BEITRÄGE

Bereitstellung und Nutzung quantitativer Forschungsdaten in der Bildungsforschung:

Memorandum des Fachkollegiums „Erziehungswissenschaft" der DFG

Petra Stanat

Die DFG und andere Wissenschaftsorganisationen haben 2010 Grundsätze zum Umgang mit Forschungsdaten vorgelegt, die eine gezielte Archivierung und verstärkte Nachnutzung solcher Daten fordern. Der Wissenschaftsrat regte mit seinen im August 2012 vorgelegten „Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020"[1] insbesondere auch in den Sozialwissenschaften einen Auf- und Ausbau geeigneter Infrastrukturen an. Verschiedene Initiativen greifen diese Thematik auf. So hat etwa die DFG 2013 in ihrer Infrastrukturförderung ein Programm eingerichtet, in dem Konzepte für die (Weiter-)Entwicklung fachspezifischer und bedarfsorientierter Infrastrukturen für einen verbesserten Umgang mit Forschungsdaten gefördert werden.[2]

Gerade auch in der Empirischen Bildungsforschung werden (quantitative) Forschungsdaten in großem Umfang erhoben. Dabei ist das Forschungsfeld der Bildung zum einen durch seine Komplexität charakterisiert, die es in der Regel erforderlich macht, einen breiten Kranz an Merkmalen zu erheben, um Forschungsfragen angemessen bearbeiten zu können. Zum anderen ist die Empirische Bildungsforschung aber auch durch einen zunehmend schwieriger werdenden Zugang zum Feld (z.B. aufgrund von Problemen bei der Genehmigung von Schuluntersuchungen durch die Kultusministerien und die eingeschränkte Teilnahmebereitschaft von Schulen aufgrund von wahrgenommener Überlastung durch zahlreiche Datenerhebungen) gekennzeichnet. Datenerhebungen im Bereich der Empirischen Bildungsforschung sind ohne Zweifel nicht nur oft mit erheblichem Aufwand verbunden, sondern sie lie-

1 Veröffentlicht am 13. Juli 2012, URL: http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf.

2 http://www.dfg.de/foerderung/programme/infrastruktur/lis/lis_foerderungsangebote/forschungsdaten/index.htm.

fern auch ein Analysepotenzial, das sich nur durch die Eröffnung der Möglichkeit für Nachnutzungen der Datensätze umfassend ausschöpfen lässt.

Inzwischen besteht in der *Scientific Community* und unter den Forschungsförderern weitgehende Einigkeit darüber, dass das Potenzial von Daten der Empirischen Bildungsforschung bislang nicht umfassend genutzt wird und die Datensätze daher auch für Re- und Sekundäranalysen durch andere als die Primärforscherinnen und -forscher nutzerfreundlich zur Verfügung gestellt werden sollten. Dies würde auch weiterführende Kooperationen zwischen verschiedenen Forschungsdisziplinen unterstützen, die sich mit Bildung beschäftigen. Die Umsetzung des Ziels einer wissenschaftlichen Nachnutzung der erhobenen Daten ist jedoch weiterhin mit einer Reihe von Herausforderungen verbunden. Diese wurden in einem Rundgespräch der DFG am 13. März 2014 in Berlin erörtert.[3] Das Ziel des Rundgesprächs bestand darin, die Rolle der Bereitstellung und Nutzung von Forschungsdaten für die gegenwärtige und zukünftige Bildungsforschung herauszuarbeiten. Im Sinne der DFG-Forschungsförderung setzte das Rundgespräch dabei primär an den Bedürfnissen der Forschenden – der primären Datenproduzenten wie auch der potenziellen Nachnutzer – an.

Um zu möglichst konkreten Ergebnissen kommen zu können, lag das Hauptaugenmerk des Rundgesprächs zunächst auf quantitativen Datensätzen, da für deren Bereitstellung bereits Infrastrukturen existieren (z.B. das FDZ am IQB in Berlin, das GESIS-Datenarchiv in Köln oder das FDZ PsychData in Trier). Darüber hinaus wäre es wichtig, Fragen der Bereitstellung von qualitativen Datensätzen der Bildungsforschung zu behandeln, für die ebenfalls ein Bedarf der Nachnutzung bestehen dürfte (z.B. Transkriptionen von Befragungen der Biographieforschung, Zeitzeugeninterviews, systematische Aufzeichnungen von Beobachtungen), die durch bestehende Infrastrukturen aber bislang nur in Ansätzen abgedeckt werden (z.B. durch das FDZ Bildung am DIPF). Aufgrund der Komplexität dieser Fragen (insbesondere auch im Hinblick auf Fragen des Datenschutzes) wäre hierfür jedoch eine eigene Initiative erforderlich.

Mit dem vorliegenden Memorandum werden zentrale Ergebnisse des Rundgesprächs vom 13. März 2014 zusammengefasst und Empfehlungen abgeleitet, die auf eine Optimierung der Bereitstellung und Nutzung quantitativer Forschungsdaten in der Bildungsforschung abzielen.

In der Diskussion herrschte Einigkeit darüber, dass letztlich ein Kulturwandel erforderlich ist, der sich sowohl auf die Bereitstellung als auch auf die Nutzung von Datensätzen bezieht.

