Reinhard Altenhöner und Claudia Oellers (Hrsg.)

LANGZEITARCHIVIERUNG VON FORSCHUNGSDATEN

STANDARDS UND DISZIPLINSPEZIFISCHE LÖSUNGEN

SCIVERO

Langzeitarchivierung von Forschungsdaten

Standards und disziplinspezifische Lösungen

Reinhard Altenhöner
Claudia Oellers
(Herausgeber)

SCIVERO

ISBN 978-3-944417-00-4
1. Auflage 2012

Dieses Buch entstand im Nachgang eines vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) gemeinsam mit dem deutschen Kompetenznetzwerk zur digitalen Langzeitarchivierung nestor und dem GESIS-Leibniz-Institut für Sozialwissenschaften in der Deutschen Nationalbibliothek in Frankfurt veranstalteten Workshops zum Thema „Archivierung sozial- und wirtschaftswissenschaftlicher Datenbestände". Dank und Anerkennung gelten ausdrücklich den beteiligten Personen und Institutionen, insbesondere dem Bundesministerium für Bildung und Forschung (BMBF), welches die Aktivitäten des RatSWD finanziert und unterstützt.

Lektorat: Dr. Gabriele Rolf-Engel
Korrekturen: Simon Wolff
Covergestaltung und Satz: Sören Schumann

Inhalt

Vorwort

Immer mehr für die wissenschaftliche Sekundärnutzung potenziell interessante Daten sind in digitaler und damit leicht speicherbarer und weitergebbarer Form verfügbar. Auf dieser Datengrundlage lassen sich neue Forschungsfragen bearbeiten, aber auch – und dies ist ein konstituierendes Merkmal von Wissenschaftlichkeit – bereits erstellte Analysen replizieren. Damit diese wertvollen Datensätze nicht verloren gehen, sind nachhaltige Konzepte der Langzeitarchivierung erforderlich.

Der im September 2011 gemeinsam vom *Rat für Sozial- und Wirtschaftsdaten (RatSWD)*, dem *Kompetenznetzwerk Langzeitarchivierung nestor* und dem *Leibniz Institut für Sozialwissenschaften GESIS* in der Deutschen Nationalbibliothek in Frankfurt veranstaltete Workshop zur *„Archivierung sozial- und wirtschaftswissenschaftlicher Datenbestände"* hatte zum Ziel, sich über „Best Practice" im Bereich der Langzeitarchivierung auszutauschen und die verschiedenen Akteure und Initiativen zusammenzubringen.

Welche Infrastrukturen sind erforderlich, damit Daten auch zukünftig auffindbar und analysierbar sind? Welche Zeithorizonte meinen wir, wenn von Langzeitverfügbarkeit die Rede ist? Wer entscheidet, welche Daten aufgehoben werden sollen? Wie lassen sich personenbezogene Daten sichern? Und nicht zuletzt: Wer sind in diesem Prozess die verantwortlichen Akteure?

Das vorliegende Buch, welches die Ergebnisse des Workshops zusammenfasst, geht über einen bloßen Konferenzbericht hinaus und gibt einen umfassenden Überblick über bestehende Standards zur Archivierung, aber auch über disziplinspezifische Besonderheiten und daraus resultierende Anforderungen an Infrastrukturen und Policies. Die Rahmenbedingungen von Forschungsinfrastrukturen haben sich in den letzten Jahren grundlegend verändert: Durch das Internet, neue Möglichkeiten der Digitalisierung und höhere Rechnerkapazitäten wurden die Sicherung, die Bereitstellung und auch die Archivierung von Forschungsdaten erleichtert. An transparenten, benutzerfreundlichen und koordinierten Strukturen fehlt es jedoch hier und dort. Prozeduren und Richtlinien sind bislang ungenügend etabliert. Am Beginn eines Forschungsprojektes wird noch immer zu selten an Datenmanagementpläne gedacht, welche auch die Langzeitarchivierung im Sinne der „guten wissenschaftlichen Praxis" umfassen. Langzeitarchivierung ist nur als arbeitsteiliges Konzept realisierbar, die Verständigung über Verantwortlichkeiten und Zuständigkeiten ist noch nicht abgeschlossen.

Vor diesem Hintergrund hoffen wir, dass hier eine Basis für weiterführende Diskussionen in einem sehr wichtigen und zentralen Bereich des Forschungsdatenmanagements vorgelegt wird.

Dank in diesem Zusammenhang gilt natürlich in erster Linie den Autorinnen und Autoren, von denen die meisten bereits als Referentinnen und Referenten im Rahmen des Workshops mit ihren Präsentationen wie Diskussionsbeiträgen zu einer gelungenen Veranstaltung und somit auch überhaupt zu der Idee beigetragen haben, die Ergebnisse in einem Band zusammen zu tragen und zu veröffentlichen.

Nicht zuletzt haben zum Erscheinen dieses Bandes vor allem beigetragen Dr. Gabriele Rolf-Engel, die kompetent und engagiert das Lektorat durchgeführt hat, Simon Wolff, verantwortlich für Korrekturen und Vereinheitlichungen und Sören Schumann, verantwortlich für Satz und Layout. Ihnen allen sei an dieser Stelle herzlich gedankt!

Claudia Oellers (RatSWD)
Natascha Schumann (DNB, seit Mitte 2012 GESIS)
Brigitte Hausstein (GESIS)

Herausforderungen der Archivierung sozial-, verhaltens- und wirtschaftswissenschaftlicher Datenbestände

Reinhard Altenhöner und Gert G. Wagner

Der Umfang von im Prinzip mehrfach analysierbaren beziehungsweise re-analysierbaren Forschungsdaten nimmt generell beständig zu und somit auch die Herausforderung, die Daten langfristig für die Wissenschaft verfügbar zu machen und zu archivieren. Der Zugang zu Forschungsdaten für Sekundäranalysen entspricht nicht nur den Regeln guter wissenschaftlicher Praxis, sondern stellt eine Voraussetzung für gleichermaßen innovative wie kostengünstige Wissenschaft dar. Die zunehmende Relevanz eines persistenten Zugangs zu Forschungsdaten wird auch dadurch sichtbar, dass – erleichtert durch die technischen Möglichkeiten der digitalen Publikation – Umfang und Dichte der direkten Verknüpfungen von wissenschaftlichen Publikationen zu den jeweiligen Forschungsdatenbeständen wachsen und damit der nahtlose Rückgriff auf die Quellen, auf denen eine wissenschaftliche Publikation aufsetzt, an Bedeutung gewinnt.

Damit Forschungsdaten für Sekundäranalysen dauerhaft zur Verfügung stehen, müssen diese archiviert und für potentielle Nutzer auffindbar sein. Dies umfasst neben der bloßen Substanzerhaltung (Technik und Software veralten bekanntlich schnell) auch die Frage, wie die oftmals komplexen Datensätze von späteren Nutzern richtig interpretiert werden können.

Entsprechende Aufgabenstellungen im Bereich der Zusammenführung und Erschließung und bei der Entwicklung geeigneter technischer Werkzeuge sind bislang breit verteilt und unterschiedlich weitreichend geregelt. Während beispielsweise für die unmittelbare Verbreitung in der wissenschaftlichen Community publizierte Materialien wie Zeitschriftenaufsätze und Bücher – seien sie digital oder als Print produziert – in den Sammelauftrag der Deutschen Nationalbibliothek fallen, ist die Situation bei Forschungsdaten sehr viel offener: Hier gibt es neben den Selbst- oder Förderverpflichtungen, die die Bewahrung und die öffentliche Zugreifbarkeit für einen begrenzten Zeitraum von 10 Jahren durch die betreuende Einrichtung statuieren, überregionale, meist fachlich definierte Einrichtungen, welche Archivierungsaufgaben übernehmen und in unterschiedlicher Tiefe und Kompetenz die Verantwortung für einzelne Ausschnitte aus der Gesamtmenge entstehender Daten übernehmen.

Dieses Nebeneinander von Sammlungsprinzipien wie der Gattung der Publikation (eine Veröffentlichung in einer Zeitschrift oder ein Set von Forschungsda-

ten) oder der fachlichen Abgrenzung (ein fachliches Repository gegenüber einer allgemein sammelnden Einrichtung wie zum Beispiel auch einem Hochschul-Repository) erschwert nicht nur aus der Nutzungsperspektive den Zugriff auf eng miteinander zusammenhängendes Material (wo recherchiere ich?), sondern stellt auch für die beteiligten Einrichtungen in zunehmendem Maße ein Problem dar, denn Such- und Verknüpfungsstrukturen müssen übergreifend funktionieren und vor allem dauerhaft verfügbar sein (also nicht nur die digitalen Objekte selbst).

So wie die materialspezifischen Aufteilungen immer weniger greifen (was schon daran deutlich wird, dass auch Verlage dazu übergehen, eigene Angebote zur Speicherung und Bereitstellung zumindest der aggregierten, für die Publikation wesentlichen Mess- und Datenreihen anzubieten), ist auch die fachliche Auftrennung von Beständen immer schwieriger zu vermitteln und durchzuhalten. Umso wichtiger wird es, gemeinsame Vorgehensweisen für die Erschließung und für die Sicherung der Langzeitverfügbarkeit des Materials zu verfolgen. Nur unter diesen Bedingungen kann ein/e forschende/r Wissenschaftler/in alle miteinander zusammenhängenden Materialien herbeiziehen, Literatur rezipieren, Erhebungsreihen sichten und ggf. mit geeigneten Werkzeugen validieren, dann aber auch Vergleichsreihen bilden und Schritt um Schritt in eine eigene Forschungsarbeit überwechseln. Eine wesentliche Voraussetzung dafür ist, dass die Forschungsdaten regelmäßig auf ihre Benutzbarkeit hin überprüft werden, ggf. verlustfrei und dokumentiert in eine mit einem moderneren Werkzeug nutzbare Form migriert werden oder aber geeignete Softwarewerkzeuge bereitstehen, die in einer aktuellen technischen Umgebung ablauffähig sind, aber den Zugang und die Arbeit mit „alten" Daten (hier sind 10 Jahre oft schon eine relevante Hürde) erlauben.

Was für die Kooperation und das Ziel einheitlicher Verzeichnungs- und Verweisungsstrukturen klar formuliert werden kann, gilt also in ähnlicher Weise für die Frage, wie denn die einzelnen archivierten Materialgruppen zugänglich gehalten werden. Hier sind sparten- und institutionenübergreifende Kooperationsstrukturen gefordert, die es Mitwirkenden auch auf pragmatischer Ebene erleichtern, Know-how und Expertise wechselseitig zur Verfügung zu halten.

Es gibt mittlerweile – variierend von Disziplin zu Disziplin – eine Vielzahl an Ansätzen und Initiativen, die den Prozess der Archivierung unterstützen und Standards bereitstellen. Was allerdings wie gezeigt fehlt, ist eine einheitliche Strategie, in welcher Form Forschungsdaten archiviert, zugänglich, auffindbar, aber auch zitiert werden. Ohne die Möglichkeit der Zitation und der damit verbundenen wissenschaftlichen Reputation fehlen wichtige Anreizstrukturen für die Wissenschaftler, diese Daten für die Nachnutzung bereit zu stellen. Dies wird besonders in den Sozial- und Wirtschaftswissenschaften deutlich. In diesen Disziplinen gibt es seit langem Datenarchive (länger als zum Beispiel in den Verhaltenswissenschaften), aber viele Datensätze sind nach wie vor nicht zugänglich

oder nur schwer zu finden. Hier bedarf es gemeinsamer Festlegungen auf einheitliche Beschreibungs- und Datenaustauschstrukturen, also der Formulierung von allen Beteiligten aktiv getragener Vereinbarungen, die sich zudem an international gängigen Standards orientieren sollten.

Ein großes Problem ist die angemessene Würdigung der Datenproduktion und -archivierung in den wissenschaftlichen Fachdisziplinen. Auch dieses Problem ist in den Sozial- und Wirtschaftswissenschaften besonders deutlich, da es dort – trotz längerer Tradition – für Nachwuchswissenschaftler riskant ist, sich der Datenproduktion und -archivierung zu widmen, denn mit dieser zeitaufwendigen Arbeit sind keine wissenschaftlichen Lorbeeren zu ernten. Deswegen besteht im Zeitalter der Explosion von Forschungsdaten eine der großen Aufgaben darin, innerhalb der Fachdisziplinen die Datenproduktion und -archivierung so zu würdigen, dass Spitzen-Nachwuchskräfte bereit sind, sich in diesem Feld zu engagieren. Das Verbesserung von Zitierbarkeit und Auffindbarkeit von Forschungsdaten und weltweite eindeutige *„Researcher Identifier"*, also eineindeutige Identifikationsnummern für einzelne Forscher und ihre Ergebnisse (Daten und Schriften), könnten hier entscheidend weiterhelfen. Denn erst derartige Referenzsysteme würden es im Zeitalter der Messung von Forschungsleistung ermöglichen, Datenproduktion und -archivierung zitierbar und damit messbar zu machen.

Die Bedeutung der Zitierbarkeit von Forschungsdaten und ihrer Produzenten verweist unmittelbar auf die Bedeutung von Bibliotheken für die Welt der Forschungsdaten. Zugleich sind Bibliotheken Spezialisten für Langzeitarchivierung. Insofern war es geradezu überfällig, dass mit dem Workshop, der diesem Sammelband zugrunde liegt, endlich Datenproduzenten (aus dem Bereich der Sozial-, Verhaltens- und Wirtschaftswissenschaften, einschließlich der amtlichen Statistik), Datenarchive und Bibliotheken zu einem intensiven Gedankenaustausch zusammengebracht wurden. Angeregt durch fachlich eher enger ausgelegte Vorträge ergaben sich weitgespannte Diskussionen im Kreis der heterogenen Teilnehmer. Eine wesentliche Erkenntnis der Überlegungen war, dass Archive (gleich welcher Art) nur so gut sein können wie die Qualität ihrer Zusammenarbeit mit den Datenproduzenten. Das Know-how wie die Daten entstanden sind, gepaart mit Methodenwissen zu ihrer Verzeichnung und Beschreibung, öffnet das Tor zu einer nachhaltigen Archivierung, deren Qualität sich in der Benutzbarkeit des Datensets auch für den Wissenschaftler und die Wissenschaftlerin aus der Nachbardisziplin erweist.

Zwar beziehen sich die Beiträge dieses Bandes ausschließlich auf Archiv- und Bibliotheksfragen. Sie sind aber vor dem Hintergrund der Diskussion dieser Veranstaltung und der Brisanz der damit verknüpften Fragen sowohl für Datenproduzenten als auch für Datenarchive in den Sozial-, Verhaltens- und Wirtschaftswissenschaften gleichermaßen von Interesse.

Der vorliegende Sammelband, der die Vorträge eines gemeinsamen Workshops von RatSWD, nestor und GESIS im September 2011 in der Deutschen Nationalbibliothek in Frankfurt vereint, gibt so einen Überblick über bestehende Archiv- und Bibliotheksstandards und liefert einen Beitrag zur Diskussion über Voraussetzungen zur Archivierung sozial- und wirtschaftswissenschaftlicher Datenbestände. Er richtet sich somit gleichermaßen an Fachbibliotheken, Archive, Infrastruktureinrichtungen und amtliche Statistik sowie Wissenschaftler. Kurzum: Der vorliegende Band richtet sich an alle, die im weitesten Sinne mit der Verfügbarmachung von Forschungsdaten betraut sind.

Der Sammelband gliedert sich in drei Teile. Er beginnt mit einem einführenden Teil, in dem grundlegende Begriffe erklärt werden sowie einer Einführung in das OAIS-Referenzmodell (der ISO-Standard, Open Archival Information System'), dem maßgeblichen Bezugssystem für die Langzeitarchivierung gegeben wird. Aber auch die Spezifika sozial- und wirtschaftswissenschaftlicher Daten und die damit verbundenen Anforderungen an die Archivierung werden berücksichtigt.

Im zweiten Teil werden aktuelle Standards dargestellt und eine Übersicht gegeben über vertrauenswürdige digitale Archive, Metadatenstandards und Systeme der persistenten Identifizierung als zentrale Bestandteile einer langfristigen und nachhaltigen Archivierung.

Im abschließenden dritten Teil steht die Anwendung im Vordergrund; hier kommen verschiedene Einrichtungen aus unterschiedlichen Teildisziplinen, den Sozial-, Verhaltens- und Wirtschaftswissenschaften sowie den Geowissenschaften und der Klimaforschung zu Wort und stellen ihre jeweiligen Konzepte der Archivierung vor.

A KONZEPTE

Datenmanagement und Data Sharing: Erfahrungen in den Sozial- und Wirtschaftswissenschaften[1]

Denis Huschka, Claudia Oellers, Notburga Ott und Gert G. Wagner

Die Menge der für Forschungszwecke zur Verfügung stehenden Daten vergrößert sich beständig (King 2011). Jedoch werden unter der Bezechung Daten in den verschiedenen wissenschaftlichen Disziplinen ganz unterschiedliche Dinge gefasst. Aus dem Lateinischen kommend bezeichnet ein Datum zunächst einmal etwas „Gegebenes". In den Geowissenschaften können Daten Eisbohrkerne sein, aber auch numerische Geokoordinaten. In den Geschichtswissenschaften können Daten das Format alter Dokumente haben. In der Medizin können es auch biologische Proben oder Laborwerte sein. In den quantitativ empirisch arbeitenden Sozial-, Verhaltens- und Wirtschaftswissenschaften ist das „gängige" Format der einschlägigen Daten das von Zahlen als Teil von Datenmatrizen oder Tabellen.

Die unterschiedlichen Phänotypien von Forschungsdaten erfordern spezifische Datenmanagementstrategien. Oft beschreiben die Daten die Ausprägung einer Eigenschaft eines Individuums oder einer Organisation, wie zum Beispiel einer Firma. In diesen, insbesondere in der Medizin, den Sozial-, Verhaltens- und Wirtschaftswissenschaften vorkommenden Fällen spricht man von personenbeziehbaren oder firmenbeziehbaren Daten, bei deren Be- und Verarbeitung sich automatisch Fragen des Datenschutzes und der Forschungsethik stellen. Auch dies hat Auswirkungen auf das Forschungsdatenmanagement und die Zugänglichkeit dieser Art von Daten.

Obgleich im Bereich der Sozial-, Verhaltens- und Wirtschaftswissenschaften in Deutschland datenschutzrechtliche Notwendigkeiten die gemeinsame Nutzung (sozusagen das Teilen von Daten – „data sharing") erschweren, nimmt Deutschland eine Vorreiterrolle hinsichtlich des Auf- und Ausbaus einer sozial- und wirtschaftswissenschaftlichen Forschungsdateninfrastruktur ein (vgl. Solga und Wagner 2007, Habich et al. 2010, Bender et al. 2008). Der Zugang zu einschlägigen Daten hat sich in den vergangenen Jahren für die Wissenschaft deutlich verbessert. Neben den klassischen Datenarchiven (zum Beispiel dem GESIS Datenarchiv für Sozialwissenschaften – vormals Zentralarchiv für empirische Sozialforschung an der Universität Köln) sind alle vom Rat für Sozial- und Wirtschaftsdaten (RatSWD) akkreditierten Forschungsdatenzentren (FDZ) und Datenservicezentren (DSZ) Teil dieser Forschungsinfrastruktur. Die Forschungs-

1 Der Beitrag ist zuerst erschienen in: Büttner, S./Hobohm, H.-C. und Müller, L. (Hrsg.) (2011): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock+Herchen Verlag.

datenzentren und Datenservicezentren als institutionalisierte Orte des data sharing ermöglichen nicht nur den Zugang zu Daten, sondern bieten darüber hinaus einen Service um die Daten herum an. Ein solcher Service ist wegen der komplexen Strukturen vieler Datensätze und der jeweils beschränkten Aussagekraft der Daten (Reichweite, Validität und Reliabilität), welche durch die Operationalisierungen der Erhebungen bedingt sind, nötig und kann am besten von denen geleistet werden, die die Daten produzieren. In den Verhaltenswissenschaften ist eine solche Tradition des data sharing noch wenig ausgeprägt. Dies beginnt sich zu ändern (vgl. Weichselgartner 2011).

So positiv die Entwicklungen hin zu mehr Datenverfügbarkeit im Bereich der Sozial- und Wirtschaftswissenschaften und zuletzt auch in den Verhaltenswissenschaften zu bewerten sind, so aktuell ist aber auch die Frage, wie man die Daten im Rahmen einer geordneten und transparenten Infrastruktur zur Verfügung stellen kann und wie man den Zugang selbst transparent und nutzerfreundlich regelt.

Für innovative Forschung wird es zunehmend wichtiger, multi- und interdisziplinär zu arbeiten. Georeferenzierte Daten, Biomarker, Transaktionsdaten oder auch Datensätze privater Firmen stellen relativ neue und besonders reizvolle Datenquellen dar, durch deren Verknüpfung mit „herkömmlichen" sozialwissenschaftlichen Daten sich innovative Fragestellungen beantworten lassen. Auch die digitale Verfügbarkeit von Daten sowie die technologischen Möglichkeiten im Umgang mit den digitalen Daten (zum Beispiel durch persistente Identifikatoren und verbesserte Computertechnik und -leistungsfähigkeit) sind aus Sicht der Wissenschaft Chance und Herausforderung an ein systematisches Datenmanagement zugleich. Eine besondere Bedeutung wird in Zukunft deshalb der Organisation der Informationen über die Daten zukommen, also der Beschreibung der Inhalte, Qualität, Analysepotenziale, Aussagekraft und insbesondere auch der Verknüpfungsmöglichkeiten zwischen Datensätzen. Es reicht also nicht, jeden einzelnen Datensatz verfügbar zu machen. Für eine breite Nutzung in der Wissenschaft ist ein „Einstiegsportal" notwendig, in welchem ein an einem bestimmten Thema interessierter Forscher alle erforderlichen Informationen über möglichst alle relevanten zur Verfügung stehenden Datensätze finden kann. Wohlgemerkt: ein solches Portal soll und kann nicht die Daten selbst vorhalten. Dies ist, wie wir unten ausführen, aus rechtlichen Gründen nicht möglich und aus Servicegründen auch gar nicht wünschenswert. Ein solches Portal sollte lediglich die nicht zu unterschätzende Funktion eines Informationsbrokers übernehmen.

Data sharing

In den Sozial- und Wirtschaftswissenschaften hat sich in den vergangenen Jahren eine Kultur des Teilens von Daten (data sharing) durchgesetzt. Teilen ist deswegen leicht möglich, weil die mehrfache Nutzung der Daten diese nicht zerstört (wie das zum Beispiel bei Biomaterial oder Bohrkernen der Fall ist). Das systematische Argument für data sharing ist, dass nur die Möglichkeit von Re-Analysen veröffentlichter Ergebnisse diese zu wissenschaftlichen Erkenntnissen macht. Denn Wissenschaft bedeutet, dass Ergebnisse nachprüfbar sind. Hinzu kommt die praktische Überlegung, dass Daten, welche im Rahmen öffentlicher, beispielsweise durch Forschungsförderung finanzierter Unterfangen entstehen, für die breite Forschung zur Verfügung gestellt werden sollen und nicht durch einen einzelnen Forscher monopolisiert werden dürfen (der ggf. nur Re-Analysen zur Prüfung von Ergebnissen erlaubt).

Die Überprüfbarkeit von Forschungsergebnissen durch Re-Analysen gehört zu den formalisierten Kriterien guter wissenschaftlicher Praxis, die von der Deutschen Forschungsgemeinschaft (DFG 1998) erarbeitet wurden. Inzwischen wird beispielsweise in der Ökonomie vermehrt einer von wissenschaftlichen Zeitschriften gestellten Anforderung entsprochen, neben der eigentlichen Publikation auch die zugrundeliegenden Datensätze zu veröffentlichen bzw. im Falle von datenschutzrechtlich sensiblen Daten in geschützten Bereichen zugänglich zu machen.

Die Ermöglichung einer Nachnutzung der Daten durch deren Übermittlung an geeignete Datenarchive oder andere Orte ist seit langem auch Bestandteil der Förderrichtlinien der Deutschen Forschungsgemeinschaft (DFG 2010) und der entsprechenden Förderprogramme des Bundesministeriums für Bildung und Forschung (BMBF). Die konsequente Umsetzung dieser Verpflichtung ist freilich in den verschiedenen wissenschaftlichen Disziplinen unterschiedlich.

Öffentlich finanziert entstehen Daten auch im Rahmen der Politiksteuerung und durch die amtliche Statistik (vgl. Hahlen 2009) und im Rahmen der Verwaltung als sog. prozessproduzierte Datensätze wie beispielsweise die Daten der Bundesagentur für Arbeit oder der Sozialversicherungen. Auch in diesen Bereichen hat sich inzwischen eine Kultur des data sharing durchgesetzt. Viele Ressortforschungseinrichtungen und die Statistischen Ämter verfügen heute über Forschungsdatenzentren, welche den Zugang zu den jeweiligen Daten ermöglichen. Diese Entwicklungen wurden maßgeblich durch den RatSWD angestoßen, dessen Arbeit inzwischen als Modell für weitere Wissenschaftsbereiche dient (vgl. Kommission Zukunft der Informationsinfrastruktur 2011, Wissenschaftsrat 2011).

Ein weiteres Argument für data sharing basiert auf der Erkenntnis der Datenproduzenten, dass eine Sekundärnutzung von Daten wissenschaftliche Vorteile bringt. Data sharing ermöglicht wissenschaftlich wertvolle Rückkopplungsprozesse, so dass die Datenproduzenten die Qualität ihrer Daten und die Effektivität ihrer Datenerhebungen und -analysen erhöhen können, wenn sie in intensivem Austausch mit der Forschung stehen. Aber auch die Forschungsergebnisse der Datenproduzenten werden durch eine intensive externe Auswertung bekannter wie damit auch deren Reputation.

Damit Forschungsdaten im Rahmen einer Sekundärnutzung richtig verwendet werden können, ist eine gute Dokumentation der Daten Voraussetzung. Diese Arbeit am Datensatz erfolgt bislang in der Regel ohne entsprechende Würdigung durch die Scientific Community, also die Gemeinschaft aller Forschenden. Dadurch ist es gerade für Spitzenforscher relativ unattraktiv, Zeit und Energie in die Erhebung von qualitativ hochwertigen Daten, deren Dokumentation und Nachnutzung zu investieren. Datensätze werden in der Regel nicht im Literaturverzeichnis von Veröffentlichungen zitiert und entsprechend erhält der Datenproduzent keine Zitate. Aber Zitate sind die Währung, mit der Wissenschaftlerinnen und Wissenschaftler innerhalb der Scientific Community entlohnt werden. Eine Verbesserung der „Belohnungsstrukturen" für diese Arbeiten trüge somit zu einer Verbesserung der Datenverfügbarkeit bei. Durch die Kennung eines Datensatzes mit einem persistenten Identifikator (zum Beispiel in Form eines Digital Object Identifiers (DOI)) in Verbindung mit einer Autorenidentifikation könnte die wissenschaftliche Arbeit an der Produktion eines Datensatzes kenntlich und zitierfähig gemacht werden (vgl. GESIS 2011).

Trotz aller Fortschritte im Bereich des data sharing besteht weiterhin eine deutliche Diskrepanz zwischen der Forderung nach einem freien Zugang insbesondere zu öffentlich finanzierten Daten auf der einen Seite, sowie Vorbehalten und Unsicherheiten die eigenen Daten zu teilen auf der anderen Seite. Aus Studien weiß man, dass die Gründe warum Daten – und dies trifft vor allem auf Daten aus kleineren wissenschaftlichen Erhebungen zu – nicht zur Weiternutzung bereitgestellt werden, vielfältig sind: Sie reichen von banaler Ressourcenknappheit – eine ordentliche Dokumentation der Daten erfordert zeitliche und personelle Ressourcen – bis hin zu Unsicherheiten über die Frage, wem die Daten eigentlich als Eigentümer gehören und der daraus resultierenden nicht geklärten Verantwortlichkeit (vgl. PARSE.Insight 2010, Feijen 2011).

Es sind also neben rechtlichen Fragen vor allem Bemühungen nötig, um das Weitergeben von Daten inklusive einer notwendigen Dokumentation der Daten so einfach und ressourcensparend wie möglich zu gestalten. Auf der technischen Ebene gibt es hier seit langem entsprechende Entwicklungen: Die Data Documentation Alliance bemüht sich um einen internationalen Standard bei der Beschreibung (Dokumentation) von Daten der Sozial-, Verhaltens- und Wirtschaftsfor-

schung (vgl. DDI Alliance 2009). Inzwischen sehen sich Datenarchive wie die GESIS zunehmend als Dienstleister und bieten umfangreiche Serviceleistungen und Hilfestellungen.

Neben ressourcenökonomischen Überlegungen können aber auch forschungsökonomische Überlegungen ausschlaggebend für die zu beobachtende Zurückhaltung mancher Forscher und mancher Disziplinen beim data sharing sein; beispielsweise die Befürchtung, dass sich eine Veröffentlichung des Datensatzes nachteilig auf die eigene wissenschaftliche Karriere auswirken kann. Piwowar et al. (2007) konnten jedoch unlängst in einer Studie nachweisen, dass das Teilen von Daten mit höheren Zitationsraten verbunden ist.

Ein oft vorgebrachtes Argument gegen data sharing ist das des Datenschutzes. Personenbeziehbare Daten (aber auch Daten der Wirtschaftsforschung, welche Branchen- oder Firmengeheimnisse beinhalten), die im Rahmen von wissenschaftlichen Erhebungen und Interviews oder auch klinischen Studien erhoben werden, sind in den meisten Fällen datenrechtlich sensitiv. Hier gilt es, die Daten selbst und deren Weitergabe (technisch) so zu organisieren, dass allen Daten- und Persönlichkeitsschutzaspekten in perfekter Weise Rechnung getragen wird. Datenschutz ist jedoch niemals ein grundsätzliches Argument gegen das data sharing.

Um den in Anfängen bereits begonnenen Paradigmenwandel im Bereich des *data sharing* erfolgreich weiterzubefördern, ist ein Dialog zwischen Wissenschaft, Wissenschaftsförderern, Datenschützern und wissenschaftlichen Verlagen notwendig. Die Aufgabe der Forschungsförderer wird es dabei sein, mehr als bisher auf die Erstellung und Umsetzung von Datenmanagement- und Datenverwertungsplänen als Bestandteil ihrer Förderpolitik zu achten (vgl. Winkler-Nees 2011). Ein solcher Dialog sollte in geeigneter Weise durch Gremien wie den RatSWD koordiniert werden, welche sich auch der besonderen Aufgabe der Bündelung der Interessen der Wissenschaft gegenüber Datenproduzenten und Politik widmen sollten. Weitere Herausforderungen bestehen in der Etablierung und Weiterentwicklung einer Kultur des data sharing, beispielsweise durch die Schaffung von Anreizsystemen zur Würdigung der Arbeit an Datensätzen. Neue Arten von Daten (beispielsweise Biomarker oder Geomarker) und deren Verknüpfbarkeit mit herkömmlichen Surveydaten stellen den Datenschutz vor immer neue Herausforderungen.

Data Access

Sozial-, verhaltens- und wirtschaftswissenschaftliche Daten weisen oft Charakteristika auf, welche rechtliche und forschungsethische Überlegungen notwendig machen. Weiterhin sind sie aufgrund ihrer in vielen Fällen komplexen Strukturen schwierig zu handhaben. Beide Aspekte erfordern eine besondere Organisation des Datenzugangs, d.h. der Forschungsdateninfrastruktur.

Rechtliche Aspekte

Die bereits angedeutete Komplexität der rechtlichen und forschungsethischen Fragen, welche – mit gutem Grund – den Zugang zu sensiblen Daten, insbesondere im Bereich der Wirtschafts- und Sozialwissenschaften, einschränken, macht Überlegungen darüber notwendig, wie der Zugang zu Daten in gleichzeitig effizienter, aber rechtlich und forschungsethisch einwandfreier Form organisiert werden kann. Im Prinzip gilt: Je gehaltvoller die Daten, desto interessanter sind sie für die Wissenschaft, aber desto sensibler sind sie auch. Hinlänglich anonymisierte – d.h. zusammengefasste und vergröberte Daten - bieten einen umfangreichen Datenschutz, jedoch zunehmend begrenzte Auswertbarkeit. Für viele Fragestellungen sind aggregierte Daten oder Individualdaten in anonymisierter Form völlig ausreichend. Solche Daten werden bereits heute als Public Use Files oder für die universitäre Ausbildung als sogenannte CAMPUS[2] Files durch viele öffentliche Datenproduzenten angeboten. Andere Fragestellungen verlangen jedoch nach Individualdaten, die zusätzlich mit weiteren Merkmalen, beispielsweise über das Wohnumfeld der Befragten oder Daten aus biologischen Proben der Befragten verknüpft werden. Hierdurch steigt das Deanonymisierungsrisiko und ethische Erwägungen müssen angestellt werden.

Wenngleich es hier keine generelle Lösung geben kann, bietet sich ein kontinuierlicher Austausch der Datenproduzenten über jeweilige technische Neuerungen und rechtliche Entwicklungen an. Generell gilt, dass der Daten- und Persönlichkeitsschutz durch die Anwendung entsprechender Vorkehrungen strikt und umfassend entlang der Gesetze eingehalten werden muss, dies jedoch niemals ein Argument dafür sein kann, Daten nicht zugänglich zu machen. Allerdings erschweren diese Besonderheiten die Umsetzung eines einfachen Zugangs zu den Daten, was durch angepasste technische und infrastrukturelle Lösungen, d.h. durch ein intelligentes Datenmanagement, überwunden werden kann. Viele Produzenten sensibler Daten, besonders jene der amtlichen Statistik und der Ressortforschung, können ihre Daten nicht in herkömmliche Archive geben und so einen Zugang für die Forschung ermöglichen. Die praktikable Lösung ist das

2 http://www.forschungsdatenzentrum.de/campus-file.asp [10.08.2011]

Angebot eigener Zugangswege, deren Konformität mit den jeweiligen Gesetzen kontinuierlich geprüft und gewährleistet werden kann.

