

datorium – ein neuer Service für Archivierung und Zugang zu sozialwissenschaftlichen Forschungsdaten

Wolfgang Zenk-Möltgen, Monika Linne

Datenarchiv für Sozialwissenschaften
GESIS - Leibniz Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8
50667 Köln
wolfgang.zenk-moeltgen@gesis.org
monika.linne@gesis.org

Abstract: Forschungsdaten in den Sozialwissenschaften werden trotz zahlreicher Anstrengungen noch nicht ausreichend archiviert und zugänglich gemacht. Das GESIS Datenarchiv führt daher mit datorium einen neuen Service für Wissenschaftlerinnen und Wissenschaftler ein, ihre Forschungsdaten auf einfache Weise zu archivieren und Anderen zugänglich zu machen. Die Konzeption und Umsetzung des Serviceangebots von datorium wird in zwei Phasen vollzogen: Zunächst wird der Zugang zu den Standard-Langzeitarchivierungsprozessen des Datenarchivs über datorium vereinfacht. In einem zweiten Schritt wird datorium als einfache Möglichkeit des data sharings ausgebaut, die neben einer bitstream preservation auch ein Angebot für eine optionale Langzeitarchivierung der Daten durch GESIS enthält. Im Beitrag werden die Funktionen von datorium, die Software für die Implementierung und die Einbettung in die Arbeitsprozesse des Datenarchivs zur Langzeitarchivierung und die Distribution dargestellt. Ziel des GESIS Datenarchivs ist es, mehr Forschungsdaten mit einer größeren thematischen Breite für die wissenschaftliche Verwendung bereit zu stellen und zu archivieren.

1 Einleitung

Das bisherige Dienstleistungsangebot des GESIS Datenarchivs wird im Jahr 2013 um das digitale Daten-Repository datorium erweitert. Dieses erlaubt es, dass Forschungsdaten selbst hochgeladen, beschrieben und weitergegeben werden können, enthält als Basissicherung eine bitstream preservation und bietet eine Anbindung an die Langzeitsicherungsangebote des GESIS Datenarchivs. datorium stellt so eine komplementäre Ergänzung zu den bisherigen Sicherungs- und Distributionsangeboten des Archivs dar. Auf diesem Wege soll die Kultur des data sharings, die das Archiv seit über 50 Jahren unterstützt und fördert, weiter vorangetrieben werden und die thematische Breite und Anzahl der durch das Archiv geförderten Nutzungen von Forschungsdaten erhöht werden. Damit folgt das Datenarchiv auch den vielen internationalen und nationalen Empfehlungen zur Verbesserung der Archivierung und des Zugangs zu öffentlich finanzierten Forschungsdaten [Wi2012, Eu2010, Oe2013].

GESIS reagiert damit auf eine sich verändernde Datenlandschaft, in der Wissenschaftlerinnen und Wissenschaftler schnelle und flexible Werkzeuge fordern, die es ihnen insbesondere erlauben, ihre Forschungsergebnisse zu publizieren und mit Anderen zu teilen.

Ein wesentlicher Bestandteil der Arbeit des Datenarchivs besteht in der Generierung von Metadaten [ZH2012]. Sie sind zentral für die Auffindbarkeit, die Nutzbarkeit und die Langzeitarchivierung der archivierten Daten. Mit dem Archiv sollen Wissenschaftlerinnen und Wissenschaftler die Möglichkeit erhalten, ihre Daten, Metadaten und relevanten Dokumente eigenständig in das Repositorium einzupflegen sowie dazugehörige Publikationen zu verlinken. Darüber hinaus können sie selbst definieren, welchen Nutzergruppen sie den Zugriff auf ihre Daten erlauben. Den Forscherinnen und Forschern wird dadurch einerseits ermöglicht, ihre Forschungsergebnisse kostenfrei zu archivieren und andererseits zeitnah Anderen zur Verfügung zu stellen. Darüber hinaus erhalten sie die Möglichkeit, dass die Forschungsdaten auf einfache Weise in die bestehende Langzeitarchivierung des GESIS Datenarchivs übernommen werden können. Ein solches Angebot kann daher dazu dienen, der Profession Zugang zu bisher unerschlossenen Forschungsprojekten zu ermöglichen.

2 Datenarchivierung

Der Prozess der Archivierung sozialwissenschaftlicher Daten im Verständnis der GESIS-Abteilung Datenarchiv für Sozialwissenschaften umfasst die Schritte Akquisition, Dateneingang, Datenaufbereitung und Dokumentation, Datenbereitstellung und Langzeitarchivierung. Die derzeitige Praxis der Datenaufbereitung und Dokumentation lässt sich dabei grob in eine Standardarchivierung und eine Added-value-Archivierung unterteilen [Ma2012]. Für die Langzeitarchivierung versteht sich das Archiv nach dem OAIS-Referenzmodell als Organisation, die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit der digitalen Forschungsdaten für ihre Nutzbarkeit durch die Zielgruppe der Sozialwissenschaften übernommen hat [vgl. DS2010].

Im Rahmen der gegenwärtigen Standardarchivierung durch das Datenarchiv (siehe Abb. 1) durchläuft jede Studie nach Dateneingang zunächst eine ausführliche Dateneingangskontrolle [Ma2012]. Dieser Prozess beinhaltet u.a. eine technische Kontrolle bzgl. der verwendeten Formate, Lesbarkeit oder Virenfreiheit der Daten. Darüber hinaus findet eine Überprüfung auf Vollständigkeit und Nutzbarkeit der Daten, Erhebungsinstrumente sowie der Dokumente statt. Außerdem wird kontrolliert, ob die Daten mit dem beschriebenen Projekt übereinstimmen und ob die Daten konsistent sind (Wild Codes, fehlende Werte, Gewichtung korrekt, question routing, Vercodungsfehler, Dopplungsfehler etc.). Ein weiterer wichtiger Arbeitsprozess der Standardarchivierung ist außerdem die Datenschutzkontrolle, in der z.B. die Quelle der Erhebung oder Datenschutzverletzungen überprüft werden.

