Aufbau des da|ra-Datennachweissystems: Hintergründe, Herangehensweise und Lessons learned

Friedrich, Tanja; Engels, Martina; Nasshoven, Verena

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Friedrich, T., Engels, M., & Nasshoven, V. (2014). *Aufbau des da|ra-Datennachweissystems: Hintergründe, Herangehensweise und Lessons learned.* (GESIS-Technical Reports, 2014/12). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-402739

gesis Leibniz-Institut für Sozialwissenschaften

TECHNICAL *Reports* 2014|12

Aufbau des da|ra-Datennachweissystems

Hintergründe, Herangehensweise und Lessons learned

Tanja Friedrich, Martina Engels, Verena Nasshoven

GESIS-Technical Reports 2014|12

Aufbau des da|ra-Datennachweissystems

Hintergründe, Herangehensweise und Lessons learned

Tanja Friedrich, Martina Engels, Verena Nasshoven

GESIS – Leibniz-Institut für Sozialwissenschaften 2014

GESIS-Technical Reports

GESIS – Leibniz-Institut für Sozialwissenschaften
Datenarchiv für Sozialwissenschaften (DAS)
Unter Sachsenhausen 6-8, 50667 Köln
50667 Köln
Telefon: (0221) 476 94 - 557
Telefax: (0221) 476 94 - 199
E-Mail: tanja.friedrich@gesis.org

ISSN: 1868-9043 (Print)
ISSN: 1868-9051 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Inhalt

Zusammenfassung

Im Rahmen des DFG-geförderten Projekts „da|ra – Aufbau einer Registrierungsagentur für sozialwissenschaftliche Forschungsdaten" wurde durch GESIS in einem eigens hierfür vorgesehenen Arbeitspaket ein Nachweissystem für nicht DOI®-registrierte Forschungsdaten aus den Sozialwissenschaften konzipiert und aufgebaut. Ziel und Zweck des Datennachweissystems ist es, Sekundärforscherinnen und -forschern eine zentrale Anlaufstelle für die Recherche sozialwissenschaftlicher Forschungsdaten zu bieten.

Hierzu wurden Organisationen und Institutionen im Umfeld der Produktion sozialwissenschaftlicher Forschungsdaten im Hinblick auf ihr Datenangebot evaluiert und Nachweise zu ihren Beständen in die da|ra-Datenbank eingebracht. Zusammen mit den in der Datenbank enthaltenen Metadaten DOI-registrierter Ressourcen sind die recherchierten Datennachweise integriert recherchierbar.

Sowohl bei der Recherche der Datenquellen und des Datenangebots, als auch bei der Erfassung der Nachweise in der Datenbank stellten sich diverse Herausforderungen, die im Projektverlauf adressiert wurden, und aus denen wertvolle Schlüsse für den Betrieb und die Weiterentwicklung des Systems gezogen wurden. Sowohl die Vorgehensweise, als auch die Herausforderungen und *lessons learned* werden im vorliegenden Bericht beschrieben.

Aus der Projektarbeit ist ein stabiles, flexibles Informationssystem mit einem Grundstock an Datennachweisen entstanden, das in einer weiteren Projektphase zu einem zentralen Suchnetzwerk für sozialwissenschaftliche Forschungsdaten weiterentwickelt werden soll.

1 Hintergrund: Datenregistrierung und Datennachweis mit da|ra

Die wissenschaftliche Arbeit mit Daten hat in den vergangenen Jahren disziplinübergreifend weiter an Bedeutung gewonnen. Ein datenintensives Forschungsparadigma[1] bestimmt weite Teile des Forschungsprozesses, von der Studienplanung über die Informationsrecherche bis hin zur Analyse und Publikation von Ergebnissen. Zentrale Infrastruktureinrichtungen unterstützen Wissenschaftlerinnen und Wissenschaftler[2] entlang dieses gesamten Prozesses mit unterschiedlichen Informationsdiensten. Im Bereich der Sozialwissenschaften erbringt GESIS seit Jahrzehnten entsprechende Dienstleistungen, zu denen seit einigen Jahren auch die Registrierung von Forschungsdaten zum Zweck der permanenten Identifikation mit Digital Object Indentifiers (DOI) gehört. Datenproduzenten erhalten durch die von GESIS und der ZBW betriebene Registrierungsagentur da|ra die Möglichkeit, ihre Daten über DOI-Namen dauerhaft und einheitlich zitierbar zumachen.[3] Dies erleichtert nicht nur die Zitation und Identifizierung von Daten in unterschiedlichen Kontexten, sondern ermöglicht auch diverse infometrische Analysen sowie die Verknüpfung von Forschungsliteratur und -daten. Insbesondere aber stellt die Registrierung über da|ra eine zentrale **Dokumentation von Forschungsdaten** dar, die nicht nur die Sichtbarkeit, sondern auch die Nachnutzbarkeit der Daten befördert. Diese Dokumentation der Forschungsdaten erfolgt durch die Aufnahme beschreibender Metadaten zu den registrierten Forschungsdaten in die zentrale da|ra-Datenbank.[4] Alle registrierten Forschungsdaten sind darüber recherchierbar.[5] Im Juli 2014 zählte die Datenbank knapp 300.000 Einträge.

da|ra bietet mit dieser Datenbank nicht nur Datenpoduzenten eine Möglichkeit, ihre Daten zu publizieren, sondern vor allem auch Sekundärforschern eine zentrale Anlaufstelle für die **Recherche relevanter Forschungsdaten**. Für die eigene Forschung geeignete Daten zu finden, ist nach wie vor ein aufwendiger Vorgang, der die Suche in breit verteilten Angeboten erfordert. In der da|ra-Datenbank konnten bereits viele dieser Informationen integriert werden, insbesondere zu Daten aus großen Umfrageprogrammen. Über diesen Kernbestand hinaus, der beispielsweise die Daten in den bedeutenden Archiven GESIS und ICPSR umfasst, existieren aber unzählige potentiell nachnutzbare Daten der empirischen Sozialforschung, die in kleineren Projekten entstehen und allein aufgrund ihrer schieren Masse nicht den Weg zur DOI-Registrierung und damit in die da|ra-Datenbank finden. Daher wurden im Zuge des Aufbaus von da|ra auch solche Datenquellen recherchiert und in der Datenbank nachgewiesen, die für eine DOI-Registrierung (derzeit) nicht infrage kommen. Dass diese Aufgabe nie vollständig bewältigt werden kann, liegt auf der Hand. Die vorgenommenen Arbeiten, über die im Folgenden ausführlich berichtet wird, haben daher Pilotcharakter und dienten in erster Linie dazu, ein genaueres Bild von der Datenlage zu zeichnen und Herausforderungen bei der Integration von Nachweisen verteilter Daten zu identifizieren. Außerdem wurde ein Grundstock von Datennachweisen in die Datenbank eingebracht, um Forscher zu motivieren, ihre Daten künftig selbständig im System nachzuweisen. In einem Folgeprojekt ist geplant, durch verbreitete Selbstmeldung von Daten und durch weitgehend automatisierte Anreicherung der Datenbank sowie der Suchfunktion das Datennachweissystem sukzessive auszubauen.

Der vorliegende Bericht beschreibt zunächst die Situation der verteilten Forschungsdatenlandschaft in den Sozialwissenschaften, wie sie sich zu Beginn des Projektes darstellte und das Ziel einer zentralen

---

1 Vgl. Hey et al. 2009.

2 Zur Erleichterung der Lesbarkeit werden im Folgenden in der Regel männliche Personenbezeichnungen verwendet. Die weibliche Sprachform ist jeweils impliziert.

3 Das Projekt „da|ra – Aufbau einer Registrierungsagentur für sozialwissenschaftliche Forschungsdaten" wurde von 2011 bis 2014 von der DFG gefördert.

4 Das da|ra-Metadatenschema ist ausführlich beschrieben in Hausstein et al. 2014.

5 Sucheinstieg unter: http://www.da-ra.de.

Suchmöglichkeit (Abschnitt 2). Im darauffolgenden Abschnitt (3 Recherche und Erfassung) werden die Vorgehensweise bei der Datenquellenrecherche sowie die Kriterien für die Aufnahme in das Datennachweissystem erläutert. Abschnitt 4, Herausforderungen und *lessons learned*, beschreibt konkrete Probleme, die sich bei der Recherche und der Erfassung der Datennachweise stellten und in welcher Weise diese das weitere Vorgehen beeinflusst haben. Der Bericht schließt mit einem Ausblick auf die weitere Arbeit am Datennachweissystem unter Gesichtspunkten der Effizienz und Nachhaltigkeit (Abschnitt 5).

2 Die Suche nach sozialwissenschaftlichen Daten

2.1 Ist-Situation: Verteilte Datenlandschaft

Datenintensive Forschung spielt inzwischen in allen Wissenschaftsdisziplinen eine Rolle. Insbesondere die empirische Sozialforschung gehört zu den Feldern, in denen seit langem **Forschungsdaten zur Sekundäranalyse** nicht nur archiviert und bereitgestellt, sondern eben für diesen Zweck produziert werden (Quandt/Mauer 2012, 63). In der Regel handelt es sich dabei um Daten aus groß angelegten Umfrageprogrammen, mit denen potentiell verschiedenste Fragestellungen bearbeitet werden können (sogenannte Mehrthemenumfragen), wie beispielsweise die in regelmäßigen Abständen seit 1980 durchgeführte Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS[6]. Die in Deutschland erstellten Erhebungen dieser Art sind inzwischen im Wesentlichen alle durch die DOI-Registrierung in der da|ra-Datenbank nachgewiesen und recherchierbar. Darüber hinaus existieren jedoch weitere verteilte Datenquellen, auf die Nutzer nur nach intensiver Internetrecherche, aufgrund anderer Informationsquellen oder nach persönlicher Empfehlung stoßen können. Mit dem Aufbau eines umfassenden Nachweissystems für Daten sozialwissenschaftlicher Forschung wird im Rahmen des da|ra-Projekts das Ziel verfolgt, auch diese Bestände zentral recherchierbar zu machen.