Im Folgenden wird nach einer einleitenden Beschreibung des Hintergrunds (Abschnitt 1) der Diskussion auf drei zentrale Aspekte eingegangen,

3 Teilnehmerliste siehe am Ende des Dokuments

die im Rundgespräch erörtert worden sind. Diese beziehen sich auf *Fragen der Identifikation und Aufbereitung von Datensätzen* (Abschnitt 2), die für Re- und Sekundäranalysen zur Verfügung gestellt werden sollten; auf die *Übergabe von Daten durch Datenproduzenten an Infrastrukturen* (Abschnitt 3), die für die Bereitstellung von Forschungsdaten zur Verfügung stehen (Forschungsdatenzentren);und auf die *Nutzung der verfügbaren Datensätze* (Abschnitt 4) durch Wissenschaftlerinnen und Wissenschaftler.

1 Hintergrund

Die Bildungsforschung gehört zu den Wissenschaftsbereichen, in denen die Anzahl von Projekten, die Forschungsdaten erheben, in den letzten 15 Jahren sprunghaft angestiegen ist. Neben den großen internationalen und nationalen Schulleistungsuntersuchungen, die zunächst primär dem Bildungsmonitoring dienen (z.B. PISA, IGLU/PIRLS, IQB-Ländervergleichsstudien), werden zahlreiche Quer- und Längsschnittstudien durchgeführt, die sich mit Bildungsprozessen sowie deren Bedingungen und Erträgen beschäftigen. In der Vergangenheit wurden diese Untersuchungen zumeist geplant und durchgeführt, ohne eine spätere Bereitstellung der Daten vorzusehen. Entsprechend waren auch Re- und Sekundäranalysen bereits vorhandener Daten zumindest in der erziehungswissenschaftlich und psychologisch orientierten Bildungsforschung bislang eine Seltenheit. Diese Situation ändert sich allmählich, wie etwa die steigenden Zahlen der in den einschlägigen Forschungsdatenzentren[4] verfügbaren Datensätze und der darauf bezogenen Nutzungsanträge zeigen. Eine herausgehobene Rolle spielt dabei das Nationale Bildungspanel (NEPS), das 2009 zunächst als BMBF-Projekt seine Arbeit aufnahm und seit dem 1. Januar 2014 im Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) weitergeführt wird. Als Forschungsinfrastruktureinrichtung zielt das NEPS – wie auch das bereits seit 1984 existierende Sozio-oekonomische Panel (SOEP) – explizit darauf ab, Forschungsdaten zu generieren und diese der Scientific Community in nutzerfreundlich aufbereiteter Form zur Verfügung zu stellen. Jenseits solcher Infrastruktureinrichtungen besteht in Bezug auf die Bereitstellung und Nutzung vorhandener Forschungsdaten jedoch noch Optimierungsbedarf. Dieser soll im Folgenden skizziert werden.

4 Im vorliegenden Memorandum bezieht sich die Bezeichnung „Forschungsdatenzentrum" bzw. „FDZ" grundsätzlich nur auf Einrichtungen, die vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) als solche akkreditiert worden sind.

2 Identifikation und Aufbereitung von Datensätzen, die für Re- und Sekundäranalysen zur Verfügung gestellt werden sollten

Es gibt nicht „die“ Forschungsdaten, sondern ein Kontinuum von hoch aufwändigen Längsschnitt- und Querschnittsdaten aus *Large-Scale-Assessments*, die die Untersuchung einer Vielzahl von Forschungsfragen erlauben, bis hin zu sehr spezifischen Datensätzen, die auf die Beantwortung nur einer bestimmten (z.B. experimentellen) Fragestellung zugeschnitten sind. Gleichzeitig ist der Aufwand, der mit ihrer Bereitstellung verbunden ist, bei spezialisierten Datensätzen oft genau so groß wie bei vielseitig nutzbaren Datensätzen aus den *Large-Scale-Assessments*, für die mittlerweile Standards der Berichtlegung und Dokumentation etabliert sind. Dies wirft (nicht nur in der Empirischen Bildungsforschung) die grundsätzliche Frage auf, welche Datensätze in welcher Form zur Verfügung gestellt werden sollten, um Re- und Sekundäranalysen zu ermöglichen. Dabei sind verschiedene Grade der Bereitstellung und Dokumentation zu unterscheiden.

Für *alle* Datensätze, die im Rahmen von Forschungsprojekten erhoben werden, gelten die Grundsätze der guten wissenschaftlichen Praxis der DFG (Deutsche Forschungsgemeinschaft, 2013)[5], die mit Empfehlung 7 die folgende Vorgabe machen: „Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden“ (Deutsche Forschungsgemeinschaft 2013, S. 21). Diese Form der Datensicherung dient zunächst dazu, die Überprüfbarkeit von publizierten Ergebnissen zu gewährleisten. Dafür ist eine grundlegende Dokumentation der Daten erforderlich, die es wissenschaftlichen Kolleginnen und Kollegen potenziell ermöglicht, die durchgeführten Analysen nachzuvollziehen und zu replizieren.