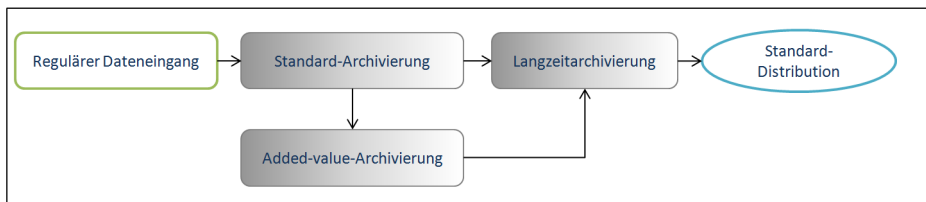


Abbildung 1: Workflow der gegenwärtigen Datenarchivierung

Im Anschluss an die Dateneingangskontrolle und nach eventuellen Fehlerkorrekturen in Absprache mit den Datengeberinnen und Datengebern findet eine Überführung der zu archivierenden Daten und Dokumente in geeignete Langzeitsicherungs- und Distributionsformate statt. Im weiteren Verlauf wird eine Studienbeschreibung erstellt, welche die inhaltlichen, methodischen und technischen Charakteristika der Studie enthält. In diesem Zuge werden außerdem eine Studiennummer und eine DOI für die aktuellste Version der Studiendokumentation vergeben [Ze2012]. Nach einer Bearbeitung der wichtigsten Metadatenfelder erfolgt die Sicherung im Archivspeicher und die Publikation der Studie über den Datenbestandskatalog. Die Publikation von Metadatenfeldern, deren Bearbeitung aufgrund einer höheren Komplexität länger dauert, kann zu einem späteren Zeitpunkt erfolgen. Zur Langzeitarchivierung gehört für das Archiv neben dem Erhalt der physischen und logischen Informationen der Studie auch der Prozess der Beobachtung aktueller Speichertechnik und -formate, sowie die Anpassung der digitalen Objekte durch Migration oder Emulation [Ma2012].

Für ausgewählte Studien oder besondere Studienkollektionen bietet das Datenarchiv den Service einer Added-value-Archivierung an (vgl. Abb. 1). Die Studien durchlaufen zunächst alle Arbeitsprozesse der Standardarchivierung und erhalten im Anschluss eine besondere Datenaufbereitung (Datenbereinigung; Standardisierung; zeit- und/oder ländervergleichende Integration/Kumulation; Harmonisierung; Ergänzung mit Kontextdaten/Aggregatdaten). Darüber hinaus findet eine umfassende Produktion von strukturierten Metadaten auf Variablenebene statt (z.B. vollständige Frage- und Antworttexte – teilweise multilingual; Intervieweranweisungen; besondere Anmerkungen zur Dokumentation; abweichende Länder- oder Wellenspezifika), deren Publikation über Online-Portale wie z.B. ZACAT oder Variable Overviews erfolgt [BZ2011]. Außerdem werden im Rahmen der Added-value-Archivierung weitere Kontextinformationen (z.B. themenspezifische oder vergleichbare Fragen; Trendvariablen) hinzugefügt sowie Codebücher, Datenhandbücher, Variablenreports oder Methodenberichte generiert.

In beiden Fällen, also sowohl innerhalb der Standardarchivierung als auch der Added-value-Archivierung erfolgt die Studiendokumentation und Langzeitsicherung durch das Archiv, welches die Verantwortung für den Erhalt der langfristigen Nutzbarkeit und Interpretierbarkeit der ihm anvertrauten digitalen Objekte übernimmt. Das GESIS Datenarchiv hat dafür eine Preservation Policy verabschiedet und bereitet sich auf eine Zertifizierung als vertrauenswürdiges digitales Archiv im Rahmen des „Data Seal of Approval“ [DS2010] vor. Die Überprüfung und Aufbereitung der Daten nach den neuesten Standards sowie Dokumentation und Sicherung von dazugehörigen

Studienmaterialien sind der Garant für spätere Sekundäranalysen und für die Vergleichbarkeit mit anderen Studien.

3 Bestand

Zahlreiche Forscherinnen und Forscher aus Institutionen der akademischen sowie der kommerziellen Markt- und Meinungsforschung nutzen das GESIS Datenarchiv für Sozialwissenschaften zur Archivierung und Distribution ihrer Studien. Zurzeit sind ca. 5.100 Studien von insgesamt etwa 2.200 Primärforschenden von fast 2.400 Institutionen im Bestand des Archivs. Davon sind die Forschungsdaten von knapp 2.900 Studien nach einer Registrierung direkt im Download für Sekundärnutzende verfügbar. Alle weiteren Studien sind über ein Warenkorbsystem bestellbar oder unter speziellen Bedingungen zu nutzen. Dies eröffnet Wissenschaftlerinnen und Wissenschaftlern die Möglichkeit, ohne eigenen Erhebungsaufwand empirisch zu arbeiten und auf Bestehendem aufzubauen. Vorhandenes Wissen wird so optimal und kostengünstig genutzt.

