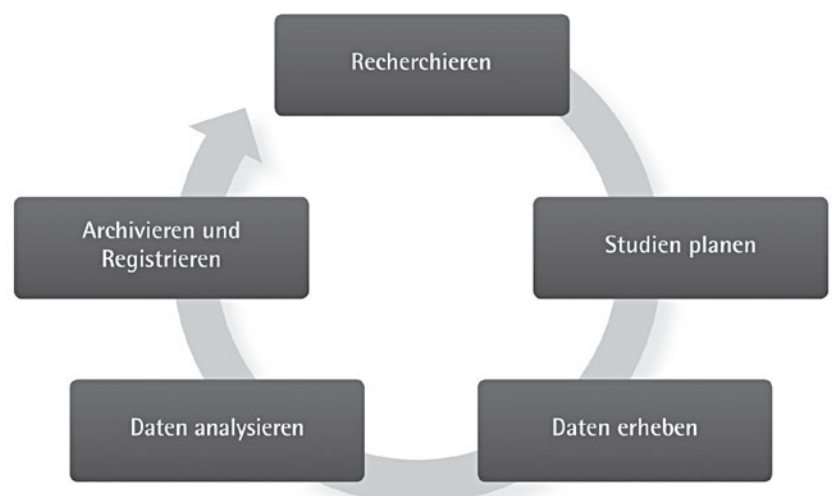


Abb. 4: Forschungsdaten-Lebenszyklus

nem erheblichen Ressourcenaufwand eine Deanonymisierung ermöglichen, der statistische Informationsgehalt jedoch erhalten bleibt. Diese Datensätze werden ausschließlich zu wissenschaftlichen Zwecken bereitgestellt. Einige Datensätze werden auch nur vor Ort, an speziellen Gastwissenschaftler-Arbeitsplätzen, zugänglich gemacht.²⁸

GESIS ist die größte Infrastruktureinrichtung für die Sozialwissenschaften in Deutschland. Das Angebot von GESIS orientiert sich an den Bedürfnissen seiner Nutzerinnen und Nutzer. Um diesen in allen Phasen des Forschungsdatenzyklus die bestmögliche Unterstützung anbieten zu können, bündelt GESIS seine Expertise auf den Gebieten der Produktion, Verwendung, Archivierung und Bereitstellung von Forschungsdaten.

Das GESIS-Datenarchiv bietet seit mehr als 50 Jahren Service zu soziologischen und politikwissenschaftlichen Forschungsdaten. Diese werden nach klar definierten Anforderungen akquiriert und bedarfsorientiert gemäß international anerkannten Standards aufbereitet, archiviert und der wissenschaftlich in-

teressierten Öffentlichkeit zugänglich gemacht. Das GESIS-Datenarchiv archiviert in erster Linie Umfragedaten und Daten der historischen Statistik. Nachdem die Daten von den Datenproduzenten an das Archiv übergeben wurden und diese u.a. auf Vollständigkeit und Richtigkeit geprüft wurden, werden sie nach standardisierten Verfahren aufbereitet. Das bedeutet, es wird eine detaillierte Studienbeschreibung erstellt und die Studie (d.h. die Daten und deren Dokumentation) wird im Datenbestandskatalog (DBK) veröffentlicht.

Die Datenproduzenten legen die Zugangsbestimmungen zu ihren Daten fest. Es stehen verschiedene Nutzungskategorien zur Verfügung: Die Daten sind für alle Interessierten zugänglich oder stehen nur für wissenschaftliche Forschungszwecke zur Verfügung. In einigen Fällen wird eine Genehmigung zur Nutzung der Daten für eine Publikation benötigt und in anderen Fällen muss der Datenproduzent seine Zustimmung zur Nutzung überhaupt geben.