Mit den zitierten Grundsätzen der DFG ist die *Aufbewahrung* der Primärdaten klar geregelt. Davon wird die *Nutzung* unterschieden, die zunächst primär den Forschenden zusteht, die sie erheben (Deutsche Forschungsgemeinschaft 2013, S. 22). Die Sicherung und Bereitstellung von Datensätzen, die im vorliegenden Memorandum im Zentrum stehen, gehen darüber hinaus. Nach den Grundsätzen zum Umgang mit Forschungsdaten dienen sie „nicht nur der Prüfung früherer Ergebnisse, sondern in hohem Maße auch der Erzielung künftiger Ergebnisse. Sie bildet eine strategische Aufgabe, zu der Wissenschaft, Politik und andere Teile der Gesellschaft gemeinsam beitragen

5 http://www.dgfe.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf

müssen" (Allianz der deutschen Wissenschaftsorganisationen 2010, S. 1).[6] Der Fokus liegt hier auf der sekundären Nutzung bzw. Nachnutzung von Forschungsdaten.

Da vorab nicht eindeutig entscheidbar ist, ob ein Datensatz in naher oder ferner Zukunft für die Nachnutzung durch andere Wissenschaftlerinnen und Wissenschaftler interessant sein könnte, erscheint es prinzipiell wünschenswert, mit öffentlichen Mitteln erhobene Forschungsdaten nach einer angemessenen Frist (siehe Abschnitt 3) *grundsätzlich* der *Scientific Community* zur Verfügung zu stellen. Erfahrungen aus dem Ausland zeigen jedoch, dass bei einem solchen Vorgehen letztlich nur ein Bruchteil der mit hohem Aufwand aufbereiteten Datensätze tatsächlich nachgenutzt wird. Daher sollte ein Verfahren für die Identifikation von Datensätzen gefunden werden, die mit einiger Wahrscheinlichkeit für eine Nachnutzung von Interesse sein könnten und entsprechend über ein Forschungsdatenzentrum zur Verfügung gestellt werden sollten. Bei Datensätzen, die zukünftig im Rahmen eines noch zu beantragenden Forschungsprojekts erhoben werden sollen, wird man dabei anders vorgehen müssen als bei bereits vorliegenden Datensätzen.

Für *zukünftig im Rahmen von Forschungsprojekten zu erhebende Datensätze* erscheint folgendes Vorgehen sinnvoll und umsetzbar:

- Bei Forschungsprojekten, die eine Erhebung von Datensätzen beinhalten, sollten bereits in der Planungs- und Antragsphase Überlegungen dazu angestellt werden, ob ein Nachnutzungspotenzial besteht und welche Konsequenzen dies später für die Form der Bereitstellung hat.
- In Abhängigkeit davon sollten Projektanträge, die bei der DFG eingereicht werden, eine Aussage darüber enthalten, in welcher Form die Datensätze bereitgestellt werden. Diese Angaben sind bereits in den Leitfäden für Projektanträge vorgesehen, die bei der DFG eingereicht werden, und sollten in Zukunft konkreter dargestellt werden. Dabei ist anzugeben und zu begründen, welche der folgenden Formen der Bereitstellung vorgesehen bzw. nicht vorgesehen ist:

  a) Archivierung und auf Nachfrage nutzerfreundliche Bereitstellung direkt durch den Datenproduzenten zwecks Prüfung publizierter Ergebnisse (im Sinne guter wissenschaftlicher Praxis).
  b) Archivierung, erweiterte Dokumentation (Codebuch) und auf Nachfrage nutzerfreundliche Bereitstellung direkt durch die Datenproduzenten zwecks weiterführender wissenschaftlicher Analysen.
  c) Gut dokumentierte Übergabe an ein Forschungsdatenzentrum zwecks Archivierung und allgemeiner Bereitstellung an die Scientific Community nach den Regularien des jeweiligen FDZ.

6 http://www.allianzinitiative.de/fileadmin/user_upload/redakteur/grundsaetze_Forschungsdaten_2010.pdf

- Die Begründung für die Wahl der Bereitstellungform kann nur mit Blick auf das konkrete Vorhaben erfolgen; hierfür lassen sich keine allgemein anwendbaren Kriterien spezifizieren. Das Potenzial einer Nachnutzung ist dabei stets gegen den erhöhten Aufwand der Datenaufbereitung abzuwägen.
- Im Falle der Bereitstellung durch die/den Datenproduzenten selbst ist in den Projektanträgen auszuführen, auf welche Weise die langfristige Datensicherung und Datenbereitstellung auch nach Auslaufen der Projektförderung und bei einem Wechsel der institutionellen Anbindung langfristig garantiert werden kann. Dies beinhaltet auch die Bereitschaft, auf Nachfrage weitere Informationen zur Verfügung zu stellen, die für Re- bzw. Sekundäranalysen erforderlich sind.
- Im Falle einer geplanten Bereitstellung von Daten zu Zwecken der Nachnutzung sind weitergehende Überlegungen darüber erforderlich, zu welchem Zeitpunkt und in welcher Form diese bereitgestellt werden sollen. Dies kann z.B. in Form eines Datenmanagement-Plans geschehen.[7] Dabei kann im Antrag auch auf eine Beratung durch ein einschlägiges Forschungsdatenzentrum Bezug genommen werden. Zugleich sollten bei der Antragstellung angemessene Ressourcen für die Aufbereitung und Archivierung der Daten und eine entsprechende Projektlaufzeit eingeplant werden. Dies kann im Rahmen des beantragten Projekts auch die anteilige Finanzierung von Stellen umfassen, die für die Aufbereitung von Datensätzen zwecks Archivierung und Nachnutzung eingerichtet werden müssen (z.B. „forschungstechnische Assistenten"). Entsprechende Antragsmöglichkeiten bei der DFG bestehen.
- Die Begründung der Antragstellenden für die geplante Form der Bereitstellung der Forschungsdaten und die dafür veranschlagten Kosten werden im Rahmen des Begutachtungs- und Entscheidungsverfahrens geprüft. Ergeben sich aus der Begutachtung Hinweise darauf, dass die Wahrscheinlichkeit einer Nachnutzung größer ist, als ihr der Antrag Rechnung trägt, könnten die Antragstellenden von der DFG im Einzelfall um zusätzliche Angaben bis hin zu einem Datenmanagement-Plan gebeten werden, bevor eine Entscheidung getroffen wird.
- Wichtig ist ferner, bereits vor den geplanten Erhebungen datenschutzrechtliche und ethische Fragen zu klären, die für die spätere Bereitstellung der Datensätze relevant sind (z.B. Information der Befragten). Für bereits erhobene Datensätze sind folgende Verfahrensweisen denkbar:
- Produzenten von bereits erhobenen Daten, die ein hohes Nachnutzungspotenzial aufweisen, sollten die Möglichkeit erhalten, Ressourcen für de-