Beim Aufbau eines integrierten Nachweises des Datenbestands bietet sich eine Orientierung an **Organisationen und Institutionen im Umfeld der Produktion sozialwissenschaftlicher Forschungsdaten** in Deutschland an. GESIS erscheint in diesem Zusammenhang als natürlicher Ausgangspunkt, allerdings sind alle im GESIS Datenarchiv vorgehaltenen Daten bereits über die DOI-Registrierung in der da|ra-Datenbank enthalten. Einzelne GESIS-Datenbestände können dennoch als Kandidaten für den Datennachweis identifiziert werden, etwa Mikrozensusdaten aus den 1960er Jahren. Ein weiterer, wichtiger Anlaufpunkt für die Suche nach sozialwissenschaftlich relevanten Daten sind die Forschungsdatenzentren (FDZ) des Rats für Sozial- und Wirtschaftsdaten (RatSWD).[7] Diese konnten jedoch ebenfalls bereits fast alle für eine DOI-Registrierung gewonnen werden, wodurch die dort vorgehaltenen Daten in weiten Teilen auch über da|ra suchbar sind. Einzelne FDZ haben allerdings (noch) nicht alle ihre Daten registrieren lassen. Ein weiterer, interessanter Bestand sind die den Forschungsberichten der Ressortforschungseinrichtungen der Bundesministerien zugrundeliegenden Daten, die bislang nur zum Teil veröffentlicht werden, ebenso wie die Daten der statistischen Ämter des Bundes und der Länder. Parteien, Stiftungen, Gewerkschaften und ähnliche Einrichtungen sind ebenfalls interessante Datenproduzenten. Sie arbeiten häufig mit Instituten der kommerziellen Marktforschung zusammen, an deren Bestände schwer heranzukommen ist. Schließlich werden in den zahlreichen Institutionen der universitären und außeruniversitären Forschung datenintensive Projekte durchgeführt, die man als den *long tail* der Datenproduktion in den Sozialwissenschaften bezeichnen kann. Auch diese Projekte haben grundsätzlich die Möglichkeit, ihre Daten an eine Institution wie das GESIS Datenarchiv zu übergeben. In der Praxis werden in kleinen Projekten erzeugte Daten aber regelmäßig lokal gespeichert und entziehen sich dadurch der Recherche. Gerade diese Bestände sind schlecht bis gar nicht erschlossen, bilden aber einen großen Anteil der produzierten Forschungsdaten. Einen ungefähren Eindruck von der stetig wachsenden Masse entsprechender Daten vermittelt die GESIS-Datenbank SOFIS[8], in der im Januar 2014 über 50.000 theoretische und empirische Forschungsprojekte universitärer und außeruni-

[6] http://www.gesis.org/allbus.

[7] http://www.ratswd.de/forschungsdaten/fdz.

[8] http://sofis.gesis.org/sofiswiki/Hauptseite.

versitärer Einrichtungen im deutschsprachigen Raum erfasst waren, und die jährlich um 5.000 bis 7.000 weitere Projektbeschreibungen anwächst.[9]

Eine **systematische Suche nach Forschungsdaten** zu einer bestimmten Forschungsfrage ist angesichts dieser verteilten Datenlandschaft sehr aufwendig. Problematisch ist nicht nur, dass interessierte Sekundärnutzer verschiedenste Quellen konsultieren müssen, sondern auch, dass sie sich mit diversen Möglichkeiten der Datenbeschreibung, -verfügbarkeit und -bereitstellung auseinandersetzen müssen. Hier gleicht keine Datenquelle der anderen: In einigen wenigen Fällen gibt es Datenkataloge oder ausführliche Datensatzbeschreibungen in Listenform, vielfach wird das Vorhandensein von Daten aber nur aus Publikationen (z.B. Forschungsberichten) implizit deutlich. Ein weiteres Problem besteht im Informationsverlust durch die Dynamik des Internets, denn wenn Daten nicht systematisch dokumentiert werden, erscheinen Hinweise darauf zum Beispiel nur in temporären Ankündigungen auf der Webseite des Datenproduzenten. Für Forscher, die Analysen und Vergleiche im Zeitverlauf anstellen wollen, sind Informationen über ältere Studien oder Erhebungswellen jedoch von zentraler Wichtigkeit.

2.2 Ziel: Zentrale Suchmöglichkeit für sozialwissenschaftliche Forschungsdaten

Die möglichst einfache Auffindbarkeit von Daten, die zur Nachnutzung geeignet sind, liegt nicht nur im Interesse einzelner Forscher, sondern trägt auch zur Effizienz der Forschungsfinanzierung bei, insofern Mehrfacherhebungen vermieden werden können. Außerdem werden kumulative Forschungsprozesse unterstützt, wenn Datenveröffentlichung zu darauf aufbauenden neuen Erhebungen und vergleichenden Analysen führt. Eine zentrale Suchmöglichkeit für sozialwissenschaftliche Forschungsdaten befördert gerade in Zeiten immer schneller anwachsender Datenbestände die wissenschaftliche Wertschöpfung insgesamt.

Eine zentrale Suchmöglichkeit ist daher wünschenswert, allerdings stellen sich bei der Realisierung dieses Ziels diverse Herausforderungen, die im hier beschriebenen Projekt berücksichtigt werden mussten. Zunächst wurden **technische Fragen unter der Prämisse der Nutzerfreundlichkeit** adressiert. Die zentrale da|ra-Datenbank und die darauf aufbauende Suche wurden zunächst im Hinblick auf das Registrierungssystem konzipiert, das heißt in erster Linie in Orientierung an den Bedürfnissen der Datenanbieter (sogenannte Publikationsagenten), deren Daten über die DOI-Vergabe in der Datenbank nachgewiesen werden. Das Metadatenschema, auf dem die Erfassung dieser Nachweise beruht, ist vor allem auf den Mechanismus der persistenten Identifikation der Daten ausgerichtet. Für die nicht zu registrierenden Daten, die im hier beschriebenen Projekt in die Datenbank aufgenommen werden sollten, spielten einige der für die DOI-Registrierung relevanten Metadatenelemente keine oder eine andere Rolle. Allen voran zählt hierzu das Element für den DOI, das bei Nachweisen nicht registrierter Daten fehlt, während es für die Registrierung ein Pflichtelement ist. Umgekehrt ist es beispielsweise für die Erfassung nicht-registrierter Datenbestände wichtig angeben zu können, dass ein bestimmter Datensatz zwar erzeugt wurde, seine Verfügbarkeit aber unbekannt ist.[10] Dieser Fall tritt nicht selten ein, zum Beispiel, weil datenproduzierende Institutionen organisatorische Veränderungen erfahren, die sich auch auf die Verfügbarkeit von Daten auswirken können. So kommt es vor, dass Daten bestimmter Langzeitstudien aus einzelnen Jahren vorhanden sind, aus anderen Jahren aber nicht. Diese wichtigen Informationen zu Studienkollektionen können bisher nicht systematisch recherchiert werden. Auch was die Erfassung der Nachweise im System angeht, gelten tendenziell andere Voraussetzungen für registrierte und nicht-registrierte Daten. Zu registrierten Daten existiert in den meisten Fällen bereits

[9] http://www.gesis.org/unser-angebot/recherchieren/sofis/.

[10] Weitere im Metadatenschema vorgesehene Verfügbarkeitsoptionen sind „Download", „Lieferbar", „Vor-Ort-Nutzung" und „Nicht verfügbar".

eine ausführliche Dokumentation, die über die da|ra-Pflichtfelder hinaus noch weitere Metadatenelemente enthält, und somit bereits eine sehr gute Beschreibung und Auffindbarkeit der Daten ermöglicht. Meistens werden diese Beschreibungen auch nicht einzeln über die Erfassungsmaske in das System eingebracht, sondern über einen Import der Informationen, zum Beispiel aus einem Datenkatalog. Im Falle der nicht-registrierten *long-tail*-Daten gibt es in den meisten Fällen keine Kataloge oder Datenbanken, die vergleichbar ausführliche und strukturierte Informationen zu den Daten bieten. Daher ist hier die Einzelbeschreibung der Datennachweise über die Erfassungsmaske der Standardweg. In vielen Fällen stellt der Eintrag im da|ra-Datennachweissystem die erste und einzige systematische Beschreibung eines Datensatzes dar. Deshalb ist es besonders wünschenswert, dass über die Pflichtangaben Titel, Primärforscher und URL hinaus auch möglichst ausführliche inhaltliche Informationen zu den Datensätzen erfasst werden. Eine Erfassungsmaske, die eine einfache, im Hinblick auf eine optimale Suche aber auch weitestgehend standardisierte Erfassung von beschreibenden Metadaten ermöglicht, ist daher insbesondere für die nicht-registrierten Forschungsdaten wichtig.

Neben der Klärung technischer Voraussetzungen stand vor allem die **Zusammenstellung und Auswertung relevanter Datenquellen** im Mittelpunkt der Arbeiten. Wie bereits beschrieben, wurde hier ein institutions- und organisationsorientierter Ansatz gewählt, beruhend auf den Kriterien der fachlichen Zuordnung sowie der Qualität und Bedeutung der in den Einrichtungen zu erwartenden Daten. Die Vorgehensweise bei der Recherche und Zusammenstellung der Datenquellen wird im folgenden Kapitel detailliert berichtet.

3 Recherche und Erfassung

Mit dem Ziel einer Zusammenstellung sozialwissenschaftlich relevanter Datenquellen wurden zu Beginn der Projektarbeit verschiedene Listen zur Erfassung von Datenanbietern erstellt. Ausgehend von diesen Listen wurden in umfangreichen manuellen Recherchen einzelne Datensätze identifiziert und ebenfalls in Listen dokumentiert. Durch diese Vorgehensweise konnte innerhalb weniger Monate ein vielfältiges Bild der sozialwissenschaftlichen Forschungsdatenlandschaft in Deutschland gezeichnet werden. Diese Erhebung bildete die Grundlage für die Erfassung von etwa 1.000 Studien im Datennachweissystem.

3.1 Kriterien für die Recherche

Das Projekt hat sich zum Ziel gesetzt, ein zentrales Nachweissystem für Daten der empirischen Sozialforschung aufzubauen. In dieser Zielsetzung sind **die wesentlichen Kriterien für relevante Datenanbieter und aufzunehmende Datensätze** bereits enthalten: zum einen das inhaltliche Kriterium der sozialwissenschaftlichen Relevanz und zum anderen das methodische Kriterium der Empirie. Im Laufe der Recherchen zeigte sich zunehmend, dass die Masse der vorhandenen Daten eine strikte Orientierung an diesen Kriterien unabdingbar macht, will man dem Datenangebot auch nur annähernd gerecht werden. Aus diesen Gründen beschränkte sich die Recherche außerdem auf Daten, die in Deutschland erhoben wurden; internationale Studien wurden nur berücksichtigt, wenn Deutschland als Erhebungsland vertreten war. Zudem wurde auf die Erfassung rein qualitativer Studien verzichtet, ebenso auf reine Sekundäranalysen.