Bevor Nutzerinnen und Nutzer die Daten erhalten, müssen sie sich kostenlos registrieren. Auch die Be-

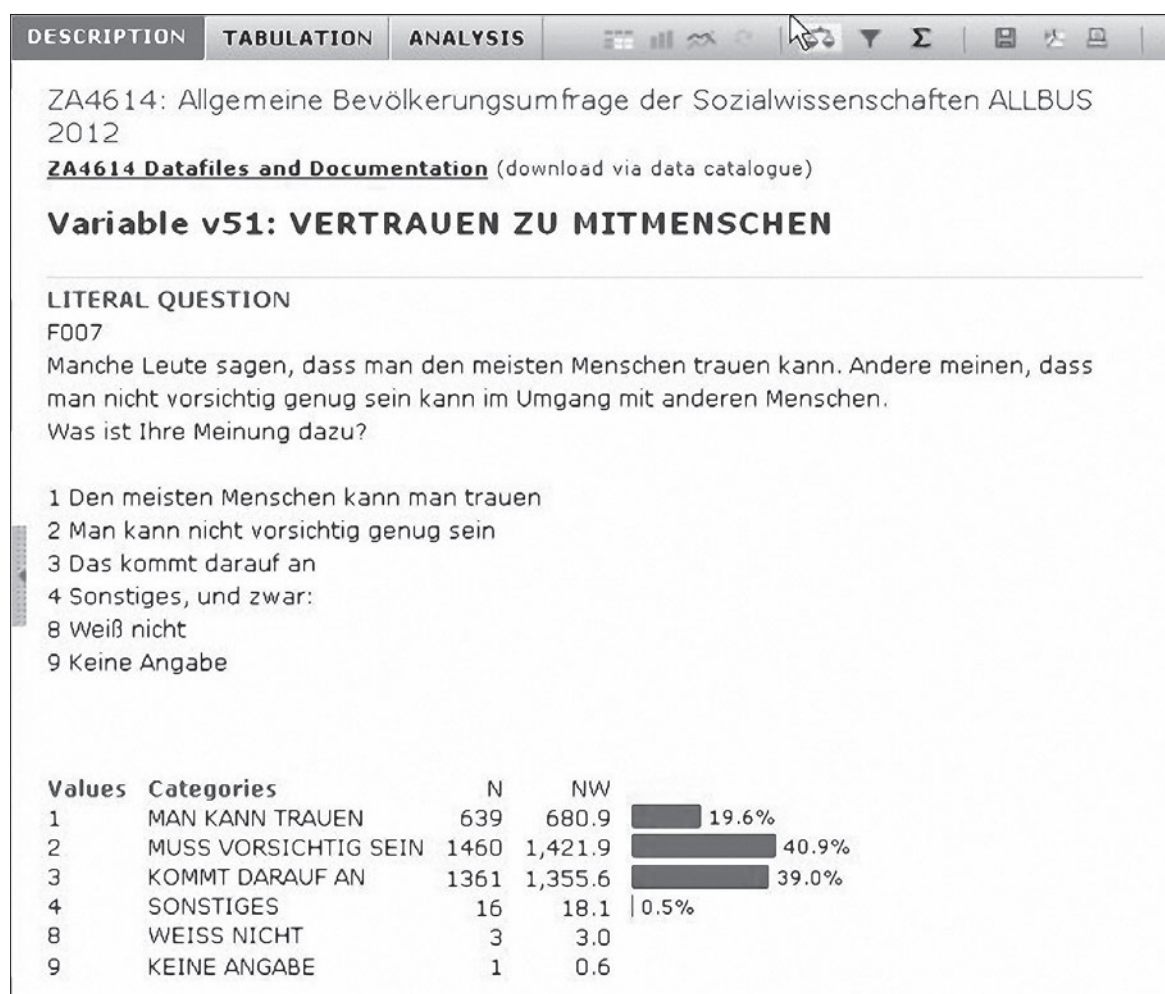


Abb. 5: Beispiel einer Frage aus der ALLBUS 2012 mit Häufigkeiten im Portal ZACAT

reistellung der Daten ist im Regelfall kostenfrei, es sei denn, die Daten werden auf Datenträgern versendet, dann entstehen Bearbeitungsgebühren.

GESIS bietet verschiedene Recherchesysteme an, in denen die Daten nachgewiesen sind. Zunächst ist der Datenbestandskatalog²⁹ zu nennen, in dem die Beschreibungen und Metadaten aller Studien eingetragen sind. Das Portal ZACAT³⁰ bietet ausgewählte Online-Studien mit extensiver Suchfunktion auf Studien- und Variablenebene, eine Online-Analyse-Funktion, einfachen Zugriff auf Daten, Dokumentationen und dazugehörige Materialien. Für die Recherche in Codebüchern und Fragetexten von ausgewählten Studien steht der CodebookExplorer³¹ und QBase zur Verfügung. Für Historische Studien und Zeitreihen kann die Anwendung HISTAT³² genutzt werden. Darüber hinaus besteht ein Service zur Nutzung amtlicher Mikrodaten³³ sowie Online-Informationen zum Bereich sozialer Indikatoren.