7 Jones, S. (2011): How to Develop a Data Management and Sharing Plan. DCC How-to Guides. Edinburgh: Digital Curation Centre. www.dcc.ac.uk/resources/how-guides/develop-data-plan.

ren Aufbereitung und spätere Übergabe an ein FDZ zu beantragen. Dies kann im Sinne einer Übergangslösung dann in Betracht kommen, wenn Fragen der Nachnutzung von Datensätzen zu Projektbeginn noch nicht bedacht wurden. Entsprechende Nachanträge wären dabei an diejenige Fördereinrichtung (z. B. die DFG) zu richten, die die Erhebung finanziert hat. Dabei sollte vorab sichergestellt werden, dass ein FDZ bereit ist, die Daten in sein Programm aufzunehmen und bereitzustellen. Gleichzeitig sollten Forschungsfragen skizziert werden, die mit den Daten bearbeitet werden könnten.

- Weiterhin sollte für Forschungsanträge zur Bearbeitung substanzieller Fragestellungen, die eine Nutzung bereits existierender, von anderen Datenproduzenten generierter Daten erfordern, die Möglichkeit bestehen, Ressourcen für die Aufbereitung des Datensatzes zwecks späterer Übergabe an ein FDZ zu beantragen.
- In beiden Fällen sollten sich die Antragsteller dazu verpflichten, die Datensätze bis zu einem vorab festgelegten Zeitpunkt an das FDZ zu übergeben. Ferner sollte die Möglichkeit bestehen, dass an den Projektanträgen auch das jeweilige FDZ, das die Daten bereitstellen soll, in geeigneter Weise beteiligt ist, um bei umfangreichen Anforderungen an die Datenaufbereitung unterstützen zu können.
- Die reine Aufbereitung existierender Datensätze ohne Projektzusammenhang und ohne das unmittelbare Ziel, substanzielle Fragestellungen zu bearbeiten, wird durch die DFG nicht gefördert.

3 Übergabe von Daten durch Datenproduzenten an Forschungsdatenzentren

Datensätze, die im Rahmen von Forschungsinfrastruktureinrichtungen erhoben werden, wie etwa die Daten des NEPS, werden grundsätzlich zeitnah der *Scientific Community* zur Verfügung gestellt. Auch die zügige Bereitstellung der großen Schulleistungsstudien (PISA, IGLU/PIRLS, TIMSS, IQB-Ländervergleichsstudien) verläuft inzwischen weitgehend reibungslos. Bei vielen anderen Datensätzen, für die ebenfalls Interesse an einer Nachnutzung bestehen dürfte, ist es dagegen deutlich schwieriger, ihre Übergabe an ein Forschungsdatenzentrum zu erreichen. Um hier den erforderlichen Kulturwandel zu unterstützen, sind Richtlinien erforderlich, die sowohl die Interessen der Datenproduzenten als auch die Interessen der Datennutzer in ausgewogener Weise berücksichtigen. Gleichzeitig sollten die Universitäten und wissenschaftlichen Einrichtungen, in denen die Daten erhoben werden, verstärkt darauf achten, dass diese Richtlinien von ihren Wissenschaftlerinnen und Wissenschaftlern angewendet werden.

3.1. Fristen für die Datenübergabe

Die Produzenten von Datensätzen, die eine hohe Qualität und großes Nachnutzungspotenzial aufweisen, haben in der Regel ein hohes Maß an Kreativität und Arbeit in deren Planung, Erhebung, Aufbereitung und Dokumentation investiert. Oft sind zudem Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler an den Vorhaben beteiligt, die in der Anfangsphase überwiegend mit Projektarbeit beschäftigt waren, um sich anschließend auf der Grundlage der erhobenen Daten wissenschaftlich weiterqualifizieren zu können. Daher ist das Anliegen von Datenproduzenten berechtigt, die selbst erhobenen Daten im laufenden Projekt zunächst eigenständig auswerten zu können, bevor diese an andere Forschende oder an ein FDZ übergeben werden.[8] Lediglich bei Infrastrukturvorhaben, die explizit zur Erhebung und Bereitstellung von Forschungsdaten eingerichtet worden sind, kann erwartet werden, dass die Datensätze so schnell wie möglich der *Scientific Community* zur Verfügung gestellt werden, wobei auch bei diesen Projekten dafür Sorge getragen muss, dass sich die an der Generierung der Daten maßgeblich beteiligten Nachwuchswissenschaftlerinnen und -wissenschaftler weiterqualifizieren können.