3.2 Allgemeines Vorgehen

Aus einer intensiven Analyse der Datenlandschaft zu Beginn des Projekts wurden zunächst Organisationen und Institutionen im Umfeld der Produktion sozialwissenschaftlicher Forschungsdaten identifiziert (s. o. 2.1). Im Einzelnen waren dies:

- GESIS
- Forschungsdatenzentren (FDZ) des RatSWD
- Behörden und Institutionen des Bundes
- Statistische Ämter des Bundes und der Länder
- Parteien, Stiftungen, Gewerkschaften
- Institute der kommerziellen Marktforschung
- Universitäre und außeruniversitäre Forschungsinstitute

Bis auf Parteien, Gewerkschaften und die Institute der kommerziellen Marktforschung wurden schließlich alle diese **Datenquellen intensiv auf ihr Datenangebot hin evaluiert**. Die Einschränkung war erneut der überwältigenden Masse von Daten geschuldet und insofern pragmatisch bestimmt; darüber hinaus zeigte sich im Laufe der Recherchen, dass Daten aus diesen Quellen ohnehin kaum der Sekundärforschung zur Verfügung gestellt werden.

Zur Evaluation des Datenangebots der verbliebenen Datenanbieter wurden in erster Linie Informationen auf den Webseiten der Institutionen herangezogen. Es wurde dabei gezielt nach Informationen zu relevanten Studien bzw. Datensätzen gesucht. Hierzu wurde eine Excel-Tabelle angelegt, in der die

gefundenen Hinweise auf Datensätze gesammelt und nach einem sehr einfachen Schema dokumentiert wurden (s. Tabelle 1). Die Tabelle enthielt die Pflichtfelder des da|ra-Metadatenschemas und einige optionale Felder, die für den Nachweis von Daten mit unklarer Verfügbarkeit und Verantwortlichkeit wichtig erschienen. Die Orientierung am Metadatenschema war notwendig, da die Nachweise zu einem späteren Zeitpunkt in die da|ra-Datenbank übertragen wurden, die auf diesem Schema basiert.

Tabelle 1: Excel-Tabelle für die Recherche

Felder
Titel
Weitere Titel
Primärforscher
Datenanbieter (mit Kontakt)
Veröffentlichungsdatum
Verfügbarkeit
URL
Version
ID
Referenzzeitraum
Datenerhebung (Institut o. Person)
Datensatz (Einheiten, Typ)

Bei der Recherche auf den Webseiten der Institutionen zeigte sich die ganze **Heterogenität des verteilten Datenangebots**, mit der Datensuchende üblicherweise konfrontiert sind. Manche Institutionen präsentieren ihre Datenbestände ausführlich dokumentiert in einem eigens dafür eingerichteten Bereich ihrer Internetseiten, sodass das Datenangebot relativ schnell und einfach in die Excel-Liste übernommen werden konnte. In den meisten Fällen aber waren Hinweise auf Datenbestände unstrukturiert und verteilt auf den Webseiten zu finden, etwa in Studienankündigungen/-besprechungen, Veröffentlichungen, Berichten, Flyern u.ä. Diese Materialien auszuwerten gestaltete sich wesentlich aufwendiger, war aber aufgrund des reduzierten Schemas der Excel-Tabelle gut zu leisten. Von den Webseiten der Institutionen aus konnte darüber hinaus Hinweisen zu Kooperationspartnern gefolgt werden, um Informationen zu weiteren Daten zu erhalten.

In der beschriebenen Art und Weise wurden die Webseiten der FDZ, der Bundeseinrichtungen und der statistischen Ämter evaluiert. Für die Recherche nach Forschungsdaten, die in den unzähligen Instituten der universitären und außeruniversitären Forschung erhoben wurden, wurde ein projektorientierter Rechercheweg gewählt. Hierzu wurde auf die Informationen zu empirischen Forschungsprojekten in der SOFIS-Datenbank zurückgegriffen (s. o. 2.1). Gleichzeitig wurde die Excel-Tabelle zur Erfassung der Datennachweise entsprechend des aktualisierten da|ra-Metadatenschemas um einige Felder erweitert (s. Tabelle 2).

Tabelle 2: erweiterte Excel-Tabelle für die Recherche

Felder	
Titel	Grundgesamtheit
Weitere Titel	Auswahlverfahren
Typ des weiteren Titels	Referenzzeitraum (formal)
Primärforscher – Person	Referenzzeitraum (frei)
Affiliation	Zeitl. Dimension (kontr.)
Primärforscher – Institution	Zeitl. Dimension (frei)
Datenanbieter	Frequenz
URL	Datenerhebung
Version	Erhebungsverfahren
Sprache	Erhebungsverfahren (frei)
Veröffentlichungsdatum	Typ der Einheiten
Identifier	Anzahl der Einheiten
Typ des Identifiers	Anzahl der Variablen
Klassifikation intern	Typ der Daten
Vokabular Klassifikation intern	Dateiname
Klassifikation extern	Datenformat
Vokabular Klassifikation extern	Größe
Schlagwörter (kontr.)	Verfügbarkeit (kontr.)
Vokabular des Schlagworts	Verfügbarkeit (frei)
Schlagwort (frei)	Rechte
Beschreibung	Relation
Art der Beschreibung	Art der Relation
Geogr. Raum (kontr.)	Typ des Identifiers
Geograph. Raum (frei)	Unstrukturierte Literaturerfassung

Im März 2013 war die Erfassungsmaske der da|ra-Datenbank so weit entwickelt, dass die Informationen aus den Excel-Tabellen dort erfasst werden konnten. Noch fehlende Informationen, die in der ersten Tabelle nicht berücksichtigt worden waren, wurden dabei nachträglich recherchiert. Bei dieser Gelegenheit konnten auch neue Datensätze im Angebot der Datenproduzenten identifiziert werden, zum Beispiel neu veröffentlichte Wellen zu Längsschnittstudien.

3.3 Recherche im Einzelnen

GESIS: FDZ German Microdata Lab und FDZ Internationale Umfrageprogramme

Natürliche Anlaufstelle für nachnutzbare Daten aus den sozialwissenschaftlichen Disziplinen mit Deutschlandbezug ist GESIS. Seit über 50 Jahren archiviert und vertreibt das GESIS Datenarchiv Daten aus sozialwissenschaftlichen Studien, vornehmlich Umfragedaten. Der Archivbestand ist zentral über den Datenbestandskatalog (DBK)[11] recherchierbar, der inzwischen etwa 5.500 Datensätze nachweist. Seit 2010 erhalten alle DBK-Studien durch da|ra eine DOI-Registrierung. Dabei werden die Katalogdaten (Metadaten) aus dem DBK an die da|ra-Datenbank übermittelt, wo sie zusammen mit allen anderen registrierten Studien durchsuchbar sind. Eine erneute Erfassung dieser Einträge war im Rahmen des Aufbaus des Datennachweises folglich nicht notwendig. Allerdings finden sich im GESIS-Angebot einige wenige Daten, die (noch) nicht im DBK nachgewiesen sind sowie ausführliche Dokumentationen

[11] https://dbk.gesis.org/dbksearch/index.asp?db=d.

zu Daten anderer, vornehmlich internationaler Anbieter. Die nicht im DBK berücksichtigten Daten sind im Wesentlichen die des Forschungsdatenzentrums German Microdata Lab (FDZ GML). Die Dokumentationen internationaler Studien finden sich im Serviceangebot des FDZ Internationale Umfrageprogramme. Beide FDZ bieten auf der GESIS-Webseite ausführliche Informationen zu den Daten (s. *Abbildung* ***1***). Der Aufbau des Datennachweises bot nun die Gelegenheit, **diese Informationen auch in formalisierter, standardisierter Weise in einem zentralen Katalog anzubieten**. Aus dem Internetangebot des FDZ GML wurden vor allem ausführliche Dokumentationen zu (Mikro-)Zensusdaten, Europäischen Mikrodaten und DDR-Mikrodaten gewonnen. Die sehr umfangreiche Übersicht international vergleichender Umfrageprogramme des FDZ Internationale Umfrageprogramme bietet Informationen zu über 60 Studien und Studienprogrammen, die ebenfalls für den Datennachweis gesichtet wurden. Viele dieser Studien sind ohnehin bei GESIS archiviert und mussten daher für das Nachweissystem nicht berücksichtigt werden. Die übrigen, nicht im DBK nachgewiesenen Studien wurden für eine Erfassung im Datennachweissystem in die Excel-Liste aufgenommen.

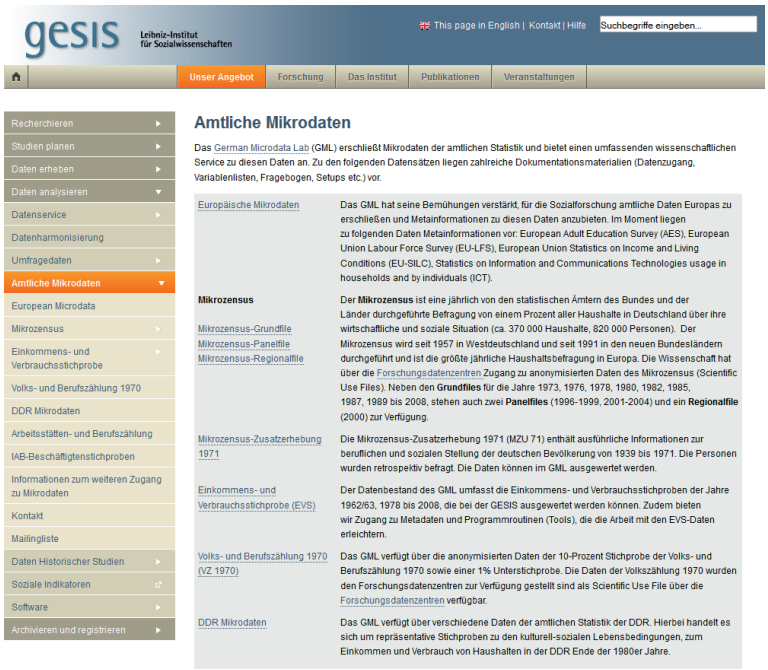

Abbildung 1: Beschreibung der Daten des FDZ GML auf der GESIS-Webseite

RatSWD

Seiner Aufgabe, zur Erschließung und besseren Nutzung von Forschungsdaten beizutragen, kommt der Rat für Sozial- und Wirtschaftsdaten (RatSWD) unter anderem durch die Akkreditierung von Forschungsdatenzentren (FDZ) und Datenservicezentren (DSZ) nach. Auf seiner Webseite beschreibt der RatSWD das Angebot dieser Datenzentren und verweist zur genaueren Recherche von Daten auf deren jeweiliges Internetangebot. Für die Datenrecherche im Rahmen des Datennachweisprojekts wurden die Internetseiten der Datenzentren systematisch nach Beschreibungen der Daten durchsucht. Die vier bei GESIS angesiedelten FDZ konnten dabei unberücksichtigt bleiben, da deren Angebot entweder über den DBK des Datenarchivs bereits in der da|ra-Datenbank nachgewiesen ist oder – im Falle der GML-Daten und der Daten internationaler vergleichender Umfrageprogramme – bereits bei der Evaluation des GESIS-Web-Angebots berücksichtigt wurde. Die Aufbereitung und Darstellung der Informationen der übrigen vom RatSWD akkreditierten Datenzentren variiert stark, jedoch bieten die Seiten **insgesamt ausführliche Informationen, die in der Regel leicht zu finden und sehr strukturiert sind**. Ins-