Ausgewählte und stark nachgefragte Daten werden besonders aufbereitet. Beispielsweise werden Variablen harmonisiert, damit sie über Länder oder Zeit vergleichbar mit weiteren Erhebungen sind. Auch ergeben sich so neue Analysepotentiale. Die Daten aus folgenden großen nationalen und internationalen Umfragen werden gesondert aufbereitet und können entsprechend recherchiert werden:

- ALLBUS: Allgemeine Bevölkerungsumfrage³⁴,
- ISSP: International Social Survey³⁵,
- ESS: European Social Survey³⁶,
- EVS: European Values Study³⁷,
- Eurobarometer: Umfrageprogramm der Europäischen Kommission zur Einstellung der Bürger zur Europäischen Einigung und ihren Institutionen³⁸,
- Wahlstudien³⁹.

Ein neuer Service von GESIS, datorium, bietet kleineren Forschungsprojekten und Einzelpersonen die Möglichkeit, ihre sozialwissenschaftlichen Forschungsdaten in einem Daten-Repository zu speichern und sie anderen Forscherinnen und Forschern zur Verfügung zu stellen. datorium bietet die physische Sicherung (Bitstream Preservation) für die Daten an, weitere Maßnahmen zur langfristigen Erhaltung, wie beispielsweise Formatmigrationen, werden nicht durchgeführt. Die Forschungsdaten werden von den Wissenschaftlerinnen und Wissenschaftlern selbst mit Metadaten versehen. Das geschieht über ein Online-Formular. Darüber hinaus werden auch über Institutionelle Repositorien an Hochschulen sozialwissenschaftliche Forschungsdaten angeboten. Ein Beispiel hierfür ist das Mannheim Research Data Repository MADATA⁴⁰.

Datenmanagement

Datenmanagement wird ein zunehmend wichtigeres Thema in den Sozialwissenschaften. Bislang ist dieser Bereich noch nicht systematisch in der Ausbildung verankert.

»Ein strukturiertes Konzept zum Datenmanagement, das in die methodisch-theoretische Planung und Durchführung einer empirischen Untersuchung eingebettet ist, kann den Projektalltag erheblich entlasten. Es hilft zugleich, neuen Anforderungen von Förderorganisationen zum Umgang mit Forschungsdaten im Rahmen von Projektanträgen frühzeitig und effizient gerecht zu werden.«⁴¹

Die Deutsche Forschungsgemeinschaft (DFG) hat in ihre »Empfehlungen zur Sicherung guter wissenschaftlicher Praxis« den Umgang mit Forschungsdaten und das Thema Datenmanagement mit aufgenommen. Hier wird dazu aufgefordert, »Primärdaten als Grundlage für Veröffentlichungen [...] auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre« aufzubewahren.⁴²

Darüber hinaus heißt es im Leitfaden zur Antragstellung der DFG: »Wenn aus Projektmitteln systematisch (Mess-)Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch – sofern vorhanden – die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien.«⁴³

Langzeitarchivierung

»Unter einem digitalen Langzeitarchiv wird [...] eine Organisation (...) verstanden, die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit digitaler Objekte sowie für ihre Interpretierbarkeit zum Zwecke der Nutzung durch eine bestimmte Zielgruppe übernommen hat.«⁴⁴

Langzeitarchivierung bedeutet nicht nur die physische Speicherung von Daten, sondern vielmehr die Erhaltung der langfristigen Nutzbarkeit der Daten. Damit ist gemeint, dass die Daten korrekt interpretierbar und authentisch bleiben. Nutzerinnen und Nutzer müssen sich darauf verlassen können, dass die Daten unverändert und korrekt sind. Denn technischer Fortschritt und der damit einhergehende Wandel von Datenträgern, Hard- und Software etc. erfordern andere Erhaltungsmaßnahmen als es bei analogen Daten der Fall ist. Auch Datenformate ändern sich im Laufe der Zeit. Bei der Erhaltung digitaler Daten spielen verschiedene Faktoren eine Rolle, die jeweils miteinander