Komplexitätsaspekte

Ein Charakteristikum sozial-, verhaltens- und wirtschaftswissenschaftlicher Daten ist deren Vielfältigkeit und deren oft hypothesenbezogene Entstehung. Die Spannweite reicht von einfachen Tabellen, in denen Makrodaten als Zahlenkolonnen dargestellt werden, über Interviewtranskripte und daraus gewonnenen qualitativen Daten, bis hin zu komplizierten Längsschnittdatensätzen, die aus sich fortlaufend verändernden und erweiternden Datenbanken bestehen, in denen mehrere Tausend Einzelitems für mehrere Tausend Personen über die Zeit verknüpfbar gespeichert sind. Voraussetzung für die Nutzung verschiedener Datensätze sind nicht nur Investitionen in eine adäquate Statistik- und Methodenausbildung und ein „Erlernen" des Umgangs mit den Besonderheiten (insbesondere der Messkonzepte) eines bestimmten Datensatzes auf Seiten der Nutzer, sondern vor allem auch ein geeignetes Serviceangebot von Seiten der Datenproduzenten. Dieser Service kann nur sehr begrenzt durch die „herkömmlichen" Datenarchive geleistet werden, auch hier sind alternative Lösungen gefragt, da Forschungsdaten oft nur mit Hilfe von Zusatzwissen (Metadaten) sinnvoll interpretierbar sind.

Beispielsweise werden Messverfahren und Skalen auf der Basis von Annahmen entwickelt, in der Hoffnung, sie mögen messen was beabsichtigt ist. Selbst scheinbar eindeutige Daten, wie die des Haushaltseinkommens sind komplexe Konstrukte: So macht es einen Unterschied, ob man neben den Gehältern der Haushaltsmitglieder auch Einkünfte durch Mieten oder Kapitalerträge zum Haushaltseinkommen hinzuzählt. Auch den zur Schätzung fehlender Angaben verwendeten Imputationsverfahren liegen komplexe Annahmen zu Grunde. Neben einer zu liefernden möglichst standardisierten, aber die Daten vollständig beschreibenden Dokumentation besteht oftmals ein Bedarf an intensiver fachlicher Beratung der Sekundärnutzer. Diese Beratungsleistung kann jedoch in der Regel nur durch die Datenproduzenten selbst und nicht von Archiven oder Bibliotheken geleistet werden.

Vor diesem Hintergrund einer sehr komplexen und mit unterschiedlichen Anforderungen an Datenschutz und Service zu charakterisierenden Datenlandschaft haben sich in den Sozial-, Verhaltens- und Wirtschaftswissenschaften verschiedene Akteure und Modelle etabliert, welche den Zugang zu Daten ermöglichen und ein den jeweiligen Bedürfnissen entsprechendes Niveau an Service bieten.

Modell I: Datenzugang über disziplinenspezifische oder themenspezifische zentrale Datenarchive

Archive, in denen in der Regel disziplinen- oder themenspezifische Datensätze gesammelt werden, stellen für Wissenschaftler oftmals eine erste Anlaufstelle bei der Suche nach geeigneten Daten für ihr jeweiliges Forschungsvorhaben dar. Hier können sie Unterstützung bei Recherche und Datenzugang sowie gelegentlich auch bei der Analyse der Daten (Methodenfragen) erhalten.

Auf der anderen Seite stellen Datenarchive für die Datenproduzenten eine komfortable Möglichkeit dar, ihre Daten sichtbar, auffindbar und somit für die wissenschaftliche Nachnutzung verfügbar zu machen. Hierzu gehört der Zugang zu den eigentlichen Forschungsdaten wie zu den dazugehörenden Dokumentationen, den sogenannten Metadaten (Informationen über Daten). Durch entsprechende Nutzerverträge können darüber hinaus basale datenschutzrechtliche Aspekte bei der Weitergabe Berücksichtigung finden.

Aufgabe von Archiven ist es, eine technologisch adäquate und nutzerorientierte Bereitstellung und Archivierung der Daten zu ermöglichen. Da die Archive aber nicht die Produzenten der Daten sind, ist eine diesbezügliche Zusammenarbeit mit den Datenproduzenten notwendig, welche für die Qualität der Daten verantwortlich zeichnen. Archive sollten durch fachliche Beratung und Unterstützungsleistungen bei der teilweise sehr anspruchsvollen und zeitintensiven Dokumentation und Aufbereitung der Daten, bei der oftmals auch Fragen der Anonymisierung eine zentrale Rolle spielen, aktive Partner der Datenproduzenten sein. Eine weitere Serviceleistung der Archive sollte in der Organisation und Sicherstellung der eineindeutigen Zitierfähigkeit inklusive der Verknüpfung mit den „Autoren“ der Daten bestehen.

Neben (informations)fachlichen Expertisen und Serviceangeboten verfügen Archive über die technologischen Möglichkeiten der (Langzeit-)Archivierung von Datensätzen, d.h. der Sicherstellung der physischen Existenz und Verfügbarkeit der Daten über lange Zeiträume. So komfortabel und leistungsfähig die elektronische Datenverarbeitung ist, so unhinterfragt und gefährlich ist sie auch: CDs, DVDs und Festplatten sind sehr anfällig für Fehler und Zerstörung. Während historisch genutzte Hollerithsysteme mit Lochkarten teilweise auch heute noch rekonstruierbar sind, reicht ein Kratzer, ein Computercrash oder ein Computervirus, um Datenbestände unter Umständen unwiederbringlich zu vernichten. Die Langzeitarchivierung ist eine in ihrer Wichtigkeit unterschätzte Aufgabe, die von Archiven am besten erbracht werden kann.

Zusammenfassung: Modell Datenarchiv

Systematik: Der Datenproduzent gibt seine Daten und deren Dokumentation in standardisierter Form an ein Archiv weiter.

Vorteile: Das Archiv kümmert sich um Zugang, Distribution, Vertragsangelegenheiten, Langzeitverfügbarkeit der Daten und bietet den Sekundärforschern bei der Auswertung der Daten einen basalen Service um die Daten herum. Dieses Modell ist insbesondere geeignet für im Rahmen von Forschungsprojekten entstandene Datensätze, in denen Wissenschaftler zeitlich begrenzt als Datenproduzenten fungieren und durch die Archivierung deren dauerhafte Verfügbarkeit sichergestellt wird.

Nachteile: Ein Archiv kann den Service um die Daten herum nur im begrenzten Maße leisten – in der Regel können inhaltliche Fragen nicht beantwortet werden. Es erfolgt bislang faktisch keine systematische Sammlung und Verknüpfung von bereits mit denselben Daten gefertigten Analysen und Papieren. Zu dieser Frage sollte die Zusammenarbeit mit den Forschungsbibliotheken und Verlagen angeregt und intensiviert werden. Archive können hier koordinierend fungieren. Ein weiterer Nachteil besteht darin, dass datenrechtlich hoch sensible Daten nicht ohne weiteres in allgemeinen Archiven gespeichert und verarbeitet werden dürfen.

Herausforderungen: Durch die Entwicklung und Verbesserung der Standards bei der Weitergabe von Daten und deren Beschreibung durch Metadaten verbessert sich die Zugänglichkeit und die Benutzerfreundlichkeit der Daten. Die dauerhafte Herstellung eines Links zwischen Datenproduzenten, Bibliotheken und Verlagen schafft die Voraussetzungen für eine adäquate Würdigung der Arbeit an den Daten und die umfangreiche Bereitstellung von Analysen mit den Daten.

Modell II: Zugang zu den Daten und Serviceleistungen durch Forschungsdatenzentren[3]

Eine zweite – in jüngerer Vergangenheit erfolgreich implementierte Variante des Datenzugangs – besteht im Angebot der Forschungsdatenzentren. Dieses Modell scheint sich insbesondere für potente Datenproduzenten zu bewähren und etabliert zu haben, die dauerhaft Daten zur Verfügung stellen (zum Beispiel statistische Ämter) und/oder besonders komplizierte Datensatzstrukturen anbieten (zum Beispiel prospektive Längsschnitterhebungen) und für die deshalb eine

3 Die Aufgabenfelder von Forschungsdatenzentren und Datenservicezentren lassen sich heute, auf der Basis der gemachten Erfahrungen nicht mehr eindeutig trennen. Im Folgenden beziehen wir uns auf Forschungsdatenzentren und Datenservicezentren gleichermaßen, ohne letztere immer zu nennen. Der Begriff Forschungsdatenzentrum scheint sich auch international durchzusetzen.

enge Verbindung zwischen Datenproduzent und Datennutzer wünschenswert ist. Auch die Einhaltung des Datenschutzes kann in „eigenen" Forschungsdatenzentren durch die Datenproduzenten oft einfacher gewährleistet werden.

In den Sozial-, Verhaltens- und Wirtschaftswissenschaften haben sich, ausgehend von einer Empfehlung der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) aus dem Jahr 2001, in den Folgejahren die ersten vier Forschungsdatenzentren und zwei Datenservicezentren gegründet: das Forschungsdatenzentrum des Statistischen Bundesamtes, das Forschungsdatenzentrum der Statistischen Ämter der Länder, das Forschungsdatenzentrum der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung und das Forschungsdatenzentrum der Rentenversicherung, das Servicezentrum für Mikrodaten des Leibniz-Instituts für Sozialwissenschaften (GESIS/MISSY), das Internationale Datenservicezentrum des Forschungsinstituts zur Zukunft der Arbeit (IZA). Ziel dieser Datenzentren war und ist es, die jeweiligen amtlichen Daten einer wissenschaftlichen Verwendung zur Verfügung zu stellen. Dies war bis dahin aufgrund der Vorgaben des Bundesdatenschutzgesetzes, des Statistikgesetzes und Sozialgesetzbuches bezüglich der zum großen Teil personenbeziehbaren Daten nicht ohne weiteres möglich. In der Zwischenzeit sind zu den genannten sechs Datenzentren eine ganze Reihe weiterer Forschungsdatenzentren hinzugekommen, die über den RatSWD akkreditiert und organisiert werden (http://www.ratswd.de/dat/fdz.php). Derzeit (Stand Sommer 2011) gibt es neunzehn vom RatSWD akkreditierte Datenzentren. Auch Daten, die für eine wissenschaftliche Nachnutzung anfänglich nur schwer zugänglich waren, wie es zum Beispiel im Bereich der Bildungsdaten der Fall ist, konnten auf diese Weise erschlossen werden.

Anders als bei Datenarchiven ist zentrales Merkmal der Forschungsdatenzentren der wissenschaftlich unterstützende inhaltliche Service um die Daten herum, der nur erbringbar ist, weil die das Forschungsdatenzentrum betreibenden Datenproduzenten in der Regel die besten Experten im Umgang mit den eigenen Daten sind. Ein zentraler Aspekt der Akkreditierungsrichtlinien des RatSWD für Forschungsdatenzentren und Datenservicezentren ist, dass in diesen wissenschaftlich gearbeitet wird und somit der Service für externe Wissenschaftler von Wissenschaftlern geleistet wird.

Obwohl die Forschungsdatenzentren über einen heterogenen Hintergrund verfügen, lässt sich mittlerweile berechtigt von einer gemeinsamen Forschungsdateninfrastruktur sprechen, welche unter dem Dach des RatSWD koordiniert wird. Das Akkreditierungsmodell des RatSWD bietet dabei eine Qualitätssicherung der prozeduralen Mechanismen. Die Koordination findet unter anderem ihren Ausdruck in der Festlegung gemeinsamer Kriterien und Standards als Antwort auf gemeinsame rechtliche und organisatorische Voraussetzungen, welche das Modell Datenarchiv ausschließen. Auch die Weiterentwicklung von Verfah-

ren des on-site und des gesicherten Fernrechnens, um sensible Daten unter strikter Einhaltung von datenschutzrechtlichen Vorgaben zur Verfügung zu stellen, oder auch die Erstellung von Skalenhandbüchern, um Vergleichbarkeit und Verknüpfbarkeit von Daten darzustellen und zu ermöglichen, sind aktuelle Felder der Zusammenarbeit.

Zusammenfassung: Modell Forschungsdatenzentren

Für Datenproduzenten, die aufgrund der Komplexität, der Menge und/oder der Datenschutzsensibilität ihre Daten nicht über Archive zur Verfügung stellen, findet sich im Modell des Forschungsdatenzentrums eine Möglichkeit, ihre Daten systematisch und unter Einhaltung aller rechtlichen Bestimmungen für die Forschung zu öffnen. Die Daten bleiben beim Datenproduzenten, er hat jederzeit die volle Kontrolle und kann so darüber wachen, dass alle Restriktionen jederzeit eingehalten werden. Der Nutzer der Daten hat direkten Kontakt zu Fachkollegen beim Datenproduzenten und erhält konkrete und kompetente Hilfe bei der Auswertung der Daten. Der Datenproduzent bleibt dadurch mit den Entwicklungen der Wissenschaft verbunden und kann durch eine formalisierte Rückkopplung mit außenstehenden Nutzern die Qualität der Daten; die Messmechanismen, Datenerhebungen und Aufbereitungen kontinuierlich verbessern. Auch hat der Datenproduzent in der Regel ein Interesse daran, Publikationen und Analysen zu sammeln, die auf den eigenen Daten beruhen. Somit können themen- und datenzentrierte Wissensdatenbanken entstehen.

Nachteile: Für die Datenproduzenten ist die Einrichtung von Forschungsdatenzentren vor allem in der Einführungsphase ressourcen- und kostenintensiv. Auch ist das Datenangebot in den Datenzentren in der Regel auf die „eigenen" Datensätze begrenzt, was zu einer dezentralen Verfügbarkeit von Datensätzen – unter Umständen sogar zum selben Forschungsgegenstand – führt; es gibt faktisch keinen zentralen Anlaufpunkt oder Ansprechpartner. Derzeit stellen sich deshalb die Zugangswege, Dokumentationen und Verknüpfungsmöglichkeiten der Daten etwas unübersichtlich dar.

Im Feld der Forschungsdatenzentren muss durch mehr Koordination, Transparenz und Abstimmung eine Verbesserung des Nutzerservices erreicht werden. Der RatSWD versucht dies durch die Schaffung einer Austauschplattform der Forschungsdateninfrastruktur zu befördern. Auch die Schaffung eines gemeinsamen Portals als „Tor zur gesamten Datenwelt" einer Disziplin inklusive der Verknüpfungen mit angrenzenden Disziplinen (beispielsweise Sozialdaten mit Biodaten und Geodaten) ist im Gespräch.

Zusammenfassung und zukünftige Entwicklungen

In den Sozial- und Wirtschaftswissenschaften hat sich in den vergangenen Jahren eine Kultur des Teilens von Daten (data sharing) durchgesetzt. Das heißt, es sind zunehmend interessante Daten für Forschungszwecke verfügbar; die Herausforderung besteht heute in der Organisation dieser Datenwelt. Archive und Datenzentren fungieren als etablierte Orte des Datenzugangs und werden den unterschiedlichen Anforderungen an Datenschutz und der Erbringung von Serviceleistungen um die Daten herum gerecht. Zusammen bilden sie eine funktionierende Forschungsinfrastruktur, die durchaus einen Modellcharakter aufweist.

Die Etablierung eines Portals, das Nutzern und insbesondere potentiellen Nutzern einen Überblick über, und einfache Zugangsmöglichkeiten zu sozial-, verhaltens- und wirtschaftswissenschaftlichen Forschungsdaten (einschließlich der Daten der amtlichen Statistik) anbietet, ist ein naheliegender nächster Schritt beim Ausbau der Forschungsinfrastruktur für die Sozial-, Verhaltens- und Wirtschaftswissenschaften in Deutschland. Zugleich sollte ein solches Portal die Zitation von Datenquellen und ihren Produzenten befördern.

Wie ein solches Portal gestaltet werden kann, sollte zügig von den verschiedenen Stakeholdern im Bereich der Archivierung diskutiert werden, also Fachbibliotheken, Archiven und Forschungsdatenzentren. Zu klären sind Fragen der langfristigen und sicheren Archivierung sowie des laufenden Services, der für vielgenutzte und noch im Wachsen begriffene Datensätze notwendig ist, um die Nutzung zu unterstützen. Diskutiert werden sollte auch, wie in diesem Zusammenhang die Anerkennung der „Produktion“ von Forschungsdaten als wissenschaftliche Leistung durch Referenzierbarkeit/Zitierbarkeit und persistente Identifikatoren für Daten, Datenproduzenten und Forscher verbessert werden kann. Denn nur wenn die Produktion von Forschungsdaten als wissenschaftliche Leistung voll anerkannt wird, werden ihre Qualität und Verfügbarkeit steigen.

Literatur

Bender, S./Himmelreicher, R./Zühlke, S. and Zwick, M. (2009): Improvement of Access to Data from the Official Statistics. Working Paper No. 118. RatSWD Working Paper Series August 2009. http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_118.pdf [09.08.2011]

DFG (Deutsche Forschungsgemeinschaft) (1998): Sicherung guter wissenschaftlicher Praxis. Denkschrift. Empfehlungen der Kommission „Selbstkontrolle der Wissenschaft". Weinheim: Wiley-VCH. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf [09.08.2011]

DFG (Deutsche Forschungsgemeinschaft) (2010): Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für Antragstellung und ergänzenden Leitfäden für die Antragstellung von Projekten mit Verwertungspotenzial, für die Antragstellung von Projekten im Rahmen einer Kooperation mit Entwicklungsländern. (DFG Vordruck 1.02-8/10). http://www.dfg.de/download/programme/emmy_noether_programm/antragstellung/1_02/1_02.pdf [09.08.2011]

DDA (Data Documentation Alliance) (2009): What is DDI? http://www.ddialliance.org/what [09.08.2011]

Feijen, M. (2011): What Researchers want. Utrecht: SURF Foundation (February 2011). http://www.surffoundation.nl/nl/publicaties/Documents/What_researchers_want.pdf [09.08.2011]

GESIS Leibniz-Institut für Sozialwissenschaften, da|ra Registrierungsagentur für sozialwissenschaftliche Daten (2011): Über da|ra. http://www.gesis.org/dara/home/ueber-dara/ [09.08.2011]

Habich, R./Himmelreicher, R.K. und Huschka, D. (2010): Zur Entwicklung der Dateninfrastruktur in Deutschland. Working Paper No. 157. RatSWD Working Paper Series September 2010. http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_157.pdf [09.08.2011]

Hahlen, J. (2009): Zur Rolle der amtlichen Statistik für eine evidenzbasierte Wirtschaftsforschung und -politik. Wirtschaft und Statistik (10). Wiesbaden: Statistisches Bundesamt, 1021-1030. http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Querschnittsveroeffentlichungen/WirtschaftStatistik/Gastbeitraege/Wirtschaftsforschung102009,property=file.pdf [09.08.2011]

King, G. (2011): Ensuring the Data Rich Future of the Social Sciences. Science 331, 719-721.

Kommission Zukunft der Informationsinfrastruktur (2011): Gesamtkonzept für die Informationsinfrastruktur in Deutschland. Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder. http://www.leibniz-gemeinschaft.de/?nid=infrastr [09.08.2011]

PARSE.Insight (2010): Insight into digital preservation of research output in Europe. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf [09.08.2011]

Piwowar, H.A./Day, R.S. and Fridsma, D.B. (2007): Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2 (3), e308.

Solga, H. und Wagner, G.G. (2007): Eine moderne Dateninfrastruktur für eine exzellente Forschung und Politikberatung – Bericht über die Arbeit des Rates für Sozial- und Wirtschaftsdaten in seiner ersten Berufungsperiode (2004-2006). Working Paper No. 1. RatSWD Working Paper Series 2007. http://www.ratswd.de/download/RatSWD_WP_2007/RatSWD_WP_01.pdf [09.08.2011]

Weichselgartner, E. (2011): Disziplinenspezifische Aspekte des Archivierens von Forschungsdaten am Beispiel der Psychologie. Working Paper No. 179. RatSWD Working Paper Series May 2011. http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_179.pdf [12.07.2011]

Winkler-Nees, S. (2011): Anforderungen an wissenschaftliche Informationsinfrastrukturen. Working Paper No. 180. RatSWD Working Paper Series June 2011. http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_180.pdf [09.08.2011]

Wissenschaftsrat (2011): Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften. Drs. 10465-11 vom 28.01.2011. Berlin: Wissenschaftsrat. http://www.wissenschaftsrat.de/download/archiv/10464-11.pdf [09.08.2011]

Einführung in die digitale Langzeitarchivierung

Natascha Schumann

In fast allen Bereichen des öffentlichen, kulturellen und wissenschaftlichen Lebens, einschließlich des privaten Bereichs ist eine Zunahme an der Erstellung und des Gebrauchs digitaler Objekte zu beobachten. Die Verbreitung und Vorhaltung der Daten in digitaler Form steigt nicht nur quantitativ an, sondern ersetzt in vielen Bereichen bisher genutzte analoge Formate.

Im Bereich der wissenschaftlichen Publikationen kann man diesen Wandel gut nachvollziehen. So besteht seit ungefähr 15 Jahren die Möglichkeit, Dissertationen und Habilitationen in elektronischer Form zu publizieren. Abgesehen von den finanziellen Vorteilen für die Autorinnen und Autoren, bietet diese Veröffentlichungsvariante zudem einen schnellen und weltweiten Zugriff auf aktuelle wissenschaftliche Ergebnisse. Mehr als 100.000 Online-Dissertationen aus dem deutschsprachigen Raum sind im Katalog der Deutschen Nationalbibliothek nachgewiesen. Darüber hinaus wächst die Anzahl der Netzpublikationen generell und der E-Books ebenfalls stetig an, so sind an der Deutschen Nationalbibliothek mittlerweile mehr als eine halbe Million elektronischer Publikationen (Online-Dissertation, Online-Artikel und E-Books) verzeichnet. Ebenso werden in zahlreichen Archiven Vorgänge zunehmend in elektronischen Akten verwaltet.

Im wissenschaftlichen Bereich sind elektronische Publikationen und Daten gar nicht mehr wegzudenken. Neben den genannten Publikationen entstehen enorme Datenmengen allein aus den unterschiedlichen Forschungsgebieten. Nicht nur in der Klimaforschung fallen enorme Klimadaten an, auch in weiteren naturwissenschaftlichen Fächern werden bei Experimenten große Datenmengen generiert. In den Sozial- und Wirtschaftswissenschaften werden große Mengen unterschiedlicher Forschungsdaten bei quantitativen und qualitativen Studien erzeugt.

Es entstehen also in allen Bereichen digitale Daten, die für unterschiedlich lange Zeiträume bewahrt werden müssen. Das stellt uns vor neue Herausforderungen, denn die bewährten Methoden der Bestandserhaltung für analoge Objekte können nicht einfach auf digitale Formate übertragen werden. Abgesehen von der grundsätzlich anderen Struktur digitaler Anwendungen, zum Beispiel in Bezug auf Verknüpfungsmöglichkeiten zu anderen Objekten, gibt es bei der Erhaltung digitaler Daten mehrere Dinge zu beachten: Digitale Daten sind an einen Datenträger gebunden. Diese unterliegen einem schnellen technologischen Wandel, man denke zum Beispiel an die verschiedenen Formate der letzten zwei

Jahrzehnte, von der Floppy-Disk zum USB-Stick. Werden Datenträgerformate obsolet, ändert sich entsprechend auch die Hardware, das bedeutet zum Beispiel, eine Diskette kann rein physisch nicht mehr gelesen werden, wenn der Rechner kein Diskettenlaufwerk mehr besitzt. Darüber hinaus ist es auch nötig, dass das Betriebssystem mit dem Datenträger umzugehen weiß und die entsprechende Software noch verfügbar ist. Ein weiteres Problem liegt darin, dass Datenträger über die Zeit beschädigt werden können und dann ebenfalls nicht mehr korrekt angezeigt werden können.

Bei der Erhaltung digitaler Daten spielen also verschiedene Faktoren eine Rolle, die jeweils miteinander verknüpft sind. Das bedeutet, verschiedene Komponenten entscheiden darüber, ob ein digitales Objekt in Zukunft nicht nur lesbar, sondern auch korrekt interpretierbar bleibt.

Die herkömmlichen Strategien zur Erhaltung schriftlichen Kulturguts sind über einen sehr langen Zeitraum entstanden und erprobt worden. So war in früheren Zeiten der Aufwand für die Erstellung eines gedruckten Buches größer als heute die Erstellung eines Textes in digitaler Form. Dafür war und ist die Nutzbarkeit doch relativ einfach zu gewährleisten. Werden Faktoren wie beispielsweise säurefreies Papier, die richtige Luftfeuchtigkeit und Raumtemperatur berücksichtigt, kann ein Buch über mehrere Jahrhunderte erhalten und interpretierbar bleiben, ohne dass das Objekt selbst verändert werden müsste. Bei digitalen Objekten hingegen sieht das etwas anders aus. Der Aufwand zur Erstellung eines digitalen Textes ist heutzutage relativ gering, die Sicherstellung der Verfügbarkeit und der richtigen Interpretierbarkeit ist jedoch sehr viel komplexer als bei gedruckten Medien. Es muss immer wieder überprüft werden, ob die digitalen Objekte mit gängigen technologischen Mitteln noch funktionstüchtig sind. Das setzt voraus, dass es Mechanismen geben muss, mit denen das (sich ständig weiterentwickelnde) Marktgeschehen beobachtet werden kann. Nur so ist es möglich zu entscheiden, welche Formate mittelfristig nicht weiter unterstützt werden, um dann entsprechende Maßnahmen zur Sicherung der Daten einleiten zu können.

Im Bereich der digitalen Langzeitarchivierung wird in der Regel auf das OAIS-Modell Bezug genommen. Das Open Archival Information System[1] gilt als Referenzmodell und wurde Ende der 1990er Jahre von der NASA in Zusammenarbeit mit verschiedenen Raumfahrtorganisationen entwickelt. Es handelt sich um ein sehr generisches Modell, welches verschiedene Komponenten der Langzeitarchivierung berücksichtigt. Dabei werden nicht nur technische Aspekte behandelt, sondern auch die organisatorischen Bedingungen einbezogen. Es wird nicht vorgegeben, wie die einzelnen Schritte in konkreten Anwendungen umgesetzt werden.

1 http://public.ccsds.org/publications/archive/650x0b1.pdf
Deutsche Übersetzung: http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2012051101

Die wesentlichen Aspekte des Modells bestehen darin, dass ein Produzent ein Objekt erstellt und es als SIP (=Submission Information Package) in das digitale Archiv abliefert. Der Vorgang der Übernahme in das Archiv wird Ingest genannt. Dieses Paket wird im Archiv mit weiteren Informationen angereichert, die im Hinblick auf die langfristige Verfügbarkeit des Objektes von Bedeutung sind. Dabei handelt es sich um verschiedene Metadaten, zum Beispiel solche, die Angaben zum Format samt Version oder zu technischen Aspekten enthalten. Das nun erweiterte Objekt wird AIP (Archival Information Package) genannt. Innerhalb des Archivs sind weitere Prozesse definiert, darunter die Verwaltung des Objektes, das Datenmanagement und die Archivierungs-/Speicherfunktionen. Als zusätzliche Aufgabe ist das Preservation Planning genannt. Damit ist ein Bündel von Maßnahmen gemeint, die es erlauben, die richtigen Entscheidungen im Hinblick auf die Erhaltung der digitalen Bestände zu treffen. Dazu zählen unter anderem die Beobachtung der technologischen Entwicklungen, zum Beispiel in Form von Technology Watch, das Sammeln und Pflegen von Informationen zu Formaten sowie zu Soft- und Hardware. Ebenfalls zu diesem Bereich gehört die Definition dessen, was für verschiedene Gruppen von Objekten die wichtigsten Eigenschaften (auch „Signifikante Eigenschaften“ genannt) sind, die es auf jeden Fall zu erhalten gilt. Für den Zugriff der Objekte durch Nutzer wird auf Grundlage des SIP ein gesondertes Paket generiert, das DIP (=Dissemination Information Package). Das ist im Regelfall die jeweils aktuellste Version eines Objektes, welches mit gängigen Systemen genutzt werden kann.

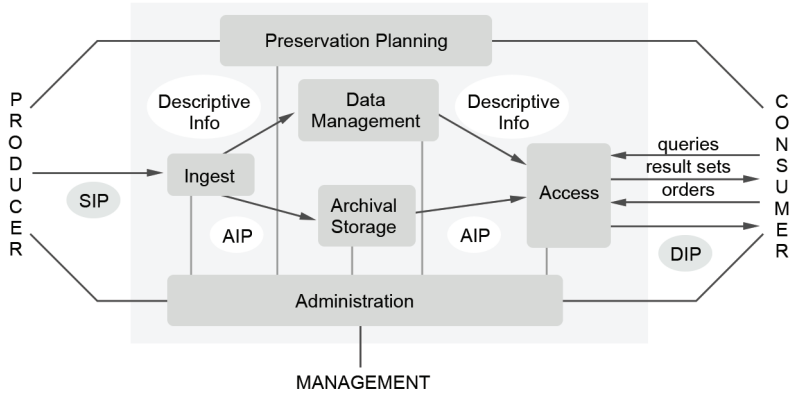


Abbildung 1: OAIS-Referenzmodell

Für ein digitales Archiv spielen die Frage der Authentizität und der Integrität von Objekten eine zentrale Rolle. Durch die Tatsache, dass zunehmend mehr Dokumente digital vorliegen, stellt sich die Frage, wie gewährleistet werden kann, dass diese Dokumente auch über einen längeren Zeitraum unverändert bleiben.

Das bedeutet, für digitale Dokumente müssen Kriterien eingehalten werden, die gewährleisten, dass Informationen und Dokumente echt sind, im Sinne, dass sie weder beschädigt noch manipuliert worden sind. Sowohl im Hinblick auf die Berechnung der Rente als auch beispielsweise in Bezug auf Grundstücksgrenzen bedarf es verlässlicher Daten, die auch zukünftig die Informationen wiedergeben, die sie bei der Erstellung enthalten haben. Darüber hinaus müssen diese Daten auch korrekt interpretierbar sein. Ein weiterer Punkt ist die Zugänglichkeit zu den Objekten, die durch geeignete Maßnahmen wie beispielsweise entsprechende Bereitstellungssysteme gewährleistet sein muss.

Um all diese Anforderungen erfüllen zu können, gibt es verschiedene Strategien und Methoden. Als Grundlage für alle weiteren Maßnahmen muss zunächst die physische Datensicherung erfolgen. Die Einsen und Nullen müssen erhalten bleiben. Diese Sicherung nennt sich Bitstream Preservation.

Digitalisierung wird teilweise als Erhaltungsstrategie bezeichnet. Dabei muss jedoch berücksichtigt werden, dass dies in der Regel im Zusammenhang mit analogen Objekten geschieht, die durch Digitalisierung erhalten werden sollen. Zum einen bietet dies den Vorteil, dass beispielsweise sehr wertvolle und von Zerfall bedrohte Werke so weiterhin Nutzern zugänglich gemacht werden können. Zum anderen können dadurch sehr viel mehr Nutzer erreicht werden. In Bezug auf digitale Langzeitarchivierung muss deutlich gemacht werden, dass Digitalisierung keine Erhaltungsstrategie darstellt. Denn Digitalisate bedürfen der gleichen Maßnahmen zur Erhaltung wie digital entstandene („digital born“) Objekte. Die derzeit am meisten genutzten Erhaltungsstrategien sind Migration und Emulation, die alternativ oder je nach digitaler Sammlung auch ergänzend eingesetzt werden.

Die vor allem bei Textdokumenten gängigste Methode, um diese über einen möglichst langen Zeitraum verfügbar zu halten, ist die Migration. Darunter versteht man die regelmäßige Konvertierung von Objekten in jeweils aktuelle Datenformate. Das kann sowohl bedeuten, dass in ein komplett anderes Format konvertiert wird, zum Beispiel von einem Worddokument in ein PDF. Aber eine Migration kann auch bedeuten, dass innerhalb eines Formats in eine neuere Version kopiert wird. Der Vorteil der Migration liegt darin, dass die digitalen Objekte immer in einem aktuell zugänglichen Format vorliegen, das mit gängiger Software lesbar ist. Der Nachteil liegt darin, dass eine Migration durchaus die Gefahr von Datenverlusten beinhalten kann. Darüber hinaus unterliegen auch die „neuen“ Formate, in die migriert wurde, einem Alterungsprozess. Das bedeutet, dass in regelmäßigen Abständen überprüft werden muss, ob das Format noch unterstützt wird und dann gegebenenfalls eine neue Migration in die Wege geleitet werden muss. Insgesamt wird es im Lebenszyklus eines digitalen Objektes mehrere Migrationen geben. Um eine Migration durchführen zu können, ist es notwendig, entsprechende Informationen über das vorliegende For-

mat zu besitzen. Daher werden bereits bei der Ablieferung in das digitale Archiv (Ingest-Prozess) Metadaten extrahiert und erfasst. Dafür stehen verschiedene Tools zur Verfügung, die diesen Prozess automatisch durchführen. Dadurch wird ermöglicht, dass bei einer anstehenden Migration alle Objekte eines bestimmten Formates mit der Version xy identifiziert und dann in das neue Format migriert werden können. Je komplexer ein Dateiformat ist, desto aufwändiger stellt sich die Migration dar. In einigen Fällen ist sie daher nicht das geeignete Mittel zur Erhaltung der Daten.