gesamt konnten bei den Datenzentren etwa 300 relevante Studien identifiziert werden. Die meisten Studien, etwa die Hälfte der Gesamtzahl, fanden sich im Forschungsdatenzentrum des Statistischen Bundesamts (FDZ-Bund). Einige weitere beim FDZ-Bund identifizierte statistische Daten wurden allerdings zunächst nicht in das Datennachweissystem aufgenommen, weil die Beschreibung zum Beispiel von reinen Tabellenbänden auf Grundlage des da|ra-Metadatenschemas nicht optimal möglich war. Durch eine besonders ausführliche Beschreibung der Daten tat sich beispielsweise das Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (IAB) hervor.

Schmollers Jahrbuch, European Data Watch und Working Paper Series

Neben seiner Aufgabe in Bezug auf die Etablierung von Datenservicezentren gibt der RatSWD **diverse Schriften** heraus, die ebenfalls eine interessante Quelle für Daten der empirischen Sozialforschung darstellen. Namentlich handelt es sich um die Zeitschrift Schmollers Jahrbuch, die darin enthaltene Publikation European Data Watch und die RatSWD Working Paper Series, die allesamt im Hinblick auf interessante Datenbestände evaluiert wurden. Die Recherche in den Artikeln von Schmollers Jahrbuch erwies sich als wenig ergiebig, da es sich bei den ermittelten Studien entweder um Sekundäranalysen oder bereits in der da|ra-Datenbank über die Registrierung erfasste Daten handelte. Wenig überraschend war die Recherche im Bereich European Data Watch erfolgreicher, da hier gezielt neue Datenquellen für Forschung und Lehre besprochen werden. Daraus konnten einige für das Datennachweissystem relevante Datensätze identifiziert werden. Die Working Paper Series erwies sich dagegen als wenig ergiebig, da hier kaum eigene Erhebungen behandelt werden.

Weitere nationale und internationale Institutionen

Von Experten des GESIS Datenarchivs wurde bereits im Vorfeld der Recherchearbeit eine Liste derjenigen internationalen und in Deutschland ansässigen Institutionen angelegt, **die erfahrungsgemäß über Datenbestände mit Bezug zu Deutschland verfügen**. Diese Liste schloss auch die durch den RatSWD akkreditierten Datenzentren ein, deren Bestände bereits evaluiert wurden. Darüber hinaus enthielt die Zusammenstellung viele Hinweise auf kommerzielle Markt- und Meinungsforschungsinstitute, deren Datenangebot allein über die öffentlich zugänglichen Informationen auf den Webseiten jedoch schwer beurteilt werden konnte; häufig blieben essentielle Informationen zur Verfügbarkeit der Daten, zu Primärforschern oder zur Datenerhebung unklar. Es wurde daraufhin entschieden, zu diesem Zeitpunkt keine weiteren Recherchen in Richtung kommerzieller Anbieter anzustrengen. Da bei diesen Instituten jedoch ohne Zweifel interessante Datenbestände vorliegen, besteht innerhalb des Projekts auch weiterhin Interesse, auch diese Daten nachzuweisen – allerdings wird hierzu wohl eine andere, kooperative Vorgehensweise notwendig sein. Auch die ermittelten internationalen Institutionen boten bei der Recherche ein interessantes Angebot, allerdings waren die Einrichtungen entweder schon durch die Registrierung der Daten über da|ra in der Datenbank erfasst oder boten keine Daten mit Deutschlandbezug.

Behörden und Institutionen des Bundes

Zahlreiche Behörden und Institutionen des Bundes, wie zum Beispiel die Ressortforschungseinrichtungen der Bundesministerien, erfüllen Forschungs- und Entwicklungsaufgaben und erheben im Rahmen ihrer Arbeit Daten. Haben diese Einrichtungen einen sozialwissenschaftlichen Bezug, sind sie eine weitere interessante Datenquelle für das Datennachweissystem. Um das Angebot der zahlreichen Institutionen auszuwerten, wurde zunächst auf Basis der Webseiten des Bundes[12] eine Übersicht der relevanten Bundesanstalten, -ämter und -institute erstellt. Direkt aussortiert wurden vollkommen fachfremde

[12] http://www.bund.de/DE/Behoerden/behoerden_node.html.

Einrichtungen; die übrigen wurden in alphabetischer Reihenfolge systematisch aufgrund ihres Internetangebots beurteilt. Aus insgesamt 900 Institutionen, zu denen zum Beispiel auch politische Stiftungen gehören, konnten schließlich 150 Einrichtungen mit sozial- oder wirtschaftswissenschaftlicher Ausrichtung identifiziert werden; bei ca. 50 Einrichtungen wurden Hinweise auf empirische Studien gefunden. Die gefundenen Informationen wurden in einer eigenen Excel-Tabelle dokumentiert. Es zeigte sich, dass die Ressortforschungseinrichtungen zahlreiche relevante empirische Studien durchführen, **zu den Daten allerdings in der Regel keinen direkten Zugang anbieten** bzw. keine Angaben zur Verfügbarkeit der Daten machen. Die Informationen zu den Studien, Daten und Methoden sind jedoch in verschiedensten zum Download zur Verfügung stehenden Veröffentlichungen (Ergebnisberichte, Tabellenbände etc.) ausführlich beschrieben. Für das Datennachweissystem wurde daher entschieden, auf das Datenangebot der Ressortforschungseinrichtungen auf der Grundlage dieser Veröffentlichungen zu verweisen. Die Informationen erscheinen zu nützlich und wertvoll, als dass auf eine Aufnahme in das Datennachweissystem mangels direkten Datenzugangs verzichtet werden sollte. Hinzukommt die Problematik der Dynamik von Webangeboten, die befürchten lässt, dass nicht datenbankgebundene Information und Dokumentation nach einiger Zeit verloren geht. Bei den Recherchen wurde insbesondere eine **überwältigende Anzahl an Statistiken** identifiziert, die zwar inhaltlich interessant sind, aus Kapazitätsgründen aber systematisch unberücksichtigt bleiben mussten (hierzu zählen vor allem Statistiken des Bundeskriminalamtes und des Bundesinstituts für Bevölkerungsforschung). Ein Nachweis dieser Daten zu einem späteren Zeitpunkt und mit einem automatisierten Verfahren bleibt aber von Interesse. Insgesamt erwies sich die Suche nach Studien auf den Webseiten der meisten Institutionen als relativ unproblematisch, da die Seiten durchweg übersichtlich gestaltet und viele relevante Publikationen verlinkt waren oder direkt zum Download zur Verfügung standen. Die Zusammenstellung von Metadaten zu den Studien war jedoch insgesamt aufwendig, da die Informationen häufig aus Broschüren, Informationsblättern, (Jahres-)Berichten oder Flyern zusammengetragen werden mussten. Bei etwa einem Drittel der Institute gestaltete sich die Suche eher schwierig, da weder Studien- noch Publikationsübersichten zur Verfügung standen (dies war zum Beispiel bei der Antidiskriminierungsstelle des Bundes der Fall, beim Bundesamt für Bevölkerungsschutz und Katastrophenhilfe, beim Bundesamt für Migration und Flüchtlinge und beim Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz). In diesen Fällen wurden anhand der Webseitensuche mit Stichworten wie „empirisch", „Erhebung", „quantitativ" oder „sozial*" nach Daten, Studien oder Publikationen gesucht. Diese unsystematische Herangehensweise machte es geradezu unmöglich einzuschätzen, wie viele Studien vorhanden sind und ob sie für das Datennachweissystem von Bedeutung sind. Einen besonderen Recherchefall stellte das Bundesinstitut für Sportwissenschaften dar. Das Institut betreibt elaborierte Datenbanken zu Forschungsprojekten, Monographien, Zeitschriften und weiteren Medien. Eine gezielte Suche nach Daten war hier allerdings nicht ohne weiteres möglich. Allein die Suche mit dem Stichwort „empirisch" lieferte 118 Treffer, was einen relevanten Forschungsdatenbestand vermuten lässt. Allerdings stellte sich bei der weiteren Untersuchung dieser Einträge heraus, dass durch die thematische Ausrichtung (Sport) viele Studien für ein sozialwissenschaftliches Datennachweissystem eher nicht infrage kommen. Hilfreich war es daher, bei der Suche nach Daten auf die angebotenen Publikationen zurückzugreifen, woraus schließlich 10 Studien für das Datennachweissystem identifiziert werden konnten. Ein ähnlich umfangreiches Datenangebot, allerdings in Form von Statistiken und Tabellenbänden, boten das Bundesinstitut für Bevölkerungsforschung, das Bundeskriminalamt und das WSI Gender Datenportal der Hans-Böckler-Stiftung an, die allerdings sowohl wegen der Quantität als auch wegen der Art der Daten nicht in das Datennachweissystem aufgenommen wurden.

Aus den Einrichtungen des Bundes wurden insgesamt etwa 250 Studien in das Datennachweissystem eingebracht.

SOFIS

Die bislang letzte Informationsquelle für den Aufbau des Datennachweissystems war das von GESIS betriebene Sozialwissenschaftliche Informationssystem (SOFIS), in dem seit den 1980er Jahren Projekte der sozialwissenschaftlichen universitären und außeruniversitären Forschung dokumentiert werden (s. *Abbildung **2***).