Allgemein sollten Zeit- und Ressourcenpläne für Projekte, die aufwändige Datenerhebungen beinhalten, so angelegt sein, dass die wissenschaftlichen Mitarbeiterinnen und Mitarbeiter sowie PIs die Gelegenheit erhalten, ihre Qualifizierungs- und Publikationspläne umzusetzen.

Welche Fristen für die Übergabe von Datensätzen an Forschungsdatenzentren angemessen sind, lässt sich nicht allgemein festlegen. Dies hängt unter anderem von der Art des Datensatzes und dem Umfang der daran gekoppelten Forschungsvorhaben ab. Auf der Grundlage von Regelungen existierender Forschungsdatenzentren erscheinen folgende grobe Richtlinien angemessen:

- International scheint sich als Standard die Erwartung zu etablieren, wonach die Übergabe des Gesamtdatensatzes einer Studie etwa zwei Jahre nach Erhebung der Daten (bzw. nach Übergabe der Datensätze an die Primärwissenschaftlerinnen und -wissenschaftler durch ein beauftragtes Erhebungsinstitut) erfolgt. Bei längerfristigen Längsschnitterhebungen bezieht sich diese Frist auf die jeweilige Erhebungswelle. In gut begründeten Ausnahmen, die sich vor allem auf unzumutbare Härten für Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftler beziehen

8 So heißt es in den Grundätzen guter wissenschaftlicher Praxis: „Die Nutzung [der Primärdaten] steht insbesondere dem/den Forscher(n) zu, die sie erheben. Im Rahmen eines laufenden Forschungsprojekts entscheiden auch die Nutzungsberechtigten (gegebenenfalls nach Maßgabe datenschutzrechtlicher Bestimmungen), ob Dritte Zugang zu den Daten erhalten sollen“ (Deutsche Forschungsgemeinschaft 2013, S. 22).

können, sollte es möglich sein, die Frist zu verlängern. Diese sollte jedoch in keinem Fall drei bis vier Jahre überschreiten.

- Darüber hinaus sollte es möglich sein, für den Zeitraum von in der Regel zwölf Monaten die Nachnutzung von Daten zur Bearbeitung von solchen Forschungsfragen zu sperren, die zu dem Zeitpunkt noch im Rahmen von Qualifikationsarbeiten bearbeitet werden (dies sehen z.B. die Regelungen des FDZ am IQB vor). Bei Forschungsanträgen, die eine Nachnutzung der erhobenen Datensätze vorsehen, wäre ein an diesen Fristen orientierter Zeit- und Ressourcenplan für die Bereitstellung zu erstellen und zu begründen.

3.2. Datendokumentation

Die Dokumentation von Datensätzen kann unterschiedlich aufwändig gestaltet werden. Diese reicht von Rohdaten mit nachvollziehbaren Labels und einem Codebuch bis hin zu Datensätzen, die neben Rohdaten auch Skalen und Metadaten umfassen und zu denen detaillierte technische Berichte und Skalenhandbücher vorliegen (also Meta- und Paradaten).

Um sowohl Datenproduzenten als auch Datennutzern gegenüber Transparenz zu schaffen, ist es wünschenswert, Standards und Beispiele der Datendokumentation zur Verfügung zu stellen. Hierfür wären die Forschungsdatenzentren, die Bildungsdaten anbieten, die kompetenten Akteure.

3.3. Zitation von Datensätzen

Bei der Generierung von Datensätzen, die für die wissenschaftliche Nachnutzung potenziell interessant sind, handelt es sich um eine wissenschaftliche Leistung, die bislang oft unzureichend gewürdigt wird.

Es muss selbstverständliche Praxis werden, dass die Urheber der Datensätze im Rahmen von Publikationen in geeigneter Weise zitiert werden. Auch die aktive Einbeziehung von Datenproduzenten als Ko-Autoren bei der Publikation von Ergebnissen, die auf den von ihnen generierten Datensätzen basieren, kann unter Umständen angemessen sein, sofern es sich nicht nur um eine „Ehrenautorenschaft“ handelt, die den Regeln guter wissenschaftlicher Praxis widersprechen würde.