Eine weitere Strategie zur Langzeitarchivierung ist die Emulation. Hierbei wird nicht das digitale Objekt selbst verändert, wie bei der Migration, sondern es wird eine Umgebung geschaffen, die das Objekt in adäquater Weise darstellen kann. Das geschieht, indem die alte Software-Umgebung auf neuer Hardware und in neuer Systemumgebung imitiert (=emuliert) wird. Das ist ein sehr aufwändiger (Programmierungs-) Prozess. Im Gegensatz zur Migration wird dieser Aufwand allerdings nicht für jedes einzelne Objekt betrieben, sondern für Sammlungen von digitalen Objekten gleichen Typs. Die Originaldaten werden nicht verändert und es drohen daher keine Datenverluste. Vor allem für komplexe Datenarten werden Emulatoren entwickelt, zum Beispiel im Bereich der Computerspiele, wo das „Look & Feel" eine große Rolle spielt. Allerdings ist auf die lange Sicht auch bei der Emulation zu beachten, dass die Emulatoren selbst dem technologischen Wandel unterliegen und daher von Zeit zu Zeit ebenfalls migriert (also aktualisiert) werden müssen.

Welche Erhaltungsstrategie man wählt, ist abhängig von der Art der Objekte, die langfristig erhalten werden sollen, und davon, was die wichtigsten zu bewahrenden Teile eines Objektes sind. In der Fach-Community spricht man dabei von „Signifikanten Eigenschaften", also denjenigen Eigenschaften eines Objektes, die unbedingt erhalten werden müssen. Es lässt sich natürlich nicht exakt voraussagen, was zukünftige Nutzer erwarten, aber man kann sich dieser Frage – je nach Disziplin – annähern. So ist es wahrscheinlich, dass beispielsweise bei Kunstzeitschriften das Layout und die Farbgestaltung als bewahrenswert betrachtet werden, da es sich dabei um relevante Kriterien für eine zukünftige Analyse handeln mag. Hingegen ist bei wissenschaftlichen Texten das Layout im Gegensatz zum Inhalt möglicherweise nicht von so großer Bedeutung. Weitergehende Informationen zum Konzept der „Signifikanten Eigenschaften" und seiner Umsetzung finden sich im nestor-Leitfaden *"Digitale Bestandserhaltung"*.[2]

Metadaten tragen in unterschiedlicher Art und Weise zur langfristigen Erhaltung von digitalen Objekten bei. Neben inhalterschließenden Angaben sind für die Langzeitarchivierung vor allem Daten zum Format von großer Bedeu-

2 nestor-Arbeitsgruppe Digitale Bestandserhaltung (Hrsg.) (2011): Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung. Version 1.0. nestor-materialien 15. Frankfurt am Main - 77 S. URN: urn:nbn:de:0008-2011101804 http://nbn-resolving.de/urn:nbn:de:0008-2011101804

tung. In welchem „Originalformat“ wurde das Objekt in das Archiv abgeliefert? In welchem Format soll es jeweils aktuell den Nutzern präsentiert werden? In welchem Format wird das Objekt archiviert? Für diese genannten Funktionen müssen nicht zwangsläufig die gleichen Formate genutzt werden. Zusätzlich werden Metadaten zu technischen Aspekten erhoben. Dabei geht es dann unter anderem darum, welche Hardware benutzt wurde, bei Bilddokumenten zum Beispiel welche Digitalkamera (Marke, Typenbezeichnung) zur Erstellung genutzt wurde. Daraus lassen sich dann später im Zweifelsfall Rückschlüsse ziehen, die für die weitere Bearbeitung notwendig sind. Gleiches gilt auch für die verwendete Software.

Nach der Durchführung einer Migration werden die Daten der dadurch vorgenommenen Änderung ebenfalls in den Metadaten dokumentiert, damit diese Schritte langfristig nachvollziehbar bleiben.

Im digitalen Bereich spielen Nutzungs- und Verwertungsrechte eine große Rolle. Diese regeln, zu welchen Bedingungen Dokumente genutzt werden dürfen, ob sie für alle frei zugänglich im Netz zur Verfügung stehen oder nur autorisierten Nutzern im Lesesaal angezeigt werden dürfen. Diese Informationen werden ebenfalls als Metadaten mit dem Objekt gespeichert.

Eines der gängigsten Metadatenformate, das im Bereich Langzeitarchivierung genutzt wird, ist *„Preservation Metadata: Implementation Strategies“* (PREMIS)[3], das an der Library of Congress entwickelt wurde. Außerdem ist *„Metadata Encoding and Transmission Standard“* (METS)[4] ein wichtiger Standard, ein Format zur Beschreibung von digitalen Objekten.

An der Deutschen Nationalbibliothek wurden die Langzeitarchivierungsmetadaten für elektronische Ressourcen (LMER)[5] entwickelt.

Damit gewährleistet ist, dass digitale Objekte auffindbar bleiben, ist es notwendig, sie mit einer stabilen Adresse zu versehen. Denn oftmals verschwinden digitale Objekte im Netz und stattdessen erscheint „404 Seite nicht gefunden“ auf dem Bildschirm. URLs können sich sehr schnell ändern, daher ist es notwendig, eine stabile und eindeutige Adresse für digitale Dokumente zu vergeben. Diesen Zweck erfüllen Persistent Identifier (PI). Ähnlich wie eine ISBN eine Identifizierung gedruckter Bücher ermöglicht, fungieren PI im digitalen Bereich. Es gibt unterschiedliche Systeme, die im Prinzip aber ähnlich aufgebaut sind und eine ähnliche Funktionsweise haben. Im Verlagswesen sind Digital Object Identifier (DOI)[6] sehr verbreitet, während im Bibliotheksbereich Uniform Resource

3 http://www.oclc.org/research/activities/past/orprojects/pmwg/default.htm
Deutsche Übersetzung: http://www.loc.gov/standards/premis/understanding_premis_german.pdf
4 http://www.loc.gov/standards/mets/
5 http://nbn-resolving.de/urn:nbn:de:1111-2005041102
6 http://www.doi.org

Names (URN)[7] in größerem Maße genutzt werden. Das Prinzip von Persistent Identifier funktioniert folgendermaßen: Für ein digitales Dokument wird eine eindeutige Kennung generiert und diesem zugeordnet. Diese wird dann in einer Datenbank hinterlegt, zusammen mit einer oder mehreren gültigen URLs. Über einen sogenannten Resolvingprozess kann dann das Objekt angesteuert werden. Sollte eine URL nicht aufrufbar sein, wird die jeweils nächste in der Datenbank gelistete aufgerufen.

Ein weiterer wichtiger Aspekt bei der digitalen Langzeitarchivierung ist die Vertrauenswürdigkeit. Denn auch für digitale Objekte muss sichergestellt sein, dass sie authentisch, echt und unbeschädigt sind. Das bedeutet, der Betreiber eines digitalen Archivs, in dem Dokumente aufbewahrt werden, muss gewisse Kriterien erfüllen, damit Vertrauenswürdigkeit hergestellt wird. Die nestor - Arbeitsgruppe Vertrauenswürdige Archive hat – in Kooperation und Absprache mit weiteren Initiativen – eine Zusammenstellung verschiedener Kriterien erarbeitet, den Kriterienkatalog Vertrauenswürdige Archive. Dieser behandelt ganz unterschiedliche Bereiche und Prozesse, die bei der langfristigen Erhaltung digitaler Objekte eine Rolle spielen. Technische Aspekte spielen dabei eine untergeordnete Rolle, im Vordergrund stehen organisatorische Fragen und der Umgang mit den Objekten. Der nestor-Kriterienkatalog Vertrauenswürdige Archive ist Grundlage für die DIN-Norm 31644. Derzeit wird die Umsetzung des Verfahrens testweise erprobt. Diese Zertifizierungsaktivitäten entstehen jedoch nicht separat auf nationaler Ebene, sondern sind eingebettet in einen europäischen Prozess. Das European Framework for Audit and Certification sieht ein dreistufiges Verfahren zur Zertifizierung von digitalen Archiven vor: Die erste Stufe (Basic Certification) sieht einen Selbst-Audit-Prozess des Data Seal of Approval (DSA) vor. Die zweite Stufe baut darauf auf, das bedeutet, zusätzlich zum DSA wird ein Self-Assessment entweder nach der DIN-Norm 31644 oder nach der ISO-Norm 16363 durchgeführt (Extended Certification). Als dritte Stufe wird dann, wiederum zusätzlich zur Basic Certification eine Fremdzertifizierung nach DIN 31644 oder ISO 16363 durchgeführt.[8]

Neben technischen und organisatorischen Fragen spielen rechtliche Aspekte eine große Rolle bei der digitalen Langzeitarchivierung. Unterschiedliche Bereiche und Gesetze sind davon betroffen. Kulturbewahrende Einrichtungen haben unterschiedliche Mandate in Bezug auf die Archivierung, sowohl im analogen als auch im digitalen Bereich.

Im Bibliotheksbereich gibt es inzwischen einige mit einem gesetzlichen Auftrag zur Sammlung und langfristigen Erhaltung von digitalen Objekten. Auf nationaler Ebene ist seit 2006 das Gesetz über die Deutsche Nationalbibliothek[9]

7 http://www.dnb.de/DE/Netzpublikationen/URNService/urnservice_node.html

8 http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

9 http://www.gesetze-im-internet.de/dnbg/index.html

in Kraft, in welchem explizit auch die langfristige Bewahrung nicht körperlicher Medien geregelt ist. Auf Länderebene gibt es inzwischen ebenfalls einige elektronische Pflichtexemplargesetze. Damit sind die rechtlichen Grundlagen für die digitale Langzeitarchivierung geschaffen. Trotzdem ist bislang nicht abschließend geklärt, ob Veränderungen an einem Objekt, wie sie beispielsweise durch eine Migration entstehen, erlaubt sind oder nicht. Stellen diese eine Kopie dar, dann könnte es problematisch in Bezug auf das Urheberrecht werden.

Auch hinsichtlich der Bereitstellung digitaler Objekte ist die juristische Seite noch ungeklärt, beziehungsweise wird diese teilweise restriktiv gehandhabt.[10] Darüber hinaus bestehen auch ganz praktische Probleme: Viele digitale Objekte sind mit einem Kopierschutz versehen oder werden mit einem Digital-Rights-Management-System verwaltet. Solche Maßnahmen verhindern eine effektive Langzeitarchivierung, weil die Objekte nicht verändert werden können. Zunehmend geraten auch Forschungsdaten aus unterschiedlichen Gründen in den Fokus der Langzeitarchivierung. Zum einen müssen Forschungsergebnisse nachvollziehbar und vor allem nachprüfbar sein, das bedeutet, sie müssen entsprechend aufbereitet werden. Zum anderen sind viele Erhebungen/Forschungsvorhaben entweder generell nicht reproduzierbar oder dies ist mit einem enormen Kostenaufwand verbunden. Daher macht es Sinn, Datensätze nachnutzen zu können. Dazu müssen nicht nur die Daten selbst, sondern auch Dokumentationen und andere Informationen, die dafür notwendig sind, zugänglich gemacht werden. In den Sozialwissenschaften kommt noch hinzu, dass viele Studien personenbezogene Daten enthalten, und in diesen Fällen sind zusätzlich noch datenschutzrechtliche Belange zu beachten.

Digitale Langzeitarchivierung betrifft sehr viele und sehr unterschiedliche Einrichtungen. Zunächst sind da die Gedächtnisorganisationen zu nennen, deren Auftrag in der Bewahrung des kulturellen Erbes besteht. Dieser ist zwar nicht immer explizit auf digitale Werke ausgedehnt, meint diese aber in einigen Einrichtungen mit. Bibliotheken, Museen und Archive müssen sich mit den Herausforderungen der digitalen Langzeitarchivierung auseinandersetzen und entsprechende Maßnahmen für ihre Sammlungen ergreifen. Dabei stehen die einzelnen Bereiche oder Communities auch vor jeweils sehr spezifischen Fragestellungen. So erfordert die langfristige Bewahrung multimedialer Kunst andere Strategien als die Langzeitarchivierung von Forschungsdaten oder die Erhaltung von Texten. Trotzdem gibt es viele Aspekte, die für alle wenn nicht gleich so doch ähnlich sind. Außerdem ist es sinnvoll, Erfahrungen auszutauschen und voneinander zu lernen.

Auch Wissenschaftlerinnen und Wissenschaftler sind in unterschiedlichen Rollen mit der Frage der langfristigen Verfügbarkeit von Daten konfrontiert: Zum einen produzieren sie Daten für ihre Forschungsvorhaben. Zum anderen

10 Siehe nestor-Stellungnahme http://files.dnb.de/nestor/berichte/nestor-Stellungnahme_AG-Recht.pdf

sind sie auch die Nutzerinnen und Nutzer von Daten. Die Frage nach der Langzeitarchivierung von Forschungsdaten wird innerhalb der wissenschaftlichen Communities zunehmend intensiver diskutiert und wirft unter anderem auch die Frage nach den jeweiligen Rollen und Verantwortlichkeiten auf. Welche Institutionen sind für die Speicherung welcher Daten zuständig? Ist jedes einzelne Forschungsinstitut für die eigenen Daten verantwortlich oder ein Institutional Repository? Gibt es ein Datenzentrum für die jeweilige Wissenschaftsdisziplin? Die Deutsche Forschungsgemeinschaft (DFG) hat in ihren Grundsätzen zur *„Sicherung der guten wissenschaftlichen Praxis"*[11] gefordert, dass wissenschaftliche Rohdaten zehn Jahre lang aufbewahrt werden sollen. Inzwischen gibt es Überlegungen, einen Datenmanagement-Plan als obligatorischen Bestandteil eines Förderantrags zu etablieren, ohne den keine Bewilligung erfolgen soll. Ein solcher Datenmanagement-Plan legt schon vor Beginn des Projekts fest, wie und wo die entstehenden Forschungsdaten langfristig aufbewahrt werden sollen.

Für Wirtschaftsunternehmen ist die revisionssichere Bewahrung von bestimmten Daten von zentraler Bedeutung. Zwar müssen nicht alle Daten „für immer" archiviert werden, aber doch für festgelegte Zeiträume. Auch viele Bereiche der Politik und Verwaltung sind zunehmend digital organisiert und müssen entsprechende Maßnahmen zur Sicherung der entstehenden Daten ergreifen. Gleiches gilt für das Gesundheitswesen.

Letztendlich kommen fast alle Bereiche des Lebens zunehmend mit digitalen Daten in Berührung und müssen sich entsprechend damit auseinandersetzen, wie diese langfristig verfügbar bleiben. Es müssen nicht alle digital erzeugten Objekte für immer und ewig archiviert werden, aber es bedarf einer Festlegung, welche Daten wie lange und von wem bewahrt werden sollen.

Zusammenfassend lässt sich sagen, dass die Bestandserhaltung von analogen und digitalen Objekten sehr unterschiedliche Maßnahmen erfordert. Digitale Objekte sind an Datenträger gebunden, die spezielle Soft- und Hardware - Umgebungen benötigen, und allen gemein ist, dass sie schnell veralten. Grundsätzlich müssen mehrere Komponenten beachtet werden und verschiedene Maßnahmen zur Langzeitarchivierung eingeleitet werden. Vor allem aber handelt es sich bei der digitalen Langzeitarchivierung um einen fortlaufenden Prozess.

nestor – Kompetenznetzwerk digitale Langzeitarchivierung

Die Herausforderungen der digitalen Langzeitarchivierung sind vielfältig und können von Community zu Community variieren. Dennoch erscheint es sinnvoll, Erfahrungen und Informationen auszutauschen.

11 http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Das Kompetenznetzwerk nestor bietet in diesem Sinne auf der einen Seite Informationen zu allen relevanten Aspekten der digitalen Langzeitarchivierung in Form von Materialien, Berichten, Checklisten etc. über seine Webseite[12] für alle Interessierten an. Auf der anderen Seite ist nestor durch seine Struktur mit einem festen Kreis an Partnern und einer großen Anzahl an Experten, die sich in den unterschiedlichen nestor-Arbeitsgruppen engagieren und austauschen, eine zuverlässige und innovative nationale Einrichtung im Bereich der digitalen Langzeitarchivierung.

nestor bündelt das vorhandene Know-how und die Kompetenzen im Bereich der digitalen Langzeitarchivierung in Deutschland. Derzeit sind 12 Einrichtungen aus den Bereichen Archiv, Museum, Bibliothek und Universität Partner von nestor.

Als community-übergreifendes Netzwerk setzt und bearbeitet nestor aktuelle und strukturbildende Themen, die gemeinsam erarbeiteten Erkenntnisse werden untereinander und mit anderen Stakeholdern geteilt. Neben technischen Entwicklungen sind besonders organisatorische Fragen im Fokus von nestor. Alle wichtigen Akteure auf nationaler Ebene sind in oder mit nestor vernetzt.

Abbildung 2: nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

12 www.langzeitarchivierung.de

Überblick über das OAIS-Referenzmodell

Sabine Schrimpf

Einführung

Das OAIS-Referenzmodell beschreibt die Grundlagen eines digitalen Langzeitarchivs. Dabei geht es ausdrücklich über den technischen Aufbau des Archivs hinaus. : Das *„Offene Archiv-Informations-System"* (OAIS) im Sinne des OAIS-Referenzmodells ist eine Organisation bestehend *„aus Menschen und Systemen [...], das die Verantwortung übernommen hat, Information zu erhalten und sie einer vorgesehenen Zielgruppe zugänglich zu machen"*[1]. Das Referenzmodell definiert zentrale Funktionsbereiche und Verantwortlichkeiten, enthält Funktions- und Informationsmodelle und trifft wichtige terminologische Festlegungen. Es ist neutral gegenüber Datentypen und -formaten, Systemarchitekturen und Institutionstypen. Damit bietet das OAIS-Modell die wesentliche Grundlage, über Spartengrenzen hinweg über den Aufbau von und über Abläufe in Langzeitarchiven zu kommunizieren.

Hintergrund

Das OAIS-Referenzmodell wurde 2002 vom Consultative Committee of Space Data Systems (CCSDS) als Empfehlung[2] und 2003 als ISO-Standard 14721:2003[3] veröffentlicht. Damit hat es seinen Ursprung in der internationalen Weltraumforschung. Im CCSDS arbeiten führende Weltraumorganisationen zusammen, um gemeinsame Methoden und Standards für ihre Kommunikations- und Datensysteme zu erarbeiten. Die internationale Zusammenarbeit und die gemeinsame Nutzung von Forschungsinfrastrukturen machen solche Festlegungen erforderlich. An der Erarbeitung des OAIS-Modells im CCSDS seit Mitte der 1990er Jahre waren allerdings auch Archiv- und Bibliotheksvertreter wie die US National Archives and Records Administration (NARA) und die Research Libraries Group

1 Referenzmodell für ein Offenes Archiv - Informations - System – Deutsche Übersetzung (nestor - Materialien 16). – Frankfurt am Main, 2012. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2012051101

2 http://public.ccsds.org/publications/archive/650x0b1.pdf

3 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683. Während der ISO-Standard kostenpflichtig ist, ist die – inhaltidentische – Version als PDF auf der Homepage des CCSDS frei verfügbar.

(RLG) beteiligt. Insgesamt war der Erarbeitungsprozess des Standards von Offenheit und Einbeziehung geprägt: Allein 18 Workshops fanden dazu in den USA statt, außerdem wurden fünf offene Treffen veranstaltet, drei davon in Europa, um Input und Feedback der interessierten Öffentlichkeit zu sammeln.[4]

Mit der Veröffentlichung als ISO-Standard erlangte das Referenzmodell international normierende Wirkung. Es fand weltweit Anwendung in den ersten Langzeitarchivierungsinitiativen von Gedächtnisorganisationen, so zum Beispiel in den Projekten CEDARS (CURL Exemplars in Digital Archivs), NEDLIB (Networked European Deposit Library) und PANDORA (Preserving and Accession Networked Documentary Resources of Australia) und prägte die Entwicklung der ersten Generation von Langzeitarchivierungssystemen mit (zum Beispiel an der Australischen Nationalbibliothek, der Nationalbibliothek der Niederlande, in der Deutschen Nationalbibliothek und dem Bundesarchiv).

2006 wurde der von CCSDS und ISO vorgesehene Review-Prozess des Standards eingeleitet, der von einer öffentlichen Kommentierungsphase begleitet wurde. Eine überarbeitete Version des Standards wurde 2009 im CCSDS veröffentlicht.[5] Die Veröffentlichung dieser Fassung als ISO-Norm steht Anfang 2012 noch aus. Sie unterscheidet sich von der ursprünglichen Fassung lediglich in Details, so dass die folgende Einführung im Prinzip für beide Versionen gültig ist. Abbildungen und Zitate, die in diesem Artikel verwendet werden, stammen aus der mit der ISO-Norm inhaltsidentischen CCSDS-Version von 2002 .

Die ISO 14721:2003 ist bislang nicht in das deutsche Normenwerk übernommen worden. Eine Arbeitsgruppe des Kompetenznetzwerks für digitale Langzeitarchivierung, nestor, hat allerdings eine deutsche Übersetzung, die auf der CCSDS-Version von 2009 basiert, erarbeitet. Sie wurde im Sommer 2012 veröffentlicht.[6] In diesem Artikel werden die dort festgelegten deutschen Begrifflichkeiten verwendet.

Der Aufbau des Referenzmodells

Das OAIS-Referenzmodell gliedert sich in sechs Abschnitte und einige informatorische Anhänge. Abschnitt 1 besteht aus einer Einführung, die auf Zweck, Aufgabenstellung und Anwendbarkeit des Standards eingeht und Konformitätsan-

4 Lee, C.A. (2009): Open Archival Information System (OAIS) Reference Model. In: Bates, M.J. and Niles Maack, M. (Eds.): Encyclopedia of Library and Information Sciences, Third Edition. Boca Raton, FL: CRC Press, 4020-4030. http://ils.unc.edu/callee/p4020-lee.pdf

5 http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

6 nestor - Arbeitsgruppe OAIS - Übersetzung/Terminologie (Hrsg): Referenzmodell für ein Offenes Archiv - Informations - System – Deutsche Übersetzung (nestor - Materialien 16). – Frankfurt am Main, 2012. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2012051101 (abgerufen: 19.11.2012)

forderungen festlegt. Der Abschnitt enthält außerdem einen Leitfaden für die Entwicklung weiterer, auf dem OAIS-Referenzmodell aufbauender Standards und eine umfangreiche Liste mit Begriffsdefinitionen.

In Abschnitt 2 werden die Hauptkonzepte vorgestellt, die für ein OAIS (also ein Archiv-Informations-System) unmittelbar relevant sind, darunter die Rollen derjenigen, mit denen das OAIS in Beziehung steht, die wichtigsten Interaktionen zwischen dem OAIS und diesen Gruppen sowie der Informationsbegriff, der dem OAIS-Modell zugrunde liegt.

Abschnitt 3 definiert die Verantwortlichkeiten eines OAIS sowie Mechanismen, mit denen das OAIS diese Verantwortlichkeiten erfüllen kann. Abschnitt 4 enthält detaillierte Modelle für alle Funktionsbereiche des OAIS (Funktionsmodell) und Modelle für die unterschiedlichen Informationstypen, die im OAIS verwaltet werden (Informationsmodell).

In Abschnitt 5 werden einige Erhaltungsstrategien vorgestellt. Dabei geht es einerseits um die Erhaltung der archivierten Information selbst und andererseits um die Erhaltung von Zugriffsdiensten, mit denen archivierte Informationen benutzt werden.

Abschnitt 6 behandelt schließlich das Thema der Archivinteroperabilität und stellt verschiedene Möglichkeiten für die Zusammenarbeit von Archiven vor, um gemeinsam bessere oder kosteneffektivere Dienste realisieren zu können.

In den – nicht normativen, sondern ausschließlich informatorischen Anhängen – werden einige Beispielimplementierungen von Langzeitarchiven vorgestellt (Annex A) und Beziehungen zu anderen, inhaltlich nahestehenden Standards aufgezeigt (Annex B). Annex C enthält eine kurze Einführung in die *„Unified Modeling Language“* (UML), die zur Beschreibung des Informationsmodells in Abschnitt 4 verwendet wird. Annex D verweist auf weiterführende Literatur, Annex E macht einen Umsetzungsvorschlag für eine spezielle Softwareanwendung und Annex F (in der CCSDS-Version von 2009 nicht mehr enthalten) zeigt die Gesamtsicht aller in Abschnitt 4 vorgestellten Funktionsmodelle.

OAIS-Hauptkonzepte

Die Umgebung des OAIS

Es wird davon ausgegangen, dass der Archivbegriff sich im digitalen Zeitalter gegenüber dem klassischen Archivbegriff des analogen Zeitalters erweitert. Neben traditionellen Archiven stehen staatliche Institutionen, Privatunternehmen und Non-Profit-Organisationen zunehmend in der Pflicht, sich aktiv an den Bemühungen zur Langzeiterhaltung ihrer Informationen beteiligen zu müssen. Die OAIS-Begrifflichkeiten und Konzepte sollen auf alle denkbaren organisato-

rischen Modelle beim Aufbau einer Langzeiterhaltungslösung anwendbar sein. Auch wenn Produzenten- und Archivrolle in die Zuständigkeit ein und derselben Institution fallen können, werden die Rollen konzeptionell klar getrennt (Abbildung 1).

Die Rolle des Produzenten (Producer) wird von denjenigen Personengruppen oder Systemen eingenommen, die die zu archivierende Information herstellen und zur Archivierung an das OAIS übergeben. Die Rolle des Endnutzers (Consumer) spielen diejenigen Personengruppen oder Systeme, die auf archivierte Information zugreifen und sie benutzen wollen. Die Rolle des Managements wird von dem Personenkreis ausgefüllt, der das OAIS, sei es als eigenständiges Archiv oder als Teil einer größeren Organisation, in seiner Zielsetzung verantwortet.

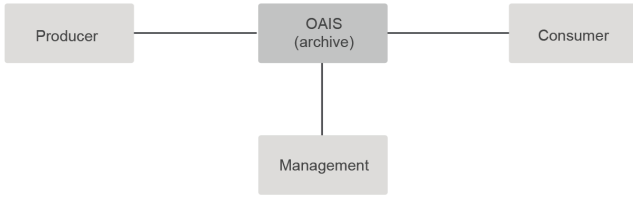


Abbildung 1: Umgebungsmodell eines OAIS (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

OAIS Information

Information wird als Wissen verstanden, das in Form von Daten oder Datenobjekten (Data Objects) ausgetauscht werden kann. Damit aus Daten bedeutungsvolle Information werden kann, wird Repräsentationsinformation (Representation Information) benötigt, die den Datenstrom zu Informationsobjekten (Information Objects) zusammensetzt.

Im Fokus aller Aktivitäten des OAIS steht die Erhaltung der Inhaltsinformation (Content Information). Konzeptionell wird sie zusammen mit Erhaltungsmetadaten (Preservation Description Information) zu einem Informationspaket „gepackt" und mit deskriptiven Informationen verknüpft, die die Wiederauffindbarkeit des Informationspakets gewährleisten. Außerdem wird konzeptionell

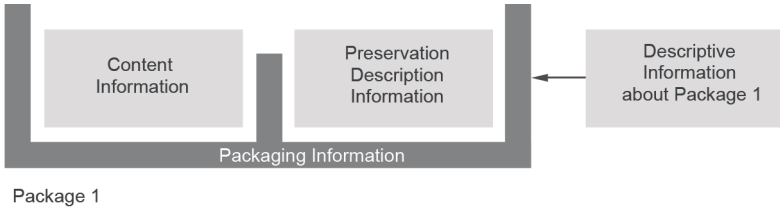


Abbildung 2: Konzepte und Beziehungen innerhalb eines Informationspakets (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

zwischen drei unterschiedlichen Arten von Informationspaketen unterschieden, und zwar dem Übergabeinformationspaket (Submission Information Package, SIP), das vom Produzenten an das OAIS übergeben wird, dem Archivinformationspaket (Archival Information Package, AIP), das vom OAIS aufbewahrt wird, und dem Auslieferungsinformationspaket (Dissemination Information Package, DIP), das vom OAIS an den Endnutzer ausgegeben wird. Diese Unterscheidung wird der Tatsache gerecht, dass das SIP noch nicht alle Informationen enthalten muss, die im AIP archiviert werden. Typischerweise wird das SIP im OAIS noch mit Metadaten angereichert, bevor das AIP „gepackt" wird. Dem Endnutzer wiederum müssen nicht notwendigerweise alle Informationen, die im AIP archiviert werden, zur Verfügung gestellt werden, damit er die Inhaltsinformation nutzen kann.

Außenbeziehungen des OAIS

Datenflüsse finden vor allem zwischen Produzent und OAIS (Übergabe bzw. Übernahme des SIP) und zwischen OAIS und Endnutzer (Auslieferung des AIP) statt. Die Beziehungen des OAIS zu beiden Gruppen können mehr oder weniger formalisiert sein. Das OAIS-Referenzmodell legt für die Ausgestaltung der Interaktionen zwischen Produzent und OAIS zwei Konzepte fest: die Übergabevereinbarung (Submission Agreement), in der festgelegt ist, welche Informationen in welchen Formaten und welchen Zeitabständen übergeben werden, und die Datenübergabesitzung (Data Submission Session), in der die Informationen vom Produzenten an das OAIS übermittelt werden. Zur Gestaltung der Interaktion mit dem Endnutzer ist das Konzept der Recherchesitzung (Search Session) vorgesehen, in der der Nutzer durch Suche in der Erschließungsinformation (Descriptive Information) Bestände seines Interesses ermittelt. Diese fordert er mittels einer ereignisbasierten Bestellung (*Event Based Order*) oder einer „Spontan-Bestellung" (*Adhoc Order*) vom OAIS an. Eine Bestellvereinbarung (Order Agreement) wird abgeschlossen und die Informationen während einer Datenauslieferungssitzung (Data Dissemination Session) an den Endnutzer übergeben.

Verantwortlichkeiten des OAIS

Dem OAIS wird eine Reihe von Aufgaben und Verantwortlichkeiten zugeschrieben:

- Informationsübernahme vom Produzenten
- Sich genügend Kontrolle über die übernommenen Daten verschaffen, um deren Langzeiterhaltung sicherzustellen
- Die „vorgesehene Zielgruppe" (*Designated Community*) bestimmen, die die archivierten Informationen in der Zukunft nutzen wird

- Sicherstellen, dass die zu archivierende Information aus sich heraus (d.h. ohne spezielle Hilfsmittel) für diese Zielgruppe verständlich ist
- Etablierten und dokumentierten Policies und Verfahren folgen, um die archivierten Informationen gegen alle vorstellbaren Gefahren zu schützen
- Der vorgesehenen Zielgruppe die archivierten Informationen zur Verfügung stellen

Außerdem werden eine Reihe von Mechanismen beschrieben, mit denen das OAIS seine Aufgaben und Verantwortlichkeiten erfüllen kann:

- Das OAIS sollte Kriterien festgelegt haben, nach denen es Informationen zur Übernahme auswählt bzw. annimmt. Mit den Produzenten sollte darüber verhandelt werden, welche Metadaten (Erschließungsinformation sowie Erhaltungsmetadaten) übernommen werden können.
- Das OAIS kann durch seinen institutionellen Auftrag oder durch Gesetze (zum Beispiel Archivgesetz, nationales Pflichtexemplar) berechtigt sein, Langzeiterhaltungsmaßnahmen an den zur Erhaltung übernommenen Informationen durchzuführen. Verfügt das Archiv nicht über diese Rechte, sollte es sie sich bei der Informationsübernahme vom Produzenten einräumen lassen.
- Zur Bestimmung der vorgesehenen Zielgruppe muss das OAIS das Grundwissen seiner Endnutzer berücksichtigen, zum Beispiel ob und in welchem Maß sie mit bestimmter Wissenschaftsterminologie vertraut sind.
- Das Grundwissen und die gängigen Arbeitsmittel der vorgesehenen Zielgruppe bestimmen die Anforderungen an die Verstehbarkeit der archivierten Information, die das OAIS erfüllen muss. Ändern sich im Lauf der Zeit das Grundwissen und die Arbeitsmittel der vorgesehenen Zielgruppe, oder gar die vorgesehene Zielgruppe selbst, dann ändern sich ggf. auch die Anforderungen daran, was nötig ist, um die Information verstehbar zu machen.
- Das OAIS sollte in all seinen Aktionen festgelegten Verfahren folgen. Zum Beispiel müssen Erhaltungsmaßnahmen genau überwacht und dokumentiert werden. Auch die vorgesehene Zielgruppe bzw. Änderungen in ihr sollten beobachtet werden, genauso wie die allgemeine Technologieentwicklung. Das Archiv sollte auch für unvorhergesehene Ereignisse vorsorgen, zum Beispiel indem es über Notfall-Pläne verfügt.
- Wie genau und mit welchen Suchhilfen das OAIS seine Informationen zur Verfügung stellt, wird sich von OAIS zu OAIS, je nach archivierten Beständen und vorgesehener Zielgruppe, unterscheiden. In allen Fällen aber muss das OAIS beim Bereitstellen der Information etwaige Zugriffsbeschränkungen und Datenschutzbestimmungen berücksichtigen. Das OAIS sollte seine Zugriffspolicies veröffentlichen, damit sie für seine Nutzer nachvollziehbar sind.

Funktionsmodell

Die abstrahierte Sicht auf die Funktionseinheiten eines OAIS gehört zu den am meisten zitierten Abbildungen im Bereich der digitalen Langzeitarchivierung: Das OAIS-Funktionsmodell unterscheidet sechs Funktionseinheiten, die im Standard jeweils detailliert und hier im Überblick vorgestellt werden:

In der *Funktionseinheit „Übernahme"* (Ingest) wird die Abgabe des Übergabeinformationspakets (SIP) vom Produzenten an das OAIS organisiert. Dort erfolgt außerdem die Vorverarbeitung des SIPs für die anschließende Übergabe an den Archivspeicher.