Zurzeit archiviert das Datenarchiv für Sozialwissenschaften digitale Datensätze aus allen Bereichen der Sozialwissenschaften, wenn die Studie Aussagen über die deutsche Bevölkerung oder über Teile von ihr erlaubt, die Untersuchung von deutschen Forschenden durchgeführt wurde, unabhängig davon, ob sich die Untersuchung auf Deutschland bezieht oder nicht, oder wenn die Studie ganz allgemein für die sozialwissenschaftliche Gemeinde von Interesse sein könnte [Ma2012]. Für eine vollständige Archivierung benötigt das Datenarchiv von den Primärforschenden alle Materialien, die für eine Sekundäranalyse notwendig sind. Dies umfasst mindestens die Daten, wenn möglich aufbereitet für die direkte Verwendung in einer Statistiksoftware, außerdem das Erhebungsinstrument und eine methodische Beschreibung.

Durchschnittlich wurden mit diesen Aufnahmekriterien innerhalb der letzten Jahre etwa 170 neue Studien pro Jahr vom Datenarchiv akquiriert, was zum Stand von jetzt ca. 5.100 veröffentlichten Studien geführt hat. Im Vergleich dazu kann man sehen, dass die DFG seit 1999 insgesamt 87.898 Projekte gefördert hat, von denen alleine 15.303 aus den Geistes- und Sozialwissenschaften stammen [Df2013]. Ebenso kann man von etwa 6.000 bis 7.000 neuen oder aktualisierten Forschungsprojekten alleine in Deutschland, Österreich und der Schweiz ausgehen, die jährlich in das sozialwissenschaftliche Forschungsinformationssystem SOFIS aufgenommen werden [Ko2012: 40] und aktuell zu einer Zahl von 51.585 Projekten geführt haben [So2013]. Auch wenn diese Zahlen nicht nur empirische Datenerhebungsprojekte umfassen, wird hier deutlich, dass das Akquisitionsvolumen für Forschungsdaten im Datenarchiv klare Steigerungsmöglichkeiten aufweist. Hier kann datorium einen Beitrag leisten, indem die Aufnahme eines Forschungsprojektes wesentlich einfacher gestaltet wird. Darüber hinaus können mit datorium auch andere Projektarten, die über die bisher von GESIS hauptsächlich archivierten Studien hinaus gehen, archiviert werden, wie etwa Dissertationsprojekte oder kleinere eigenfinanzierte Projekte.

Mit dem Angebot von datorium verfolgt GESIS demnach das Ziel, das Akquisitionsvolumen zu erhöhen und auch solche Forschungsprojekte zu erfassen,

welche über die o.g. Archivierungskriterien des Datenarchivs hinausgehen. Sowohl die Anzahl der erfassten Studien, der Datengebenden, als auch der Daten- und Dokumentendownloads durch Nutzerinnen und Nutzer sollte durch ein solches Repositorium erheblich gesteigert werden können. Dies bedeutet, dass sich auf beiden Seiten, sowohl bei den Datengebenden als auch bei den Datennutzenden, möglicherweise neue Nutzungsprofile erschließen. Dies soll insbesondere durch die geplanten sozialen Vernetzungsmöglichkeiten innerhalb von datorium unterstützt werden, so dass sich neue produktive Kontakte mit synergetischen Effekten zwischen Wissenschaftlerinnen und Wissenschaftlern ergeben können. Darüber hinaus wird – durch die niedrigschwellige Möglichkeit für eine zunächst einfache Sicherung der Forschungsdaten und die Option zur Übernahme in die Langzeitsicherung des Datenarchivs – bei Forschenden ein größeres Bewusstsein für die Notwendigkeit der digitalen Langzeitarchivierung geschaffen.

Wissensproduktion ist ein Prozess, der darauf basiert, dass neue Erkenntnisse schnell verbreitet werden, damit sie von anderen Forscherinnen und Forschern weiterverwendet werden können. Wesentlich ist deshalb die Weitergabe von Forschungsergebnissen in der Forschungsgemeinschaft zum Zweck des Erkenntnisgewinns. Studien konnten zeigen, dass die Bereitstellung von Forschungsdaten auch die Zitationen der Publikationen über die Forschungsergebnisse von Forschenden erhöht [PDF2007]. Dieser Reputationsgewinn ist ein wesentlicher Anreiz für Forschende zur Archivierung und Bereitstellung ihrer Daten.

Darüber hinaus erleichtert die Publikation von Forschungserkenntnissen der wissenschaftlichen Gemeinde das Auffinden und die Zitation von relevanten wissenschaftlichen Forschungsergebnissen. In diesem Sinne eignet sich ein niedrigschwelliges Webangebot in Form des datorium-Portals besonders gut, um die Verbreitung und Nutzung von wissenschaftlichen Arbeiten in einem größeren Spektrum als bisher zu unterstützen und zu erleichtern. Mit dem Angebot von datorium durch das GESIS Datenarchiv werden die Forscherinnen und Forscher daher auch bei der Beachtung von Empfehlungen zu Open Access unterstützt [Re2013].

4 Vorgehen

Zunächst wurden für das Projekt datorium ein Konzept und die Spezifikationen für den Betrieb des Repositoriums erarbeitet. Dabei wurde die Implementierung in zwei Phasen geplant: Phase I (s. Abb. 2) ermöglicht es Wissenschaftlerinnen und Wissenschaftlern, ihre Forschungsdaten über datorium selbständig einzupflegen und dazugehörige Dokumente hochzuladen. Die technische Implementierung dieses Datenzugangs von Forschungsprojekten in das Datenarchiv steht seit Juni 2013 zur Verfügung. Die Daten durchlaufen daraufhin die auch bisher bei der Standardarchivierung durchgeführten Prozesse durch das Datenarchiv und werden in Abhängigkeit des Studieninhalts und der Datenaufbereitung in entsprechenden Online-Portalen publiziert (Datenbestandskatalog, ZACAT, CESSDA Data Portal, Variable Overview, HISTAT).