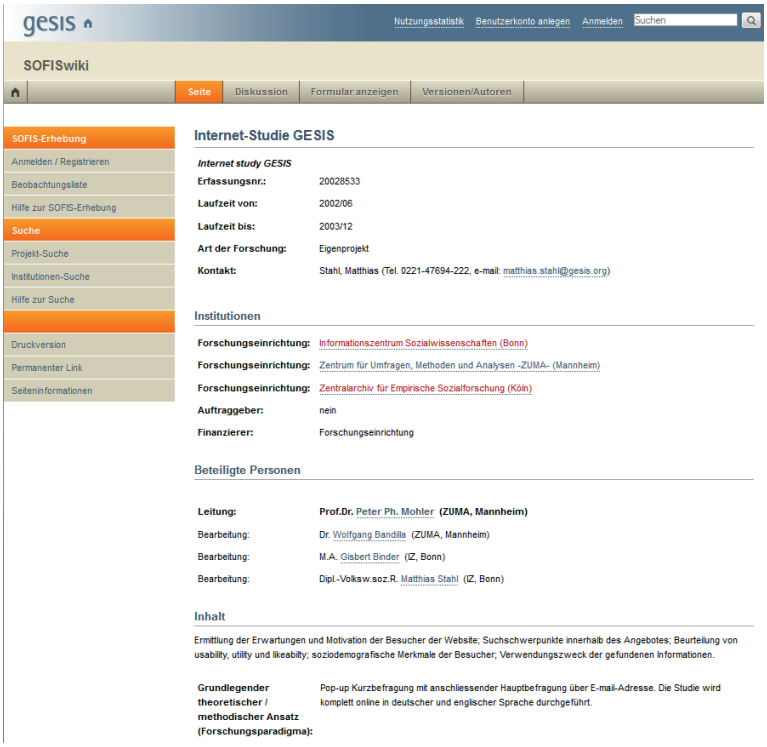

Abbildung 2: Beispiel einer Projektbeschreibung in SOFIS

Auch kleine Institute, Forschungsgruppen und Einzelforscher können hier ihre Projekte und Vorhaben eintragen und so auf ihre Arbeit hinweisen. Da in SOFIS empirische Forschung aus der ganzen Bandbreite der Forschungslandschaft repräsentiert ist, bietet diese Datenbank für den Aufbau eines zentralen Datennachweissystems eine Anlaufstelle von unschätzbarem Wert. **Daten empirischer Forschungsprojekte universitärer und außeruniversitärer Institute** könnten aufgrund ihrer Masse und verteilten sowie heterogenen Dokumentation ansonsten nicht berücksichtigt werden. Primäre Herausforderung bei der Evaluation der SOFIS-Datenbank war die mit über 50.000 Projekten erwartungsgemäß hohe Zahl von Einträgen. Da die Datenbank auch theoretische Projekte und qualitative Forschung nachweist, wurde zunächst ein Ausschluss dieser Einträge angestrebt. Es wurde daher ein Datenbankexport aller Einträge durchgeführt, die im Feld „Methode" die Angabe „empirisch-quantitativ" enthielten. Zudem wurde eine zeitliche Grenze gezogen, sodass nur Studien die seit dem Jahr 2000 durchgeführt wurden, in den Export eingeflossen sind. Wie sich im weiteren Verlauf der Evaluation zeigte, war insbesondere die zeitliche Einschränkung sinnvoll, da weiterführende Informationen zu Studien, die vor über 10 Jahren durchgeführt wurden, im Internet nur noch in Einzelfällen zu finden sind. Aus dem Datenbank-Export wurde eine Excel-Liste erstellt, die trotz der Einschränkungen immer noch 11.000 Projekte enthielt. Eine kritische Prüfung der Projekte in der Liste wurde im Folgenden von zwei Personen begonnen, wobei jeweils weitere Kriterien in die Entscheidung für oder gegen eine Aufnahme in das Datennachweissystem angelegt wurden. Insbesondere wurden solche Projekte direkt gestrichen, die sich nicht auf Deutschland bezogen oder deren fachliche Ausrichtung nur peripher sozialwissenschaftlich war. Auch reine Sekundäranalysen konnten identifiziert und von der Liste ausgeschlossen

werden. Diese Kriterien konnten mangels entsprechender Dokumentation leider nicht automatisiert beim Export aus der SOFIS-Datenbank berücksichtigt werden. Zudem fehlten in den exportierten Einträgen oftmals wesentliche Informationen, die aufgrund der Pflichtfeldvorgabe für die Aufnahme in das Datennachweissystem essentiell sind – ganz abgesehen von der zentralen Information über die Verfügbarkeit der Daten. Die Vorgehensweise bei der daher notwendigen Prüfung der Projekte bestand im Wesentlichen aus Internetrecherchen, wobei als primäre Anhaltspunkte die in SOFIS angegebenen Forschungsinstitute und Primärforscher herangezogen wurden. Bei der Prüfung stellte sich unter anderem heraus, dass viele Studien, die von Forschern an SOFIS als „empirisch-quantitativ" gemeldet worden waren, tatsächlich auf qualitativen Methoden beruhten, weshalb sie für das Datennachweissystem schließlich doch nicht infrage kamen. Die Recherchen waren wesentlich aufwendiger als alle vorhergehenden, da hier eine beispiellose Heterogenität in der Informationsaufbereitung vorlag und vor allem zu länger zurückliegenden Projekten im Internet keine Informationen mehr zu finden waren. In der Zeit, die für den Aufbau des Datennachweissystems zu Verfügung stand, konnte die Liste der 11.000 Projekte auf diese Weise nicht annähernd vollständig evaluiert werden. Lediglich 800 Einträge konnten bislang nachrecherchiert werden. Von diesen erwiesen sich allerdings mehr als die Hälfte der Projekte (ca. 420) als relevant, weshalb sie in das Datennachweissystem aufgenommen wurden.

3.4 Erfassung der Nachweise im Datennachweissystem

Insgesamt konnten in der Projektlaufzeit über 1110 relevante Studien identifiziert und in der da|ra-Datenbank erfasst werden. Die Erfassung der Nachweise erfolgte in **zwei Schritten**. Für eine recherchierte Studie wurde im Datennachweissystem zuerst ein Eintrag anhand der recherchierten und in der Excel-Tabelle erfassten Pflichtfelder angelegt (s. *Abbildung **3***). Weitere optionale Informationen aus der Excel-Tabelle wurden, soweit vorhanden, ergänzt. Da die Recherche bereits einige Zeit zurücklag, wurde das Datenangebot auch noch einmal im Internet überprüft, woraus sich in vielen Fällen noch einmal Änderungen und Ergänzungen ergaben. Im Falle von Längsschnittstudien waren beispielsweise nicht selten weitere Erhebungswellen hinzugekommen. In einem zweiten Schritt wurden die angelegten Einträge jeweils noch einmal anhand von Qualitätskriterien überprüft und nach evtl. notwendigen Korrekturen online gestellt.

Die Prüfkriterien zielen im Sinne einer optimalen Durchsuchbarkeit und Darstellung der Datenbankinhalte auf eine **möglichst große Homogenität** der Einträge ab. Dies bedeutet vor allem, dass bei der Erfassung auf **größtmögliche Standardisierung** zu achten ist. Als interne Regel wurde in diesem Zusammenhang festgelegt, dass Angaben in kontrollierter Form der Verwendung der Freitextfelder immer vorzuziehen sind. Beispielsweise sollte bei der Erfassung des geographischen Raums (Feld 18 im Metadatenschema, vgl. Hausstein et al. 2014) bevorzugt aus der kontrollierten Liste gemäß ISO 3166-2/3 (Feld 18.1) gewählt werden. Nur, falls eine weitere Spezifizierung notwendig ist (etwa „ohne West-Berlin") kann diese im Freitextfeld (18.2) ergänzend vorgenommen werden. Ähnliche Regeln wurden für die anderen Felder vereinbart, die sowohl kontrollierte, als auch freie Einträge erlauben. Eine weitere Regelung betrifft die Erfassung von Personennamen und Institutionenbezeichnungen. Hier wurde vereinbart, im Sinne einer einheitlichen Schreibweise über die Datennachweise hinweg auf die Ansetzung der von der Deutschen Nationalbibliothek (DNB) geführten Gemeinsamen Normdatei (GND) zurückzugreifen. Im Falle der Institutionenbezeichnungen wurden außerdem gängige – und damit für die Suche relevante – Abkürzungen in Klammern ergänzt, zum Beispiel „Institut für Angewandte Sozialwissenschaft (INFAS)". Schließlich wurde beschlossen, die persistenten Identifikatoren der jeweiligen GND-Einträge mit zu erfassen, um eine künftige Verknüpfung mit den GND-Daten zu erleichtern. Grundsätzlich wurde auch vereinbart, **möglichst ausführliche Beschreibungen** zu erstellen, d.h. möglichst alle auf der Webseite oder in Veröffentlichungen des Datenanbieters zur Verfügung stehenden Informationen auch zu erfassen. In Bezug auf die inhaltliche Beschreibung sollten zu allen erfassten Studien mindestens eine Klasse aus der ZA-Klassifikation und einige Schlagwörter aus dem Thesaurus

Sozialwissenschaften vergeben werden. Hintergrund dieser ausführlichen Beschreibung ist die Tatsache, dass die im Datennachweissystem erfassten Studien in der Regel nicht in anderen Katalogen detailliert beschrieben und durchsuchbar sind. Durch die standardisierte, ausführliche Beschreibung im da|ra-Datennachweissystem werden viele Studien erst systematisch recherchierbar. Außerdem erleichtert eine weitestgehende Standardisierung die Einbindung der Inhalte in das Semantic Web und ermöglicht damit eine künftige Integration unserer Datenbank mit ähnlichen Diensten. Ein besonderes Anliegen war in diesem Zusammenhang die Aussicht auf eine bessere Verknüpfung von Forschungsdaten und Forschungsliteratur. Um diesem entgegenzukommen, wurden bei der Erfassung auch auf den Webseiten angegebene Primärpublikationen im dafür vorgesehenen Metadatenbereich (32.1 Strukturierte Erfassung von Publikationen) eingetragen.