Die *Funktionseinheit „Archivspeicher"* (Archival Storage) beinhaltet die Speicherung der (bei der Übernahme aus SIPs erzeugten) AIPs, inklusive Monitoring, Umkopieren und Fehlerkontrolle und Ressourcen zur Notfallwiederherstellung der Daten.

Die *Funktionseinheit „Datenverwaltung"* (Data Management) enthält Dienste zur Verwaltung der Metadaten der Archivbestände, also der Daten, die die Archivbestände beschreiben, und über die sie identifiziert werden können. Die Datenverwaltung beinhaltet auch administrative Informationen zur Verwaltung des Archivs.

In der *Funktionseinheit „Erhaltungsplanung"* (Preservation Planning) werden Erhaltungsstrategien wie Migration und Emulation für die Archivbestände erarbeitet, der Technologiefortschritt und die vorgesehene Zielgruppe beobachtet und konkrete Erhaltungsmaßnahmen geplant.

Die *Funktionseinheit „Administration"* stellt Dienste zur Verwaltung des gesamten Betriebs des Archivsystems zur Verfügung. Hier werden zum Beispiel Übergabevereinbarungen mit den Produzenten ausgehandelt, der Archivbetrieb überwacht, und Standards und Policies des Archivsystems festgelegt und gepflegt.

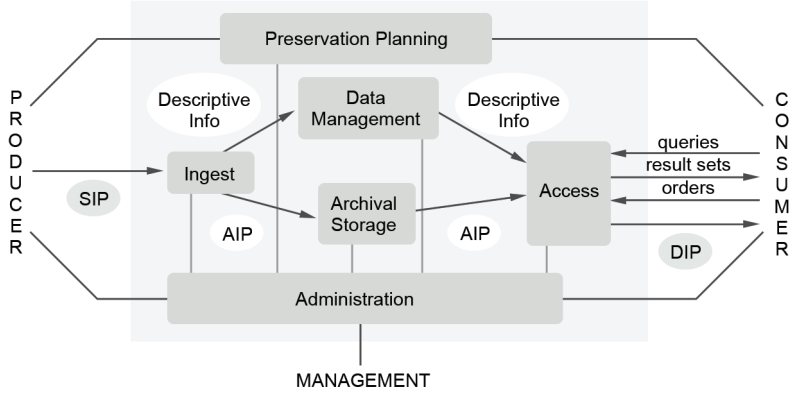


Abbildung 3: OAIS Funktionseinheiten (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

In der *Funktionseinheit „Zugriff"* (Access) werden die Endnutzer bei der Recherche von Beständen des OAIS unterstützt. Bei einem Zugriff fordert diese Funktionseinheit – unter Wahrung etwaiger Zugriffsbeschränkungen – die Informationen aus dem „Archivspeicher" an und liefert sie dem Endnutzer aus.

Als Einzeldienste, losgelöst von den sechs Funktionseinheiten, werden im OAIS-Referenzmodell außerdem noch Betriebssystem-Dienste benannt, Netzwerkdienste und Sicherheitsdienste wie Authentifizierung und Zugriffskontrolle.

Informationsmodell

Um Information im Sinne des OAIS-Modells „langfristig" erhalten zu können, also über technologische Veränderungen hinweg bis in die unbestimmte Zukunft, muss das OAIS über den reinen Informationsinhalt hinaus weitere Informationen speichern. Das Informationsmodell definiert die unterschiedlichen Informationsarten und stellt Konzepte und Modelle vor, wie diese Informationen innerhalb des OAIS organisiert werden können.

Informationsobjekt

Grundlegend für das OAIS-Informationsmodell ist das Konzept eines zusammengesetzten Informationsobjekts. Damit wird reflektiert, dass zur Anzeige digitaler Daten immer weitere Information benötigt wird. Das OAIS-Informationsobjekt setzt sich aus dem Datenobjekt selbst und aus Repräsentationsinformation zusammen, die benötigt wird, um das Datenobjekt interpretieren und anzeigen zu können. Das Datenobjekt könnte zum Beispiel der Datenstrom sein, der die Inhalte einer PDF-Datei codiert. Die Repräsentationsinformation beinhaltet dann das Wissen darüber, dass die Daten im PDF-Format codiert sind und zur Anzeige ein PDF-Viewer gebraucht wird.

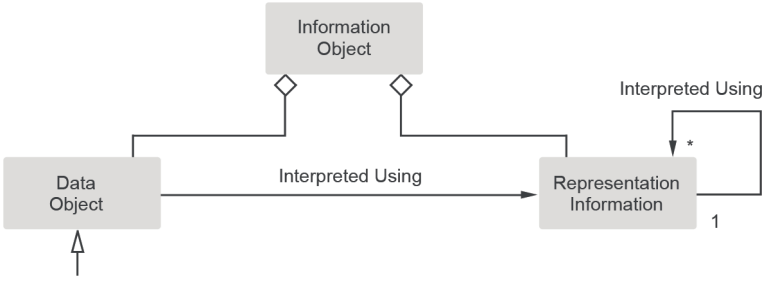


Abbildung 4: Ausschnitt aus der Abbildung „Informationsobjekt" (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

Über die Repräsentationsinformation hinaus sieht das OAIS noch weitere Informationsarten vor, darunter die Inhaltsinformation selbst (Content Information), die das eigentliche Ziel aller Erhaltungsbemühungen ist. Sie wird begleitet von Erhaltungsmetadaten (Preservation Description Information), die Informationen über Herkunft und Kontext der Inhaltsinformation enthalten und alle an der Inhaltsinformation vorgenommenen Erhaltungsmaßnahmen dokumentieren. Die Verpackungsinformation (Packaging Information) bestimmt die einzelnen Komponenten des Informationspakets und ihr Verhältnis zueinander. Erschließungsinformation (Descriptive Information) dient der Wiederauffindbarkeit des Informationsobjekts. Abbildung 5 ist in dieser Hinsicht möglicherweise missverständlich. Obwohl es so aussieht, als ob all diese Arten von Information zu einem Informationsobjekt gehören, ist tatsächlich aber gemeint, dass jede Informationsart als eigenes Informationsobjekt (nach dem in Abbildung 4 verdeutlichten Prinzip) betrachtet werden kann.

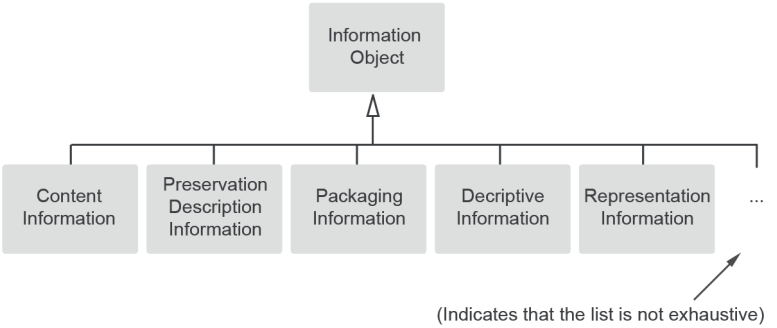


Abbildung 5: Taxonomie der Informationsobjekte (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

Informationspaket

Das OAIS-Referenzmodell unterscheidet zwischen drei Arten von Informationspaketen, dem SIP, AIP und DIP. Ihnen allen ist gemeinsam, dass sie sich aus Inhaltsinformation und Erhaltungsmetadaten zusammensetzen, wie bereits oben in Abbildung 2 gezeigt . Das komplexeste Informationspaket ist das AIP, in dessen Gesamtsicht einige Konzepte zusammengeführt werden:

Das Informationspaket (in diesem Fall das AIP) wird von der Paketbeschreibung (Package Description) beschrieben und von der Verpackungsinformation (Packaging Information) begrenzt. Es enthält die Informationsobjekte selbst, Inhaltsinformation (Content Information) und Erhaltungsmetadaten (Preservation Description Information). Die Erhaltungsmetadaten setzen sich wiederum aus fünf Klassen von Metadaten zusammen:

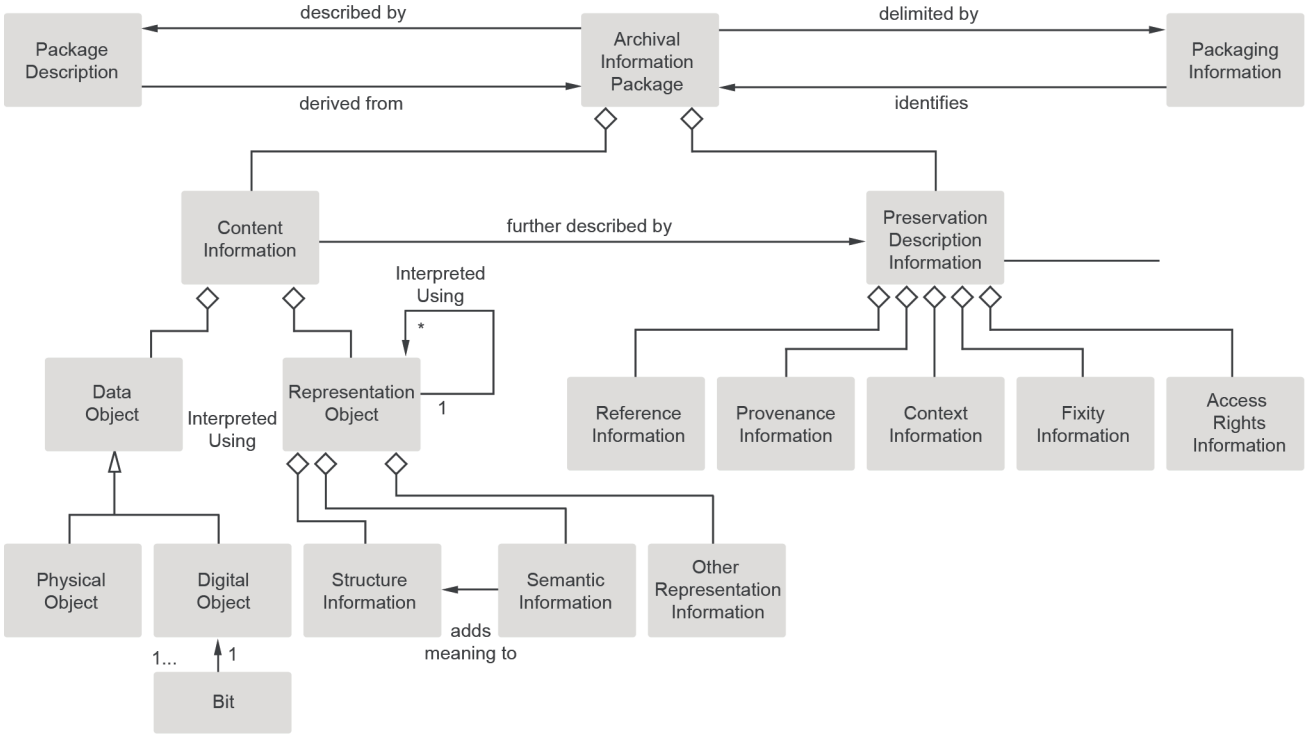


Abbildung 6: Detaillierte Ansicht eines Archivinformationspakets (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

- Referenzinformation (Reference Information) identifiziert die Inhaltsinformation des Informationspakets, zum Beispiel mittels eines Persistenten Identifikators.
- Provenienzinformation (Provenance Information) dokumentiert alle Maßnahmen, die an der Inhaltsinformation seit ihrem Entstehen durch wen vorgenommen wurden.
- Kontextinformation (Context Information) enthält erläuternde Information über die Inhaltsinformation, zum Beispiel Hilfedateien oder einen Sprachcode.
- Persistenzinformation (Fixity Information) weist zum Beispiel mit Hilfe digitaler Signaturen, Prüfsummen, oder anderen Authentizitätsindikatoren nach, dass an der Inhaltsinformation keine undokumentierten Änderungen durchgeführt wurden.
- Information über Zugriffsrechte (Access Rights Information) hält Zugriffs-, Verbreitungs- und Lizenzbedingungen vor, die für die Inhaltsinformation gelten.

Transformation von Informationspaketen

In diesem Abschnitt geht es darum, wie sich SIPs zu AIPs und AIPs zu DIPs verhalten bzw. welche Transformationen die Informationspakete bei der Übergabe vom Produzenten an das OAIS und bei der Auslieferung vom OAIS an den Endnutzer erfahren.

In der Übergabevereinbarung, die zwischen Produzent und OAIS geschlossen wird (siehe auch oben unter OAIS-Hauptkonzepte), ist idealerweise festgehalten, wie die SIPs, die der Produzent an das OAIS abliefert, gestaltet sind, d.h. welche Information enthalten ist und in welchen Formaten sie der Produzent an das OAIS übergibt. In der Funktionseinheit „Übernahme" im OAIS kann das OAIS erhaltene SIPs so umgestalten, dass sie seinen Archivstrukturen am besten entsprechen. Dabei kann es die Zuordnung 1 SIP = 1 AIP beibehalten. Es kann aber auch mehrere SIPs zu einem AIP zusammenpacken, indem es zum Beispiel einmal im Monat wöchentliche Datenlieferungen eines Produzenten zusammenfasst und in einem AIP archiviert. Genauso ist denkbar, dass das OAIS ein SIP eines Produzenten in mehrere AIPs unterteilt, zum Beispiel um sie sukzessive zu archivieren.

Bei der Übergabe der AIPs an die Funktionseinheit *„Archivspeicher"* werden die Paketbeschreibungen vom AIP getrennt und in die Funktionseinheit *„Datenverwaltung"* überführt, wo sie für Recherchen zur Verfügung stehen und die Wiederauffindbarkeit der AIPs gewährleisten.

Der Endnutzer startet Recherchen nach Beständen des OAIS in Findmitteln in der Funktionseinheit *„Zugriff"*. Über diese Funktionseinheit greift er mittelbar auf die Datenverwaltung zu, in der seine Suchanfragen in den dort gespeicherten Paketbeschreibungen durchgeführt werden. Fordert der Endnutzer Bestände aus dem OAIS an, werden ihm diese aus dem Archivspeicher als DIPs bereitgestellt. Im einfachsten Fall beinhaltet das DIP eine genaue Kopie des AIPs und die dazugehörige, aus der Datenverwaltung angeforderte, Paketbeschreibung.

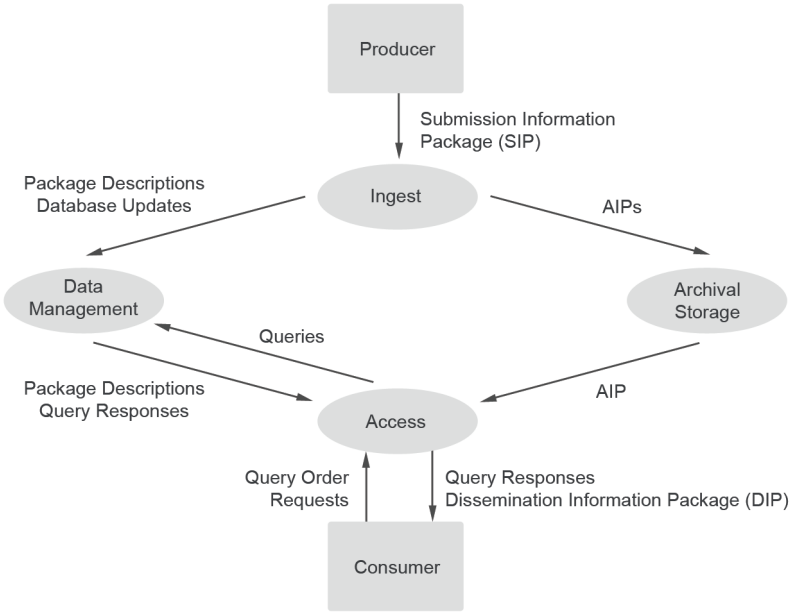


Abbildung 7: Datenflüsse in einem OAIS im Überblick (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

Erhaltungsmöglichkeiten

Das OAIS-Referenzmodell behandelt zwei Erhaltungsstrategien, die Migration und die Emulation. Den Beschreibungen der Erhaltungsstrategien merkt man am deutlichsten an, dass das OAIS-Modell im datenintensiven Raum- und Luftfahrtkontext entstanden ist, sie sind eher auf Daten- denn auf Dokument-Bestände zugeschnitten.

Bei der Migration steht zunächst der Aspekt der Datenträgermigration im Vordergrund. Als erste Migrationsvariante wird die *„Auffrischung"* (Refreshment) dargestellt, bei der AIPs unverändert von einem Datenträger auf einen anderen kopiert werden. Es folgt die *„Replikation"* (Replication), bei der es sich um eine Auffrischung handelt, bei der sich aber auch die Ablagestrukturen der AIPs verändern könnten. Deswegen müssen nach einer Replikation die Beziehungen zwischen Archivspeicher und Datenverwaltung aktualisiert werden. Bei der *Umverpackung* („Repackaging") können noch Änderungen an der Verpackungsinformation hinzukommen.
Die vierte Migrationsvariante, *„Transformation"* (Transformation) entspricht dem, was allgemein als *„Formatmigration"* bekannt ist. Hier wird im Referenzmodell wiederum zwischen *„reversibler"* und *„irreversibler"* Transformation

unterschieden. Bei der reversiblen Transformation wird zum Beispiel ein verwendeter Zeichencode gegen einen anderen ausgetauscht, zum Beispiel ASCII gegen UNICODE. Der Austausch könnte rückgängig gemacht werden, weil die Beziehung zwischen den Zeichencodes eindeutig ist. Eine irreversible Transformation erfolgt zum Beispiel bei der Kompression in ein speicherärmeres Format, die nicht rückgängig gemacht werden kann.

Für Emulationen beschreibt das OAIS-Referenzmodell drei Varianten: Die Einrichtung und Freigabe einer Programmierschnittstelle (API), mit der die vorgesehene Zielgruppe direkt auf die Archivbestände zugreifen kann, die Erhaltung des Look and Feel von Zugriffsprogrammen und zuletzt die Freigabe des Quellcodes der Zugriffsprogramme für die vorgesehene Zielgruppe beziehungsweise, wenn es sich um ein proprietäres Zugriffsprogramm handelt, um dessen Hinterlegung im OAIS.

Archivinteroperabilität

Abschließend werden einige vorstellbare Interaktions- und Kooperationsformen zwischen OAIS-Archiven beschrieben, bei denen sich die beteiligten Archive jeweils auf gemeinsame Standards einigen müssen. Gründe zur Kooperation können zum Beispiel darin bestehen, Endnutzern gemeinsame Findmittel über mehrere Archive hinweg anbieten zu wollen, Produzenten die Ablieferung zu erleichtern, indem ein einheitliches SIP-Schema verabredet wird, oder Anschaffungskosten für Hard- und Software durch geteilte Nutzung zu reduzieren. Es werden Kooperationsmodelle für mehrere voneinander völlig unabhängige Archive, für durch Vereinbarungen lose miteinander verbundene Archive, für formalisierte Archivverbünde (Föderationen) und für Archive mit geteilten Funktionsbereichen skizziert. Voneinander völlig unabhängige Archive können sich zum Beispiel auf einheitliche Standards für ihre SIPs, DIPs oder die angebotenen Findmittel verständigen wollen. Engere Absprachen, die in Vereinbarungen festgehalten werden sollten, sind nötig, wenn Archive Daten austauschen wollen oder müssen. Zum Beispiel kann es nötig sein, die Ausgestaltung von SIPs und DIPS aufeinander abzustimmen, wenn ein Archiv von einem anderen Daten bezieht. Wenn Archive überlassende Produzenten- oder Endnutzergruppen haben, kann es für sie in Frage kommen, ihre Übergabe- und Auslieferungsschnittstellen im Interesse der Nutzer zu vereinheitlichen.

Archivverbünde werden als endnutzerorientiert beschrieben. Im Mittelpunkt ihrer Darstellung steht daher die Bereitstellung ihrer Bestände mittels eines oder mehrerer gemeinsamer Findmittel. Archive mit geteilten Funktionsbereichen nutzen ganze Archivkomplexe wie etwa den Archivspeicher oder Infrastruktur zur Übernahme gemeinsam. In solchen Fällen müssen die internen Strukturen der Archive gründlich aufeinander abgestimmt sein.

Schlussbemerkung

Kaum ein Aufsatz zur digitalen Langzeitarchivierung kommt ohne Bezug auf das OAIS-Referenzmodell aus. Es stellt eine Begriffs- und Konzeptwelt zur Verfügung, mit der die Verständigung über digitale Langzeitarchivierung auch über Sparten- und Disziplingrenzen hinweg möglich wird. Das hohe Abstraktionsniveau des Referenzmodells ist dabei seine Stärke und Schwäche zugleich. Häufig wird die Kritik geäußert, dass das Modell zu abstrakt ist und unterschiedliche Anwender zu so verschiedenen Auslegungen kommen können, dass sie untereinander kaum mehr vergleichbar sind. Unter dem „Deckmantel" der einheitlichen Terminologie können so fundamental verschiedene Umsetzungen und Missverständnisse entstehen.

Dennoch existiert mit dem OAIS-Referenzmodell eine wichtige Basis, um über digitale Langzeitarchivierung und digitale Langzeitarchive sprechen zu können. Die Rollenbeschreibungen können den an der Archivierung beteiligten Einrichtungen dazu dienen, ihre konkreten Aufgaben und Verantwortlichkeiten zu definieren und voneinander abzugrenzen. Die Funktions- und Informationsmodelle geben wichtige Hinweise zu Implementierungsansätzen.

Weitere Standards schließen an die Begriffswelt und Konzepte des OAIS-Modells an, darunter der *„Producer-Archive Interface Methodology Abstract Standard"*[7] (PAIMAS), der die Interaktion zwischen Produzent und Archiv während des Ingest-Prozesses in zahlreiche Einzelprozesse zerlegt, oder der Kriterienkatalog *„Audit and Certification of Trustworthy Repositories"*[8]. Beide Standards wurden frei in den deutschsprachigen Raum übertragen, vergleiche zum ersten den nestor-Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv[9] und zum zweiten den nestor-Kriterienkatalog für vertrauenswürdige digitale Langzeitarchive[10] bzw. die darauf basierende DIN-Norm 31644.

Die Bedeutung des OAIS-Referenzmodells als zentraler Standard in der digitalen Langzeitarchivierung ist daher gerechtfertigt. Seine Rezeption sollte sich nicht auf ein oberflächliches Verständnis der sechs Funktionsbereiche beschränken. Für Archivbetreiber und Institutionen, die den Aufbau von Archivsystemen planen, ist die nähere Beschäftigung mit den Konzepten und Modellen des OAIS-Referenzmodells nach wie vor relevant und lohnend.

7 http://public.ccsds.org/publications/archive/651x0m1.pdf, auch veröffentlicht als ISO 20652:2006.
8 http://public.ccsds.org/publications/archive/652x0m1.pdf
9 http://files.d-nb.de/nestor/materialien/nestor_mat_10.pdf
10 http://files.d-nb.de/nestor/materialien/nestor_mat_08.pdf

B Standards

DIN Norm 31644 „Kriterien für vertrauenswürdige digitale Langzeitarchive“: Zielsetzung, Genese und Perspektiven

Christian Keitel

Im April 2012 erschien die DIN Norm 31644: Vertrauenswürdige digitale Langzeitarchive. Was beabsichtigten die Autorinnen und Autoren der Norm, wie verlief ihre Entstehungsgeschichte und welche Inhalte bietet sie? Was sind die weiteren Perspektiven zum Einsatz der Norm und der verwandten Kriterienkataloge? Diesen Fragen soll hier nachgegangen werden.

Zielsetzung der Norm

Weshalb überhaupt diese Norm? Zentraler Bezugspunkt sind die digitalen Langzeitarchive. Es ist eine Binsenweisheit, dass sich digitale Objekte sehr leicht ändern lassen. Diese Änderungen lassen sich oft nur sehr schwer nachweisen. Die leichte Veränderbarkeit digitaler Objekte stellt nun die Archive, die diese Objekte sehr lange Zeit aufbewahren wollen, vor ein grundsätzliches Problem. Wie können sie ihre Nutzer davon überzeugen, dass sie die Objekte inhaltlich nicht geändert haben? DIN 31644 unterstützt digitale Archive bei der Beantwortung dieser Frage und adressiert dadurch das Grundproblem der digitalen Langzeitarchivierung: Digitale Langzeitarchive verlieren ihre Kernaufgabe, wenn sie in den Augen ihrer Nutzer keine glaubwürdigen Objekte vorlegen können[1].

Nun scheint sich das Problem der Glaubwürdigkeit mit einer unterschiedlichen Dringlichkeit zu stellen. Klassische Archive können darauf vertrauen, dass ihnen bereits seit Jahrhunderten eine besondere Glaubwürdigkeit zugesprochen wird. Wenn sie etwas übernommen haben, ist es dem unmittelbaren inhaltlichen Interesse derjenigen Personen und Einrichtungen entzogen, die das Dokument erstellt und damit gearbeitet haben. Archivaren ist es inhaltlich nicht wichtig, ob in einem Dokument bestimmte Inhalte stehen oder eben deren Gegenteil, ob also beispielsweise die Person X den Auftrag für ein millionenschweres Bauvorhaben vergeben hat oder die Person Y. Nicht zuletzt hat sich dieses Vertrauen in der frühneuzeitlichen Rechtsfigur des Ius Archivi (Schäfer 1999: insbesondere

1 Open Archival Information System (OAIS): 1-1. http://public.ccsds.org/publications/archive/650x0b1.pdf. Der im Internet kostenfrei abrufbare Standard ist inhaltlich identisch mit der ISO-Norm 14721. Alle Internetadressen wurden am 16. April 2012 überprüft.

169-171) niedergeschlagen. Danach genießen Archivalien grundsätzlich hohe Glaubwürdigkeit vor Gericht. Heute ist das Ius Archivi kein geltendes materielles Recht mehr, sinngemäß kommt es aber nach wie vor zur Anwendung, da die Glaubwürdigkeit der von klassischen Archiven vorgelegten Unterlagen in aller Regel vor Gericht nicht bezweifelt wird. Auch im angelsächsischen Raum kommt der unbroken custody im Archiv eine besondere Glaubwürdigkeit zu[2]. Andere Einrichtungen wie zum Beispiel Forschungsdatenzentren können sich, wenn sie ein Archiv aufbauen, nicht auf vergleichbare Traditionen berufen. Es verwundert daher wenig, wenn gerade von den neuen digitalen Archiven[3] verstärkt Normen nachgefragt werden, die bei der Feststellung ihrer Vertrauenswürdigkeit behilflich sein können.

Die zentrale Zielsetzung für alle Archive ist es, benutzt zu werden. Ob die Nutzer den vorgelegten Objekten Glauben und Vertrauen schenken, kann von keiner Norm der Welt festgeschrieben werden. Dies ist Ansichtssache der Nutzer. Jedoch muss dieser Gedanke noch weiter entwickelt werden. Zunächst wird ein kritischer Nutzer mit den Methoden der Quellenkritik, also mit der Diplomatik versuchen, die Glaubwürdigkeit der einzelnen Unterlagen zu befestigen oder auch zu erschüttern. Discrimen veri ac falsi, die wahren und falschen Dokumente unterscheiden, so hat bereits Daniel von Papenbroeck, einer der Urväter der Diplomatik, deren Aufgaben beschrieben (vgl. Boyle 1992: 83). Weshalb dann DIN 31644, müssten die Fragen der Glaubwürdigkeit nicht auf der Ebene der Objekte statt der Einrichtung selbst beantwortet werden? Welche Aufgaben kommen dabei noch den Archiven zu?

Auch in Zukunft sollen einzelne interessierte Nutzer die Möglichkeit haben, Wahres vom Falschen zu unterscheiden. Selbst diese Nutzer werden aber schon aus Zeitgründen kaum die Möglichkeit haben, bei allen vorgelegten Objekten eine eingehende diplomatische Untersuchung vorzunehmen. Sie müssen sich daher, ebenso wie alle anderen Nutzer, grundsätzlich darauf verlassen können, dass das Archiv die ihm anvertrauten Inhalte unverändert erhalten hat. Dieses Vertrauen müssen sich die neuen digitalen Archive erst noch erwerben. Aber auch die klassischen Archive müssen darauf sehen, das ihnen geschenkte Vertrauen nicht zu verlieren. Beide Fälle werden von der DIN Norm 31644 adressiert.

2 Vgl. zum Beispiel Jenkinson (1922: 14): "In any case, given an unbroken custody, the possibility of forgery is practically nil."

3 Die neuen digitalen Archive können als eigenständige Gruppe von Einrichtungen verstanden werden, die nun im Sinne von OAIS gezwungen sind, archivische Tätigkeiten vorzunehmen, die bislang den klassischen Archiven vorbehalten waren (siehe OAIS: 2-1: „These organizations are finding, or will find, that they need to take on the information preservation functions typically associated with traditional archives because digital information is easily lost or corrupted."). Siehe hierzu Keitel (2011).

Vertrauenswürdigkeit definiert die Norm wie folgt:

> *„Ein digitales Langzeitarchiv ist vertrauenswürdig: wenn es gemäß seinen Zielen und Spezifikationen zum Informationserhalt über lange Zeiträume hinweg operiert und seine Nutzer, Produzenten, Betreiber, Partner ihm dieses zutrauen."*

Diese Definition spiegelt bereits einige wesentliche Konzepte der Norm wider. In erster Linie sollen Informationen (nicht Daten) erhalten werden. Zweitens werden die Ziele und Spezifikationen nicht von der Norm vorgegeben, sondern vom Archiv festgelegt, es kann also verschiedene Umsetzungen der Norm geben. Drittens werden „lange Zeiträume" in den Blick genommen. Der zentrale Standard zur digitalen Archivierung OAIS[4] umschreibt diese langen Zeiträume als *„über die Lebensdauer der heutigen Hard- und Software hinaus"*[5], eine Definition, die auch von der DIN Norm 31644 übernommen wurde (DIN 31644: 2012-04: 4). Viertens werden neben den Nutzern auch noch weitere Gruppen genannt, für die Vertrauenswürdigkeit eine Rolle spielen dürfte. Ein Produzent der Daten und Dokumente sollte davon ausgehen können, dass diese nach Abgabe ans Archiv in ihrer Aussagekraft nicht geschmälert werden. Der Betreiber des Archivs sollte feststellen können, ob er das Richtige tut. Auch alle anderen Partner des Archivs sollten ihm Vertrauen schenken können. Die Definition der Vertrauenswürdigkeit integriert daher einige zentrale Konzepte der Norm, die weiter unten näher ausgeführt werden.

Das eingangs beschriebene Problem der leichten Veränderbarkeit digitaler Daten zwingt dazu, über die Glaubwürdigkeit archivierter digitaler Dokumente nachzudenken. Darüber hinaus stellen die besonderen Eigenschaften digital gespeicherter Informationen gerade digitale Langzeitarchive vor weitere Probleme. Sie müssen die digitalen Objekte bei ihrer Archivierung verändern, um sie überhaupt erhalten zu können. Regelmäßig müssen die Daten auf andere, neuere Datenträger umkopiert werden, da es keine wirklich brauchbaren dauerhaften digitalen Datenträger gibt[6]. Zwar ist es vorteilhaft, digitale Dateien vollständig und verlustfrei auf neue Datenträger kopieren zu können. Bei Archivalien auf Papier oder Pergament ist dies nicht möglich. Zugleich verschärft sich eben

4 Siehe Fußnote 1.

5 Ebd.: 1-11. „Long Term: A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future."

6 Zwar gibt es immer wieder Berichte über haltbare Spezialdatenträger. Hier stellt sich aber bereits nach wenigen Jahren die Frage, ob es überhaupt noch Laufwerke gibt, um die Spezialdatenträger abspielen zu können. Seit über zehn Jahren haben die digitalen Langzeitarchive sowohl in Deutschland als auch international die Suche nach einem wirklich haltbaren digitalen Datenträger aufgegeben. Es erscheint ausreichend, die Daten regelmäßig auf neue Datenträger zu überführen.

dadurch das Glaubwürdigkeitsproblem, da eine Überprüfung der Glaubwürdigkeit nicht mehr wie bei analogen Archivalien am Datenträger festmachen kann (beispielweise Wasserzeichen in Papier). Schließlich sind auch die heutigen Dateiformate nicht von Dauer. Die meisten Archive speichern die Inhalte daher regelmäßig in neueren Dateiformaten ab (Migrationsstrategie) und verändern insofern ihre digitalen Archivalien. Manche Archive verändern die Archivalien zwar nicht, spielen sie aber mit zusätzlicher Software, sogenannten Emulatoren, ab (Emulationsstrategie). Auch hierbei wird jedoch die von Menschen wahrnehmbare Ausgabe der Information auf Monitoren, Lautsprechern oder in einer anderen, sinnlich wahrnehmbaren Form, verändert[7], und gerade diese Information ist wie erwähnt das eigentliche Ziel der Erhaltung. Unabhängig von der Erhaltungsstrategie werden daher im Laufe der Archivierung sowohl die Datenträger als auch die Dateiformate oder die verarbeitende Software geändert. Die Glaubwürdigkeit digitaler Archivalien ist daher auf verschiedenen Ebenen gefährdet.