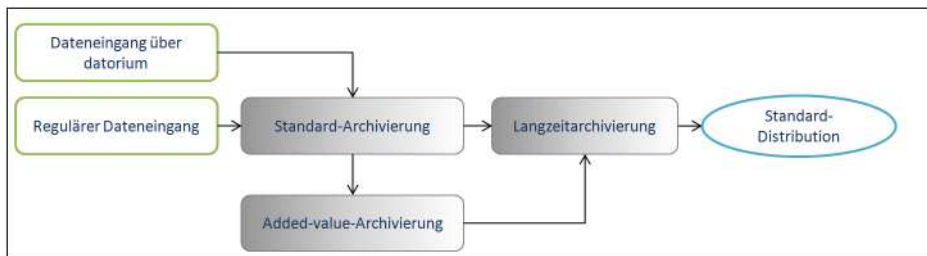


Abbildung 1: Workflow datorium und regulärer Dateneingang – Phase I

In dieser Phase I der Implementierung von datorium wird der Dateneingang interessierten Datengeberinnen und Datengebern zur Verfügung gestellt. Diese können den neuen Zugang nutzen, um auf einfache Weise ihre Daten für die Archivierung und Distribution an das Datenarchiv zu übergeben. Zusätzlich können sie die Nutzungsfreundlichkeit testen und tragen damit auch zu einer Weiterentwicklung des Angebots bei.

Im weiteren Projektverlauf wird in einer Phase II (s. Abb. 3) das Angebot von datorium einen eigenen Datenbestand enthalten, der nicht durch die Standardprozesse des Archivs behandelt wird. Nach Implementierung dieser Phase II durchlaufen alle weiteren und neu hochgeladenen Forschungsdaten im datorium-Portal einen neuen Review-Prozess durch das Datenarchiv. In diesem wird entschieden, ob die Forschungsdaten – falls durch die Datengebenden gewünscht – im Standardarchiv archiviert werden oder ob sie ohne eine weitere Bearbeitung durch das Datenarchiv in datorium gehalten und dort publiziert werden. Für Studien in datorium ist eine Basissicherung als bitstream preservation vorgesehen, denn diese deckt die unmittelbaren Bedürfnisse nach Sicherung und Bereitstellung ab. Wenn Wissenschaftlerinnen und Wissenschaftler eine echte Langzeitarchivierung durch das Datenarchiv wünschen, können sie dies bei der Einstellung ihrer Daten und Metadaten in datorium angeben. Eine Entscheidung über diesen Wunsch nach Aufnahme in das Standardarchiv kann allerdings nur im Review-Prozess durch das Datenarchiv getroffen werden, da eine Übernahme der Daten in die Standardarchivierung mit Ressourcenaufwand verbunden ist.

Darüber hinaus können die Nutzenden in datorium auch bestimmen, ob ihre Daten publiziert oder ohne Sichtbarkeit nach außen lediglich archiviert und langfristig gesichert werden sollen. Im Fall der Archivierung ohne Publikation findet keine Datenaufbereitung durch das Datenarchiv statt.

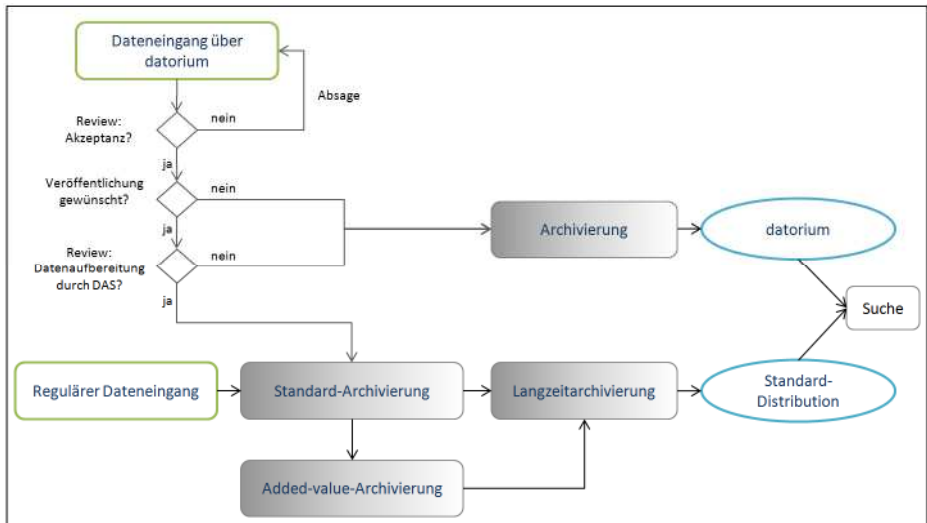


Abbildung 3: Workflow datorium und regulärer Dateneingang – Phase II

Für Datennutzende sollen in Phase II alle Bestände des Datenarchivs gemeinsam durchsuchbar gemacht werden, so dass nur ein Zugangspunkt angelaufen werden muss. Dazu wird die Suche im Datenbestandskatalog so ausgebaut werden, dass sie die bisherigen Bestände des Datenarchivs und die Bestände in datorium umfasst (s. Kap. 9). Für den Zugang zu den Daten durch Dritte werden im Standardarchiv weiterhin die bisherigen Regeln gelten. Für datorium können die Datengebenden selbst die Zugänglichkeit ihrer Forschungsdaten bestimmen.