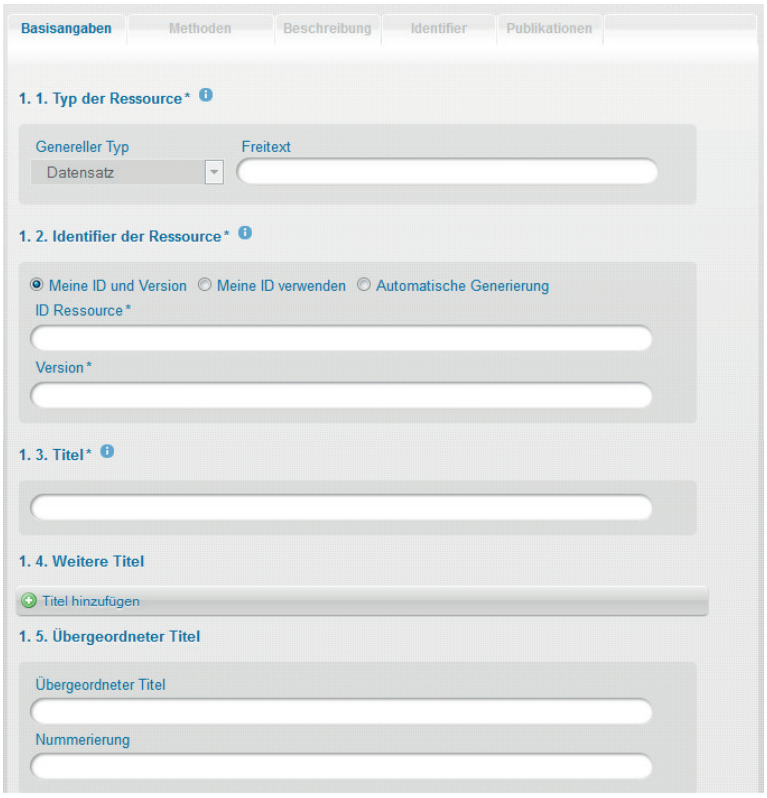

Abbildung 3: Ausschnitt aus der da|ra-Erfassungsmaske

Während der Erfassung wurden das Metadatenschema und die Erfassungsmaske kontinuierlich weiterentwickelt. Die manuelle Erfassung der Datennachweise diente bei diesen Weiterentwicklungen auch als Test. Durch diverse Herausforderungen, die während der Erfassung der Datennachweise auftraten, konnten einige technische Probleme identifiziert und bearbeitet sowie das Metadatenschema angepasst werden.

4 Herausforderungen und *lessons learned*

Die Recherche und Zusammenstellung der Forschungsdaten sowie die Erfassung der Metadaten im Datennachweissystem haben zu wertvollen Einsichten geführt, sowohl im Hinblick auf die zu erwartenden Daten und Quellen, als auch auf die technische Umsetzung der Erfassung. Im Folgenden werden spezifische Herausforderungen genauer beschrieben und Erkenntnisse sowie Lösungsansätze als *lessons learned* beschrieben.

4.1 Heterogene und fehlende Informationen

Die beschriebene Heterogenität der Informationsangebote macht vor allem deutlich, wie wichtig die Arbeit an einem zentralen Datennachweissystem ist. So haben die Recherchen einerseits erahnen lassen, wie umfangreich der Bestand sozialwissenschaftlich relevanter Forschungsdaten ist, und andererseits gezeigt, dass **jedes Institut eine eigene Vorgehensweise bei der Beschreibung und Vermittlung seiner Daten verfolgt**. Der im Rahmen des Projekts betriebene Aufwand zur Integration der Informationen zu unterschiedlichen Datenbeständen ist eben jener Aufwand, den Forscher derzeit bewältigen müssen, wenn sie auf der Suche nach geeigneten Daten für ihre Forschung sind. Neben der Heterogenität in der Darstellung der Datenbestände sind vor allem die eingeschränkte Dokumentation und fehlende Informationen ein Problem für interessierte Nutzer. Eine in diesem Zusammenhang stets wiederkehrende Situation ist das Fehlen von Informationen zur Verfügbarkeit und Nachnutzbarkeit der Daten. In der Praxis führte dies beim Aufbau des Datennachweissystems dazu, dass viele Studien zwar beschrieben, aber nur mit dem Vermerk „Verfügbarkeit: Unbekannt" aufgenommen werden konnten.

Besonders gut und ausführlich sind in der Regel die Daten der vom RatSWD akkreditierten Datenzentren beschrieben. Da die Datenzentren die Aufgabe erfüllen, Daten aufzubereiten und **zur Nachnutzung zur Verfügung zu stellen**, ist dies nicht weiter verwunderlich. Positiv fielen darüber hinaus auch einige Einrichtungen des Bundes auf, etwa waren die Informationen zu Studien des Bundesministeriums für Bildung und Forschung (BMBF) und des Bundesministeriums für Familie, Senioren, Frauen und Jugend (BMFSFJ) besonders transparent und übersichtlich beschrieben. Je kleiner jedoch die Bundesinstitute sind, desto dürftiger fällt die Information zu den dort vorhandenen Forschungsdaten aus. Kleine Institutionen führen insgesamt auch seltener eigene empirische Studien durch und verfügen über weniger ausführliche Internetseiten. Generell sind die Webseiten der Bundesinstitute – anders als die Internetangebote der FDZ – nicht in erster Linie auf die Vermittlung von Informationen an Sekundärnutzer von Forschungsdaten ausgerichtet, sondern auf das Informationsbedürfnis der Bürger. Deshalb ist es sehr einfach, dort allgemeinverständlich aufbereitetes Material zu Studienergebnissen zu finden, weniger aber Detailangaben zu den erhobenen Daten.

Die größte Heterogenität im Informationsangebot fand sich erwartungsgemäß bei der Auswertung der SOFIS-Datenbank. Die Internetrecherche nach dort eingetragenen Projekten führte zu den unterschiedlichsten Webseiten mit jeweils eigener Informationsdarstellung. Eine wesentliche Erkenntnis aus der Auswertung der SOFIS-Daten war, dass hier nicht nur **verbreitet Informationen fehlten**, sondern teilweise im System auch irreführend angegeben waren; insbesondere wurden zahlreiche Projekte in SOFIS als „empirisch-quantitativ" dargestellt, obwohl, wie sich nach Recherchen herausstellte, tatsächlich keine quantitativen Methoden angewandt wurden. Vielfach gab es auch Unstimmigkeiten innerhalb der in der Projektbeschreibung enthaltenen Informationen, zum Beispiel zwischen der Studienbeschreibung und den nachfolgenden Angaben zu Erhebungsmethoden und der zeitlichen Dimension. Im Hinblick auf eine künftig wünschenswerte automatisierte Integration von Informationen anderer Datenbanken in das Datennachweissystem bedeutet dies, dass selbst anhand von Metadatenschemata und kontrolliertem Vokabular erfasste Informationen nur bedingt verlässlich sind.

4.2 Datenarten, Methodik und Fachgebiete

Wie bereits beschrieben, wurden bei den Recherchen neben Umfragedaten bei einigen Institutionen auch weitere interessante Forschungsdaten gefunden. Aufgrund ihrer Menge fielen hier insbesondere Statistiken des Bundeskriminalamtes und des Bundesinstituts für Bevölkerungsforschung auf, jedoch auch Daten aus dem Statistischen Bundesamt. Prinzipiell ist die **Erfassung verschiedener Datenarten** im Datennachweissystem angelegt, so kann im Freitextfeld 26.4 „Typ der Daten" nach Belieben die Datenart spezifiziert werden. Allerdings stellte sich bei der Recherche und Erfassung der Statistiken heraus, dass die weiteren Metadatenfelder, insbesondere die zur Spezifikation des Datensatzes (Feld 26), für eine vergleichbar detaillierte Beschreibung von statistischen Daten (z.B. in Form von Tabellenbänden) wenig geeignet sind. Unter Berücksichtigung der Masse an gefundenen Statistiken erschien die Option, diese mit den eingeschränkten Möglichkeiten des Metadatenschemas zu erfassen, zu diesem Zeitpunkt nicht als angemessene Herangehensweise. Kollektionen dieses Ausmaßes sollten in vergleichbarer Qualität wie die Umfragedaten erfasst werden können, weshalb auf die Erfassung eines Großteils der identifizierten Statistiken zunächst verzichtet wurde. Inzwischen wurde das da|ra-Metadatenschema weiterentwickelt, unter anderem besteht jetzt im Rahmen der DOI-Registrierung die Möglichkeit, auch andere Ressourcentypen als Datensätze in der Datenbank zu beschreiben. Die neu hinzugekommenen Ressourcentypen sind Sammlung, Text, Video, Bild, Audio und Interaktive Ressource. Entsprechend wurden auch andere Felder im Metadatenschema ergänzt, die eine genaue Beschreibung dieser Ressourcentypen ermöglichen. Für eine (vielleicht auch automatisierte) Aufnahme von Statistiken im großen Stil ist eine entsprechende Anpassung des Metadatenschemas denkbar.

Prinzipiell erlaubt das Metadatenschema darüber hinaus auch die Erfassung von qualitativen Studien oder von Forschungsdaten, die **mit anderen in den Sozialwissenschaften angewandten Methoden** erhoben wurden, zum Beispiel Daten aus Simulationen, Beobachtungen oder Experimenten (s. kontrolliertes Vokabular zu den Erhebungsmethoden in Hausstein et al. 2014, 27-29). Die Gesamtausrichtung des Metadatenschemas begünstigt jedoch nach wie vor die Erfassung empirisch-quantitativer Forschungsdaten. Bei den Recherchen zum Aufbau des Datennachweissystems wurde daher an diesem ursprünglich aus pragmatischen Gründen aufgestellten Kriterium festgehalten. In Zukunft ist allerdings auch hier eine Weiterentwicklung des Metadatenschemas denkbar, zum Beispiel in einem ersten Schritt durch eine Erweiterung des kontrollierten Vokabulars zu den Erhebungsmethoden in Anlehnung an das Vokabular der Data Documentation Initiative (DDI).[13] Außerdem besteht das Vorhaben, einen Versuch des Nachweises von Social-Media-Daten zu starten.

Eine weitere Herausforderung zeigte sich in der **Beschränkung des Datennachweissystems auf die sozialwissenschaftlichen Disziplinen**. Die Frage nach der inhaltlichen Relevanz stellte sich während der Recherche immer wieder. Wie andere Disziplinbegriffe, ist auch der der Sozialwissenschaften nicht exakt abgrenzbar und Entwicklungen unterworfen. Gemeinsam ist den zahlreichen Subdisziplinen das menschliche Zusammenleben als Forschungsschwerpunkt (Quandt/Mauer 2012, 61). Anhand dieses Kriteriums eine Entscheidung zur Aufnahme oder Ablehnung von Studien zu fällen, war nicht immer einfach. Um die Entscheidungsprozesse besser zu kontrollieren, wurde daher entschieden, die Klassifikation Sozialwissenschaften[14] bei der Beurteilung der Relevanz als Hilfsmittel heranzuziehen (s. *Abbildung 4*). Konnte eine Studie nicht oder nur schwer innerhalb der Klassifikation verortet werden, wurde auf eine Aufnahme in das Datennachweissystem verzichtet. Besonders in den Bereichen Betriebswirtschaftslehre, Medizin und Psychologie waren die Forschungsfragen oft so spezifisch, dass die Nähe zu sozialwissenschaftlichen Forschungsfeldern nicht mehr erkennbar war.