Internationale Diskussion

Lässt man alle diese Gefährdungen Revue passieren, dann verwundert es nicht, dass bereits 1996 das Glaubwürdigkeitsproblem der digitalen Langzeitarchivierung thematisiert wurde. Eine vom Online Computer Library Center (OCLC) eingesetzte Expertengruppe stellte fest:

> *"A process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information." (Task Force on Archiving of Digital Information 1996: 40)*

Empfohlen wurde daher:

> *"Institute a dialogue among the appropriate organizations and individuals on the standards, criteria and mechanisms needed to certify repositories of digital information as archives." (Ebd.: 42)*

Diese Empfehlungen führten 2002 zu dem von der Research Library Group RLG zusammen mit der OCLC veröffentlichten Report *„Trusted Repositories Attributes & Responsibilities"*[8]. Hier werden konkrete Anforderungen an ein digitales Langzeitarchiv benannt und aufgeschlüsselt. Der Report strebt eine Compliance mit dem zentralen Standard für die digitale Langzeitarchivierung, Reference Model

7 Vgl. das Performance Model des Australischen Nationalarchivs (Heslop et al. 2002). Siehe auch nestor (2011).

8 http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf

for an Open Archival Information System (OAIS) an. Er umschreibt die für Fragen der Vertrauenswürdigkeit relevanten Bereiche und benennt die Punkte, an denen die Forschung ansetzen sollte. Die abschließend gegebenen sieben Empfehlungen sind noch eher abstrakter Natur. So lauten die ersten beiden Empfehlungen *„Develop a framework and process to support the certification of digital repositories."* und *„Research and create tools to identify the attributes of digital materials that must be preserved."* So nennt dieser Report zwar viele zu beachtende Aspekte, er formuliert aber noch keine eindeutige Kriterien.

Auf den Report von RLG und OCLC von 2002 bauen alle Kriterienkataloge im engeren Sinne auf. Außerdem wurde auf derselben Basis als alternatives Verfahren die Methode von DRAMBORA entwickelt[9]. Die Kriterienkataloge lehnen sich an den Standard OAIS an und definieren eine abgeschlossene Liste benannter Kriterien. Dagegen unterstützt DRAMBORA die Archive dabei, ergebnisoffen Risiken aufzufinden und über eine Einstufung beherrschbar zu machen. Die DIN-Norm 31644 zählt zur Gruppe der Kriterienkataloge, die im Zentrum der folgenden Betrachtung stehen sollen.

Ein erster Kriterienkatalog im engeren Sinne wurde 2005 von der National Archives and Records Administration (NARA), also dem amerikanischen Nationalarchiv und der RLG mit der *„Audit Checklist for the Certification of Trusted Digital Repositories"* zur öffentlichen Kommentierung ins Internet gestellt[10]. Erstmals werden hier einzelne Kriterien benannt. Entsprechend konkret beginnen die meisten Kriterien mit *„Repository has/commits/is..."*, zum Beispiel *„A 3.4 Repository has a documented history of the changes to its operations, procedures, software, and hardware, traceable to its preservation strategies where appropriate."*[11]

Die 86 Kriterien verteilen sich auf einzelne Abschnitte:

Bereich	Kriterien
A Organization	*21*
B Repository Functions, Processes, & Procedures	*36*
C The Designated Community & The Usability Of Information	*10*
D Technologies & Technical Infrastructure	*19*
Gesamt	86

Tabelle 1: Audit Checklist for the Certification of Trusted Digital Repositories

9 Zu DRAMBORA siehe http://www.repositoryaudit.eu/. Zur Abgrenzung zwischen beiden Ansätzen siehe Ross et al. (2008: 34-36). Prägnant auch Ruusalepp (2010).

10 http://www.rebiun.org/opencms/opencms/handle404?exporturi=/export/docReb/audit_cheklist.pdf

11 Ebd.: 10.

Im Kern des Katalogs stehen die von OAIS beschriebenen Funktionsbereiche (Teil B). Weitere Abschnitte behandeln die Organisation (A), die zu erwartenden Nutzer (C) und die Technik (D). Der Vorteil dieser Gliederung bestand in der engen Anlehnung an OAIS.
Bereits am 10. Dezember 2004 hatte sich in München erstmals die nestor-Arbeitsgruppe *„Vertrauenswürdige Archive – Zertifizierung"* getroffen, um an einem eigenen Kriterienkatalog zu arbeiten. Das Protokoll vermerkt als Zielsetzung der Gruppe:

> *„Die Arbeitsgruppe soll aufbauend auf internationalen Vorarbeiten – vor allem des Berichtes „Trusted Digital Repositories: Attributes and Responsabilities" der RLG/OCLC-WG – Spezifika und Anforderungen an digitale Depots diskutieren, Kriterien für die Evaluation von digitalen Depots im Hinblick auf die Langzeitarchivierungsbeschaffenheit definieren sowie ein Zertifizierungsmodell erarbeiten und dieses in die Praxis überführen."*[12]

Der Begriff des Depots zeigt, dass es zunächst noch sehr stark um IT-Systeme ging, eine Zielsetzung, die im Laufe der Zeit und in Anlehnung an OAIS von dem des digitalen Langzeitarchivs abgelöst wurde. OAIS versteht unter digitalen Langzeitarchiven eine durch Menschen und IT-Systeme gebildete Einheit. Im ursprünglichen Depotbegriff spiegelt sich dagegen der Bezug auf ein zweites Referenzdokument wider, das sogenannte DINI-Zertifikat, das erstmals 2004 veröffentlicht worden war[13].

Rückblickend wird deutlich, dass sehr unterschiedliche Gruppen und Personen bei der Erarbeitung des Kriterienkatalogs beteiligt waren. Ganz am Anfang stand ein Fragebogen, den die nestor-AG an zahlreiche digitale Archive versandte. Auch war die Arbeitsgruppe selbst mit Mitarbeiterinnen und Mitarbeitern aus ganz unterschiedlichen Einrichtungen besetzt. Sie arbeiteten zunächst in den klassischen Gedächtnisinstitutionen wie Bibliotheken, Archiven und Museen, aber auch Vertreter aus Universitäten und Forschungsdatenzentren wirkten in der AG mit. Neben den klassischen Gedächtnisinstitutionen sollte sich der Katalog an alle Einrichtungen wenden, die digitale Archive betrieben oder aufbauen wollten. Auf verschiedenen Workshops wurden daher die Ergebnisse vorgetragen und weitergehende Wünsche aus den einzelnen Communities aufgenommen.

Bereits im ersten Jahr wurde auch die Zielsetzung für diesen Katalog erweitert. Zunächst sollte er nun einzelnen Einrichtungen helfen, ihr digitales Archiv aufzubauen und auch bestehende Schwachstellen zu finden. Das Ziel der Zerti-

12 Protokoll der ersten Sitzung, unveröffentlicht.
13 http://www.dini.de/dini-zertifikat

fizierung wurde beibehalten, konnte aber in den ersten Jahren zunächst nicht weiter konkretisiert werden.

Es erschien sinnvoll, die Zahl der Kriterien der Audit Checklist zu reduzieren und inhaltlich sich überlappende Kriterien zusammenzufassen. Die AG beschloss, eine eigene, abweichende Gliederung zu erstellen. 2006 wurde der nestor-Kriterienkatalog vertrauenswürdige digitale Langzeitarchive erstmals zur öffentlichen Kommentierung ins Internet gestellt und auch ins Englische übersetzt[14]. Die Aspekte mancher Kriterien wurden dabei durch weitere Kriterien konkretisiert (Kriterien 2. Kategorie).

Abschnitt	Hauptkriterien	Kriterien 2. Kategorie
A Organisatorischer Rahmen	*5*	*16*
B Umgang mit Objekten	*7*	*22*
C Infrastruktur und Sicherheit	*2*	*2*
Gesamt	14	40

In der zweiten, ebenfalls deutsch und englisch veröffentlichten Fassung von 2008 bzw. 2009 (englische Version) wurden vor allem Begriffe geklärt und Kriterien eindeutiger formuliert. Die Zahl und Gliederung der Kriterien blieb unverändert.

Die inhaltliche Gliederung des Kriterienkatalogs wurde von der zweiten Version des amerikanischen Katalogs übernommen, der 2007 unter dem Titel *„Trustworthy Repositories Audit & Certification: Criteria and Checklist"* veröffentlicht und unter dem Kurznamen TRAC bekannt wurde[15].

Abschnitt	Kriterien
A. Organizational Infrastructure	*24*
B. Digital Object Management	*44*
C. Technologies, Technical Infrastructure & Security	*16*
Gesamt	84

Tabelle 3: Trustworthy Repositories Audit & Certification: Criteria and Checklist

Die Arbeit an TRAC wurde von der fast gleichnamigen, personell aber anders besetzten Gruppe RAC – Repositories Audit and Certification – beim Consultative Committee for Space Data Systems (CCSDS) fortgesetzt. Diese Gruppe ver-

14 Der Kriterienkatalog ist in seiner zweiten, verbesserten Version abrufbar unter http://nbn-resolving.de/urn:nbn:de:0008-2008021802 und http://nbn-resolving.de/urn:nbn:de:0008-2010030806 (englische Fassung).

15 http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

öffentlichte *„Audit And Certification Of Trustworthy Digital Repositories"*[16], die später in die ISO-Norm 16363: 2012 eingingen. Hauptziel dieser Norm ist eine third party certification. Dagegen tritt die mögliche Funktion derartiger Kataloge, bei der Konstruktion eines digitalen Langzeitarchivs beizutragen, in den Hintergrund[17].

Abschnitt	Hauptkriterien	Kriterien 2. Kategorie	Kriterien 3. Kategorie
Organizational Infrastructure	*15*	*10*	*0*
Digital Object Management	*29*	*26*	*5*
Infrastructure And Security Risk Management	*6*	*7*	*11*
Gesamt	50	43	16

Tabelle 4: Audit And Certification Of Trustworthy Digital Repositories/ISO 16363

Die Norm umfasst 50 Hauptkriterien, 43 Unterkriterien der ersten Kategorie und 16 Unterkriterien der zweiten Kategorie („Unterunterkriterien").[18] Sie ist damit wesentlich komplexer aufgebaut als die DIN 31644.

Als dritter Kriterienkatalog wurde 2007 von den niederländischen Kollegen der Data Seal of Approval veröffentlicht[19]. Seine Guidelines enthalten insgesamt 16 nicht weiter untergliederte Kriterien und ein Angebot zur Überprüfung: *„There is no audit, no certification: just a review on the basis of trust."* Die Kriterien lehnen sich an die Kriterien der zuvor veröffentlichten Kataloge an und können als Minimalset verstanden werden[20]. Beispielsweise wird in Kriterium 12 definiert: *„The data repository ensures the authenticity of the digital objects and the metadata."* DIN 31644 behandelt das Thema in den Kriterien 17 bis 19. In ISO 16363 wird die Authentizität in den Kriterien 4.1.1.1, 4.2.6.3, 4.4.2 und vor allem in Kriterium 4.6.2[21] behandelt: *„The repository shall follow policies and procedures that enable the dissemination of digital objects that are traceable to the originals, with evidence supporting their authenticity."*

Ende 2007 lagen daher drei verschiedene Kriterienkataloge vor, die sich in ihrer Zielsetzung, der Zahl der Kriterien (und damit einhergehend auch des

16 http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206520R1/Attachments/652x0r1.pdf (Redbook vom Oktober 2009).

17 Ebd.: 1-1.

18 Diese Unterteilung lässt sich leicht durch ein Beispiel aus der Geographie erläutern: 1. Deutschland; 1.1 Baden-Württemberg; 1.1.1 Stuttgart und 1.1.1.1 Stuttgart-Bad Cannstatt.

19 http://www.datasealofapproval.org

20 "The guidelines of the Data Seal of Approval can be seen as a minimum set distilled from the above proposals." (Harmsen 2008: 17)

21 Die Nummerierung der Kriterien beginnt mit der Nummer des Kapitels, in dem sie behandelt werden.

zu erwartenden Aufwands bei der Anwendung des Katalogs) und nicht zuletzt durch die jeweiligen Kontexte ihrer Entstehung unterschieden. Neben dem eher schmalen Data Seal of Approval standen TRAC, aus dem sich später die ISO-Norm 16363 und der nestor-Kriterienkatalog, aus dem sich DIN 31644 entwickeln sollte. Mehrfach wurde die Erarbeitung einer gemeinsamen Norm diskutiert. Da die drei Kriterienkataloge aber zahlreiche interne Querbezüge haben und vielfach mit ihren jeweiligen Communities abgestimmt worden waren, hätte dies ein sehr aufwändiges Projekt erfordert, das den Vertretern der einzelnen Initiativen bislang nicht umsetzbar erschien.

Aufbau und Inhalte von DIN 31644

2008 endete die Förderung des nestor-Projekts. nestor wurde nun als Kooperationsverbund fortgeführt. Es erschien schwierig, die aufwändigen Arbeiten an dem Kriterienkatalog vertrauenswürdige digitale Langzeitarchive bei nestor fortzuführen. Zugleich wurde von Seiten des Deutschen Instituts für Normung, kurz DIN, der Wunsch geäußert, daraus eine Norm zu entwickeln. Beim Normausschuss *„Schriftgutverwaltung und Langzeitverfügbarkeit digitaler Informationsobjekte"* (NABD 15) konstituierte sich daher eine Arbeitsgruppe für die Weiterentwicklung des nestor-Kriterienkatalogs zur DIN-Norm 31644. Die Arbeitsgruppe nahm einen guten Teil der Mitglieder aus der nestor-AG, aber auch neue Mitglieder auf. Die Kriterien wurden einer erneuten Revision unterzogen, dabei wurde die Unterscheidung zwischen Hauptkriterien und Kriterien 2. Kategorie nicht übernommen. Die Norm kennt nun 34 Kriterien. Ihre Terminologie wurde vereinheitlicht und mit den zeitgleich in Entstehung begriffenen DIN-Normen 31645 und 31646 abgeglichen[22]. Im Herbst 2010 konnte eine Version zur öffentlichen Kommentierung ins Internet gestellt werden. Am 21. März 2011 lud die AG alle Kommentatoren ins Hauptstaatsarchiv Stuttgart zu einer Einspruchssitzung ein. Hier nahmen Kommentatoren und AG einvernehmlich weitere Änderungen an dem Normtext vor, der schließlich im April 2012 veröffentlicht wurde.

DIN 31644 richtet sich an alle *„Einrichtungen, die digital gespeicherte Informationen langfristig erhalten"* möchten[23]. Der Anwendungsbereich ist damit weiter gefasst als beim nestor-Kriterienkatalog. Er schließt in dieser Hinsicht unmittelbar an die Definition von OAIS an, das sich ebenfalls als allgemeiner Standard für alle derartige Einrichtungen begreift. Da es sehr unterschiedliche

22 Auch diese beiden Normen gehen auf Vorarbeiten von nestor zurück. DIN 31645, Leitfaden zur Informationsübernahme in digitale Langzeitarchive, basiert auf dem nestor-Leitfaden „Wege ins Archiv" und die gegenwärtig im April 2012 noch als Normentwurf einzuteilende DIN 31646 „Anforderungen an die langfristige Handhabung persistenter Identifikatoren (Persistent Identifier)" basiert auf dem nestor-Kriterienkatalog zur Überprüfung der Vertrauenswürdigkeit von PI-Systemen.

23 DIN 31644: 5.

Zielsetzungen gibt und geben kann, ist DIN 31644 ein Rahmenstandard, der keine konkrete Umsetzung vorschreibt. Von diesem Rahmenstandard sind Richtlinien und Standards zu unterscheiden, die sich auf einzelne Bereiche beziehen, zum Beispiel die Technische Richtlinie *„Beweiswerterhaltung kryptographisch signierter Dokumente"*[24] oder die Prüfkriterien des VOI (2008). Beide Ansätze sind zwar unabhängig von DIN 31644 entstanden. Sie können aber als Versuche angesehen werden, eine bereichsspezifische Antwort auf die in DIN 31644 verhandelten Fragen zu geben.[25]

DIN 31644 ist ein Rahmenstandard, der die eingangs wiedergegebene und von den Verhältnissen abhängige Definition der Vertrauenswürdigkeit reflektiert. Dementsprechend kann er sehr unterschiedlich umgesetzt werden. Sinnvolle Anwendungsfelder finden sich sowohl bei Gedächtnisinstitutionen als auch in der freien Wirtschaft. Ob nun ein Staatsarchiv elektronische Akten oder eine Autofirma die Konstruktionszeichnungen ihrer Automobile erhalten will, stets geht es darum, digital gespeicherte Information so zu erhalten, dass sie zu einem späteren Zeitpunkt wie ursprünglich vom Archiv beabsichtigt verarbeitet und genutzt werden kann. Die Norm kann daher sowohl auf Archive mit gesetzlichen Aufträgen (zum Beispiel die Archivgesetze für die staatlichen Archive oder die Pflichtexemplargesetze für die National- und Landesbibliotheken) als auch für die Anliegen der freien Wirtschaft oder für rein private Zwecke angewandt werden. Die Aufbewahrungsdauer kann nur einige wenige Jahre betragen oder auch unbeschränkt sein.

Einheitlicher Bezugspunkt der DIN-Norm ist das digitale Langzeitarchiv, das wiederum im Sinne von OAIS durch ein Zusammenspiel von Personen und technischen Systemen konstituiert wird. Gerade wegen des Zusammenspiels erschien es nicht sinnvoll, einzelne Teile isoliert zu betrachten[26]. Mit der Anwendung der Norm kann ein digitales Langzeitarchiv mehrere Ziele verfolgen. Die Norm entlastet die Benutzer von Fragen der Vertrauenswürdigkeit; sie gibt digitalen Archiven Hinweise auf mögliche Schwachstellen; sie fördert die öffentliche Vertrauenswürdigkeit digitaler Archive und ermöglicht dadurch ein weitergehendes Engagement von privaten und staatlichen Geldgebern.

Obwohl DIN 31644 in terminologischer Hinsicht an den OAIS-Standard anknüpft, erscheint diese ISO-Norm nicht als „normativer Verweis". Das bedeutet, dass OAIS zwar zum Verständnis von DIN 31644 sinnvoll herangezogen werden kann. Zur Erfüllung der DIN-Norm 31644 muss nicht zwangsläufig die OAIS

24 Bundesamt für Sicherheit in der Informationstechnik -Technische Richtlinie 03125: Vertrauenswürdige elektronische Langzeitspeicherung https://www.bsi.bund.de/ContentBSI/Publikationen/TechnischeRichtlinien/tr03125/index_htm.html

25 Ob die beiden genannten Beispiele ganz oder teilweise konform zu DIN 31644 sind, kann hier nicht weiter behandelt werden.

26 Vgl. dagegen Ruusalepp (2010): "Linking trust to services that a repository is offering is more meaningful than to a whole institution or unit within an organisation."

erfüllt werden. Ähnlich wurde mit der Gruppe von denkbaren technischen Normen verfahren. Sie werden zwar im Literaturverzeichnis genannt, sind aber keine zwingende Voraussetzung zur Erfüllung von DIN 31644. Durch diese Entscheidungen wird eine Umsetzung der Norm nicht durch zusätzliche Anforderungen erschwert.

Im zweiten Abschnitt der Norm werden insgesamt 26 Begriffe definiert. Diese Begriffe sind, wie schon erwähnt, mit DIN 31645 und dem Normentwurf zu DIN 31646 abgeglichen. Neben den bereits erwähnten Begriffen sind für das Verständnis der Norm die Begriffe des Informationsobjekts und der Repräsentation wesentlich. Sie gehen auf die grundsätzliche Verfasstheit digital gespeicherter Informationen zurück. Computer geben die gespeicherten Daten auf einem Ausgabegerät (zum Beispiel Monitor, Lautsprecher) wieder, von wo aus der Mensch sie anhand seiner Sinne als Information wahrnehmen und verarbeiten kann. Während die Daten nur vom Computer gelesen werden können, sind Informationen ein beliebiger Typ austauschbaren Wissens. Im Laufe der Erhaltung ändern sich nun die zur Verfügung stehenden Computer, weshalb das digitale Langzeitarchiv entweder weitere Software hinzuziehen oder die Daten verändern muss. Wichtig ist dabei, dass nur klar definierte und damit abgegrenzte Datenpakete erhalten werden können. Diese Pakete werden in Anlehnung an den weit rezipierten PREMIS-Standard Repräsentation genannt. Da nur eine abgegrenzte Repräsentation erhalten werden kann, ist auch die darzustellende Information begrenzt, es entsteht ein Informationsobjekt. Eine Repräsentation umfasst alles, was für die Darstellung eines Informationsobjekts notwendig ist. Wenn die in einer Repräsentation enthaltenen Dateien aufgrund ihrer veraltenden Dateiformate migriert, d.h. als neue Dateien mit anderem Dateiformat abgespeichert werden müssen, bilden die neu entstandenen Dateien zusammen mit den noch nicht zu migrierenden anderen Dateien der ersten Repräsentation eine neue Repräsentation. Ein Informationsobjekt kann daher durch mehrere Repräsentationen dargestellt werden, wobei hier die Ausgabe der Information geringfügig variieren kann[27]. Im Laufe der Zeit vermehrt sich zumindest bei der Migrationsstrategie die Zahl der zugeordneten Repräsentationen[28].

Der dritte Abschnitt beschäftigt sich mit der Realisierung und Evaluierung vertrauenswürdiger Archive. Dabei wird ein mehrstufiger Prozess angenommen, der von der Konzeption über die Planung und Spezifikation, die Umsetzung und Implementierung bis hin zur Evaluierung geht. Wichtig ist jedoch, dies nicht als starres Phasenmodell zu verstehen. Stattdessen ist davon auszugehen, dass diese Abfolge mehrmals oder immer wieder durchlaufen werden muss.

---

27 Der Kern der ausgegebenen Information muss gleich bleiben. Er wird durch die signifikanten Eigenschaften beschrieben, siehe DIN 31644 Kriterium 13.

28 Preservation Metadata Maintenance Activity (PREMIS), http://www.loc.gov/standards/premis/ und Keitel (2010).

Im vierten Abschnitt der Norm werden ihre vier Grundprinzipien erläutert. Zwar sind diese Prinzipien wegen ihres hohen Abstraktionsgrades nicht überprüfbar. Sie bilden aber die Hintergrundfolie, vor der erst die 34 Kriterien zutreffend eingeordnet werden können. Das Prinzip der Dokumentation erlaubt es schon heute, den Entwicklungsstand eines digitalen Langzeitarchivs zu beurteilen. Es ist daher eine wichtige Vorbedingung für eine mögliche Zertifizierung. Außerdem ermöglicht erst die Dokumentation der Aktivitäten des Archivs, die spätere Glaubwürdigkeit der archivierten Objekte zu überprüfen und auch die angemessenen Erhaltungsmaßnahmen festzulegen. Auch die beste Dokumentation hilft den Nutzern aber nicht bei ihren Fragen nach Glaubwürdigkeit, wenn sie nicht eingesehen werden kann. Dem kann durch das Veröffentlichen geeigneter Teile der Dokumentation, also durch Transparenz begegnet werden[29]. Aber auch nach innen ist Transparenz notwendig. Die Archivmitarbeiter müssen erfahren können, wie mit den Objekten bisher umgegangen wurde, damit sie den eingeschlagenen Weg fortsetzen können. Das dritte Prinzip der Angemessenheit ist vielleicht das zentrale Prinzip der Norm. Hier wird nicht zuletzt reflektiert, dass bei der digitalen Archivierung keine absoluten Maßstäbe möglich sind. Das Archiv muss daher zunächst seine Ziele und Aufgaben benennen und dann dementsprechend – also angemessen hierzu – handeln. Schließlich ist es auch anzustreben, dass zu jedem Kriterium möglichst klare und überprüfbare Angaben gemacht werden können. Viertes Prinzip ist daher die Bewertbarkeit.

Die 34 Kriterien werden nach einem einheitlichen Schema beschrieben. In einer ersten Zeile stehen die Nummer und der Name des Kriteriums. Darunter folgt ein erläuternder Text, unter diesem dann Beispiele einer idealen Ausprägung. Im Anhang werden noch weitere Beispiele und Literaturhinweise aufgeführt. Diese Ausführungen sind von voraussichtlich kürzerer Haltbarkeit als die im Haupttext genannten Kriterien. Sie können daher bei einer Revision der Norm leichter ersetzt beziehungsweise ergänzt werden.

Beispielhaft sei im Folgenden das Kriterium 1 wiedergegeben:

K1	Auswahl der Informationsobjekte und ihrer Repräsentationen
Kriterien für die Auswahl der Informationsobjekte und ihrer Repräsentationen für das digitale Langzeitarchiv sind festgelegt. Der Rahmen ist vorgegeben durch gesetzliche Vorgaben, den Gesamtauftrag der Institution, des Unternehmens, eigene Zielvorgaben.	
Veröffentlichte Kriterien für die Auswahl der Informationsobjekte und ihrer Repräsentationen: Sammelrichtlinien, Auswahlkriterien.	

29 Vgl. auch die sehr pointierte Gegenüberstellung von Ruusalepp (2010): "With very little transparency from audits we may become over-confident (the excess of trust) which will lead to additional risks. With too much transparency may lead to insufficient confidence (excess of diffidence) and we may miss good opportunities/services."

Die Kriterien selbst sind in drei Abschnitte gegliedert. Elementar ist der Abschnitt A Organisatorischer Rahmen. Er enthält als Kriterien:

K1: Auswahl der Informationsobjekte und ihrer Repräsentationen
K2: Verantwortung für den Erhalt
K3: Zielgruppen
K4: Zugang
K5: Interpretierbarkeit
K6: Rechtliche und vertragliche Basis
K7: Rechtskonformität
K8: Finanzierung
K9: Personal
K10: Organisation und Prozesse
K11: Erhaltungsmaßnahmen
K12: Krisen-/Nachfolgeregelung

Die Kriterien des Abschnitts B Umgang mit Informationsobjekten und deren Repräsentationen zielen auf Signifikante Eigenschaften (K 13), Integrität (K 14 – 16), Authentizität (K 17 – 19), Pakete (K 20 – 26) und Metadaten (K 27 – 32):

K13: Signifikante Eigenschaften
K14: Integrität: Aufnahmeschnittstelle
K15: Integrität: Funktionen der Archivablage
K16: Integrität: Nutzerschnittstelle
K17: Authentizität: Aufnahme
K18: Authentizität: Erhaltungsmaßnahmen
K19: Authentizität: Nutzung
K20: Technische Hoheit
K21: Transferpakete
K22: Transformation der Transferpakete in Archivpakete
K23: Archivpakete
K24: Interpretierbarkeit der Archivpakete
K25: Transformation der Archivpakete in Nutzungspakete
K26: Nutzungspakete
K27: Identifizierung
K28: Beschreibende Metadaten
K29: Strukturelle Metadaten
K30: Technische Metadaten
K31: Protokollierung der Langzeiterhaltungsmaßnahmen
K32: Administrative Metadaten

Der Abschnitt C Infrastruktur und Sicherheit enthält die beiden letzten Kriterien:

K33: IT-Infrastruktur

K34: Sicherheit

Die Kriterien sind eng miteinander verwoben. Während die ersten Kriterien vor allem die Zielsetzungen des Archivs erfragen, geht es in den folgenden Kriterien darum, ob diese Zielsetzungen auch adäquat umgesetzt wurden. Nach Kriterium 2 verpflichtet sich ein digitales Langzeitarchiv grundsätzlich, die ihm anvertrauten Objekte zu erhalten. K 11 fragt, welche Erhaltungsmaßnahmen geplant sind. Diese Maßnahmen sind unter anderem verbunden mit den festzulegenden signifikanten Eigenschaften (Kriterium 13), spezifischen Fragen zur Authentizität (K 18) und der Protokollierung (Kriterium 31).

Auf die Kriterien folgen drei Anhänge. Alle drei Anhänge sind informativ, also nicht verpflichtend bei der Umsetzung der Norm. Da ein großer Teil der Literatur ausschließlich auf Englisch verfügbar ist, bietet Anhang A eine Konkordanz der englischen und deutschen Begriffe. Anhang B zählt Beispiele für digitale Langzeitarchive und Anhang C Beispiele für einzelne Kriterien auf. Danach beschließen Literaturhinweise die DIN-Norm.

Perspektiven

Heute stehen mit der DIN-Norm 31644, der ISO-Norm 16363 und dem DATA Seal of Approval drei „Kriterienkataloge" für die Selbstevaluierung und Zertifizierung von digitalen Langzeitarchiven bereit. Die Unterschiede zwischen diesen Ansätzen liegen nicht zuletzt in den jeweiligen nationalen und sprachlichen Theoriebildungen begründet. Bereits im Januar 2007 und damit noch vor der Veröffentlichung des Data Seal of Approval hatten sich in Chicago Vertreter von vier Projekten beziehungsweise Einrichtungen getroffen, um die in dem TRAC/RAC-Projekt entstandenen Vorstellungen mit jenen aus dem nestor-Umfeld anzugleichen. Gemeinsam wurden 10 grundlegende Anforderungen an digitale Archive formuliert, die an die Größe und den Typ der jeweiligen Archive angepasst werden sollten[30].

2010 lud dann die EU Vertreter der drei Initiativen nach Luxemburg ein, um über einen gemeinsamen Rahmen für die drei Ansätze zu sprechen. Am 8. Juli 2010 unterzeichneten sie ein Memorandum of Understanding, nach dem drei verschiedene Stufen der Zertifizierung zu unterscheiden sind[31]. Die Basic

30 Beteiligt waren The Digital Curation Center, DigitalPreservationEurope, nestor und das Center for Research Libraries, http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re

31 http://www.trusteddigitalrepository.eu/Site/Welcome.html

Certification kann anhand des Data Seal of Approval vorgenommen werden. Die weitergehende Extended Certification und die durch externe Experten durchzuführende Formal Certification können alternativ durch DIN 31644 oder ISO 16363 umgesetzt werden. In anderen Worten sind DIN 31644 und ISO 16363 gleichwertig. Derzeit bereitet eine nestor-Arbeitsgruppe eine mögliche Zertifizierung nach DIN 31644 vor[32].

32 http://www.langzeitarchivierung.de/Subsites/nestor/DE/Arbeitsgruppen/AGZertifizierung.html;jsessionid=9893C25CCDCE75E69FF3ED471068D10E.prod-worker4

Literatur

Boyle, L.E. (1992): Diplomatics. In: Powell, J.M. (Ed.): Medieval Studies, 2. ed., Syracus, 82-113.

Harmsen, H. (2008): Data Seal Of Approval – Assessment And Review Of The Quality Of Operations For Research Data Repositories. Vth DLM-Forum Conference Proceedings.

Heslop, H./Davis, S. and Wilson, A. (2002): An Approach to the Preservation of Digital Records. National Archives of Australia. http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf [16.04.2012]

Jenkinson, H. (1922): A Manual Of Archive Administrations. Oxford.

Keitel, C. (2010): Das Repräsentationenmodell des Landesarchivs Baden-Württemberg. In: Wolf, S. (Hrsg.): Neue Entwicklungen und Erfahrungen im Bereich der digitalen Archivierung: von der Behördenberatung zum Digitalen Archiv. 14. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen". München. 69-82. http://www.staatsarchiv.sg.ch/home/auds/14/_jcr_content/Par/downloadlist/DownloadListPar/download_8.ocFile/Text%20Keitel.pdf [16.04.2012]

Keitel, C. (2011): Archivwissenschaft zwischen Marginalisierung und Neubeginn. Archivar 64, 33-37.

nestor (nestor-Arbeitsgruppe Digitale Bestandserhaltung) (Hrsg.) (2011): Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung. Version 1.0. nestor-materialien 15. http://files.d-nb.de/nestor/materialien/nestor_mat_15.pdf [16.04.2012]

Ross, S./McHugh, A./Hofman, H./Innocenti, P. and Ruusalepp, R. (2008): Two Words, Two Challenges: Distinguishing Audit And Certification Of Digital Archives. In: Ve conférence du DLM-Forum – Vth DLM-Forum Conference. La gestion de l'information et des archives électroniques en Europe : réalisations et nouvelles directions, Toulouse, 34-36, http://www.archivesdefrance.culture.gouv.fr/static/2768

Ruusalepp, R. (2010): Repository audit and risk profiles: trust through transparency. http://dci.ischool.utoronto.ca/Raivo.ppt [16.04.2012]

Schäfer, U. (1999): Authentizität. Vom Siegel zur digitalen Signatur. In: Schäfer, U. und Bickhoff, N. (Hrsg.): Archivierung elektronischer Unterlagen. Werkhefte der Staatlichen Archivverwaltung Baden-Württemberg Serie A Landesarchivdirektion Heft 13. Stuttgart, 165-191. Jetzt auch abrufbar unter: http://www.staatsarchiv.sg.ch/home/auds/02/_jcr_content/Par/downloadlist_5/DownloadListPar/download_9.ocFile/Text%20Schaefer%20Authentizitaet.pdf [16.04.2012]

Task Force on Archiving of Digital Information (1996): Preserving Digital Information. Report of the Task Force on Archiving of Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group. Washington D.C. http://www.clir.org/pubs/reports/pub63watersgarrett.pdf [16.04.2012]

VOI (VOI Verband Organisations- und Informationssysteme e.V.) (2008): PK-DML – Prüfkriterien für Dokumenten- und Enterprise Content Management-Lösungen, Bonn.

Metadaten für die Langzeitarchivierung

Stefan Hein

Einleitung

Der vorliegende Beitrag betrachtet die Thematik Metadaten aus dem Blickwinkel der digitalen Langzeitarchivierung. Im Hauptteil werden theoretische Konzepte wie die Einordnung in das OAIS-Referenzmodell sowie die Entwicklung und der Einsatz von Standards beleuchtet. Abschließend werden einige praktische Aspekte zur automatisierten Generierung von Metadaten sowie konkrete Anwendungsszenarien im Umfeld der Langzeitarchivierung vorgestellt.

Motivation

Zu den Kernaufgaben der digitalen Langzeitarchivierung gehören bekanntermaßen die Erfassung, die langfristige Aufbewahrung und das Sicherstellen der dauerhaften Verfügbarkeit und Zugänglichkeit von digitalen Informationen einschließlich der zu deren Darstellung beziehungsweise Wiedergabe benötigten Hilfsmittel. Mittlerweile hat sich eine Vielzahl von Disziplinen um genau diese Kernaufgaben herausgebildet. Die langfristige Aufbewahrung, im digitalen Zeitalter also die Bestandserhaltung des Bitstroms, wird in der Literatur häufig als Bitstream Preservation bezeichnet und kann mittlerweile mit etablierten Ansätzen zur redundanten Speicherung (vgl. LOCKSS 2012 und RAID-Technologie) und dem regelmäßigen Einsatz von Checksummenprüfverfahren zur Sicherung der Datenintegrität als gelöst betrachtet werden.