Auf jeden Fall ist ein innerwissenschaftlicher Kulturwandel in Bezug auf die Würdigung der Datengenerierung angezeigt, der unter anderem durch folgende Maßnahmen unterstützt werden kann:

- Eine basale Möglichkeit, Datensätze leicht auffindbar und zitierbar zu machen, besteht darin, sie mit einem Persistent Identifyer (z.B. Digital Object Identifier, DOI) zu versehen. Für die Urheber von Daten ist jedoch die Zitation von Datensatzbeschreibungen attraktiver, die möglichst in anerkannten Zeitschriften erschienen sind. Die Möglichkeit, solche

Beschreibungen in Fachzeitschriften mit Peer-Review zu veröffentlichen, ist in der Empirischen Bildungsforschung (und nicht nur dort) jedoch bislang sehr begrenzt. Es wäre wünschenswert, dass einschlägige Zeitschriften hierfür Formate vorsehen, die von Datenproduzenten genutzt werden können. Dies könnte auch ein Anreiz dafür sein, vorhandene Datensätze zur Verfügung zu stellen. Denkbar wäre zudem die Gründung eines Journals, das auf die Präsentation von Datensätzen spezialisiert ist. Ferner sollten Herausgeberinnen und Herausgeber von Zeitschriften sowie Gutachterinnen und Gutachter bei eingereichten Manuskripten stärker als bisher darauf achten, dass Datensätze angemessen gewürdigt und korrekt zitiert werden.

- Auch Fachgesellschaften sollten Wissenschaftlerinnen und Wissenschaftler verstärkt ermutigen, auf ihren Tagungen und in ihren Verbandszeitschriften Datensätze zu präsentieren, was für große Vorhaben wie das NEPS bereits geschieht.
- Ferner sollten Forschungsförderer bei Projektanträgen neben Publikationen von Forschungsergebnissen die Veröffentlichung von Datensätzen als gleichberechtigten Output von Wissenschaftlerinnen und Wissenschaftlern berücksichtigen, wie es die National Science Foundation (NSF) in den USA bereits seit 2013 tut. Darüber hinaus wird mit dem Thomson Reuters' Data Citation Index derzeit ein Impact Factor für Daten erarbeitet, der ebenfalls einen Anreiz schaffen soll, die eigenen Daten für Re-und Sekundäranalysen zur Verfügung zu stellen und im Feld bekannt zu machen.[9]
- Da gerade an der Generierung komplexer und innovativer Datensätze häufig viele Wissenschaftlerinnen und Wissenschaftler beteiligt sind, kann es für eine angemessene Würdigung ihres jeweiligen Beitrags mitunter erforderlich sein, Teildatensätze zu definieren, die jeweils separat zu zitieren sind.
- Über die Forschungsdatenzentren ist es zudem möglich, Datennutzer vertraglich dazu zu verpflichten, die von ihnen verwendeten Datensätze so zu zitieren, wie es von den Datenproduzenten vorgegeben wurde. Von der Möglichkeit, eine solche Auflage zu definieren und ggf. zu sanktionieren, machen die Datenproduzenten bislang noch nicht immer Gebrauch.

4 Nutzung verfügbarer Datensätze

In einigen Disziplinen besteht bereits eine Tradition der Nachnutzung existierender Datensätze, etwa in der Ökonomie und in den Sozialwissenschaften.

9 http://wokinfo.com/products_tools/multidisciplinary/dci/

Diese ist in der erziehungswissenschaftlich und psychologisch orientierten Bildungsforschung deutlich weniger ausgeprägt als etwa in der soziologischen Bildungsforschung und in der Bildungsökonomie. Teilweise scheint es fast als anstößig empfunden zu werden, Analysen von Daten durchzuführen, die nicht selbst erhoben worden sind. So werden Daten der Empirischen Bildungsforschung zwar häufig durch Wissenschaftlerinnen und Wissenschaftler aus der erziehungswissenschaftlich oder pädagogisch-psychologisch orientierten Bildungsforschung generiert, ihr Potenzial für Re- und Sekundäranalysen wird jedoch primär von Kolleginnen und Kollegen aus anderen Disziplinen genutzt. Gleichzeitig werden in Forschungsprojekten gelegentlich neue Daten erhoben, deren Fragestellungen sich unter Umständen mit bereits vorhandenen Datensätzen bearbeiten ließen. Vor diesem Hintergrund erscheinen mit Bezug auf die Nutzung von Forschungsdaten folgende Maßnahmen sinnvoll:

- Die Möglichkeit, bei der DFG Forschungsprojekte zu beantragen, die sich auf die Auswertung bestehender Daten beziehen, hat schon immer bestanden, wird aber von der empirischen Bildungsforschung (wie auch z.B. der Psychologie) kaum nachgefragt. Ob solche Vorhaben bewilligt werden oder nicht, hängt – wie bei jedem Projektantrag – von ihrer wissenschaftlichen Qualität ab. Ein gutes Beispiel ist das Schwerpunktprogramm 1646 „Education as a Lifelong Process", in dem alle Projekte mit NEPS-Daten arbeiten. Aber auch im Normalverfahren und bezogen auf andere Datensätze können Forschungsvorhaben beantragt werden, die auf Re- und Sekundäranalysen basieren. Über diese Möglichkeit sollten potenzielle Antragstellerinnen und Antragsteller besser informiert werden. Dabei können neben den Einrichtungen der Forschungsförderung wiederum auch die Fachverbände eine wichtige Rolle spielen, indem sie auf das Potenzial von Re- und Sekundäranalysen aufmerksam machen und sich zu dieser Art von Forschung bekennen. So könnte etwa in „Calls for Paper" für Fachtagungen betont werden, dass auch Vorträge zu Ergebnissen von Re- und Sekundäranalysen ausdrücklich erwünscht sind.
- Bei Projektanträgen, die eine Erhebung neuer Daten beinhalten, sollte grundsätzlich geprüft werden, ob nicht bereits vergleichbare Datensätze vorliegen und verfügbar sind, die sich zur Untersuchung der Forschungsfragen eignen könnten. Sofern dies der Fall ist, wäre ggf. der Nachweis zu führen, warum eine Nachnutzung nicht sinnvoll ist und neue Daten erhoben werden sollen. Dieser Punkt wäre dann auch als Gegenstand der Begutachtung einzubeziehen.
- Die Auffindbarkeit verfügbarer Datensätze in der Bildungsforschung sollte durch geeignete Maßnahmen (Vernetzung der einschlägigen FDZs, einheitliche Verschlagwortung, zentrale Clearingstelle) weiter optimiert werden. Initiativen zur gemeinsamen Präsentation bestehender Datensät-