5 Zielgruppen und Inhalte

GESIS versteht sich als Teil der globalen sozialwissenschaftlichen Forschungsgemeinschaft. Allgemein richtet sich das Angebot des Instituts primär an Forscherinnen und Forscher der empirischen Sozialforschung mit einem Schwerpunkt auf den Fachgebieten Soziologie und Politikwissenschaft sowie an solche der Sozialwissenschaften insgesamt. Weitere Zielgruppen liegen im politischen, sozialen und kommerziellen Umfeld der Sozialwissenschaften.

Ein Großteil der Datengeberinnen und Datengeber im Datenarchiv kommt zurzeit aus größeren sozialwissenschaftlichen Forschungseinrichtungen mit dem Schwerpunkt auf Langzeitstudien, wie z.B. ALLBUS, European Values Study, International Social Survey Programme, Eurobarometer, Politbarometer oder German Longitudinal Election Study. Mit dem Service des datorium-Portals wird deshalb u.a. eine Erhöhung der archivierten Einzelstudien angestrebt, indem beispielsweise Forschungsdaten von Datengeberinnen und Datengebern aus kleineren Forschungsprojekten akquiriert werden. Die thematische Breite sozialwissenschaftlicher Forschungsdaten, etwa aus der Medienforschung, der Gesundheitsforschung, Wirtschaftsforschung oder Umweltforschung soll verbessert

werden. Denkbar ist außerdem eine Ausweitung der Zielgruppe auf andere wissenschaftliche Disziplinen (wie z.B. Geistes- oder Erziehungswissenschaften), um interdisziplinäre Forschungssynergien zu unterstützen.

Das geplante Daten-Repository soll wissenschaftliche Forschungsbeiträge der genannten Zielgruppen enthalten, die bisher noch keinen Eingang in das Standardarchiv von GESIS gefunden haben. Somit kann ein solches Repository eine sinnvolle Ergänzung zu existierenden Archivbeständen darstellen. Kleinere Forschungsprojekte z.B. aus DFG-Förderung oder vergleichbaren Drittmittelprojekten sind aktuell im Bestand des Archivs tendenziell eher unterrepräsentiert. Die geförderten Projekte der DFG oder die nachgewiesenen Projekte in SOFIS bieten ein breites mögliches Erschließungspotential für datorium (s. Kap. 3).

Darüber hinaus können Forscherinnen und Forscher z.B. ihre Ergänzungen zu einem Archivdatensatz von GESIS – z.B. Kontextdaten zu einem Eurobarometer – ins Repository laden. Oder sie können ihre noch unveröffentlichten Forschungsprojekte publizieren und diesbezügliche Aufsätze, Berichte sowie Vorträge hervorheben oder aber auf die aktuelle Forschungsentwicklung ihrer Studie hinweisen, wodurch andere Forscherinnen und Forscher den Fortschritt des Projekts verfolgen können. Außerdem eignet sich datorium für die Aufnahme von Daten, die im Rahmen von Qualifizierungsarbeiten, z.B. zu Dissertationen oder Habilitationen, erstellt worden sind. Ebenfalls geplant ist die Aufnahme von Analyse-Syntaxen zur Generierung von Replikationsdatensätzen, die der scientific community für Reanalysen zur Verfügung gestellt werden sollen. Weitere Repositorysinhalte könnten Daten zu wissenschaftlichen Aufsätzen sein, die in Zeitschriften publiziert werden sollen, wenn diese Regeln für die Archivierung und Zugänglichmachung verabschiedet haben [VS2012].

Wissenschaftlerinnen und Wissenschaftler können durch diese Form der Web-Publikation Anerkennung für ihre Forschungsarbeiten mit Daten erhalten, auch wenn ihre inhaltlichen wissenschaftlichen Publikationen bisher nicht in einem etablierten Journal veröffentlicht wurden. Vorteile gegenüber einer Publikation von Forschungsdaten auf einer eigenen Homepage sind, dass die Sichtbarkeit in einem disziplinären Repository deutlich höher ist, und dass zunächst die mittelfristige, später auch die langfristige Sicherung, Verfügbarkeit und Identifizierung der Forschungsdaten gewährleistet sind.

6 Daten-Review

Die in datorium eingestellten Daten und Metadaten durchlaufen nach Fertigstellung von Phase II nicht mehr den Dokumentations- und Archivierungsprozess der Standardarchivierung (vgl. Abb. 3). Eine Prüfung dieser Daten auf Vollständigkeit, Sinnhaftigkeit, Format, Einhaltung des Datenschutzes, etc. erfolgt ab Phase II mittels eines Reviews durch GESIS innerhalb von datorium. Im Reviewprozess wird u.a. eine Studiennummer zugewiesen, die keine inhaltliche Bedeutung trägt und ähnlich wie die im Standardarchiv verwendeten ZA-Nummern zur schnellen Identifizierung einer Studie

dient. Dieser Reviewprozess soll in stark standardisierter Form nur eine grundlegende Prüfung leisten, damit er für eine große Anzahl an Studien durchführbar ist. Nach der Einführungsphase von datorium wird eine Evaluation erfolgen, ob die Einzelheiten des Reviewprozesses neu festzulegen sind, um dieses Ziel zu erreichen. Eine hochgeladene Studie wird erst dann nach außen publiziert, wenn sie im Rahmen der Dateneingangskontrolle die Kriterien des Reviews erfüllt und eine Freischaltung durch das Datenarchiv erhält.