Den genannten Kernaufgaben und damit verbundenen Herausforderungen wird auch durch die Erzeugung und Verwaltung von Metadaten begegnet. Ohne das Vorhandensein von inhaltsbeschreibenden Metadaten und robusten Identifikatoren wird der Zugang bzw. das Auffinden digitaler Information nur schwer möglich sein. Technische Metadaten hingegen eignen sich, um die zur Interpretation der Informationen notwendige Präsentation zu erhalten. Sie liefern damit erst die Hinweise, die einem digitalen Objekt mehr als nur die Bedeutung einer Sequenz von Nullen und Einsen geben. Zu den wohl wichtigsten Vertretern technischer Metadaten zählt dabei das Dateiformat. Nur die Kenntnis und Aufbewahrung dieser Informationen und der damit verbundenen Formatspezifikation

ermöglicht es, den Bitstrom auch in Zukunft in der Art zu entschlüsseln, in der er sich heute durch einen Mausklick manifestiert.
Diese Herausforderungen sind zugleich Motivation des vorliegenden Beitrages, der in der Folge die Fragen adressiert:

- Welche Informationen werden benötigt, um den Strom aus Nullen und Einsen auch in Zukunft noch interpretieren zu können?
- Wie können diese Informationen gewonnen werden?

Einordnung in OAIS

Die Einordnung in das OAIS-Referenzmodell (vgl. OAIS 2009) soll hier mit der Betrachtung seiner zentralen Informationseinheit Information Package beginnen. Die einzelnen Komponenten des Information Package werden in Abbildung 1 dargestellt:

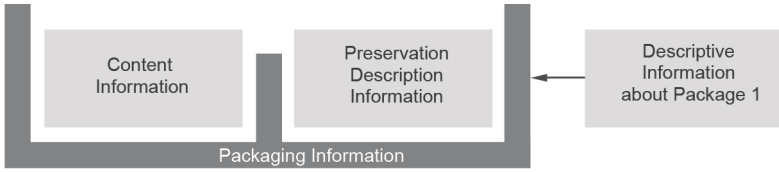


Abbildung 1: Konzepte und Beziehungen innerhalb eines Informationspakets (nach OAIS, CCSDS Draft Recommended Standard, Aug. 2009)

Die Content Information repräsentiert hier das zu archivierende Informationsobjekt. Sie beinhaltet den Datenstrom (Data Object) und die sog. Representation Information, mithilfe derer sich ein abstraktes Data Object in einer für den Nutzer verständlichen Form darstellen lässt.

Unter Preservation Description Information (PDI) werden nach dem OAIS-Referenzmodell alle Informationen verstanden, die die sichere Aufbewahrung der entsprechenden Content Information gewährleisten. Dazu gehören auch Informationen, die die Integrität der Content Information garantieren und diese in den Kontext zu anderen Objekten innerhalb des Archivs stellen. Zusätzlich lassen sich auch die Änderungshistorie sowie Funktionen zur Authentizitätssicherung integrieren.

Content Information und PDI sind innerhalb des OAIS-Referenzmodells nur konzeptionelle Komponenten einer logischen Einheit, deren physische Struktur auf einem Speichermedium völlig undefiniert bleibt. Die physische Verknüpfung dieser Informationen zum Beispiel durch Dateireferenzen wird deshalb eigens innerhalb der Packaging Information festgehalten.

Das Auffinden eines archivierten Information Package wird über zusätzliche Metadaten, den sog. Descriptive Information, ermöglicht. Sie beinhalten bibliografische Informationen zum eigentlichen Inhalt eines Information Package (Borghoff et al. 2003: 28).

Bei genauerer Betrachtung der PDI wird deutlich, dass dieses Konzept bereits einige der wichtigsten Informationsaspekte zur Langzeitarchivierung aufnimmt (siehe Abbildung 2).

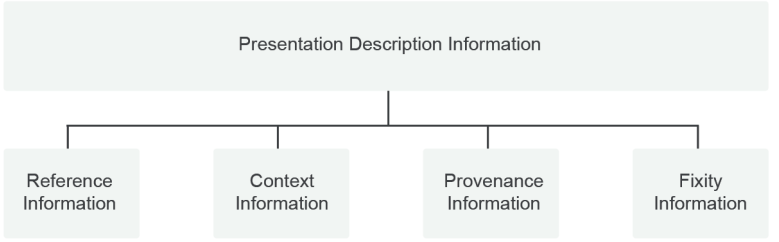


Abbildung 2: PDI

Der Container Reference Information hält die Informationen vor, die die eindeutige Identifizierbarkeit der Content Information sicherstellen. Neben bibliographischen Informationen wie Autor, Titel oder Verlag, sollen hier auch Referenzsysteme (zum Beispiel URN, DOI) angegeben werden.

Die dargestellte Tabelle dient der praktischen Veranschaulichung und zeigt einige Elemente der einzelnen PDI-Container am Beispiel eines zu archivierenden Softwarepakets.

Content Information Type	Reference	Provenance	Context	Fixity
Software Package	• Name • Author/ Originator • Version number • Serial Number	• Revision history • License holder • Registration • Copyright	• Help file • User guide • Related software • Language	• Certificate • Checksum • Encryption • CRC

Quelle: modifiziert nach [OAIS 2009: 4-31]

Die Beziehung der Content Information zu dessen Umgebung dokumentiert die Context Information. Gründe bzw. Umstände zur Entstehung der archivierten Ressource und Bezüge zu anderen Objekten sind Bestandteile dieses Containers.

Die Provenance Information beschreibt den Lebenszyklus der Content Information von seiner Entstehung an bis hin zum aktuellen Status. So liefert dieser Container Transparenz über die Herkunft bzw. die Quelle eines archivierten Objektes und trägt dabei maßgeblich zur Schaffung vertrauenswürdiger Archive bei.

Eine weitere Maßnahme zur Förderung von Vertrauen in elektronisch archivierte Objekte ist die Angabe der Fixity Information. Aspekte zur Datenintegrität und Authentizitätssicherung sollen mit diesem vierten Bestandteil der PDI Berücksichtigung finden. Vor allem sollen damit undokumentierte Änderungen an der Content Information und sich einschleichende Fehler besser erkennbar sein.

Typische Elemente dieser technischen Informationen sind beispielsweise Prüfsummen und digitale Signaturen.

Die im zweiten Kapitel angesprochenen Aspekte (zum Beispiel Zugänglichkeit, Datenintegrität) haben also auch ihre Entsprechung im Informationsmodell des OAIS-Referenzmodells. Etablierte Metadatenstandards können somit als eine praktische Umsetzung, der im OAIS-Referenzmodell definierten Informationen, verstanden werden. Das folgende Kapitel Metadatenstandards wird zeigen, dass sich diese Klassifikation in der realen Welt der Metadatenstandards jedoch nicht eins zu eins abbilden lässt. Die Vielzahl von Metadateninitiativen verfolgen in erster Linie eigene Zielstellungen in einem gegebenen Kontext. Dabei tangieren sie zumeist Elemente des OAIS-Referenzmodells und decken nur in wenigen Fällen vollständige Bereiche ab. Die Aufgabe der Entwickler von OAIS-konformen Archivsystemen sollte hingegen sein, sich wenn möglich vorhandener Metadatenkonzepte zu bedienen und diese im Sinne von OAIS zu integrieren. Wird dabei auf etablierte Metadatenstandards zurückgegriffen, fördert das die anzustrebende Interoperabilität der Bibliotheken und Archive weltweit.

An dieser Stelle sollte deutlich geworden sein, dass die Metadaten-Klassifizierung nach dem OAIS-Referenzmodell auch als Referenzmodell verstanden werden muss. Das Referenzmodell macht in diesem Sinne keine Vorgaben über die tatsächliche Implementation der Metadaten. Auf der Modellebene sind beispielsweise Representation Information alle Informationen, um den Übergang von der physischen zur logischen Ebene von digitalen Dokumenten dauerhaft sicherzustellen. In der Praxis wird versucht, dieser Anforderung durch die Generierung und zusätzliche Archivierung von technischen Metadaten nachzukommen. Diese Beziehung zwischen einem Referenzmodell und der praktischen Umsetzung wird in gleicher Weise bei der Betrachtung des in der Welt der Rechnerkommunikation anerkannten Open Systems Interconnection-Referenzmodells (OSI) und der in der Praxis etablierten Übertragungsprotokolle wie TCP/IP deutlich.

Metadatenstandards

Das vorherige Kapitel zeigte, dass sich Metadaten, je nachdem welchem Zweck sie dienen, zum Großteil in verschiedene Klassen einordnen lassen. Eine mögliche, in der Literatur vielfach verwendete Klassifikation von Metadaten liefert Gartner (2008). Hierbei werden die folgenden Klassen unterschieden, die im Verlauf dieses Beitrags auch als weitere Gliederung dienen sollen:

- deskriptive Metadaten,
- strukturelle Metadaten,
- administrative Metadaten (technische Metadaten, Rechte-Management, digitale Provenienz).

Deskriptive Metadaten entsprechen den bibliografischen Informationen (Autor, Titel usw.). Sie liefern also Informationen über den intellektuellen Inhalt einer Ressource und sind somit für das aus Nutzersicht relevante Auffinden einer Ressource erforderlich. Genau genommen ist der Begriff deskriptiv nicht ganz glücklich gewählt, denn im Sinne von „beschreibend" trifft diese Eigenschaft auf alle Metadatenklassen zu.

Strukturelle Metadaten beschreiben die interne Struktur eines archivierten Objektes, sodass die Darstellung für den Nutzer in der ursprünglich beabsichtigten Form erhalten bleiben kann. Die interne Struktur berücksichtigt sowohl die Beziehungen zwischen logischen und physischen Komponenten wie beispielsweise Dokumententeilen, als auch einzelne Dateien. Ein Buch zum Beispiel wird typischerweise in viele Seiten gegliedert. Auch die multimediale Darstellung einer aus vielen unterschiedlichen Dateien bestehenden Website ist ein Beispiel für die Notwendigkeit struktureller Metadaten.

Administrative Metadaten umfassen alle Informationen, die dem eigentlichen Archivierungsprozess, dem späteren Zugriff und der authentischen Wiedergabe der archivierten Ressourcen dienlich sind.

Die Kombination ausgewählter Bestandteile aus administrativen und strukturellen Metadaten liefert schließlich die benötigte Basis, um digitale Dokumente auch unter Berücksichtigung von rechtlichen Belangen dauerhaft zu archivieren und verfügbar zu halten. Prinzipiell bildet also die Klasse *„Langzeitarchivierungsmetadaten"* einen Querschnitt über alle genannten Metadatenklassen und fokussiert dabei die Unterstützung zur Durchführung von Erhaltungsstrategien und die Dokumentation der Auswirkungen durch deren Anwendung auf das archivierte Objekt.

Nicht alle Metadateninitiativen wie zum Beispiel das Dublin Core Metadata Element Set lassen sich ausschließlich in eine der genannten Klassen einordnen. Einige Initiativen verfolgen das Ziel, ein umfassenderes Metadatenset zu entwi-

ckeln, das mehrere Metadatenklassen berücksichtigt. PREMIS zeigt mit seinem Konzept des Erweiterungscontainers, dass das nur gelingen kann, wenn spezialisierte Metadatenstandards, wie zum Beispiel NISO Z.39.87, eingebettet werden können. Die Einordnung in nur eine der genannten Klassen wäre in diesen Fällen aufgrund der Überschneidungen sachlich unzutreffend. Die Schaffung eines allumfassenden und zugleich handhabbaren Metadatensets muss jedoch aufgrund der enormen Komplexität und Spezifität jeder einzelnen Metadatenklasse und jedes einzelnen Anwendungsfalls als nahezu unmöglich eingeschätzt werden.

Vorteile von Metadatenstandards

Im Allgemeinen definieren und vereinheitlichen Metadatenstandards den strukturellen Aufbau und Wertebereich von fachlich zusammengehörigen Metadatenelementen. Einige der hinlänglich bekannten Vorteile zur Verwendung von Standards werden hier nun kurz in Hinblick auf die Langzeitarchivierung dargestellt.

Die Verwendung von Metadatenstandards steigert im Allgemeinen die Interoperabilität. Setzen beispielsweise zwei Langzeitarchivsysteme auf den gleichen Metadatenstandard, lassen sich Informationen wie beispielsweise Katalogdaten in Form von deskriptiven Metadaten zwischen beiden leicht austauschen. Auch die Austauschbarkeit von Softwarekomponenten erhöht sich. Folgen beispielsweise die eingesetzten Tools zur Generierung von technischen Metadaten einschlägigen Standards, so ist ein Upgrade dieser Tools oder ein Austausch des Langzeitarchivsystems mit deutlich weniger Problemen (zum Beispiel durch Konvertierungsmaßnahmen) verbunden. Die Einhaltung von Standards ermöglicht ebenso die Wiederverwendbarkeit der Informationen in anderen Zusammenhängen. So ist es denkbar, strukturelle Metadaten neben den Aspekten zur Langzeitarchivierung beispielsweise auch für barrierefreie Präsentationssysteme zu nutzen.

Deskriptive Metadaten

Einer der bekanntesten und zugleich vielfach verwendeten Metadatenstandards ist das Dublin Core Metadata Element Set (DCMES). Das DCMES geht aus einer Entwicklerkonferenz aus dem Jahr 1995 in Dublin (Ohio) hervor, bei der sich die Beteiligten unter der Federführung der Dublin Core Metadata Initiative (DCMI) auf diese erweiterbare Basismenge zur Beschreibung von Dokumenten für die elektronische Archivierung einigten. Vor allem der Wunsch nach Einfachheit begründet dabei die Beschränkung auf zunächst 15 Basiselemente. Es darf angenommen werden, dass dieser Aspekt merklich zu einer breiten Akzeptanz des DCMES beigetragen hat. Auch Bibliotheken, Museen und Archive zeigten ausgeprägtes Interesse an dem Vorschlag der DCMI, der häufig als erster Ansatz der „digitalen Karteikarte“ angesehen wird (Borghoff et al. 2003: 147-148).

In der Version 1.1 sind folgende Elemente Bestandteil des mittlerweile ISO-zertifizierten DCMES (die deutsche Übersetzung nach Frodl et al. 2007):

Inhaltsbeschreibung	Technische Metadaten
• *title: Name der Ressource* • *subject: Thema der Ressource* • *coverage: Geltungsbereich* • *description: Beschreibung der Ressource* • *language: Sprache*	• *type: Art und Gattung der Ressource* • *identifier: Identifikator* • *format: Dateiformat, Datenträger oder Umfang der Ressource* • *date: Zeitangabe*

Personen und Rechte	Vernetzung
• *creator: Urheberin / Urheber* • *publisher: Verlegerin / Verleger* • *contributor: Mitwirkende / Mitwirkender* • *rights: Rechte*	• *source: Quelle* • *relation: Beziehung*

Alle Elemente sind hierbei optional, können mehrfach auftauchen und in beliebiger Reihenfolge stehen. Die hier verwendete Klassifizierung soll lediglich einen Eindruck vermitteln, welcher Anforderungsbereiche der Archivierung elektronischer Ressourcen man sich bereits mit dem DCMES bewusst war, und zeigt, dass das DCMES kein reiner deskriptiver Metadatenstandard ist. Schnell wird hierbei auch deutlich, dass spezielle Bedürfnisse mit diesem Kernset nicht abgedeckt werden können. Beispielsweise fehlt der Zeitangabe eine eindeutige Semantik. Auch das Codierungsschema der hinterlegten Werte muss erfasst werden können. Vor allem in Bezug auf die Langzeitarchivierung sollte klar sein, dass die Semantik der Elemente deutlich verfeinert werden muss.

Abhilfe schaffen Qualifier, von denen für jedes Element mehrere definiert werden können. Nach Borghoff et al. (2003: 152-155) werden für das Element date durch die DCMI beispielsweise folgende Verfeinerungen vorgeschlagen:
- created: Erstellungsdatum,
- modified: Datum der letzten Änderung.

Qualifier bieten also die Möglichkeit, das Vokabular des DCMES unter verschiedenen Gesichtspunkten zu erweitern. Eine Variante eines das DCMES enthaltenden Vokabularsets veröffentlichte die DCMI unter dem Namen DCMI Metadata Terms (DC-TERMS).

Strukturelle Metadaten

Digitale Dokumente unserer Zeit setzen sich zum Großteil aus mehreren Objekten zusammen. Angesichts des Bedarfs der Archivierung solcher Dokumente liefert der durch die Digital Library Federation (DLF) auf den Weg gebrachte

Metadata Endcoding & Transmission Standard (METS) ein geeignetes Metadatenkonzept, das sowohl die physische als auch die logische Struktur eines komplexen multimedialen Dokuments beschreibt und zudem als Containerformat für die verschiedenen Metadatenstandards konzipiert wurde (LOC 2009).

Als Containerformat dient METS vor allem der Beschreibung komplexer Dokumentstrukturen, indem es den Inhalt mit administrativen und deskriptiven Metadaten verknüpft. Die Tatsache, dass METS zu diesem Zweck die Verwendung externer Metadatenstandards wie beispielsweise Dublin Core vorsieht, macht das DLF-Konzept ausgesprochen flexibel. Die verwendeten Metadatenstandards können zum einen in das METS-Dokument integriert oder per ID-Referenz entsprechend verlinkt werden. METS erlaubt selbst für binären Dokumentinhalt, diesen Base64-kodiert einzubetten. Eine weitere Stärke von METS ist die Möglichkeit, die Granularität selbst festzulegen. Einzelne Metadatenabschnitte können sowohl das zu archivierende Dokument als Ganzes beschreiben als auch einzelnen Teilelementen eines Dokuments zugeordnet werden (Borghoff et al. 2006: 142-147).

Aufbau eines METS-Dokuments

Ein METS-Dokument besteht strukturell aus den folgenden sieben Sektionen (LOC 2009):

1. Kopfbereich (METS-Header) – enthält Metainformationen über das jeweilige METS- Dokument selbst,
2. deskriptive Metadaten – zum Beispiel Dublin Core,
3. administrative Metadaten – zum Beispiel technische Metadaten, Urheberrechte und Lizenzinformationen, Provenienz-Informationen,
4. Dateiabschnitt – eine geordnete und optional gruppierbare Liste aller zum archivierten Objekt gehörigen Dateien,

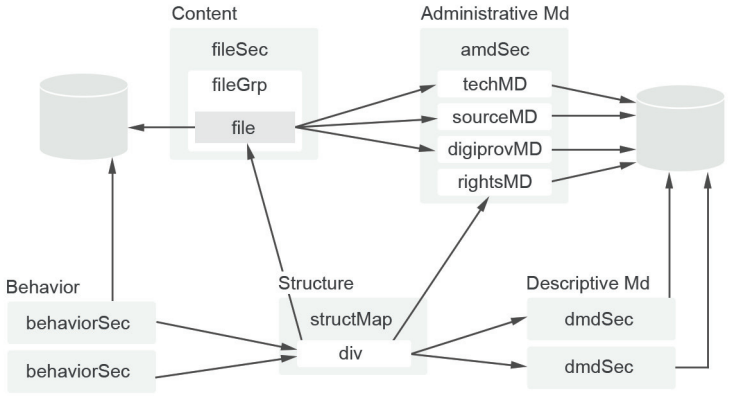


Abbildung 3: METS-Diagramm (Quelle: modifiziert nach Beaubien 2002: 17)

5. Strukturbeschreibung (Structural Map) – beschreibt den inneren Aufbau des digitalen Dokuments und verknüpft die Strukturelemente mit Dateien aus dem Dateiabschnitt und den dazugehörigen Metadaten,
6. Strukturverknüpfungen (Structural Links) – erlauben Hyperlinks zwischen den verschiedenen Komponenten eines METS-Dokuments,
7. Verhalten (Behavior) – verknüpft ausführbare Anweisungen (zum Beispiel Programmcode) – die das Verhalten des digitalen Dokuments beeinflussen – mit den Inhalten innerhalb eines METS-Objektes.

Das Zusammenspiel dieser elementaren Sektionen verdeutlicht die Abbildung 3 nach Beaubien (2002). So kann beispielsweise die Strukturbeschreibung (Structure) mit Elementen des Dateiabschnitts (Content) verbunden sein und auf deskriptive und administrative Metadatenelemente verweisen. Letztere können sich beispielsweise innerhalb einer separaten Datenbank befinden, während die einzelnen Dateien wiederum in einem Dateisystem abgelegt werden.

Administrative Metadaten

Wie oben bereits erwähnt, umfassen administrative Metadaten ein breiteres Spektrum an verschiedenen Informationen. Zwei wichtige Vertreter technischer Metadatenstandards sind textMD für Textdokumente und ANSI/NISO Z.39.87 für Bilddokumente.

textMD

Im Vergleich zu Bilddokumenten erscheinen die Langzeitarchivierung und die dafür benötigte technische Beschreibung von einfachen ASCII-kodierten Textdokumenten zunächst weniger problematisch. Nicht vergessen werden sollte jedoch, dass in fernerer Zukunft die Übersetzung von 7 Bit in ein ASCII-Symbol eines uns bekannten Alphabets nicht zwingend geläufig sein muss, wie es zurzeit der Fall ist. Auch aktuelle Anforderungen wie die originalgetreue Wiedergabe von Schriftarten, Layout und der Seitenfolge sollten zur technischen Beschreibung eines Textdokuments gehören. Die Library of Congress (LOC) stellt mit dem Schema Technical Metadata for Text (textMD) auf Basis der Vorarbeiten des New York University Digital Library Team einen weiteren Metadatenstandard zur Verfügung (LOC 2012).

Das textMD-Schema fokussiert vor allem die folgenden Aspekte zur technischen Beschreibung von digitalen Textdokumenten:

- technische Informationen der Kodierung
- Zeichenkodierung (Zeichensatz, Byte-Reihenfolge, usw.)
- Sprache
- Schriftart

- Dokumenten-Markup
- technische Anforderungen für Druck und Anzeige
- Seiten-Reihenfolge.

Das textMD-Element character_info ist beispielsweise ein Element, das Informationen über die Kodierung der Zeichen innerhalb eines Dokuments beinhaltet. Es besteht aus den Unterelementen charset, byte_order, byte_size, character_size und linebreak, die in der folgenden Tabelle beschrieben werden.

Elementname	Definition	Wert (Bsp.)
charset	*Der Name des verwendeten Zeichensatzes. Dabei sollte auf ein einheitliches Vokabular zurückgegriffen werden.*	*ANSI_X3.4-1968*
byte_order	*Die Byte-Reihenfolge, also entweder big, little oder middle.*	*little*
byte_size	*Die Größe eines Bytes in Bits.*	*8*
character_size	*Die Größe eines einzelnen Zeichens in Bytes. Dabei bezieht sich ein Byte auf die unter byte_size festgelegte Größe.*	*variable*
linebreak	*Der Mechanismus, der verwendet wird, um einen Zeilenumbruch zu markieren.*	*CR*

ANSI/NISO Z.39.87

Ein vom American National Standards Institute (ANSI) anerkannter Metadatenstandard zur Beschreibung technischer Metadaten für die Verwaltung von digitalen Bildern und Bildsammlungen ist der von der National Information Standards Organisation (NISO) entwickelte Standard NISO Z39.87. Der Standard selbst wird durch das Data Dictionary Technical Metadata for Digital Still Images beschrieben. Ein Data Dictionary ist ein Verzeichnis der Elemente eines Metadatenstandards. Es enthält die einzelnen Elementdefinitionen und darüber hinaus Informationen und Empfehlungen zu dessen Anwendung. Die erste Version des Data Dictionary orientierte sich vor allem an den Elementen der TIFF-Spezifikation. Aktuell wurde der Standard hinsichtlich der Integration von Metadaten aus Formaten wie JPEG und JPEG2000 sowie der Exif-Spezifikation erweitert. Da Metadaten innerhalb solcher Dateiformate auch automatisiert ausgelesen werden können, sind sie bequem in kompatible Metadatenstandards überführbar. Somit wird, durch NISOs Orientierung an etablierten Dateiformaten, die Erstellung von Metadatensätzen bei der Archivierung erleichtert (NISO 2006).

Ein Metadatensatz enthält nach NISO die fünf folgenden Abschnitte:

1. Basic Digital Object Information,
2. Basic Image Information,
3. Image Capture Metadata,
4. Image Assessment Metadata,
5. Change History.

Abschnitt 1 enthält allgemeine Metadatenelemente (zum Beispiel Identifier, Dateigröße), die für alle digitalen Objekte anwendbar und nicht spezifisch auf Bilddokumente ausgerichtet sind. Da dieser Abschnitt beispielsweise mit dem Container-Element Fixity auch Metadaten zu Aspekten der Langzeitarchivierung enthält, wurde bei der Definition auf eine Harmonisierung mit dem im folgenden Kapitel vorgestellten PREMIS preservation metadata element set geachtet.

Alle Informationen, die für die optische Darstellung des durch das digitale Objekt kodierten Bildes notwendig sind, werden in Abschnitt 2 festgehalten. Typische Angaben aus diesem Bereich sind die Bildauflösung und der Farbraum.

Im dritten Abschnitt finden alle Informationen Platz, die mit der Bilderzeugung in Zusammenhang stehen. Vor allem die technische Umgebung, die der Umwandlung vom analogen hin zum digitalen Bild diente, steht hier im Vordergrund. Charakteristika zur Beschreibung des ursprünglichen Filmmaterials sowie Herstellerangaben zu verwendeten Scannern und Digitalkameras werden hier hinterlegt. Auch GPS-Daten können hier untergebracht werden.

Der vierte Abschnitt enthält Informationen, die eine Bewertung der digitalen Bildqualität und Exaktheit der Ausgabe ermöglichen.

Weitere Informationen, die der Dokumentation der Verarbeitungsgeschichte eines Bildes dienen, finden sich im fünften Abschnitt wieder. Alle Verarbeitungsschritte, die zeitlich nach den in Abschnitt 3 festgehaltenen Prozessen ablaufen, sollen hier aufgenommen werden. Auch Erhaltungsstrategien wie durchgeführte Migrations-Operationen werden hier dokumentiert. Durch die versionierte Speicherung aller durch Verarbeitungsprozesse veränderten Metadatenabschnitte, findet damit die im OAIS-Referenzmodell geforderte Provenance Information Berücksichtigung.

Für das Markup entwickelte die NISO in Zusammenarbeit mit der Library of Congress und dem dort ansässigen, ebenfalls bei der Entwicklung von METS beteiligten Network Development and MARC Standards Office das XML-Schema Metadata for Images in XML (MIX) (LOC 2008). Durch diese Zusammenarbeit wird MIX auch als ein Erweiterungsschema für den Abschnitt administrative Metadaten der METS-Spezifikation empfohlen.

Langzeitarchivierungs-Metadaten

Die Initiative PREservation Metadata: Implementation Strategies (PREMIS) ist das international renommierteste Projekt, das sich für die Entwicklung eines Kernsets von Metadaten speziell für die Langzeitarchivierung einsetzt. Im Vordergrund von PREMIS stehen also die Informationen, die für die Umsetzung der Erhaltungsstrategien wie zum Beispiel Migration und Emulation erforderlich sind. Das Online Computer Library Center (OCLC) und die Research Library Group (RLG) gründeten PREMIS im Jahre 2003. In den drei Jahren zuvor legten vor allem die Projekte NEDLIB und CEDARS sowie die Vorschläge der National Library of Australia (NLA) die Grundsteine für PREMIS. Die OCLC/RLG-Arbeitsgruppe veröffentlichte auf Basis dieser Vorarbeiten bereits im Jahre 2002 eine Rahmenstruktur zur Unterstützung der Konservierung digitaler Objekte, die sich mit den bekannten Metadatentypen Content Information und PDI am OAIS-Referenzmodell orientierte (Day 2004).

Das PREMIS-Datenmodell

Wie die Abbildung 4 veranschaulicht, lassen sich bei der digitalen Archivierung fünf Entitäten identifizieren, die für sich und in Relation zueinander das PREMIS-Datenmodell definieren (PREMIS 2011a).

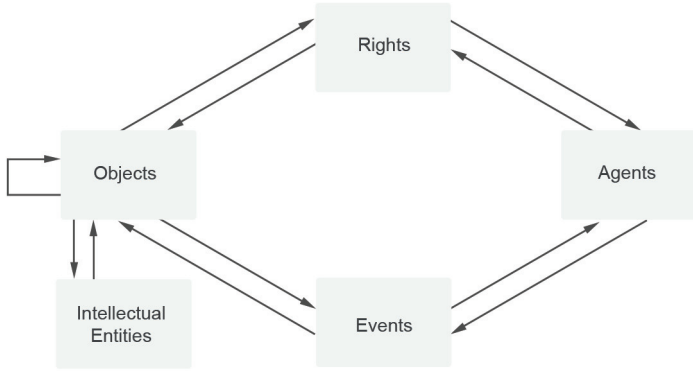


Abbildung 4: Das PREMIS-Datenmodell (Quelle: PREMIS 2011a: 5)

Eine aus zusammenhängenden Inhalten bestehende Einheit wird in diesem Modell als Intellectual Entity bezeichnet. Da beispielsweise eine Zeitschrift in der Regel aus mehreren Artikeln besteht, wird deutlich, dass eine Intellectual Entity durchaus mehrere weitere Entitäten dieses Typs enthalten kann. Auch Webseiten bestehen aus mehreren Unterseiten und können Bilder, Tondokumente und Videomaterial enthalten.

Object Entities sind diskrete Informationseinheiten in digitaler Form. Sie stehen im Zentrum der für die Planung und das Management relevanten Erhaltungsprozesse. Einer Object Entity sind die Typen file, bitstream und representation untergeordnet.

Unter einem *file* bzw. einer Datei verbirgt sich eine benannte geordnete Sequenz von Bytes, die sich typischerweise auf einem Datenträger befindet und vom Betriebssystem als solche erkannt wird. Ein bestimmter adressierbarer Teil einer Datei wird als bitstream bezeichnet. Ohne zusätzliche zum Beispiel formatspezifische Informationen ist dieser Datenstrom nicht interpretierbar. Während die Bedeutung einer Datei nach PREMIS mit der Vorstellung einer gewöhnlichen Computerdatei nahezu übereinstimmt, kann sich ein bitstream nach der Definition von PREMIS nicht über mehrere Dateien erstrecken.

Schließlich stellt der Typ representation die sichtbare Manifestation einer Intellectual Entity dar. Damit die Darstellung eines archivierten Objektes als logisch-funktionale Einheit wahrgenommen werden kann, müssen Informationen beispielsweise zur Strukturbeschreibung im besonderen Maße berücksichtigt werden. Das Beispiel einer Website zeigt hierbei, dass die Repräsentation häufig aus mehreren Dateien gebildet wird.

Event Entities stellen Ereignisse und Aktionen dar, bei denen mindestens eine Entität vom Typ Object oder Agent beteiligt ist. Die Dokumentation von Ereignissen und daraus resultierenden Veränderungen der Entitäten eines Archivsystems ist ein wesentlicher Bestandteil der bereits eingeführten Provenance Information, die für das Vertrauen in die Authentizität eines archivierten Objektes nahezu unerlässlich ist.

Die handelnden Akteure in Form von Personen, Organisationen, aber auch Softwareagenten werden nach PREMIS als *Agent Entity* bezeichnet. Sie nehmen unter der Berücksichtigung entsprechender Rechte (Rights) maßgeblich Einfluss auf den Lebenslauf der Object Entities.

Rights Entities beherbergen schließlich sämtliche Informationen über Rechte und Befugnisse, die Object und Agent Entities betreffen und die vor dem Ausführen von Aktionen (Events) geprüft werden müssen.

Das PREMIS-Data Dictionary

Überwiegend wird PREMIS mit dem durch die PREMIS Managing Agency entwickelten und mittlerweile in der Version 2.1 verfügbaren Data Dictionary assoziiert. Die in der Library of Congress angesiedelte PREMIS-Abteilung entwickelte das Data Dictionary als eine Übersetzung des eingangs bereits erwähnten Framework in implementierbare, sog. semantische Einheiten. Es enthält darüber hinaus Empfehlungen für die Erzeugung, das Management und die Nutzung von Metadaten im Kontext der Langzeitarchivierung (PREMIS 2011b).

Das Data Dictionary verfolgt das Ziel, nur die für die digitale Konservierung wirklich nötigen Metadaten einzubeziehen, und schließt dabei beispielsweise einen Großteil deskriptiver Metadaten aus, die vornehmlich der Suche durch den Nutzer dienen. Für diesen Zweck wird vielmehr auf etablierte Standards wie zum Beispiel Dublin Core verwiesen. Nach Abbildung 5 beschränkt sich das Data Dictionary auf Metadaten, die als kleinster gemeinsamer Nenner aus verschiedenen Metadatentypen gewonnen werden können. Damit entspricht dieses Konzept ebenfalls den oben in diesem Beitrag angestellten Überlegungen, Langzeitarchivierungsmetadaten als Querschnittsmetadatenklasse zu verstehen.