ze für die nationale und internationale Fachöffentlichkeit sollten ausgebaut und von den Forschungsorganisationen unterstützt werden.

- Wie bereits in Abschnitt 2 erwähnt, sollte es im Rahmen von Projektanträgen möglich sein, Ressourcen für die Aufbereitung existierender Datensätze zu beantragen, die zur Beantwortung der Forschungsfragen genutzt werden sollen und noch nicht in einem Forschungsdatenzentrum liegen. Dies kann auch das Heranspielen externer Daten, wie etwa Kontextinformationen, beinhalten. Bedingung hierfür sollte wiederum sein, dass die Datensätze nach einer angemessenen Frist an ein FDZ übergeben werden.
- Es sollte ausgelotet werden, ob und wie der Zugang zu bisher nicht oder nur sehr schwer zugänglichen Datensätzen, die ein hohes Forschungspotenzial aufweisen (wie etwa Daten der Schulstatistik, Daten aus Einschulungsuntersuchungen oder Daten länderspezifischer Bildungsmonitorings) erleichtert werden kann. Einige dieser Datensätze, die teilweise sogar im Längsschnitt vorliegen und mehrere Ebenen umfassen, weisen ein erhebliches Forschungspotenzial auf, das nicht ansatzweise ausgeschöpft wird. Es wäre daher wünschenswert, dass die Bildungsforschung und die Länder ins Gespräch darüber kommen, wie eine wissenschaftliche Nachnutzung dieser Daten ermöglicht werden kann.
- Um die internationale Zusammenarbeit in der Empirischen Bildungsforschung zu stärken, sollten besonders einschlägige Datenbestände in englischer Sprache dokumentiert werden.
- Zu Datensätzen, die in einem Forschungsdatenzentrum liegen, sind in der Regel auch die Fragebogeninstrumente verfügbar, die bei der Datenerhebung verwendet wurden. Für die Testinstrumente ist dies dagegen meistens nicht der Fall, da diese oft in zukünftigen Studien eingesetzt werden sollen und die Datenproduzenten daher die Testsicherheit durch Geheimhaltung gewährleisten müssen. Die Bearbeitung mancher Fragestellungen, etwa in der Fachdidaktik, erfordert jedoch die Kenntnis der Aufgaben. Es wäre wünschenswert, in den Forschungsdatenzentren den vertraulichen und kontrollierten Zugang zu den Testaufgaben zu ermöglichen. Allerdings wäre dies mit zusätzlichem Aufwand für die FDZs verbunden, der finanziert werden müsste.
- Die Fortbildungsangebote, die für die Nutzung der in Forschungsdatenzentren verfügbaren Datensätze angeboten werden, sollten weiterhin zur Verfügung gestellt und bei steigender Nachfrage und nachgewiesener Qualität möglichst weiter ausgebaut werden. Ferner wäre es wünschenswert, zu einschlägigen Datensätzen Public Use Files zur Verfügung zu stellen sowie darüber hinaus auch College Use Files, die im Rahmen von Lehrveranstaltungen genutzt werden können.

5 Schlussbemerkung

Der in diesem Memorandum skizzierte Kulturwandel in Bezug auf die Bereitstellung und Nutzung von Daten der Empirischen Bildungsforschung würde nicht nur zu einer umfassenderen Ausschöpfung des Potenzials von Daten führen, die mit öffentlichen Geldern und hohem Aufwand erhoben worden sind, sondern auch die Qualität der Forschung erhöhen und damit die Empirische Bildungsforschung allgemein stärken . Dabei geht es nicht nur um die Aufdeckung und Korrektur von Fehlern in den Daten, sondern vor allem auch um die Replikation von Forschungsergebnissen und um vertiefende Analysen zu bereits publizierten Befunden. Zudem wird durch die Bereitstellung von Forschungsdaten die Möglichkeit ihrer Auswertung aus der Perspektive unterschiedlicher Disziplinen eröffnet, was die multi- und interdisziplinäre Bearbeitung von Forschungsfragen unterstützt, die gerade in einem so facettenreichen und komplexen Forschungsfeld wie Bildung wichtig ist. Ferner kann durch die Bereitstellung von Forschungsdaten ihre langfristige Sicherung gewährleistet werden.