In Ausnahmefällen wird bei Erfüllung spezieller Qualitätskriterien im Reviewprozess von GESIS entschieden, ob eine Studie aus dem datorium-Portal in das Standardarchiv übernommen wird und hier den Aufbereitungsprozess und die Dokumentation der Standardarchivierung durchläuft. Dies können inhaltliche Kriterien sein, z.B. eine hohe Relevanz des Untersuchungsgegenstandes, oder auch methodische Kriterien, wie die Repräsentativität der Stichprobe oder ein innovatives Untersuchungsdesign. Für solche Studien ist dann auch die Langzeitarchivierung gegeben, also eine langfristige Sicherung und Interpretierbarkeit der Daten z.B. durch Formatmigration, die zu einer höheren Sicherheit der zukünftigen Nutzbarkeit für diese Studie führt. Da die Ressourcen des GESIS-Datenarchivs begrenzt sind, muss im Review-Prozess eine Auswahl getroffen werden. Denkbar ist bei verfügbaren Ressourcen auch, dass eine Added-value-Dokumentation einer solchen Studie durchgeführt wird. Dieses wird in solchen Ausnahmefällen mit den Datengeberinnen und Datengebern einzeln vereinbart.

7 Registrierung von DOI-Namen

Zur sicheren und dauerhaften Zitierung von Forschungsdaten bietet GESIS in Kooperation mit DataCite, der internationalen Initiative zur Unterstützung des Zugangs zu digitalen Forschungsdaten, den DOI Registrierungsservice für sozial- und wirtschaftswissenschaftliche Daten in Deutschland dalra an [HZ2011]. Ebenso wie andere persistente Identifikatoren, etwa Handle oder URN, ist der DOI-Name unveränderlich und identifiziert ein Objekt unmittelbar, also nicht lediglich eine Eigenschaft des Objekts, wie beispielsweise die Adresse, an der es platziert ist. Darüber hinaus wird ein Objekt durch den DOI-Namen mit aktuellen und strukturierten Metadaten verknüpft, wozu dann auch die Adresse oder URL gehört, an der das Objekt zu finden ist. Die Benutzung von DOIs zur Zitierung von Zeitschriftenaufsätzen ist bereits seit langem etabliert. Durch die Vergabe von DOIs für Forschungsdatensätze kann daher die Zitierung dieser Daten erleichtert werden. Dies ist auch eine zentrale Voraussetzung für eine dauerhafte Verknüpfung von Forschungsergebnissen in Form von Berichten und Publikationen und den ihnen zugrunde liegenden Primärdaten. Die Metadaten der Forschungsdaten in datorium werden in dalra zur Verfügung gestellt und werden dann dort auch recherchierbar sein [Ha2013].

Dieser Service der Datenregistrierung zur Zitierung und Verlinkung von elektronischen Ressourcen soll ebenfalls in datorium angeboten werden. In Phase I erfolgt die Vergabe einer DOI noch über die Standardarchivierung im Datenbestandskatalog mit dem Präfix „10.4232“. In Phase II wird eine direkte Registrierung der DOI-Namen aus datorium heraus mit dem Präfix „10.7802“ erfolgen, da datorium dann einen zusätzlichen Bestand

an Daten enthalten wird. Objekte dürfen aus datorium nicht gelöscht werden, sondern lediglich der Zugang zum Objekt kann beschränkt werden. Die dazugehörigen Metadaten zum Objekt bleiben also auf jeden Fall erhalten, so dass eine Zitation dieses Objekts auch später noch möglich ist.

8 Regeln für die Nutzung

Die Benutzung des Services von datorium erfolgt nach Zustimmung zu den allgemeinen Geschäftsbedingungen. Darüber hinaus können die Datengeberinnen und Datengeber in Phase II durch die Auswahl einer vorgegebenen Zugangskategorie flexibel definieren, welchen Personen oder Gruppen sie ihre Daten und hinterlegten Dokumente zugänglich machen wollen. Folgende Zugangskategorien sind vorgesehen:

- a) Freier Zugang: Die Forschungsdaten sind ohne Einschränkung für alle registrierten Nutzer und Nutzerinnen zugänglich, ohne dass eine Einwilligung der Datengebenden eingeholt werden muss.
- b) Eingeschränkter Zugang: Für den Zugang zu den Forschungsdaten muss eine Einwilligung der Datengebenden über das System eingeholt werden. Diese bestimmen eigenständig, welchen registrierten Nutzerinnen und Nutzern sie den Zugang erlauben.
- c) Kein Zugang: Datengeberinnen und Datengeber möchten Forschungsdaten ausschließlich archivieren und nicht für Andere zugänglich machen. Diese Daten werden nicht veröffentlicht.

Die Zugangsberechtigung zu eingeschränkt zugänglichen Daten für Sekundäranalysen muss bei den Datengeberinnen und Datengebern beantragt werden. Diese können eigenständig entscheiden, welchen Nutzerinnen und Nutzern sie ihre Daten zugänglich machen möchten. Für diesen Zweck wird es im System eine Möglichkeit geben, den Zugang zu diesen Daten zu beantragen. Daraufhin wird vom System eine E-Mail an die in den Anmelde Daten hinterlegte E-Mail-Adresse der Datengeberin oder des Datengebers gesendet. Diese wird auch in Kopie an den Service von datorium zugestellt, so dass eine Dokumentation der Kontaktaufnahme möglich ist. Die Angefragten können sich in datorium einloggen und die Zugangsberechtigung erteilen oder verweigern. Reagieren Datengeberinnen oder Datengeber nicht innerhalb einer definierten Zeit, nimmt GESIS Kontakt zu ihnen auf, um den Vorgang zu bearbeiten. Für den Fall, dass Datengeberinnen und Datengeber nicht mehr erreichbar sind (z.B. im Todesfall), müssen diese bereits vorher in der Vertragsvereinbarung bestimmen, welche Zugangskategorien und -bedingungen für ihre Daten gelten sollen.