Hauptgegenstand des Data Dictionary sind die bereits erwähnten semantischen Einheiten, die genau die Elemente beschreiben, die ein Archivsystem in der Lage sein sollte, zu speichern und ggf. auch zu exportieren. Bis auf die Intellecutal Entity werden somit die Attribute aller Entitäten des PREMIS-Datenmodells beschrieben. Semantische Einheiten können auch selbst Container für andere semantische Einheiten sein. Die semantische Einheit, die Informationen zur Abspielumgebung beinhaltet, ist beispielsweise Bestandteil der semantischen Einheit environment. PREMIS führte hier ebenfalls die zur Object Entity gehörende semantische Einheit significant properties ein, die in der Community mittlerweile breit diskutiert wird. Die hier abgelegten Informationen dienen der Beschreibung der zu erhaltenen „signifikanten" Eigenschaften und Charakteristika eines digitalen Objektes und sollten bereits vor dem eigentlichen Archivierungsprozess unter Berücksichtigung der Anforderungen aus der Designated Community (vgl. OAIS 2009: 1-11) festgelegt werden. Aus den nahezu 100 semantischen Einheiten sei beispielhaft auf die Einheit size eingegangen, die im Data Dictionary als Teileinheit von objectCharacteristics beschrieben wird.

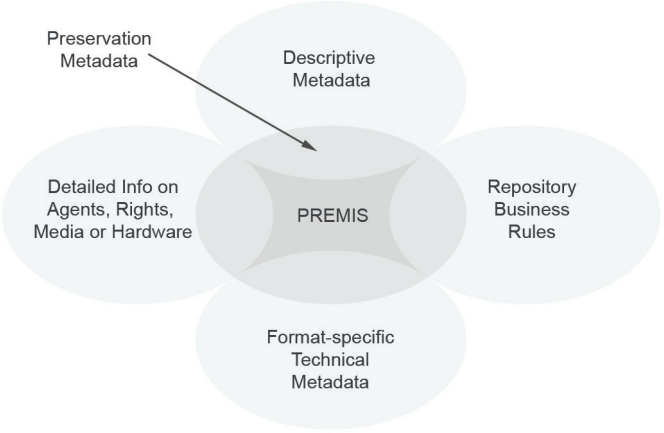


Abbildung 5: PREMIS als Teilmenge verschiedener Metadatentypen (Quelle: Caplan 2009: 5)

Die tabellarische Abbildung zeigt die Beschreibung des technischen Elements *„size"* zur Angabe der in Bytes gemessenen Speichergröße einer archivierten Datei oder eines Datenstroms. Die Begründung für die Verwendung als Bestandteil von Langzeitarchivierungs-Metadaten liefert unter anderem die Zeile *„Rationale"*. Die Speichergröße ist somit ein Indikator, um zu prüfen, ob ein Objekt auch vollständig vom Datenträger gelesen wurde, und unterstützt des Weiteren die Abspielumgebung bei der Allokation ausreichend dimensionierten Arbeitsspeichers.

Semantic Unit	**1.5.3 size**		
Semantic Components	*None*		
Definition	*The size in bytes of the file or bitstream stored in the repository.*		
Rationale	*Size is useful for ensuring the correct number of bytes from storage have been retrieved and that an application has enough room to move or process files. It might also be used when billing for storage.*		
Data constraint	*Integer*		
Object category	Representation	File	Bitstream
Applicability	*Not applicable*	*Applicable*	*Aplicable*
Examples		*2038937*	
Repeatability		*Not repeatable*	*Not repeatable*
Obligation		*Optional*	*Optional*
Creation / Maintenance notes	*Automatically obtained by the repository.*		
Usage notes	*Defining this semantic unit as size in bytes makes it unnecessary to record a unit of measurement. However, for the purpose of data exchange the unit of measurement should be stated or understood by both partners.*		

Die semantische Einheit „size" (Quelle: Caplan 2009: 13)

Generierung von Metadaten

Im Folgenden wird die einleitendende Frage aufgegriffen, wie sich Metadaten gewinnen lassen. Im Fokus soll hierbei die automatische Generierung solcher Informationen stehen. Intellektuell erschlossene bibliografische Metadaten werden somit ausgenommen, auch wenn es hierzu aktuell einige Projekte wie zum Beispiel das PETRUS-Projekt gibt, die sich dem Ziel verschrieben haben, auch diese Informationen automatisiert zu generieren (PETRUS 2012).

Um die Vorteile von Metadatenstandards wie Interoperabilität, Austauschbarkeit etc. zu erreichen, sollten die durch Tools generierten Metadaten zumindest maschinenlesbar sein, am Besten in XML kodiert werden und wenn möglich einem anerkannten Metadatenstandard folgen. Die Ablage dieser Daten sollte nach OAIS im Archivpaket (AIP) selbst und für den schnellen Zugriff im Data-Management erfolgen.

Strukturelle Metadaten

Die Entwicklung praktikabler Tools zur automatischen Generierung struktureller Metadaten aus dem Dokumenteninhalt ist Gegenstand aktueller Forschung und hängt im Falle von Digitalisaten auch von der Qualität der OCR-Texterkennung ab. Für PDF-Dateien wurde beispielsweise im Rahmen des europäischen Projekts Substaining Heritage Access through Multivalent ArchiviNg (SHAMAN) ein von Xerox entwickeltes Tool namens Xeproc vorgestellt, das Metadaten wie zum Beispiel den Titel, das Inhaltsverzeichnis und die Gliederung erkennen und als METS-Objekt zur Verfügung stellen kann (XRCE 2010).

Administrative Metadaten

Technische Metadaten als Untermenge administrativer Metadaten sind die Informationen, die sich auf Dateiebene mittlerweile sehr gut automatisiert gewinnen lassen. So zählt man schon lange die Generierung technischer Metadaten als einen fest integrierten Schritt im Ingest-Prozess. Zu nennen ist hierbei vor allem das JSTOR/Harvard Object Validation Environment (JHOVE) Tool, das unter anderem von der Universitätsbibliothek in Harvard als Open-Source Software entwickelt wurde und mittlerweile auch in der Version 2 zur Verfügung steht. JHOVE verarbeitet zwar nicht die gleiche Vielzahl von Dateiformaten wie das Tool Digital Record Object Identification (DROID), jedoch unterstützt es die Generierung von technischen Metadaten und prüft zudem auch auf Wohlgeformtheit und Validität. DROID beherrscht im Vergleich lediglich die Identifikation des Dateiformats und dessen Version (Artefactual 2009). Ein Framework für die Nutzung eines ganzen Tool-Sets wird ebenfalls unter Mitwirkung aus Harvard unter dem Namen File Information Tool Set (FITS) entwickelt. Mit diesem Framework können eine ganze Reihe von Tools angesprochen werden, darunter auch JHOVE, DROID und der NLNZ Metadata Extractor. Die Nutzung eines Tool-Sets verbreitert die Unterstützung von Dateiformaten und mindert das Fehlerrisiko bei der Identifikation und Validierung. Einen deutlichen Mehrwert bietet FITS zudem durch die leicht konfigurierbare Normalisierung der verschiedenen Tool-Outputs in das FITS-Format. Mit dem FITS-Format liegt also ein formatspezifisches Metadatenset vor, welches die verschiedenen technischen Metadatenelemente mehrerer Metadatentools vereinheitlicht und strukturell zu einem Standard zusammenführt (FITS 2011).

Einige der genannten Tools (zum Beispiel JHOVE) erlauben zudem die Erkennung von Dokumentenbeschränkungen wie sie zum Beispiel in PDF-Dateien in Form eines Zugriffsschutzes per Passwort auftauchen können. Werden Rechte-Informationen (zum Beispiel Nutzungslizenzen) definiert im Format abgelegt, wie es zum Beispiel beim ePub-Format möglich ist, so lassen sich diese Rechte-Metadaten leicht auslesen und in ein eigenes Rechte-Management überführen.

Die Objekthistorie (Provenienz), also die Dokumentation aller zum Beispiel durch Formatmigrationen motivierten Veränderungen am Objekt, lässt sich beispielsweise durch die Ablage der jeweils entsprechenden technischen Metadaten dokumentieren. Neben diesen für jede Objektversion generierten technischen Daten, sollten weitere Hintergrundinformationen zur durchgeführten Migration (zum Beispiel eingesetzte Konvertierungstools, Grund der Migration) angegeben werden.

Beispiel

Das folgende gekürzte XML-Dokument zeigt den FITS-Output nach Analyse einer Bilddatei im JPEG-Format. Zu erkennen sind hierbei vier verschiedene Bereiche sowie die Angabe, mit welchem Tool die entsprechende Information ermittelt wurde.

```
<?xml version="1.0" encoding="UTF-8"?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:xsi="http://
www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://hul.harvard.edu/
ois/xml/ns/fits/fits_output http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd"
version="0.5.0" timestamp="01.04.11 17:29">

  <identification>
    <identity format="JPEG File Interchange Format" mimetype="image/jpeg">
      <tool toolname="Jhove" toolversion="1.5" />
       ..
      <version toolname="Jhove" toolversion="1.5">1.01</version>
      <externalIdentifier toolname="Droid" toolversion="3.0"
      type="puid">fmt/43</externalIdentifier>
    </identity>
  </identification>

  <fileinfo>
    <size toolname="Jhove" toolversion="1.5">9071</size>
    <lastmodified toolname="Exiftool" toolversion="7.74" status="SINGLE_RE
    SULT">2011:02:11 15:54:50+01:00</lastmodified>
    <filename toolname="OIS File Information" toolversion="0.1" status="SINGLE_
    RESULT">bauarbeiter.jpg</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1"
    status="SINGLE_RESULT">e63f2ff66703d55ee29aaad329cf4d9b</md5checksum>
    ..
```

```
  </fileinfo>
  <filestatus>
    <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true
    </well-formed>
    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true
    </valid>
  </filestatus>
  <metadata>
    <image>
      <byteOrder toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">big
      endian</byteOrder>
      <compressionScheme toolname="Exiftool" toolversion="1.5" status="SINGLE_
      RESULT">JPEG(old-style)</compressionScheme>
      <imageWidth toolname="Jhove" toolversion="1.5">100</imageWidth>
      <imageHeight toolname="Jhove" toolversion="1.5">100</imageHeight>
      <colorSpace toolname="Jhove" toolversion="1.5" status="SINGLE_
      RESULT">YCbCr</colorSpace>
      ..
    </image>
  </metadata>
</fits>
```

Der Abschnitt identification liefert Informationen zum Dateiformat und dessen Formatversion sowie den MIME-Type und den PRONOM Unique Identifier (PUID).

FileInfo beinhaltet formatunabhängige Informationen wie zum Beispiel die Dateigröße oder eine MD5-Prüfsumme.

Die Bereiche fileStatus und metadata sind formatspezifisch. FileStatus gibt Auskunft darüber, ob die untersuchte Datei gegenüber der Formatspezifikation valide ist. Schlägt dieser Test fehl, werden entsprechende Fehlermeldungen angegeben. Der Bereich metadata enthält letztlich formatspezifische Metadaten, die vom Dateityp abhängig sind. Zu sehen ist, dass das JHOVE- und Exif-Tool die für Bilder typischen Informationen wie Auflösung, Kompressionsschema und Farbraum lieferte.

Anwendungsszenerarien

Nachdem in den vorherigen Abschnitten beschrieben wurde, welche Informationen als Metadaten der Langzeitarchivierung dienlich sein und wie diese möglichst automatisiert erfasst werden können, werden abschließend, wenn auch vereinfacht, drei konkrete Anwendungsszenarien vorgestellt. Die hier getroffenen Überlegungen basieren vornehmlich auf den Erfahrungen, die die Deutsche Nationalbibliothek (DNB) im Rahmen ihrer Beteiligung an Projekten zur Langzeitarchivierung wie zum Beispiel in kopal, DP4lib oder KEEP gewonnen hat (DNB 2012).

Anwendungsszenario Preservation Planning
Sofern die erzeugten Metadaten im Data-Management abgelegt wurden, können sie dem ebenfalls im OAIS-Referenzmodell verankerten Preservation Planning als wertvolle Datenquelle dienen. Das Data-Management kann zum Beispiel folgende Frage beantworten:

> *Wieviele GIF-Dateien der Version 87a wurden in den letzten drei Jahren archiviert?*

Ist die Anzahl beispielsweise stark rückläufig, kann das ein Indikator dafür sein, dass auch Viewer und Migrationswerkzeuge für dieses Format rückläufig sind, da der Bedarf abnimmt. Das Format wird in diesem Fall also zukünftig obsolet und die Notwendigkeit einer GIF-Migration in ein aktuelles Format sollte diskutiert werden.

Anwendungsszenario Migration
Eine Formatmigration sollte immer eine Qualitätssicherung im Anschluss beinhalten. Zumindest stichprobenartig kann beispielsweise geprüft werden, ob die signifikanten Eigenschaften erhalten wurden. Mithilfe der Fähigkeiten zur Generierung technischer Metadaten kann diese Überprüfung für bestimmte signifikante Eigenschaften dann sogar für jedes konvertierte Objekt – also nicht nur stichprobenartig – erfolgen. Angenommen die Auflösung eines Bildes wird als signifikante Eigenschaft definiert, die im Rahmen aller Maßnahmen zur Langzeitarchivierung stets erhalten werden muss: Wird aus diesem Grund bei der Qualitätssicherung jedes konvertierte Objekt mithilfe von Metadatentools im Hinblick auf die Eigenschaften imageWidth und imageHeight analysiert, so kann der Vergleich der Analyseergebnisse mit dem Original zum Erhalt der signifikanten Eigenschaft Auflösung beitragen. Denn sollten beide Werte nicht mit denen des ursprünglichen Objektes übereinstimmen, kann die Softwareumgebung zur Konvertierung oder dessen Konfiguration entsprechend angepasst werden.

Anwendungsszenario Risiko-Management
Die Generierung technischer Metadaten insbesondere die Erkennung von Dokumentenbeschränkungen und Formatvalidität bereits während des Ingest-Prozesses – also noch vor der eigentlichen Archivierung – kann ebenso als Baustein des Risiko-Managements betrachtet werden. Problembehaftete Objekte werden somit frühzeitig erkannt und können ggf. selbstständig repariert oder dem Produzenten zur Reparatur oder Entfernung von eventuell vorhandenen Sicherheitsbarrieren gemeldet werden. In der Praxis zeigt sich häufig, dass Produzent oder Autor einer Publikation bei einer späteren Erkennung dieser Probleme nur noch schwer erreichbar oder auffindbar sind. Das Objekt wäre dann zwar im Langzeitarchiv

gespeichert, der Nutzungserhalt kann damit jedoch nicht garantiert werden bzw. die Nutzung ist bereits aktuell nicht gegeben, weil beispielsweise ein Passwortschutz den Zugriff auf das Dokument verhindert. Die Deutsche Nationalbibliothek verfolgt diesen proaktiven Ansatz durch die Integration entsprechender Analyse-Tools wie dem oben erwähnten FITS-Tool in ihren Geschäftsgang für Netzpublikationen.

Schlusswort

Die Langzeitarchivierung digitaler Objekte wird trotz vergangener und aktueller Bemühungen auch zukünftig ein wichtiges Thema bleiben. Obwohl die Entwicklungen von standardisierten Metadaten, Erhaltungsstrategien und Softwaresysteme das Ziel haben, den jeweils aktuellen Anforderungen an die digitale Langzeitarchivierung gerecht zu werden, ist der technologische Wandel als „Gegenspieler" der Bemühungen zur Langzeitarchivierung nicht aufzuhalten. Die Bedürfnisse der Konsumenten und die Industrie sind Motor dieses Wandels. Der aktuelle Boom von Tablet-PCs und E-Book-Readern ist nur ein Beispiel für die Vielzahl von Geräten und Vertriebskonzepten und lässt die Fragen zum Thema Kompatibilität und Digital Rights Management (DRM) aktueller denn je erscheinen. Digitale Langzeitarchivierung ist somit nie abgeschlossen. Sie ist vielmehr eine dauerhafte Aufgabe, die eine große Herausforderung darstellt.

Die Betrachtung von Metadatenstandards führt unweigerlich auch zu der Anforderung, Metadaten in standardisierter Form in offene Dateiformate einzubetten. Allein die Funktionalität, Metadaten einbinden zu können, sichert jedoch nicht die Lieferung qualitativer Informationen. Hier muss ggf. das Bewusstsein beim Produzenten geschaffen werden, die für das Einpflegen von Metadaten nötigen Ressourcen zur Verfügung zu stellen.

Das Zusammenspiel der Metadatenstandards verschiedener Metadatenklassen und die Integrationsfähigkeit in etablierte Containerformate (zum Beispiel METS) sind ein wesentliches Kriterium, das bereits bei der Schema-Entwicklung beachtet werden sollte und zunehmend eine bedeutende Rolle für die Akzeptanz eines Standards spielen wird. Dieser Anforderung an übergreifender Interoperabilität widmet sich beispielsweise auch das Kompetenzzentrum Interoperable Metadaten (KIM). Aktuell und in Zukunft sind neben der Unterstützung auf politischer Ebene auch besonders internationale Institutionen wie zum Beispiel die Alliance for Permanent Access (APA) und im deutschsprachigen Raum das Kompetenzzentrum zur digitalen Langzeitarchivierung nestor gefragt, die Entwicklungen von Metadatenstandards weltweit zu beobachten, zu bündeln und – wenn möglich – mit lenkenden Empfehlungen zu unterstützen.

Literatur

Artefactual (Artefactual Systems Inc.) (2009): DROID, JHOVE, NLNZ Metadata Extractor. New Westminster, BC (CA): Artefactual Systems Inc. http://artefactual.com/wiki/index.php?title=DROID,_JHOVE,_NLNZ_Metadata_Extractor [13.04.2012]

Beaubien, R. (2002): METS: An Introduction. Part II. METS Mechanisms with XML. U. C. Berkeley Library Systems Office. http://www.loc.gov/standards/mets/presentations/METSIntro2.ppt [13.04.2012]

Borghoff, U.M./Rödig, P./Scheffczyk, J. und Schmitz, L. (2003): Langzeitarchivierung. Methoden zur Erhaltung digitaler Dokumente. Heidelberg: dpunkt.verlag.

Borghoff, U.M./Rödig, P./Scheffczyk, J. und Schmitz, L. (2006): Long-Term Preservation of Digital Documents. Principles and Practices. Berlin, Heidelberg: Springer Verlag.

Caplan, P. (2009): Understanding PREMIS. Library of Congress Network Development and MARC Standards Office. http://www.loc.gov/standards/premis/understanding-premis.pdf [13.04.2012]

Day, M. (2004): Preservation Metadata Initiatives: Practicality, Substainability, and Interoperability. In: Bischoff, F.M./Hofman, H. and Ross, S. (Eds.): Metadata in Preservation. Marburg: Archivschule Marburg, 91-117.

DNB (Deutsche Nationalbibliothek): Projekte. http://www.dnb.de/DE/Wir/Projekte/projekte_node.html [06.05.2012]

FITS (2011): File Information Tool Set (FITS). http://code.google.com/p/fits/ [13.04.2012]

Frodl, C./Fischer, T./Baker, T. und Rühle, S. (2007): Deutsche Übersetzung des Dublin-Core-Metadaten-Elemente-Sets. Version 1.1. Kompetenznetzwerk Interoperable Metadaten. http://nbn-resolving.de/urn:nbn:de:101:1-200911103125 [13.04.2012]

Gartner, R. (2008): Metadata for digital libraries: state of the art and future directions. Bristol: Joint Information Systems Committee.
http://www.jisc.ac.uk/media/documents/techwatch/tsw_0801pdf.pdf [13.04.2012]

LOC (Library of Congress) (2008): MIX. NISO Metadata for Images in XML Schema. http://www.loc.gov/standards/mix/ [13.04.2012]

LOC (Library of Congress) (2009): METS: An Overview & Tutorial. Library of Congress, http://www.loc.gov/standards/mets/METSOverview.v2.html [13.04.2012]

LOC (Library of Congress) (2012): textMD. Technical Metadata for Text. Official Website. http://www.loc.gov/standards/textMD/ [13.04.2012]

LOCKSS (Lots Of Copies Keep Stuff Safe) (2012). http://www.lockss.org [13.04.2012]
NISO (National Information Standards Organization) (2006): Data Dictionary – Technical Metadata for Digital Still Images. Bethesda, MD: NISO Press. http://www.niso.org/kst/reports/standards?step=2&gid%3Austring3Aiso-8859-1=&project_key%3Austring%3Aiso-8859-1=b897b0cf3e2ee526252d9f830207b3cc9f3b6c2c [13.04.2012]
OAIS (Consultative Committee for Space Data Systems) (2009): Reference Model for an Open Archival Information System (OAIS). Washington, DC: National Aeronautics and Space Administration. http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf [13.04.2012]
PETRUS (Schöning-Walter, C.) (2012): PETRUS – Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. Frankfurt am Main: Deutsche Nationalbibliothek. http://www.dnb.de/DE/Wir/Projekte/Abgeschlossen/petrus.html [13.04.2012]
PREMIS (PREMIS Editorial Committee) (2011a): Introduction and Supporting Materials from PREMIS Data Dictionary for Preservation Metadata. Version 2.1. http://www.loc.gov/standards/premis/v2/premis-report-2-1.pdf [13.04.2012]
PREMIS (PREMIS Editorial Committee) (2011b): Data Dictionary section from PREMIS Data Dictionary for Preservation Metadata. Version 2.1. http://www.loc.gov/standards/premis/v2/premis-dd-2-1.pdf [13.04.2012]
XRCE (Xerox Research Centre Europe) (2010): Xeproc: a model for XML document processing. XEROX. http://www.xrce.xerox.com/Research-Development/Services-Innovation-Laboratory/Document-Structure/Xeproc-C-a-model-for-XML-document-processing [13.04.2012]

Metadaten und die Data Documentation Initiative (DDI)

Wolfgang Zenk-Möltgen

Über DDI

Die Data Documentation Initiative (DDI)[1] ist eine Initiative, einen internationalen Standard zur Beschreibung sozialwissenschaftlicher Daten zu definieren und zu verbreiten. Dieser Standard wird in einem XML-Format (Extensible Markup Language) definiert, das sowohl für Menschen als auch für Maschinen lesbar ist. DDI hat den Anspruch, den gesamten Forschungsdaten-Lebenszyklus zu unterstützen. DDI Metadaten beziehen sich auf die Studienkonzeption, die Datenerhebung, die Datenbearbeitung und -auswertung sowie auf die Sekundärnutzung und Archivierung.

Die Konzeption und Definition der Ziele der Initiative kamen aus der Welt sozialwissenschaftlicher Datenarchive. 1995 wurde DDI als Projekt finanziert, gestartet und organisiert vom ICPSR (Inter-university Consortium for Political and Social Research, USA).[2] 2003 wurde die DDI Alliance gegründet, welche auf der Mitgliedschaft von Institutionen basiert und formalisierte Prozesse zur Weiterentwicklung der Initiative einführte. Die Gründungsmitglieder kamen aus dem Bereich sozialwissenschaftlicher Datenarchive, den Produzenten von Statistikdaten und weiteren, wie zum Beispiel von Forschungsdatenzentren, Datenerhebungsinstitutionen und einigen kommerziellen Organisationen. Heute sind 36 Institutionen aus 14 Ländern in Nordamerika, Europa und Australien Mitglieder in der DDI Alliance, zusätzlich Eurostat und die World Bank Development Data Group als internationale Organisationen (siehe Tabelle 1). Der Bereich der Anwender von DDI ist noch weiter als der der Mitglieder von DDI. Allein die World Bank setzt DDI in über 100 Statistik-Ämtern in 67 Ländern ein. Auf der Website der DDI Alliance wird mithilfe der *„DDI is being used around the world"* die weltweite Verbreitung von DDI gut dargestellt.[3]

University of Alberta, Canada

Australian Bureau of Statistics (ABS)

Australian Data Archive (ADA)

1 http://www.ddialliance.org
2 http://www.icpsr.umich.edu
3 http://www.ddialliance.org/community

University of California, Berkeley - Computer-Assisted Survey Methods Program and UCDATA
Centre for Longitudinal Studies, Institute of Education, University of London (Associate Member)
Centro De Investigaciones Sociologicas (CIS), Spain
Cornell University (CISER)
Danish Data Archive
Data Archiving and Networked Services (DANS), The Netherlands
Eurostat
Finnish Social Science Data Archive
German Institute for International Educational Research (DIPF)
German Socio-Economic Panel Study (SOEP)
GESIS - Leibniz Institute for the Social Sciences
University of Guelph
Institute for Quantitative Social Science (IQSS) at Harvard University
Institute for the Study of Labor (IZA)
Institute for Social and Economic Research (ISER)
Inter-university Consortium for Political and Social Research (ICPSR)
Massachusetts Institute of Technology (MIT)
University of Minnesota, Minnesota Population Center
Norwegian Social Science Data Service (NSD)
Open Data Foundation
Princeton University
Research Data Centre of the German Federal Employment Agency, Institute for Employment Research (IAB)
Roper Center
Stanford University
Statistics New Zealand
Survey Research Operations, University of Michigan
Swedish National Data Service (SND)
Swiss Foundation for Research in Social Sciences (FORS)
United Kingdom Data Archive
University of Toronto Scholars Portal
University of Washington, Center for Studies in Demography & Ecology (CSDE)
U.S. Bureau of Labor Statistics (Associate Member)
World Bank, Development Data Group (DECDG)

Tabelle 1: Mitglieder der DDI Alliance

Die Entwicklung und weitere Verbesserung des DDI Standards führte zu der Veröffentlichung von verschiedenen Versionen. Im Jahr 2000 wurde die DDI Version 1.0 veröffentlicht, in welcher einfache Umfragen dokumentiert werden konnten und zum Beispiel nur Mikrodaten, aber keine Aggregatdaten. Im Jahr 2003 erschienen die Versionen 2.0 und 2.1 als Erweiterung von DDI, in denen nun auch Aggregatdaten und weitere Datentypen dokumentiert werden können sowie eine Unterstützung für geographische Elemente möglich ist. Diese Version von DDI wird auch gegenwärtig noch verwendet und gepflegt und ist unter dem Namen *„DDI-Codebook"* bekannt, da sich die Dokumentation sehr stark an der Struktur eines Codebuchs für einen Datensatz orientiert.

Im Jahr 2008 erschien mit DDI 3.0 die erste *„DDI-Lifecycle"* Version, in der eine Erfassung des Daten-Lebenszyklus im Gegensatz zum Codebuch-zentrierten Modell im Mittelpunkt steht. Hier wurde der Blick auf die Erzeugung der Metadaten und ihre Wiederverwendung in den verschiedenen Stadien des Forschungsdaten-Lebenszyklus gerichtet. Zusätzlich wurde das Konzept eingeführt, dass die Metadaten-Elemente *"machine-actionable"* sein sollten, also so strukturiert, dass ein programmierter Zugriff auf sie möglich ist. Eine Erweiterung der Fragebogendokumentation wurde zur Unterstützung der Verwendung von CAI-Instrumenten (Computer Aided Interview) eingeführt. Weitere Neuerungen betrafen die Unterstützung für Datenreihen (Längsschnitt-Umfragen, Panel Studien, etc.), die Möglichkeit zum Vergleich von Metadaten (*"by design"* und *"ex-post"* möglich) und eine verbesserte Unterstützung zur Beschreibung komplexer Datensätze.

2009 wurden mit der Version DDI 3.1 etliche Fehlerkorrekturen durchgeführt, darunter auch eine neue URN-Struktur, um dauerhafte Identifikatoren aller "identifiable"-Elemente von DDI zu erhalten (siehe unten). Die Version 3.1 ist gegenwärtig die aktuelle Version von DDI-Lifecycle.

Als Update zur DDI-Codebook Variante wurde in 2012 die Version DDI 2.5 veröffentlicht. Sie soll eine Erleichterung der Migration von DDI-Codebook nach DDI-Lifecycle ermöglichen, indem zum Beispiel alle notwendigen Elemente (Pflichtelemente) von DDI-Lifecycle mit einem Gegenstück in DDI 2.5 eingeführt wurden. Dies erleichtert vor allem auch die parallele Verwendung von DDI-Codebook und DDI-Lifecycle, da die Pflichtelemente von DDI-Lifecycle bei einer Konvertierung in das DDI-Codebook Format nicht verloren gehen.

Für 2012 ist eine Veröffentlichung von DDI 3.2 angekündigt. Darin soll zum Beispiel ein Element *"DataItem"* neu eingeführt werden, das eine Wiederverwendung erlaubt, und es sollen einige Konsistenz-Fragen gelöst werden, etwa bei *"RecordRelationship"* oder bei der Verwendung von Missing Values. Weiterhin wird die URN-Struktur überarbeitet, damit ein verteilter Resolving-Mechanismus für DDI-URNs möglich wird. Zusätzlich wird die Verwendung kontrollierter Vokabulare verbessert werden.

Grundlegende DDI Metadaten

Die DDI-Codebook Version enthält Metadaten zu den vier Bereichen: Dokumentbeschreibung, Studienbeschreibung, Variablenbeschreibung und Dateibeschreibung. Die wichtigsten Elemente innerhalb dieser Bereiche sind für die Dokumentbeschreibung der Titel, die Autoren und die Beschreibung der Publikation des DDI Dokuments selbst. Für die Studienbeschreibung sind vor allem wichtig die Inhalte der Studie, Titel, Autoren und Institutionen, zeitliche und geographische Angaben zur Studie, verwendete Methoden, Grundgesamtheit und Stichprobenbeschreibung sowie Literaturhinweise. Der Bereich Variablenbeschreibung enthält als Hauptelemente die Namen, Typen und Labels zu den Variablen, die Fragen und Antwortmöglichkeiten der Interviews, verwendete Codes und ihre Häufigkeiten im Datensatz, Interviewer-Anweisungen und Filterinformationen aus dem Fragebogen und Hinweise zur Codierung oder Berechnung. Die Dateibeschreibung schließlich enthält Angaben zur Anzahl der Variablen und Fälle und Namen, Formate und Versionen der Datendateien.

Die DDI-Lifecycle Version folgt hier einem anderen Konzept, nämlich einer unabhängigen Dokumentation einzelner Stadien des Lebenszyklus und damit der Möglichkeit ihrer Wiederbenutzung. Im Folgenden werden daher die Prinzipien der Strukturierung von DDI-Lifecycle näher erläutert: der Lebenszyklus, DDI Module, Elemente, die als Maintainables, Versionables und Identifiables klassifiziert werden können, die Einführung von Schemes (pflegbare Listen). Desweiteren wird auf die Beziehungen zu anderen Standards eingegangen und es wird die Verwendung kontrollierter Vokabulare in DDI 3 genauer erläutert.

Forschungsdaten-Lebenszyklus

Die DDI Alliance hat einen Lebenszyklus für Forschungsdaten definiert, in dem die verschiedenen Phasen als Struktur für die verwendeten Module in DDI 3 dienen können (siehe Abbildung 1).

Als Phasen wurden dabei acht verschiedene identifiziert, die jedoch nicht in linearer Reihenfolge durchlaufen werden müssen. Mit der Phase Concept beginnt eine Studienkonzeption, indem Forschungsfrage und Methodik der Untersuchung festgelegt werden. Anschließend wird in der Phase Collection die Datenerhebung durchgeführt und in Phase Processing die Datenaufbereitung geleistet. Von hier ab kann entweder zur Phase Archiving weitergegangen werden, in der eine Sicherung der Forschungsdaten geleistet wird, oder aber auch direkt zur nachfolgenden Phase Distribution, welche den Datenvertrieb leistet. Von hier folgt eine Phase Discovery, die die Daten auffindbar macht, und die Phase Analysis, in der die Forschungsdaten ausgewertet werden. Hier schließt sich dann noch eine Phase Repurposing an, die eine anderweitige Verwendung

der Daten für Sekundärnutzung umfasst und dann wieder zur Phase Processing zurückführt. Dieses allgemeine Modell der Phasen des Forschungsdaten-Lebenszyklus diente als Grundlage für die Entwicklung der Module. Die Reihenfolge der Verwendung der einzelnen Phasen ist dabei aber komplett offen und nicht in der DDI Spezifikation vorgegeben. Daher sind alle anderen Pfade durch die Phasen denkbar und können in DDI auch so dokumentiert werden.

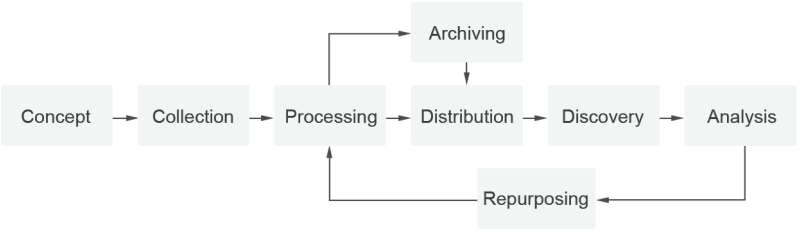


Abbildung 1: Der Forschungsdaten-Lebenszyklus der DDI Alliance

Module

Module in DDI-Lifecycle sind Gruppen von zusammengehörigen Dokumentationselementen. Manche Module beziehen sich auf das Lebenszyklusmodell, andere sind aus technischen Gründen gruppiert. Zusammengehörige Elemente sollten in einem Modul zu finden sein, was aber nicht immer möglich war. Dies hat jedoch keine Auswirkungen auf die Verwendung der Elemente, da die Gruppierung in Modulen allein der internen Strukturierung der DDI Spezifikation dient. Folgende Module sind in DDI 3.1 definiert:

DDI 3.1 Module
Archive module
Comparative module
Conceptual components module
Data collection module
Dataset module
Dublin Core Elements module
DDI profile module
Grouping module
Instance module
Logical product module
Physical data product module *(plus inline n-cube, normal n-cube, tabular n-cube module and proprietary module)*

Physical instance module
Reusable module
Study unit module

Tabelle 2: DDI 3.1 Module

Benutzung der DDI 3 Module

Für einige Module soll hier gezeigt werden, welche Metadaten-Elemente sie enthalten und wie sie verwendet werden (siehe Abbildung 2). Zunächst gibt es das Modul StudyUnit, in dem grundlegende Metadaten über eine einfache Studie enthalten sind. Hier werden zum Beispiel Informationen abgelegt, die zur Identifizierung dienen, etwa Studiennummer, Persistent Identifier oder Zitationsinformationen. Desweiteren sind hier die räumliche und zeitliche Einordnung der Studie und die abgedeckten Themen dokumentiert. Grundlegende Konzepte, die abzubildende Grundgesamtheit, eine inhaltliche Zusammenfassung und Informationen über den Zweck der Studie sowie über Forschungsanträge und ihre Finanzierung sind hier ebenfalls Gegenstand der Dokumentation.