Wichtig ist bei der Diskussion von Richtlinien, dass sowohl die Interessen der Datenproduzenten als auch die Interessen der Datennutzer respektiert und angemessen berücksichtigt werden. Es darf weder sein, dass Wissenschaftlerinnen und Wissenschaftler, die mit öffentlichen Geldern Daten erhoben haben, ihre Datensätze langfristig unter Verschluss halten, noch können Datennutzer verlangen, dass sie Zugang zu Daten erhalten, bevor deren Urheber die Möglichkeit hatten, ihre zentralen Fragestellungen zu bearbeiten. Die Erfahrungen der letzten Jahre etwa mit den Daten der großen internationalen und nationalen Studien zum Bildungsmonitoring zeigen, dass ein solcher Interessensausgleich durchaus möglich ist. Diese Prozesse sollten weiter optimiert und auf andere Datensätze, die Nachnutzungspotenzial aufweisen, ausgedehnt werden.

Petra Stanat, Prof. Dr., ist Direktorin des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin.

Teilnehmerliste DFG-Rundgespräch „Forschungsdaten in der Empirischen Bildungsforschung", Humboldt-Universität zu Berlin (IQB), 13. März 2014

Professor Dr. Cordula Artelt (Otto-Friedrich-Universität Bamberg, Lehrstuhl für Empirische Bildungsforschung)

Professor Dr. Sigrid Blömeke (Humboldt-Universität zu Berlin, Abteilung Systematische Didaktik und Unterrichtsforschung)

Dr. Susanne von Below (Bundesministerium für Bildung und Forschung, Berlin)

Dr. Edith Braun (Universität Kassel, Internationales Zentrum für Hochschulforschung Kassel, INCHER-Kassel)

Professor Dr. Hartmut Ditton (Ludwig-Maximilians-Universität München (LMU), Institut für Pädagogik, Bildungs- und Sozialisationsforschung)

Professor Bernd Fitzenberger, Ph.D. (Albert-Ludwigs-Universität Freiburg, Abteilung für Empirische Wirtschaftsforschung und Ökonometrie)

Professor Dr. Hans Gruber (Universität Regensburg, Institut für Pädagogik)

Dr. Marcel Helbig (Wissenschaftszentrum Berlin für Sozialforschung gGmbH (WZB))

Dr. Nina Jude (Deutsches Institut für Internationale Pädagogische Forschung (DIPF))

Professor Dr. Frank Kalter (Universität Mannheim, Professur für Allgemeine Soziologie)

Dr. Stefan Koch (Deutsche Forschungsgesellschaft e.V. (DFG Bonn))

Professor Dr. Olaf Köller (IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik)

Dr. Poldi Kuhl (Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen)

Professor Dr. Detlev Leutner (Universität Duisburg-Essen, Institut für Psychologie)

Professor Dr. Katharina Maag Merki (Universität Zürich, Institut für Erziehungswissenschaft)

Professor Dr. Kai Maaz (Deutsches Institut für Internationale Pädagogische Forschung (DIPF))

Norbert Maritzen (Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ), Hamburg)

Reiner Mauer (GESIS – Leibniz-Institut für Sozialwissenschaften, Köln)

Dr. Jutta von Maurice (Otto-Friedrich-Universität Bamberg, Nationales Bildungspanel NEPS)

Prof. Dr. Benjamin Nagengast (Eberhard-Karls-Universität Tübingen, Institut für Erziehungswissenschaft)

Elfriede Ohrnberger (Bayerisches Staatsministerium für Bildung und Kultus Wissenschaft und Kunst)

Professor Dr. Hans Anand Pant (Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen)

Professor Dr. Beatrice Rammstedt (GESIS – Leibniz-Institut für Sozialwissenschaften, Center for Survey Design and Methodology (CSDM))

Professor Dr. Marc Rittberger (Deutsches Institut für Internationale Pädagogische Forschung (DIPF))

Professor Dr. Hans-Günther Roßbach (Otto-Friedrich-Universität Bamberg, Lehrstuhl für Elementar- und Familienpädagogik)

Professor Dr. Josef Schrader (Eberhard-Karls-Universität Tübingen, Abteilung Erwachsenenbildung/Weiterbildung)

Professor Dr. Claudia Schuchart (Bergische Universität Wuppertal, Professur für Empirische Bildungsforschung)

Dr. Katharina Schulte (Projektträger im Deutschen Zentrum für Luft- und Raumfahrt e.V. (DLR))

Professor Dr. Knut Schwippert (Universität Hamburg, Arbeitsbereich für Interkulturelle und International Vergleichende Erziehungswissenschaft)

Teilnehmerliste DFG-Rundgespräch „Forschungsdaten in der Empirischen Bildungsforschung", Humboldt-Universität zu Berlin (IQB), 13. März 2014

Professor Petra Stanat, Ph.D. (Humboldt-Universität zu Berlin, Institut für Erziehungswissenschaften)

Professor Dr. Miriam Vock (Universität Potsdam, Department Erziehungswissenschaft)

Professor Dr. Gert G. Wagner (Deutsches Institut für Wirtschaftsforschung (DIW))

Professor Dr. Sabine Walper (Ludwig-Maximilians-Universität München (LMU), Institut für Pädagogik, Bildungs- und Sozialisationsforschung)

Privatdozent Dr. Erich Weichselgartner (Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID))

Professor Dr. Sabine Weinert (Otto-Friedrich-Universität Bamberg, Lehrstuhl Psychologie I: Entwicklung und Lernen)