9 Technische Implementierung

Für die Implementierung des datorium-Portals wurde entschieden, eine bereits existierende Repository-Software zu nutzen, bei der bereits auf Erfahrungen in der Anwendung zurückgegriffen werden konnte. Dazu wurde eine ausführliche Evaluation

verschiedener Repository-Softwarelösungen durchgeführt. Verbreitete Lösungen für ähnliche Anwendungen sind DSpace, EPrints, Fedora, MyCoRe, und OPUS, CKAN, Git, Mendeley oder Dataverse (vgl. [Op2004], [Ku2012], [Lo2006], [Za2011], [Cr2011]). Auf dieser Basis fiel die Wahl auf DSpace, eine nicht-kommerzielle Open Source Plattform zum Betrieb eines Dokumentenservers. Ausschlaggebende Faktoren für die Wahl von DSpace waren u.a. die flexible Erweiterbarkeit von Metadatenfeldern, ein auf GESIS anpassbares Rechtemanagement, die freie Definition von Workflows, die benutzungsfreundliche Programmiermöglichkeit einer multilingualen Oberfläche, sowie die positiven Erfahrungen durch den Betrieb von SSOAR. Für den Einsatz im GESIS-Datenarchiv lässt sich DSpace in die vorhandenen Mechanismen gut integrieren. Dies sind Eigenschaften, auf die bei der Verwendung von anderen Systemen hätte verzichtet werden müssen, und die die unabhängige Gestaltung des Portals und die Anpassung an neue mögliche Nutzungsanforderungen ermöglichen.

Der Schwerpunkt von DSpace liegt auf der Erfassung, Speicherung und Distribution von digitalen Ressourcen [DHD2011]. Die Software wurde vom Massachusetts Institute of Technology und der Forschungsabteilung von Hewlett Packard in Anlehnung an das OAIS-Referenzmodell entwickelt und ist architektonisch in der Lage, Strategien zur Langzeitverfügbarkeit von digitalen Ressourcen zu unterstützen. Die Verbreitung findet unter der BSD-Lizenz statt. Der Einsatz dieser Software findet primär in Universitäten, Bibliotheken und Forschungseinrichtungen statt.

Prinzipiell wird im Rahmen der technischen Implementierung berücksichtigt, dass möglichst keine neuen Suchschnittstellen erzeugt werden müssen, sondern vorhandene, z.B. des Datenbestandskatalogs, verwendet werden können. Für die Implementierung in Phase II ist eine gemeinsame Bestands-Durchsuchung des datorium-Portals und der Distributionswege des Standardarchivs geplant. Dazu wird ein Suchindex auf der Basis von Solr verwendet werden, wie er bereits bei dalra in Verwendung ist. Da das Metadatenschema von datorium auf dem Datenbestandskatalog DBK aufbaut und auch gemeinsame kontrollierte Vokabulare enthält, ist die semantische Integration der Suche möglich. Dieses stellt eine enorme Erleichterung für die Nutzenden dar, da sie mittels einer integrierten Suche in datorium und im DBK alle bei GESIS verfügbaren Forschungsdaten finden können. Darüber hinaus werden die Metadaten auch für Recherchen in dalra verfügbar sein, wo Nutzende breit gefächerte Suchanfragen durchführen und auch Bestände von anderen Institutionen finden können.

10 Ausblick

Das Angebot der Phase I von datorium zur Aufnahme von Forschungsdaten ins Archiv für die Standardarchivierung steht bereits seit Juni 2013 zur Verfügung. Dieses Angebot wird mit einzelnen Datengeberinnen und Datengebern getestet und verbessert. Gegenwärtig arbeitet GESIS an der Umsetzung der Phase II von datorium. Dabei müssen Datengeberinnen und Datengeber die alternative Möglichkeit erhalten, die Archivierung und Distribution ihrer Forschungsdaten über das datorium-Portal eigenständig und selbstbestimmt durchzuführen. Die bisherige Planung sieht vor, dass datorium zur Aufnahme von Forschungsdaten mit Phase II gegen Ende des Jahres 2013

zur Verfügung stehen soll. Zusätzlich dazu ist bereits eine Weiterentwicklung des Angebots in Zusammenarbeit mit den Einrichtungen DIW (Deutsches Institut für Wirtschaftsforschung Berlin), WZB (Wissenschaftszentrum Berlin für Sozialforschung) und ZBW (Deutsche Zentralbibliothek für Wirtschaftswissenschaften Leibniz-Informationszentrum Wirtschaft) im Rahmen eines Drittmittelanspruchs für die Jahre 2014-2016 geplant. Hiermit soll sowohl eine technisch als auch konzeptionell anspruchsvollere Entwicklung zu einer Realisierung der Anforderungen aus der Profession führen. Wenn mit diesen Arbeiten eine größere thematische Breite und eine höhere Anzahl an gesicherten und langzeitarchivierten, sowie zur Verfügung gestellten Forschungsdaten in den Sozialwissenschaften erreicht werden kann, werden damit die Empfehlungen zahlreicher Organisationen zur Verbesserung der Langzeitarchivierung und des Zugangs zu öffentlich finanzierten Forschungsdaten in den Sozialwissenschaften umgesetzt.