Recherchesysteme

gesonderte Aufbereitung

Erhaltungsmaßnahmen

der verknüpft sind. Das bedeutet, verschiedene Komponenten entscheiden darüber, ob ein digitales Objekt in Zukunft nicht nur lesbar, sondern auch korrekt interpretierbar bleibt. Eine gute Übersicht über die verschiedenen Aspekte der digitalen Langzeitarchivierung bietet das nestor-Handbuch.⁴⁵

Wie eingangs beschrieben, werden in den Sozialwissenschaften sehr unterschiedliche Daten erhoben und genutzt. Daher ist die Situation in Bezug auf die langfristige Verfügbarkeit entsprechend unterschiedlich. Im Bereich der quantitativen Daten existiert eine überschaubare Anzahl an genutzten Formaten und dazu gehöriger Software. Zumeist handelt es sich dabei um proprietäre Software. Am weitesten verbreitet sind »SPSS« und »Stata«. Aber auch Open Source Software gewinnt in diesem Bereich an Bedeutung, z. B. mit der Programmiersprache »R«. Im kommerziellen Bereich wird überwiegend mit »SAS« gearbeitet. Diese Formate sind innerhalb der sozialwissenschaftlichen Community weit verbreitet, so dass Archive lediglich mit einer überschaubaren Anzahl an verschiedenen Formaten konfrontiert sind. Daten werden den Nutzerinnen und Nutzern auch in den gängigen Formaten zur Verfügung gestellt. Allerdings sind die proprietären Formate weniger für die langfristige Sicherung der Daten geeignet, da für ihre Nutzung die entsprechende Software benötigt wird, die gegebenenfalls nicht dauerhaft vom Hersteller unterstützt wird. Daher sind (relativ) offene Formate wie bspw. PDF/A für digitale Publikationen oder TIFF für Bilddokumente zu bevorzugen.

Neben zur Langzeitarchivierung brauchbaren Formaten bedarf es weiterer Maßnahmen zur Sicherung der Daten. Die physische Sicherung (Bitstream-Preservation) ist hierbei die Grundvoraussetzung für alle weiterführenden Maßnahmen. Regelmäßige Integritätsprüfungen (u. a. mithilfe von Checksummen) sind unerlässlich. Die Daten werden auf unterschiedlichen Medien gespeichert und redundant vorgehalten. Als zusätzliche Sicherung sollten die Daten an unterschiedlichen Orten gespeichert sein und die verschiedenen Datenträger immer wieder erneuert werden (Medienrefreshment). Als Erhaltungsstrategien werden in der Regel Formatmigrationen durchgeführt.

Alle vorgenommenen Veränderungen an den Daten müssen dokumentiert werden, damit sie auch über einen längeren Zeitraum hinweg nachvollziehbar bleiben. Dies kann z. B. in einer Aufbereitungssyntax geschehen, die zudem den Bezug zwischen der Originaldatei und der archivierten Version herstellt. Neben den Originaldaten wird für Archivierungszwecke eine um Metadaten erweiterte Version erstellt, und

beide Varianten werden langfristig gesichert. Für die Nachnutzung durch Forscherinnen und Forscher werden die Daten in den oben genannten proprietären Formaten zur Verfügung gestellt. Metadaten werden also nicht nur zur inhaltlichen und methodischen Beschreibung eingesetzt, sondern auch zur Dokumentation von technischen und administrativen Aspekten.

Im Bereich der Langzeitarchivierung wird in der Regel auf das Open Archival Information System (OAIS)-Referenzmodell⁴⁶ Bezug genommen. Es beschreibt in abstrakter Form die Funktionalitäten eines digitalen Langzeitarchivs. Wenngleich die meisten sozialwissenschaftlichen Archive weitaus länger existieren als das OAIS-Modell, so sind zumindest die Hauptfunktionalitäten in allen Archiven zu finden: Die Aufnahme (Ingest) der Daten in das Archiv, die Verwaltung der Daten, die Speicherung und der Zugang.

Das OAIS-Modell ist ein Referenzmodell und dient dazu, sich auf ein gemeinsames Vokabular zu verständigen und eine generische Vorstellung eines digitalen Langzeitarchivs darzulegen. Jedes reale Archiv, nicht nur im sozialwissenschaftlichen Bereich, muss überlegen, welche Funktionalitäten und Verantwortlichkeiten aus dem OAIS-Modell für die Erfüllung der eigenen Aufgabe notwendig sind, welche nicht und welche in dem Modell gar nicht berücksichtigt werden, für das Archiv aber notwendig sind.