[13] http://www.ddialliance.org/Specification/DDI-CV/ModeOfCollection_1.0.html.

[14] http://www.gesis.org/unser-angebot/recherchieren/thesauri-und-klassifikationen/klassifikation-sozialwissenschaften/.

11001	**Allgemeines, spezielle Theorien und "Schulen", Methoden, Entwicklung und Geschichte der Sozialpolitik** Basic Research, General Concepts and History of Social Policy
11002	**Lehre und Studium, Professionalisierung und Ethik, Organisationen und Verbände der Sozialpolitik** Training, Teaching and Studying, Professional Organizations of Social Policy
11003	**soziale Sicherung** Social Security *Beispiele:* Sozialversicherung allgemein, Krankenversicherung, Pflegeversicherung, Unfallversicherung, Arbeitslosenversicherung, Rentenversicherung, Beamtenversorgung, Kriegsopferversorgung
11004	**Einkommenspolitik, Lohnpolitik, Tarifpolitik, Vermögenspolitik** Income Policy, Property Policy, Wage Policy
11005	**Arbeitswelt** Working Conditions *Beispiele:* Arbeitsbedingungen, Arbeitsschutz, Arbeitssicherheit
11006	**Gesundheitspolitik** Health Policy *Beispiele:* Gesundheit und Krankheit, öffentliches und privates Gesundheitswesen, Prävention im Gesundheitswesen, Rehabilitation
11007	**Familienpolitik, Jugendpolitik, Altenpolitik** Family Policy, Youth Policy, Policy on the Elderly
11099	**Sonstiges zur Sozialpolitik** Other Fields of Social Policy *Beispiele:* Trägerschaften spezieller Maßnahmen, Wohlfahrtsverbände
20000	**Interdisziplinäre und angewandte Gebiete der Sozialwissenschaften** Interdisciplinary and Applied Fields of the Social Sciences
20100	**Arbeitsmarkt- und Berufsforschung** Employment Research
20101	**Arbeitsmarktforschung** Labor Market Research
20102	**Berufsforschung, Berufssoziologie** Occupational Research, Occupational Sociology *Beispiele:* Berufstheorie, Berufsfelder und einzelne Berufe, Berufswahl, Qualifizierung und Professionalisierung, Berufsverlauf, Berufsberatung, berufliche Rehabilitation
20103	**Arbeitsmarktpolitik** Labor Market Policy
20200	**Frauen- und Geschlechterforschung** Women's Studies, Feminist Studies, Gender Studies

Abbildung 4: Ausschnitt aus der Klassifikation Sozialwissenschaften

4.3 Aktualität

Die Dezentralität und Dynamik des Internets führen zwangsläufig dazu, dass Webinformationen einem kontinuierlichen Veränderungsprozess unterliegen. Diese grundlegende Flexibilität bietet unbestrittene Vorteile gegenüber statischen Medien, führt aber andererseits auch zu permanentem Informationsverlust. Das Internet ist ein flüchtiges Medium, das in manchen Bereichen besser und in anderen Bereichen weniger gut Information bewahrt (Koehler 2004). Der ephemere Charakter des Webs führt nicht zuletzt zu dem bereits lange diskutierten und vielfach untersuchten **Phänomen des Linksterbens** (ebda.). Studien haben gezeigt, dass zum einen URLs bereits nach kurzer Zeit nicht mehr abrufbar und zum anderen auch bei persistierenden URLs wechselnde Webseiteninhalte eher die Regel als die Ausnahme sind (ebda.). Beide Probleme spielten auch bei der Recherche für das Datennachweissystem eine große Rolle. Insbesondere bei der Auswertung der SOFIS-Datenbank wurde festgestellt, dass dort angegebene Links zu Institutionen oder Projekten entweder in die Leere oder zu ganz anderen Inhalten führten. Erwartungsgemäß waren tote Links umso wahrscheinlicher, je älter die SOFIS-Einträge waren, aber selbst bei den jüngeren Projekten trat dieses Phänomen auf. Durch Internetrecherchen konnte in einigen dieser Fälle Informationen an anderen Orten gefunden werden; je länger die Projekte zurück-

lagen, desto unergiebiger waren jedoch auch diese Bemühungen. Da der Anspruch eines zentralen, integrierten Nachweissystems für Forschungsdaten auch das Kriterium der Aktualität beinhaltet, wurden tote Links aus SOFIS nicht in den Datennachweis übernommen. Stattdessen wurden, wenn vorhanden, nachrecherchierte Links aufgenommen oder, falls solche nicht gefunden werden konnten, auf den jeweiligen Eintrag in der SOFIS-Datenbank verlinkt. Die Information, dass Primärforscher X zu einem Zeitpunkt Y die Studie Z durchgeführt hat, kann so zwar recherchierbar gemacht werden; für an den Daten tatsächlich interessierte Personen wird sie jedoch kaum ausreichen.

Das Problem der gefährdeten Aktualität stellt sich auch in anderen Bereichen des Datennachweissystems. So werden auch im Datennachweissystem Links angelegt, die – anders als DOI-Namen – nicht persistent sind. Das heißt, selbst die nachrecherchierten Links zu den SOFIS-Einträgen werden eventuell früher oder später dem Linksterben zum Opfer fallen, genauso wie alle anderen Links zu Datenangeboten. Es ist daher geplant, in einer weiteren Entwicklungsphase einen **Linkchecker** zu implementieren, der tote Links in der Datenbank aufdeckt. Außerdem sollen Mechanismen entwickelt werden, die auf wesentliche Veränderungen im Datenangebot hinweisen. In erster Linie sind davon Längsschnittstudien betroffen, zu denen bei der Erfassung immer nur die Wellen berücksichtigt werden konnten, die bereits gelaufen waren und zu denen folglich Informationen abrufbar waren. Für diese Studien soll daher eine Routine programmiert werden, die die Redaktion des Datennachweissystems benachrichtigt, wenn eine neue Welle vorliegen könnte.

Ein weiteres Aktualitätsproblem stellt sich in Bezug auf Personennamen und Institutionsbezeichnungen. Zwar wird hier eine kontrollierte Ansetzung in Orientierung an der GND vorgenommen (s.o. 3.4); da aber **keine direkte Einbindung der GND** als kontrolliertes Vokabular in das System besteht, gibt es derzeit keine Möglichkeit, Datennachweise von Primärforschern oder Institutionen, deren Namen sich über die Zeit geändert haben, im Retrieval zusammenzufassen.

4.4 Verfügbarkeit

Das Datennachweissystem wurde aufgebaut, um eine zentrale Recherchemöglichkeit sozialwissenschaftlicher Forschungsdaten zu ermöglichen. Dabei ist es grundsätzlich auch wünschenswert, Datensätze nachzuweisen, deren Verbleib oder Verfügbarkeit unklar ist. So können sich Interessierte beispielsweise Klarheit darüber verschaffen, welche Wellen einer Längsschnittstudie tatsächlich erhoben wurden, **auch wenn möglicherweise einige davon nicht verfügbar sind**. Beim Aufbau des Datennachweissystems war es nicht immer einfach, die Verfügbarkeit von Daten anhand der im Internet vorhandenen Informationen zu bestimmen. Im Falle der FDZ und anderer Institutionen, die Daten offensiv vertreiben, trat dieses Problem zwar nicht auf, umso stärker jedoch bei den aus der SOFIS-Datenbank identifizierten Forschungsprojekten. Es muss wohl davon ausgegangen werden, dass die Daten zu diesen Projekten in der Regel nicht zur Sekundäranalyse zur Verfügung stehen. Für Interessierte ist es dennoch vorteilhaft zu erfahren, in welchen Projekten relevante Daten erhoben wurden - möglicherweise kann auch auf dem Weg einer persönlichen Kontaktaufnahme mit den Projektverantwortlichen ein Datenzugang zustande kommen. Vor diesem Hintergrund ist das Datennachweissystem in erster Linie ein umfassendes Informationssystem zu in Deutschland erhobenen Daten und weniger ein Datenzugangskanal.

4.5 Registrierte und nicht-registrierte Studien in einer Datenbank

Wie eingangs erläutert, wurde die da|ra-Datenbank zunächst zum Zweck der DOI-Registrierung entwickelt. Der Aufbau eines zentralen Suchdienstes stand zu diesem Zeitpunkt nicht im Vordergrund der Entwicklung. Folglich orientiert sich das Metadatenschema mit seinen Pflichtfeldern an den **Erforder-**

nissen der DOI-Registrierung. Dies bedeutet zum Beispiel, dass die Metadatenelemente „Veröffentlichungsdatum", „Publikationsagent", „Version" und „DOI" Pflichtangaben für Einträge in der Datenbank sind; diese Angaben sind jedoch für Nachweise nicht-registrierter Daten entweder irrelevant (Veröffentlichungsdatum), schwer ermittelbar (Version) oder nicht anwendbar (Publikationsagent, DOI). Für den Nachweis nicht-registrierter Daten wurde daher ein leicht modifiziertes Metadatenschema entwickelt, wobei sich die Modifikationen im Sinne der Homogenität der Datenbank auf Details beschränkten. Tabelle 3 zeigt die Pflichtfelder für registrierte und nicht-registrierte Ressourcen in einer Gegenüberstellung.

Tabelle 3: Pflichtfelder des da|ra-Metadatenschemas

Pflichfelder mit DOI-Registrierung	Pflichtfelder ohne DOI-Registrierung
Genereller Ressourcentyp	**n. a. (ist immer Datensatz)**
Titel	**Titel**
Creator (Person/Institution)	**Creator (Person/Institution)**
Publikationsagent	**n. a.**
Registrierungsagentur	**n. a.**
DOI	**n. a.**
URL	**URL**
Version	**Version**
Veröffentlichungsdatum	**Veröffentlichungsdatum**
Verfügbarkeit	**Verfügbarkeit**

Die für nicht-registrierte Studien nicht anwendbaren Felder („n. a." in der Tabelle) konnten für die Erfassung der Metadaten deaktiviert werden, weil sie ohnehin vom System generiert werden (Publikationsagent, Registrierungsagentur, DOI) oder eine Default-Einstellung vorgenommen werden konnte (Genereller Ressourcentyp). Dadurch müssen bei der Erfassung von Datennachweisen nur sechs Felder verpflichtend ausgefüllt werden, eventuell sogar nur vier, denn das System generiert bei fehlender Versionsangabe automatisch eine Versionierung und übernimmt bei fehlender Datumsangabe das aktuelle Datum als Veröffentlichungszeitpunkt.