Literatur

- [BZ2011] Brislinger, E.; Zenk-Möltgen, W.: Findings of the original language documentation for European Values Study (EVS) 2008. Präsentation auf der IASSIST 2011 “Data Science Professionals: A Global Community of Sharing”. Vancouver, 31.05. 2011. Online http://www.iassistdata.org/downloads/2011/2011_b3_brislinger_etal.pdf [Zugriff am 18.04.2013], 2011.
- [Cr2011] Crosas, M.: The Dataverse Network. An Open-Source Application for Sharing, Discovering and Preserving Data. In: D-Lib Magazine, Vol. 17, Nr. 1/2 doi:10.1045/january2011-crosas, 2011.
- [Df2013] DFG: GEPRIS Datenmonitor. Online <http://gepris.dfg.de/gepris/OCTOPUS/?module=gepris&task=showMonitor> [Zugriff am 18.04.2013], 2013.
- [DHD2011] Droogmans, L.; Hollister, V.; Donohue, T.: DSpace institutional repository platform. Online: <https://atmire.com/labs17/bitstream/handle/123456789/7641/0000082812f0000000267.pdf?sequence=1> [Zugriff am 19.04.2013], 2011.
- [DS2010] Dobratz, S.; Schoger, A.: Grundkonzepte der Vertrauenswürdigkeit und Sicherheit. In (Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Huth, K., Hrsg.): nestor-Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3. Kap. 5.2. urn:nbn:de:0008-2010071949, 2010.
- [Eu2010] European Union: Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data. A submission to the European Commission. Online <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf> [Zugriff am 20.2.2013], 2010.
- [Ha2013] Hausstein, B.; Quitzsch, N.; Jeude, K.; Schleinstein, N.; Zenk-Möltgen, W.: dalra Metadata Schema. Version 2.2.1. GESIS Technical Reports, 2013/03, 2013.
- [HZ2011] Hausstein, B.; Zenk-Möltgen, W.: dalra – Ein Service der GESIS für die Zitiation sozialwissenschaftlicher Daten. In (Schomburg, S.; Leggewie, C.; Lobin, H.; Puschmann, C., Hrsg.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. Beiträge der Tagung vom 20./21. September 2010, Köln, 2., ergänzte Fassung, 2011; S. 139–147.
- [Ko2012] Kolle, Ch.: Wissenschaftliche Literaturrecherche. In (Berninger, I.; Botzen, K.; Kollé, Ch.; Vogl, D.; Watteler, O.): Grundlagen sozialwissenschaftlichen Arbeitens. Verlag Barbara Budrich, Opladen & Toronto, 2012; S. 33-61.
- [Ku2012] Kunze, S. R.: Ein Personalisierungskonzept für Dataset-Repositorys am Beispiel von CKAN. In (Technische Universität Chemnitz – Fakultät für Informatik,

- Hrsg.): Studentensymposium Informatik Chemnitz 2012. Tagungsband zum 1. Studentensymposium Chemnitz vom 4. Juli 2012, Chemnitz, 2012; S. 27-37.
- [Lo2006] Loeliger, J.: Collaborating With Git. A variety of clever tools allow you to share your software. In: Linux Magazine, June 2006.
- [Ma2012] Mauer, R.: Das GESIS Datenarchiv für Sozialwissenschaften. In (Altenhöner, R.; Oellers, C., Hrsg.): Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen. Scivero, Berlin, 2012; S. 197-215.
- [Oe2013] OECD: New Data for Understanding the Human Condition: International Perspectives. OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences. Online <http://www.oecd.org/sti/scitech/new-data-for-understanding-the-human-condition.pdf> [Zugriff am 18.04.2013], 2013.
- [Op2004] Open Society Institute: A Guide To Institutional Repository Software. New York, 3rd edition. Online http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf [Zugriff am 20.02.2013], 2004.
- [PDF2007] Piwowar, H. A.; Day, R. S., Fridsma, D. B.: Sharing Detailed Research Data Is Associated with Increased Citation Rate. In: PLoS ONE, Vol. 2, Nr. 3, S. e308. doi:10.1371/journal.pone.0000308, 2007.
- [Re2013] RECODE: Policy RECommendations for Open Access to Research Data in Europe (RECODE) project. Online: <http://recodeproject.eu/> [Zugriff am 18.04.2013], 2013.
- [So2013] SOFIS: Projekte. Online <http://www.gesis.org/sofiswiki/Kategorie:Projekte> [Zugriff am 18.04.2013], 2013.
- [VS2012] Vlaeminck, S; Siegert, O.: Welche Rolle spielen Forschungsdaten eigentlich für Fachzeitschriften? Eine Analyse mit Fokus auf die Wirtschaftswissenschaften. In (RatSWD Working Paper Series) No. 210, November 2012.
- [Wi2012] Wissenschaftsrat: Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. Drs. 2359-12 vom 13.7.2012. Berlin. Online <http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf> [Zugriff am 20.02.2013], 2012.
- [Za2011] Zaugg, H.; West, R. E.; Tateishi, I.; Randall, D. L.: Mendeley: creating communities of scholarly inquiry through research collaboration. In: TechTrends, Vol. 55, Nr. 1, S. 32–36. doi: 10.1007/s11528-011-0467-y, 2011.
- [Ze2012] Zenk-Möltgen, W.: The metadata in the data catalogue DBK at the GESIS data archive. Präsentation auf dem RatSWD Workshop "Metadata and Persistent Identifiers for Social and Economic Data". Berlin, 7.-8.05.2012, Online http://www.ratswd.de/ver/docs_PID_2012/Zenk-Moeltgen_PID2012.pdf [Zugriff am 18.04.2013], 2012.
- [ZH2012] Zenk-Möltgen, W.; Habel, N.: Der GESIS Datenbestandskatalog und sein Metadatenschema, Version 1.8. GESIS Technical Reports, 2012/01, Online http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-01.pdf [Zugriff am 18.04.2013], 2012.