AUSBLICK

In den letzten Jahren wurden auf nationaler und internationaler Ebene im Hinblick auf die Vertrauenswürdigkeit digitaler Archive mehrere Verfahren zur (Selbst-)Zertifizierung entwickelt. Diese finden auch Anwendung in sozialwissenschaftlichen Archiven. So haben beispielsweise die im CESSDA Kontext zusammengeschlossenen Archive als Grundlage ihrer gemeinsamen Arbeit ein Self-Audit durchgeführt, welches sich auf die Kriterien des Data Seal of Approvals (DSA) bezieht. Auch wenn Unterschiede in der Umsetzung der jeweiligen Institution durchaus erkennbar sind, so ist dieser Prozess vor allem im Hinblick auf eine gemeinsame Vorstellung von Vertrauenswürdigkeit sehr hilfreich und wichtig. Es zeigt auch, dass sich die sozialwissenschaftlichen Datenarchive in Europa mit Fragen beschäftigen, wie sie die ihnen anvertrauten Daten am besten aufbereiten und die langfristige Nutzung sichern können.

Die Herausforderungen für die Zukunft liegen vor allem im Umgang mit neuen Formaten. Qualitative Erhebungen mit ihren unterschiedlichen Methoden, wie beispielsweise audio-visuelle Aufzeichnungen, bergen nicht nur Probleme hinsichtlich der zu erwartenden Datenmengen, sondern stellen die sozialwissenschaft-

OAIS-Referenzmodell

Vertrauenswürdigkeit

Dokumentation

liche Forschung auch in Bezug auf datenschutzrechtliche Fragen vor neue Herausforderungen.

¹ Bayer, Otto; Stölting, Erhard: Sozialwissenschaften, in: Seiffert, Helmut; Radnitzky, Gerard (Hrsg.), Handlexikon zur Wissenschaftstheorie, 2. Aufl., München 1994, S. 302–313.

² Vgl. Bortz, Jürgen; Döring, Nicola: Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler, Berlin 2006.

³ Watteler, Oliver: Recherche nach sozialwissenschaftlichen Forschungsdaten, in: Berninger, Ina et al. (Hrsg.): Grundlagen sozialwissenschaftlichen Arbeitens. Eine anwendungsorientierte Einführung, Opladen, Toronto, (2012), S. 123.

⁴ Brümmerhoff, Dieter; Grömling, Michael: Volkswirtschaftliche Gesamtrechnung, München 2011, S. 1.

⁵ Vgl. Diekmann, Andreas: Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, Reinbek 2007, S. 193–194.

⁶ Vgl. Wingerter, Christian: Der Wandel der Erwerbsformen und seine Bedeutung für die Einkommenssituation Erwerbstätiger, in: Wirtschaft und Statistik 11 (2009), S. 1080–1098 und Keller, Berndt; Seifert, Hartmut: Atypische Beschäftigungsverhältnisse. Stand und Lücken der aktuellen Diskussion, in: WSI Mitteilungen 3 (2011), S. 138–145.

⁷ Diekmann, Andreas: Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen, Reinbek 2007, S. 18–19.

⁸ Ebd., S. 33–37.

⁹ Flick, Uwe; Kardorff, Ernst von; Steinke, Ines: Qualitative Forschung. Ein Handbuch, Reinbek 2010, S. 14.

¹⁰ Vgl. Flick, Uwe; Kardorff, Ernst von; Steinke, Ines: Qualitative Forschung. Ein Handbuch, Reinbek 2010.

¹¹ Vgl. Schnell, Rainer; Hill, Paul B.; Esser, Elke: Methoden der empirischen Sozialforschung, 8., unveränderte Aufl., München, Wien, 2008.

¹² Watteler, Oliver: Recherche nach sozialwissenschaftlichen Forschungsdaten, in: Berninger, Ina et al. (Hrsg.): Grundlagen sozialwissenschaftlichen Arbeitens. Eine anwendungsorientierte Einführung, Opladen, Toronto (2012), S. 120–121.

¹³ www.ropercenter.uconn.edu/ (Stand: 08.08.2013).

¹⁴ www.gesis.org/das-institut/wissenschaftliche-abteilungen/datenarchiv-fuer-sozialwissenschaften/ (Stand: 08.08.2013).

¹⁵ www.icpsr.umich.edu/icpsrweb/landing.jsp (Stand: 08.08.2013).