Ein weiteres grundlegendes Problem in Bezug auf das gemeinsame Metadatenschema konnte jedoch bislang nicht gelöst werden: die fehlende Möglichkeit, bei der Erfassung **einen Datenanbieter anzugeben**. Ein Feld „Datenanbieter" war ursprünglich für die da|ra-Datenbank nicht vorgesehen, weil der sogenannte Publikationsagent, der die DOI-Registrierung beauftragt, auch immer gleichzeitig der Datenanbieter ist. Im Falle der Nachweise nicht-registrierter Studien gibt es keinen Publikationsagenten, der die von ihm nachgewiesenen Studien in seinem Bestand vorhält. Stattdessen ist es prinzipiell für jedermann möglich, Datennachweise in das System einzubringen, unabhängig davon, wer die Daten vorhält. In der in diesem Bericht beschriebenen Pilotphase zum Aufbau des Datennachweissystems hat beispielsweise ein Team des GESIS Datenarchivs sämtliche Einträge vorgenommen, ohne jedoch selbst die Verantwortung für die nachgewiesenen Daten zu haben. Während der Publikationsagent im Falle der registrierten Studien also automatisch auch als Datenanbieter gilt und als solcher auch in der da|ra-Datenbank suchbar ist, ist derjenige, der eine nicht registrierte Studie nachweist (sozusagen der „Nachweisagent") nicht zwangsläufig auch Datenanbieter. Für die Homogenität der Datenbank bedeutet dies, dass die Feldsuche nach Studien eines bestimmten Datenanbieters immer nur auf registrierte Ressourcen beschränkt ist. Sucht man zum Beispiel über die Feldsuche nach Studien des Datenanbieters BIBB findet man nur die 128 registrierten Studien dieses Publikationsagenten, nicht aber die wei-

teren sechs BIBB-Studien, die im Rahmen des Aufbaus des Datennachweissystems erfasst wurden. Auch die Einschränkung eines Suchergebnisses über die Facette „Datenanbieter" bezieht sich immer nur auf die Kollektion registrierter Ressourcen. In einer weiteren Entwicklungsstufe des Systems ist geplant, für diese Divergenz eine Lösung zu finden. Um aber inzwischen die Information „Datenanbieter" auch bei nicht-registrierten Studien zumindest mit aufnehmen zu können, wurde diese jeweils im Freitextfeld „Verfügbarkeit" in einem standardisierten Format erfasst („Datenanbieter: XY."). Ein homogenes Retrieval dieser Information wird aber erst mit einer Weiterentwicklung des Systems möglich sein.

Eine weitere Herausforderung, die in Bezug auf registrierte und nicht-registrierte Ressourcen künftig adressiert werden muss, ist die Möglichkeit der **Registrierung einer Ressource, die bereits ohne DOI im System enthalten ist**. Dies könnte mit dem genannten BIBB-Beispiel passieren, wenn sich der Publikationsagent entscheidet, für die sechs nicht-registrierten Studien auch eine DOI-Registrierung durchzuführen. Idealerweise sollten in einem solchen Fall nicht sechs neue Einträge zwecks Registrierung in das System eingebracht, sondern die bereits bestehenden Einträge jeweils mit DOI-Namen versehen werden. Um Situationen wie diese künftig bearbeiten zu können, wird derzeit an entsprechenden Prozessvorschriften gearbeitet, die in konkrete Entwicklungsaufgaben münden werden. Ziel ist es, eine möglichst homogene Datenbank anzubieten, in der registrierte und nicht-registrierte Ressourcen gleichberechtigt beschrieben und auffindbar gemacht werden.

5 Ausblick

Der in diesem Bericht beschriebene Aufbau eines Datennachweissystems für die Sozialwissenschaften ist als erster Schritt in der Entwicklung einer zentralen Anlaufstelle für Datensuchende zu sehen. Mit den vorhandenen Kapazitäten konnte zunächst nur eine geringe Zahl der potentiell für die Sekundärforschung interessanten Daten nachgewiesen werden. Der Sinn dieses Pilotprojekts bestand in erster Linie darin, die Praktikabilität des Systems und der Workflows zu testen sowie einen Grundstock von Datennachweisen aufzubauen, der Forscher animieren kann, ihre Daten ebenfalls bei da|ra nachzuweisen. Idealerweise soll eine breitere Praxis des Nachweisens von Daten auch zu einer vermehrten Registrierung von Daten führen – denn nur durch persistente Identifikation sind Forschungsdaten verlässlich zitierbar und auffindbar. Was im Einzelnen geplant ist, um das Datennachweissystem in diesem Sinne auszubauen und zu festigen, wird im Folgenden kurz beschrieben.

5.1 Ausbau des Datennachweissystems

Der Aufbau des Datennachweissystems wurde durch die Deutsche Forschungsgemeinschaft (DFG) unterstützt. In einer bereits beantragten zweiten Förderphase sollen verschiedene Weiterentwicklungen angegangen werden, die zum Teil in diesem Bericht bereits angesprochen wurden.

Zum einen soll die Plattform durch verstärkte **Einbindung der Forschungsdatenproduzenten** inhaltlich ausgeweitet werden. Schon jetzt ist es jedermann möglich, sich ein Nutzerkonto für das Datennachweissystem einzurichten und Nachweise eigener oder fremder Daten anzulegen. Diese Möglichkeit soll durch verschiedene Marketingmaßnahmen und zu entwickelnde Anreizsysteme breit in der sozialwissenschaftlichen Disziplin beworben und etabliert werden. Weitere Datennachweise könnten außerdem über die systematische Integration von Statistiken und anderen Datenarten in das System eingebracht werden. Hier sind auch Metadatenimporte durch Kooperationen mit datenhaltenden Instituten denkbar. Darüber hinaus sollen verstärkt **Datennachweise aus anderen Informationssystemen** entweder in die Datenbank oder in den Suchindex integriert werden. In diesem Zusammenhang ist auch eine Ausweitung auf internationale Datenangebote geplant. Ziel ist es, das da|ra-Datennachweissystem zu einem **zentralen Suchnetzwerk für sozialwissenschaftliche Forschungsdaten** weiterzuentwickeln.

Sowohl die vermehrte Beteiligung der Datenproduzenten, als auch die Automatisierung bergen die Gefahr einer größeren Heterogenität in den Beschreibungen der Datensätze. Vor diesem Hintergrund spielt die **Sicherung der Metadatenqualität** in Zukunft verstärkt eine Rolle. Auch wenn es Richtlinien zur manuellen Erfassung und Mappings zu anderen Metadatenschemata gibt, muss damit gerechnet werden, dass diese Maßnahmen nicht ausreichen, eine gleichbleibend hohe Metadatenqualität zu garantieren. Unter anderem durch den Einsatz von Suchalgorithmen soll die zu erwartende Heterogenität ausgeglichen werden. Insgesamt ist angesichts des Anwachsens der Datenbank eine **Verbesserung der Suchfunktionen** vorgesehen. Neben verschiedenen Mechanismen zur Unterstützung des Suchvorgangs sollen auch Verbindungen zwischen Nachweisen verwandter Datensätze erstellt werden, damit ein komfortables Navigieren zwischen relevanten Datennachweisen möglich ist.

5.2 Nachhaltigkeit des Datennachweissystems

Auch wenn im Rahmen des Ausbaus des Datennachweissystems diverse Automatisierungsprozesse und die stärkere Einbindung der Forschungsdatenproduzenten geplant sind, wird die **redaktionelle Betreuung** weiter mit einem gewissen Aufwand betrieben werden. Neben der Redaktion fallen außerdem technische und konzeptionelle Weiterentwicklungen des Systems an.

Zu den **Weiterentwicklungen** gehört insbesondere die stetige Anpassung und Verbesserung des Metadatenschemas, das angesichts der erwarteten Ausweitung der Datenbank ein zentrales Instrument zur Qualitätssicherung darstellt. Die Vereinbarkeit von Nachweisen registrierter und nicht-registrierter Daten spielt hierbei in Zukunft verstärkt eine Rolle.

Schließlich werden Methoden entwickelt, die an verschiedenen Stellen die **Aktualität** der erfassten Metadaten sicherstellen sollen. Ziel ist es, die Redaktionsarbeit durch eine so genannte Beobachtungssuite zu unterstützen, mit deren Hilfe zum Beispiel tote Links und neue Erhebungswellen von Längsschnittstudien identifiziert werden können.

Redaktionelle Aufgaben, Qualitätssicherung und die Weiterentwicklung des Systems werden auch künftig von der bei GESIS institutionalisierten da|ra Registrierungsagentur wahrgenommen. Die institutionelle Anbindung und die internationale Vernetzung des Datennachweissystems bilden eine verlässliche Grundlage für seinen nachhaltigen Betrieb.

Literatur

Hausstein, Brigitte/ Schleinstein, Natalija/ Koch, Ute/ Meichsner, Jana/ Becker, Kerstin/ Stahn, Lena-Luise (2014): *da|ra Metadata Schema. Version 3.0.* DOI: 10.4232/10.mdsxsd.3.0 (gesis Technical Reports 2014/07.)

Hey, Tony; Tansley, Stewart; Tolle, Kristin (Hg.) (2009): *The Fourth Paradigm. Data-Intensive Scientific Discovery.* Redmond: Microsoft Research.

Koehler, Wallace (2004): A longitudinal study of Web pages continued: a report after six years. In: *Information Research.* **9**(2), paper 174. URL: http://InformationR.net/ir/9-2/paper174.html, zuletzt aufgerufen am 16.07.2014.

Quandt, Markus; Mauer, Reiner (2012): Sozialwissenschaften. In: Neuroth, Heike/ Strathmann, Stefan/ Oßwald, Achim/ Scheffel, Regine/ Klump,Jens/ Ludwig, Jens (Hg.): *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme.* Boizenburg: Verlag Werner Hülsbusch, 61-81.