¹⁶ www.dans.knaw.nl/en (Stand: 08.08.2013).

¹⁷ <http://data-archive.ac.uk/> (Stand: 08.08.2013).

¹⁸ www.cessda.org/ (Stand: 08.08.2013).

¹⁹ <http://ec2-50-17-181-92.compute-1.amazonaws.com/wordpress/> (Stand: 08.08.2013).

²⁰ Vgl. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik 2001: »Wege zu einer besseren informationellen Infrastruktur«.

²¹ <http://ratswd.de/> (Stand: 08.08.2013).

²² www.dfg.de/foerderung/programme/listen/index.jsp?id=SFB (Stand: 08.08.2013).

²³ www.da-ra.de/ (Stand: 08.08.2013).

²⁴ www.qualiservice.org/ (Stand: 08.08.2013).

²⁵ www.fernuni-hagen.de/geschichteundbiographie/deutsches-gedaechtnis/ (Stand: 08.08.2013).

²⁶ www.ddialliance.org/ (Stand: 08.08.2013).

²⁷ Zenk-Möltgen, Wolfgang: Metadaten und die Data Document Initiative (DDI), in: Altenhöner, R. und Oellers, C. (Hrsg.): Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen (2012), S. 111.

²⁸ Vgl.: www.forschungsdatenzentrum.de/datenzugang.asp

²⁹ www.gesis.org/unser-angebot/recherchieren/datenbestands-katalog/ (Stand: 08.08.2013).

³⁰ www.gesis.org/unser-angebot/recherchieren/zacat-online-study-catalogue/ (Stand: 08.08.2013).

³¹ www.gesis.org/unser-angebot/recherchieren/codebookexplorer/ (Stand: 08.08.2013).

³² www.gesis.org/unser-angebot/daten-analysieren/daten-historischer-studien/histat/ (Stand: 08.08.2013).

³³ www.gesis.org/unser-angebot/daten-analysieren/amtliche-mikrodaten/ (Stand: 08.08.2013).

³⁴ www.gesis.org/allbus/allbus-home/ (Stand: 08.08.2013).

³⁵ www.gesis.org/issp/issp-home/ (Stand: 08.08.2013).

³⁶ www.europeansocialsurvey.org/ (Stand: 08.08.2013).

³⁷ www.gesis.org/unser-angebot/daten-analysieren/umfrage-daten/european-values-study/ (Stand: 08.08.2013).

³⁸ www.gesis.org/eurobarometer-data-service/home/ (Stand: 08.08.2013).

³⁹ www.gesis.org/wahlen/wahlen-home/ (Stand: 08.08.2013).

⁴⁰ <https://madata.bib.uni-mannheim.de/> (Stand: 08.08.2013).

⁴¹ Jensen, Uwe: Leitlinien zum Management von Forschungsdaten. Sozialwissenschaftliche Umfragedaten (2012), S. 9.

⁴² Ergänzung der Empfehlungen der Deutschen Forschungsgemeinschaft zur Sicherung guter wissenschaftlicher Praxis (2013), S. 5.

⁴³ Leitfaden für die Antragstellung, 2.4 Umgang mit den im Projekt erzielten Forschungsdaten, S. 6.

⁴⁴ nestor-Kriterienkatalog vertrauenswürdige digitale Langzeitarchive Version II / hrsg. von der nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (2008), S. 4.

⁴⁵ www.nestor.sub.uni-goettingen.de/handbuch/index.php (Stand: 08.08.2013).

⁴⁶ <http://public.ccsds.org/publications/archive/65oxom2.pdf> (Stand: 08.08.2013).

DIE VERFASSER

Natascha Schumann ist Mitarbeiterin im Datenarchiv für Sozialwissenschaften bei GESIS und dort verantwortlich für Langzeitarchivierung und Zertifizierung, GESIS – Leibniz-Institut für Sozialwissenschaften, Unter Sachsenhausen 6–8, 50667 Köln, E-Mail: natascha.schumann@gesis.org

Oliver Watteler ist Mitarbeiter im Datenarchiv für Sozialwissenschaften bei GESIS und dort verantwortlich für die Akquisition neuer Studien, GESIS – Leibniz-Institut für Sozialwissenschaften, Unter Sachsenhausen 6–8, 50667 Köln, E-Mail: oliver.watteler@gesis.org