Study Unit	**Data Collection**
Identification Coverage • *Topical* • *Temporal* • *Spatial* Conceptual Components • *Universe* • *Concept* • *Representation (optional replication)* Purpose, Abstract, Proposal, Funding	Methodology Question Scheme • *Question* • *Reponse domain* Instrument • *using Control Construct Scheme* Coding Instructions • *question to raw data* • *raw data to public file* Interviewer Instructions

Logical Product	**Archive**
Category Schemes Coding Schemes Variables NCubes Variable and NCube Groups Data Relationships	Organization or individual which has control over the metdata Livecycle events Archive specific information

Physical Data Structure	Physical Instance
Links to Data Relationships	One-to-one relationship with data file
Links to Variable or NCube Coordinate	Coverage constraints
Description of phsycial storage structure	Variable and category statistics
• *in-line, fixed, delimited or proprietary*	

Abbildung 2: Verwendung der DDI Module

Das nächste Modul Data Collection enthält Informationen zur angewandten Datenerhebungsmethode, dem Instrument – etwa dem Fragebogen oder auch anderen Messinstrumenten – mit den zugehörigen Fragen und Antwortdomänen sowie ihrer Abfolge im Fragebogen. Zusätzlich sind Interviewerанweisungen aus dem Fragebogen und Codierungsanweisungen für die Rohdaten und auch für die letztlich publizierten Datensätze Teil der Metadaten in diesem Modul.

Im Modul Logical Product werden die Metadaten zur Struktur der erhobenen Daten abgelegt: Hier sind die Listen der Antwortkategorien und der verwendeten numerischen Codes und die daraus entstehenden Variablen des Datensatzes dokumentiert. Dazu gehören auch sog. NCubes, das sind aggregierte Daten von Variablen mit mehreren (N) Dimensionen oder generell n-dimensionale Datenstrukturen. Variablen und NCubes können in Gruppen zusammengefasst und dokumentiert werden. Beziehungen zwischen ihnen können ebenfalls beschrieben werden.

Im Modul Physical Data Structure werden die physikalischen Eigenschaften der verwendeten Datenstrukturen dokumentiert, etwa ein festes, ein variables oder ein Trennzeichen-Format. Die Verbindung zu den Variablen aus dem Logical Product erfolgt über die Data Relationships. Die eigentliche Datendatei wird dann im Modul Physical Instance beschrieben. Dort besteht eine Eins-zu-eins-Relation mit einer Datei, die die Umfragedaten enthält, etwa einer SPSS- oder STATA-Datei. In diesem Modul können auch Tabellen mit statistischen Ergebnissen zu den Variablen abgelegt werden.

Das Modul Archive ist in einem sehr weiten Sinne zu verstehen, zum Beispiel sind in diesem Modul alle Informationen über beteiligte Personen und Institutionen zu finden. Dazu gehören auch die Lifecycle Events, mithilfe derer alle Prozesse aus den verschiedenen Lebenszyklus-Stadien erfasst werden können. Daneben gibt es hier auch Informationen zur Archivierung der Studie und zu den zugehörigen Katalog-Metadaten.

Neben den hier gezeigten grundlegenden Modulen zur Beschreibung von Forschungsdaten gibt es noch weitere Möglichkeiten in anderen Modulen, etwa zur Dokumentation von Konzepten im Conceptual Components Modul, zum Vergleich verschiedener Elemente im Modul Comparative oder zur Vererbung von Dokumentation durch die Benutzung des Group Moduls. Wichtig ist auch das

Element ResourcePackage aus dem Modul Group: Es dient zur Dokumentation wiederverwendbarer Elemente unabhängig von ihrem Einsatz in einer Studie, zum Beispiel für Fragen, Antwortskalen oder Variablen.

Die Elemente aus allen diesen Modulen können vielfältig miteinander vernetzt werden, indem Referenzen auf Elemente benutzt werden, die an anderer Stelle dokumentiert sind. So kann eine maximale Wiederverwendung der Dokumentationsteile in den verschiedenen Stadien des Forschungsdaten-Lebenszyklus erreicht werden.

Maintainables, Versionables und Identifiables

Die Elemente in DDI 3 können folgendermaßen klassifiziert werden: Zunächst gibt es einfache Elemente, die Metadaten für ein Objekt enthalten oder eine Referenz auf ein anderes Element (siehe Abbildung 3). Auf der nächsten Stufe gibt es die sog. Identifiables, welche zusätzlich über eine ID verfügen. Mithilfe der ID können diese Elemente eindeutig identifiziert werden, so dass auf diese Elemente eine Referenz gesetzt werden kann. Dabei gibt es zwei technische Möglichkeiten, diese ID festzulegen, entweder über ein ID-Attribut oder über eine URN (Uniform Resource Name), die den speziellen Vorgaben der DDI Spezifikation folgt (siehe unten).

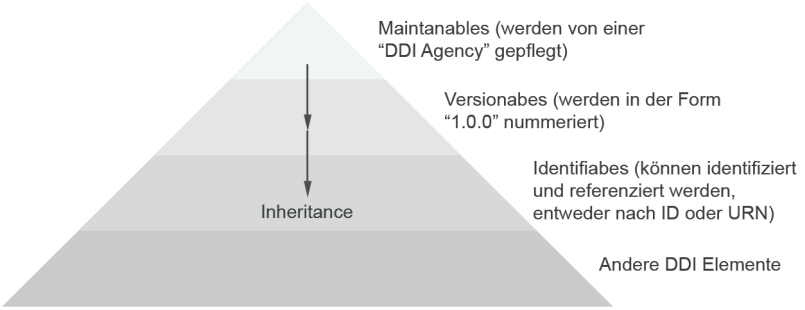


Abbildung 3: Hierarchie der Elemente

Eine weitere Gruppe von Elementen verfügt zusätzlich zur ID über eine Versionsnummer und gehört damit zu den Versionables. Diese Elemente können in verschiedenen Versionen vorliegen, die mithilfe einer dreistelligen Versionsnummer gekennzeichnet wird. Referenzen auf diese Elemente müssen auch die Versionsangabe enthalten.

Die Gruppe der Maintainables schließlich verfügt zusätzlich zur ID und der Version über das Attribut Agency zur Kennzeichnung eines Anbieters, der die Informationen des Elements pflegt. Institutionen können einen Agency-Namen bei der DDI Alliance beantragen und können mit der Verwendung dieses Namens für DDI Instanzen ihre Verantwortlichkeit für die dort enthaltenen Metadaten erklären.

Das Konzept der Schemes

Im DDI Standard stehen Schemes für Listen von Elementen eines gleichen Typs. Um die Verwaltung dieser Elemente zu vereinfachen, können sie zu Schemes zusammengefasst werden. Schemes sind in der Regel Maintainables und können daher von einer DDI Agency gepflegt und von ihr selbst oder anderen per Referenz wiederverwendet werden. Beispiele für Schemes sind das Organization Scheme im Modul Archive, das Question Scheme, Control Construct Scheme und Interviewer Instruction Scheme im Data Collection Modul, das Concept Scheme, Universe Scheme, Geographic Structure Scheme und Geographic Location Scheme im Modul Conceptual Components, das Category Scheme, Code Scheme, Variable Scheme und NCube Scheme im Modul Logical Product, das Physical Structure Scheme und Record Layout Scheme im Physical Data Product Modul.

Beziehungen von DDI zu anderen Standards

In die Entwicklung des DDI Standards sind eine ganze Reihe von Erfahrungen aus anderen Standards mit eingeflossen. Darüber hinaus können einige andere Standards auch direkt in DDI eingebunden werden. Dazu gehört zum Beispiel der Dublin Core Standard[4] zur Dokumentation grundlegender bibliographischer Zitationsinformationen und zur Dokumentation von Sammlungen und vorliegenden Formaten. Die Definition von Dublin Core wurde so eingebunden, dass diese Elemente an bestimmten Stellen direkt innerhalb von DDI verwendet werden können: Dazu dient das DCElements-Tag, das in jedem Citation-Element verwendet werden kann.

Von grundlegendem Einfluss für die Entwicklung von DDI war das OAIS Referenzmodell für Archive. Viele Elemente, die in OAIS genannt sind, finden sich in der DDI Spezifikation wieder. Daneben wurden aber auch Elemente aus METS mit einbezogen, die eine Beschreibung zum Management digitaler Objekte auf einer oberen Ebene darstellen, und aus PREMIS, einem Standard mit spezifischen Strukturen für die digitale Langzeitsicherung.

Die grundlegenden Konzepte für die Metadaten zu geographischen Informationen wurden entlang des ISO-Standards 19115 Geography (FGDC) entwickelt, der Elemente wie Shape, Boundary oder Map Image Dateien und ihre Attribute enthält. Die Struktur der Beschreibung von Konzepten lehnt sich am ISO/IEC-Standard 11179 an. Dieser sieht eine Repräsentation von Metadaten in einer Registratur (Metadata Registry) vor, die auch eine Hierarchie von Konzepten und eine detaillierte Beschreibung der Konzepte enthalten kann.

Für die Modellierung des DDI Standards in XML-Schema wurde auf Erfahrungen mit dem Standard SDMX zurückgegriffen. Dieser wurde für den Austausch und die Dokumentation von statistischen Aggregatdaten, etwa Zeitreihen oder

4 http://dublincore.org/documents/dcmi-terms/

Indikatoren entwickelt und wird bereits sehr verbreitet durch statistische Ämter und andere Statistikproduzenten eingesetzt. Die Verwendung von DDI und SDMX kann durch die ähnliche Ausrichtung und die inhaltliche Nähe gut komplementär erfolgen, was auch durch eine gemeinsame Arbeitsgruppe der Initiativen noch verbessert werden soll.

Verwendung von kontrollierten Vokabularen

Kontrollierte Vokabulare erlauben es, die Einträge von Metadaten-Elementen auf eine Liste von erlaubten Werten einzuschränken. Dadurch wird eine höhere Standardisierung erreicht, als es mit den Einträgen von Texten zur Beschreibung möglich ist. So kann zum Beispiel die Angabe einer Klassifikation mithilfe eines kontrollierten Vokabulars erfolgen. In DDI-Lifecycle ist daher die Verwendung von kontrollierten Vokabularen möglich und wird empfohlen.

Wegen der Fülle der Möglichkeiten der Anwendung von kontrollierten Vokabularen und der Breite der Themen, auf die sie sich erstrecken können, hat man sich entschieden, die kontrollierten Vokabulare nicht als Teil des DDI Standards zu formulieren, sondern als eine Empfehlung. Eine Arbeitsgruppe der DDI Alliance (CVG) hat bereits Empfehlungen zu einer Reihe von kontrollierten Vokabularen veröffentlicht. Darunter sind zum Beispiel Empfehlungen zu den Elementen LifeCycleEvent, Commonality, TimeMethod, ResponseUnit, SoftwarePackage, CharacterSet und AnalysisUnit. Weitere Elemente, an denen gearbeitet wird, sind IntendedFrequency, ModeOfDataCollection, AggregationMethods, DataType, CategoryStatistic, DateCalendar, ContributorRole, PublisherRole und KindOfData.

Am Beispiel von TimeMethod kann man sehen, dass folgende Werte dort enthalten sein können: Longitudinal (Cohort or Trend), Panel (Continuous or Interval), TimeSeries (Continuous or Discrete), CrossSectional, CrossSectionalAdHocFollowUp und Other. Hiermit werden also die Typen des Studiendesigns über Zeit festgelegt, so dass leichter Suchen durchgeführt werden können oder Gruppen von Studien gebildet werden können.

Diese kontrollierten Vokabulare werden im Format Genericode[5] veröffentlicht, eine Spezifikation von OASIS[6] zur Dokumentation und Versionierung von Codelisten. Manche Vokabulare sind auch als Hierarchien angelegt, so dass sie weitere oder engere Begriffe vorsehen. Die kontrollierten Vokabulare der DDI Alliance werden von der CVG gepflegt, so dass neuere Versionen auch auf der Website der DDI Alliance[7] zu finden sein werden.

5 http://genericode.org

6 https://www.oasis-open.org

7 http://www.ddialliance.org/controlled-vocabularies

Identifizieren von Elementen in DDI 3

Alle Identifiables in DDI 3 verfügen über ein ID-Attribut, das dazu dient, Referenzen auf diese Elemente erzeugen zu können. Mithilfe dieser Referenzen können Elemente wiederbenutzt werden, wobei sie nur einmal in DDI dokumentiert werden müssen.

Es gibt zwei technische Möglichkeiten, die ID festzulegen, entweder über das ID-Attribut oder über ein URN-Attribut (Uniform Resource Name), das den speziellen Vorgaben der DDI Spezifikation folgt. Beide Varianten sind logisch gleich und enthalten die gleichen Angaben. Lediglich die Syntax ist unterschiedlich, aber beide Varianten können ineinander überführt werden. So sprechen allein technische Gründe für die Verwendung der einen oder anderen Variante in einer konkreten Implementierung.

Bei Verwendung des ID-Attributs muss ein eindeutiger Identifier in der Form id="A1234" angegeben werden. Die Angabe zur Agency wird aus dem übergeordneten Maintainable, die Angabe zur Version wird aus dem übergeordneten Versionable geerbt. Mit diesen drei Angaben kann dann eine eindeutige Referenz an anderer Stelle auf dieses DDI Element erzeugt werden.

Bei Verwendung des URN-Attributs werden die Angaben zum Identifier, der Agency und der Version in einen String kombiniert, die den Vorgaben der URN-Spezifikation folgt. Dies wird zum Beispiel für eine Variable V100 im VariableScheme ZA1234_VarSch der GESIS für Version 1.0.0 in der Form urn="urn:ddi:de.gesis:VariableScheme.ZA1234_VarSch.1.0.0:Variable.V100.1.0.0" notiert. Diese Variante der Notation wird empfohlen, beide Varianten sind erlaubt.

Für eine effektive Wiederbenutzung von DDI Elementen wird ein Resolver Service benötigt, der diese Identifier so auflöst, dass die Lokation des DDI Elements gefunden werden kann und seine Eigenschaften dann ermittelt werden können. Die DDI Alliance arbeitet zurzeit an der Implementierung eines solchen Services, die auf der Verwendung des DNS-Systems (Domain Name System) beruht.

DDI im GESIS Datenarchiv

Das GESIS Datenarchiv nutzt heute DDI-Codebook für den Workflow bei der Daten- und Metadatenbearbeitung und für die Langzeitarchivierung der zu archivierenden Studien. Da das Datenarchiv in vielen Projekten auch bereits bei der Datenerhebung, bei der Datenaufbereitung und -pflege, und auch generell bei der Datendistribution und -analyse tätig ist, beschränkt sich die Verwendung von DDI nicht auf die Lebenszyklus-Phase der Archivierung im engeren Sinne, sondern die DDI Metadaten werden auch in der Unterstützung von Dokumentation, langfristiger Sicherung, Recherche und Datenservice für Sekundärnutzer und für

die DOI Registrierung über das Angebot da|ra[8] verwendet. DDI-Lifecycle wird gegenwärtig nur für spezielle Anwendungen, wie etwa die Unterstützung von Enhanced Publications (Verbindung von Publikationen zu den dabei benutzten Daten) oder im Projekt STARDAT verwendet, das eine Integration der Archiv-Tools auf der Basis von DDI-Lifecycle beinhaltet. Eine Migration aller Anwendungen und Metadatenbestände nach DDI-Lifecycle wird zurzeit geplant, dabei sind allerdings noch eine Reihe von Hindernissen zu überwinden, etwa eine Einführung einer effektiven Versionskontrolle für einzelne Metadaten-Elemente.

Workflow

Die Abläufe im GESIS Datenarchiv für die Archivierung, Datendokumentation, -bearbeitung, Langzeitsicherung und Distribution werden zum großen Teil durch DDI-Codebook Metadaten unterstützt (siehe Abbildung 4). Die Studienbeschreibungen im Datenbestandskatalog (DBK)[9] werden im webbasierten Programm DBKEdit in einer DDI kompatiblen relationalen Datenbank erstellt und gepflegt. Sie werden nicht nur in DBKSearch publiziert, sondern dienen auch zur Anbindung an die Datenregistrierung da|ra zur Vergabe von persistenten Identifikatoren (DOI Namen) und an das Nesstar-basierte ZACAT-Angebot[10] zur Analyse, Recherche und zum Download von archivierten Studien.

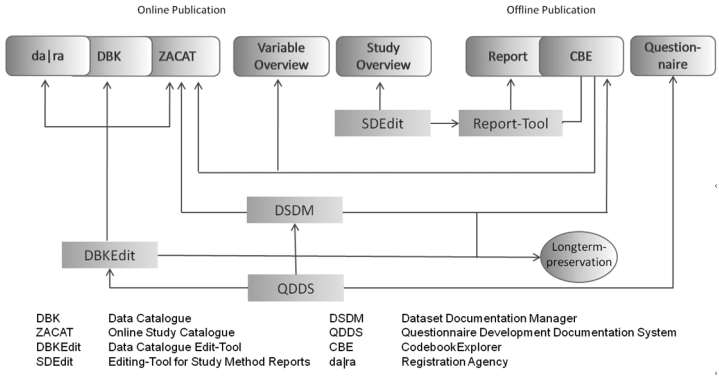


Abbildung 4: Workflow im GESIS Datenarchiv

Die Dokumentation der Studien auf der Variablenebene erfolgt mithilfe der Software Dataset Documentation Manager (DSDM) oder kann auch in der Fragebogensoftware Questionnaire Development and Documentation System (QDDS) erfolgen. DSDM exportiert die Dokumentation in DDI-Codebook Format zur Verwendung in ZACAT oder im CodebookExplorer (CBE), einer stand-alone Software

8 http://www.gesis.org/dara/
9 http://www.gesis.org/dbk/
10 http://zacat.gesis.org

zur Recherche und Analyse von Datenkollektionen. Hierdurch wiederum werden verschiedene Ausgabeformate unterstützt, die alle auf derselben Dokumentationsbasis beruhen: Ein DDI-Codebook Export unterstützt das Nesstar-System, eine CBE-Datenbank kann für einen Variable-Overview benutzt werden, welcher komplexe Datenkollektionen im Web darstellt und durchsuchbar macht, und Variable Reports können als Print- oder Online-Ausgabe für eine vollständige Datensatzdokumentation erzeugt werden.

Die Ausgabe eines speziellen Langzeitsicherungsformats wird sowohl von DBKEdit als auch von DSDM unterstützt. Dieses Format ist vollständig DDI-Codebook kompatibel und wegen seiner Lesbarkeit von Mensch und Maschinen gut als Format für eine langfristige Sicherung geeignet.

Das Beispiel Datenbestandskatalog

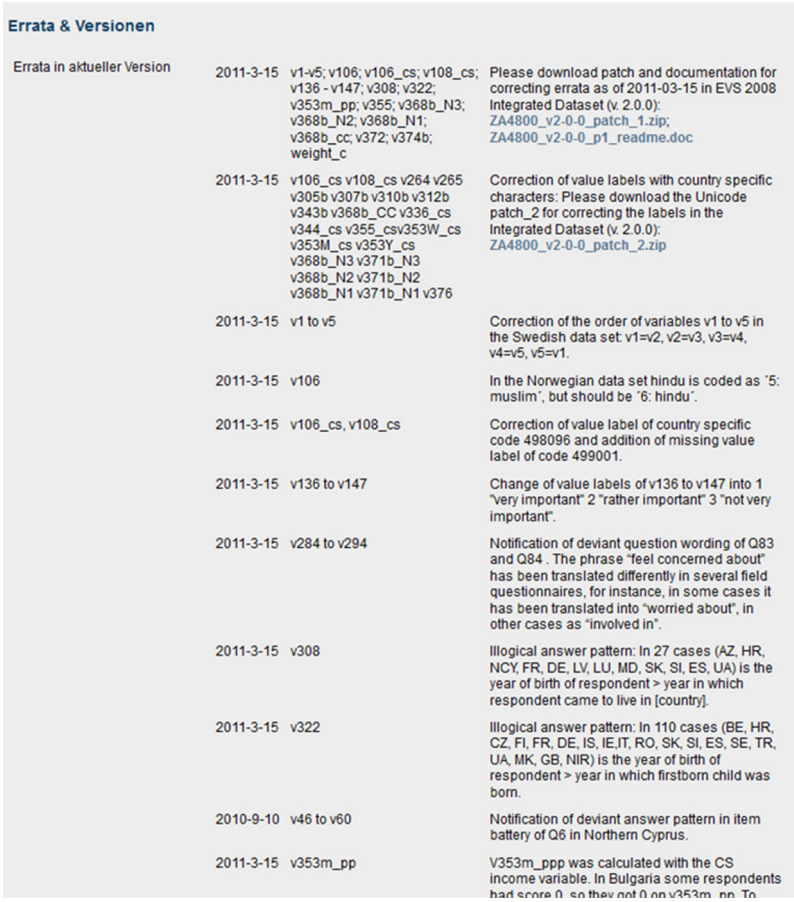
Das GESIS Datenarchiv für Sozialwissenschaften, im Jahr 1960 als Zentralarchiv für empirische Sozialforschung der Universität Köln gegründet, hat bereits seit langer Zeit unter anderem durch die Entwicklung eines standardisierten Studienbeschreibungsschemas an einer Vereinheitlichung von Metadaten gearbeitet. Zusammen mit den anderen Archiven des Verbundes CESSDA (Council of European Social Science Data Archives)[11] wurden diese Bemühungen ständig intensiviert und flossen in die Mitarbeit von vielen CESSDA-Archiven bei der DDI Alliance ein. Die standardisierten Studienbeschreibungen wurden und werden als Datenbestandskatalog (DBK) regelmäßig in gedruckter Form oder online veröffentlicht. Die Anwendung DBKEdit dient seit 2006 als relationales Datenbanksystem zur Verwaltung und Pflege der Studienbeschreibungen, die den Nutzern mithilfe der Anwendung DBKSearch zur Verfügung gestellt werden. DBKEdit leistet dabei auch die Metadaten-Produktion für die Publikation in verschiedenen Retrieval- und Distributionsplattformen (siehe oben). Seit kurzem sind diese Anwendungen als DBKfree[12] unter einer Open Source Lizenz auch für weitere Anwender verfügbar.

Im April 2010 wurde eine Versionshistorie der Datensätze als einheitliches System für alle archivierten Daten mithilfe des DBK eingeführt. Diese enthält eine eindeutige Versionsnummer, eine detaillierte Dokumentation von Errata und der Korrektur-History der Datensätze. Die Errata zur aktuellen Version werden mit Datum, einer Fehlerbeschreibung und einer Beschreibung, welche Variablen betroffen sind, versehen (siehe Abbildung 5). Die Versionsnummern werden DDI-Lifecycle konform mit Version 1.0.0 begonnen und erhöhen sich bei Major-, Minor- oder Revision-Nummer je nach Änderung im Datensatz. Dies führt zu höherer Transparenz im Laufe der Datenbearbeitung. Zusätzlich wird

11 http://www.cessda.org

12 http://info1.gesis.org/dbkfree/

auf diese Art ein einfaches Zitieren der Daten ermöglicht, da eine Version immer mit einem DOI Namen als persistenter Identifikator über das da|ra-System versehen wird. Die Zitation von Datensätzen wird den Nutzern im DBK vorgeschlagen und enthält die genaue Version zur Erleichterung von Replikationsanalysen. Die Studienbeschreibungen zur aktuellen Version des Datensatzes können in DDI-Codebook oder DDI-Lifecycle exportiert werden. Ein Export der Dokumentation der kompletten Versionshistorie im DDI-Format steht allerdings noch aus.



Errata & Versionen

Errata in aktueller Version	2011-3-15	v1-v5; v106; v106_cs; v108_cs; v136 - v147; v308; v322; v353m_pp; v355; v368b_N3; v368b_N2; v368b_N1; v368b_cc; v372; v374b; weight_c	Please download patch and documentation for correcting errata as of 2011-03-15 in EVS 2008 Integrated Dataset (v. 2.0.0): ZA4800_v2-0-0_patch_1.zip; ZA4800_v2-0-0_p1_readme.doc
	2011-3-15	v106_cs v108_cs v264 v265 v305b v307b v310b v312b v343b v368b_CC v336_cs v344_cs v355_csv353W_cs v353M_cs v353Y_cs v368b_N3 v371b_N3 v368b_N2 v371b_N2 v368b_N1 v371b_N1 v376	Correction of value labels with country specific characters: Please download the Unicode patch_2 for correcting the labels in the Integrated Dataset (v. 2.0.0): ZA4800_v2-0-0_patch_2.zip
	2011-3-15	v1 to v5	Correction of the order of variables v1 to v5 in the Swedish data set: v1=v2, v2=v3, v3=v4, v4=v5, v5=v1.
	2011-3-15	v106	In the Norwegian data set hindu is coded as ´5: muslim´, but should be ´6: hindu´.
	2011-3-15	v106_cs, v108_cs	Correction of value label of country specific code 498096 and addition of missing value label of code 499001.
	2011-3-15	v136 to v147	Change of value labels of v136 to v147 into 1 "very important" 2 "rather important" 3 "not very important".
	2011-3-15	v284 to v294	Notification of deviant question wording of Q83 and Q84 . The phrase "feel concerned about" has been translated differently in several field questionnaires, for instance, in some cases it has been translated into "worried about", in other cases as "involved in".
	2011-3-15	v308	Illogical answer pattern: In 27 cases (AZ, HR, NCY, FR, DE, LV, LU, MD, SK, SI, ES, UA) is the year of birth of respondent > year in which respondent came to live in [country].
	2011-3-15	v322	Illogical answer pattern: In 110 cases (BE, HR, CZ, FI, FR, DE, IS, IE,IT, RO, SK, SI, ES, SE, TR, UA, MK, GB, NIR) is the year of birth of respondent > year in which firstborn child was born.
	2010-9-10	v46 to v60	Notification of deviant answer pattern in item battery of Q6 in Northern Cyprus.
	2011-3-15	v353m_pp	V353m_ppp was calculated with the CS income variable. In Bulgaria some respondents had score 0, so they got 0 on v353m_pp. To

Abbildung 5: Versionshistorie im DBK

Eine weitere Standardisierung der Angaben wurde durch die Einführung einer kontrollierten Liste für Untersuchungsgebiet nach ISO3166-1 und ISO3166-2 für Nationen und sub-nationale Einheiten erreicht. Ebenfalls standardisiert wurden die Erhebungszeiträume im Format TT.MM.JJJJ kompatibel zu ISO 8601 und

der Möglichkeit Zeiträume (von-bis) anzugeben. Eine Standardisierung der bisher im Freitext erfassten Literaturangaben zu den Studien wird zurzeit durchgeführt.

Der GESIS Datenbestandskatalog enthält ebenfalls Links zum Datenzugang. Seit Anfang 2012 können viele Studien der Zugangsklasse A (frei für wissenschaftliche Verwendung) direkt im DBK über einen Download erreicht werden. Alle weiteren Studien können über ein Warenkorb-System bestellt werden. Dazu ist für Nutzer lediglich eine kostenfreie Registrierung mit Angabe des Forschungsprojekts nötig.

Standardisierung

Eine standardisierte Dokumentation, wie sie mit DDI möglich ist, erlaubt den einfachen Austausch von Metadaten und Daten zwischen den Akteuren im Forschungsdaten-Lebenszyklus. Sie führt zu einer einfacheren Übernahme in neue Systeme und Anwendungen, so dass die Wiederverwendung der Dokumentation oder einzelner Teile möglich wird. Die Standardisierung führt auch zu klareren Bedeutungen einzelner Teile der Dokumentation. Daher ist die Standardisierung für die langfristige Sicherung von Forschungsdaten und ihre Nachnutzung unerlässlich. Die Etablierung eines Standards kann natürlich nur in der Community erreicht werden, die die verwendeten Dokumentationen benutzt. Eine Standardisierung erfordert daher zunächst einen höheren Aufwand bei der Dokumentation in allen Phasen des Lebenszyklus. Darüber hinaus ist eine Standardisierung ein dauernder Prozess, da sich durch Weiterentwicklungen neue Anforderungen ergeben. Der Mehrwert durch eine Standardisierung wiegt diese Anstrengungen aber mehr als auf.

Literatur

Bauske, F. (1992): Europäische Informationsbasis über Datensätze in CESSDA-Archiven. ZA-Information 31, 109-111.

Bauske, F. (2000): Das Studienbeschreibungsschema des Zentralarchivs. ZA-Information 47, 73-80.

Blank, G. and Rasmussen, K. B. (2004): The Data Documentation Initiative. The Value and Significance of a Worldwide Standard. Social Science Computer Review 22 (3), 307-318.

Consultative Committee for Space Data Systems (2002): Reference model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1 Blue Book, Januar 2002.

Gregory, A./Heus, P. and Ryssevik, J. (2010): Metadata. In: German Data Forum (RatSWD) (Ed.): Building on Progress. Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences. Opladen & Farmington Hills, 487-508.

Hausstein, B. und Zenk-Möltgen, W. (2011): da|ra – Ein Service der GESIS für die Zitation sozialwissenschaftlicher Daten. In: Schomburg, S./Leggewie, C./Lobin, H. und Puschmann, C. (Hrsg.): Digitale Wissenschaft: Stand und Entwicklung digital vernetzter Forschung in Deutschland. Beiträge der Tagung vom 20./21. September 2010. Köln, 139-147.

Jääskeläinen, T./Moschner, M. and Wackerow, J. (2009): Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability. IASSIST Quarterly 33, 34-39.

Jensen, U./Katsanidou, A. und Zenk-Möltgen, W. (2011): Metadaten und Standards. In: Büttner, S./Hobohm, H. und Müller, L. (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock u. Herchen, 83-100.

Kramer, S./Oechtering, A. und Wackerow, J. (2009): Data Documentation Initiative (DDI): Entwicklung eines Metadatenstandards für Forschungsdaten in den Sozialwissenschaften. KIM-Technology Watch Report, September 2009.

Mochmann, E. (1979): Bericht über die IASSIST Konferenz in Ottawa. ZA-Information 4, 24-27.

Vardigan, M./Heus, P. and Thomas, W. (2008): Data Documentation Initiative: Toward a Standard for the Social Sciences. The International Journal of Digital Curation 3 (1), 107-113.

Zenk-Möltgen, W. und Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. Köln: GESIS Technical Reports 2012/01.

N.N. (1990): Neuauflage des Datenbestandskatalogs des Zentralarchivs. ZA-Information 27, 5-15.

Persistent Identifier: Versionierung, Adressierung und Referenzierung

Nicole von der Hude

Persistent Identifier

Die enorme Zunahme von digitalen Objekten, die für die Wissensgesellschaft langfristig von Bedeutung sind, hat es notwendig gemacht, neue Lösungen für eine zuverlässige und dauerhafte Versionierung, Adressierung und Referenzierung zu finden.

Da diese Anforderungen durch standortgebundene Verweise, meist in Form von URLs (Uniform Resource Locator), nicht erfüllt sind, müssen standortunabhängige Identifizierungs- und Adressierungsmechanismen angewendet werden, die eingebettet in eine internationale Struktur vertrauenswürdiger Institutionen die erforderliche Nachhaltigkeit gewährleisten.

Ein persistentes Identifizierungssystem muss somit ein digitales Objekt dauerhaft unabhängig vom Ort der Speicherung adressieren, gleichzeitig auf mehrere Speicherorte verweisen und ein digitales Objekt als Informationseinheit weltweit eindeutig identifizieren.

Gleichzeitig muss ein System bereitgestellt werden, das den dauerhaften Zugriff auf das digitale Objekt über Systemgrenzen und Systemwechsel hinaus sicherstellt: Mithilfe eines Resolving-Mechanismus kann auf eine Ressource auch dann noch zugegriffen werden, wenn sich ihr physikalischer Speicherort verändert hat. Es entsteht eine vorteilhafte Redundanz durch Umleitung des Zugriffs, wenn einzelne Speicheradressen nicht mehr gültig sind, 404-Fehlermeldungen (not found) werden vermieden.

Für den Wissenschaftler und allgemein für alle Nutzer bedeutet die Verwendung eines Persistent Identifiers eine Unterstützung der Praxis des wissenschaftlichen Arbeitens und macht die Zitierbarkeit von digitalen Objekten erst möglich. Ein Persistent Identifier gewährleistet nicht nur die eindeutige Identifizierbarkeit auf Dauer, er stellt auch sicher, dass inhaltlich unterschiedliche Versionen verschiedene individuelle Identifikatoren erhalten. Voraussetzung für den dauerhaften Zugriff ist jedoch die Langzeitarchivierung in einem vertrauenswürdi-

gen Archiv[1]: Langzeitarchivierung ohne persistente Identifikatoren ist möglich, persistente Identifikatoren ohne Langzeitarchivierung sind jedoch nicht sinnvoll.

In der Praxis existieren verschiedene Persistent Identifier Systeme nebeneinander. Die Gründe für die Wahl eines bestimmten Persistent Identifier Systems hängen unter anderem davon ab, welche Art von digitalem Objekt einen Identifier erhalten soll, in welchem Land und von welcher Institution der Identifier vergeben wird und ob ein kostenpflichtiges oder kostenfreies System präferiert wird. Die zurzeit am weitesten verbreiteten Persistent Identifier Systeme sind Handle, DOI, ARC und URN. Die Reihenfolge stellt dabei keine Wertung dar und auf die nähere Beschreibung der einzelnen Systeme wird hier verzichtet.

Rolle der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek (DNB) sieht sich als verlässlicher Partner für Forschung, Wissenschaft und Kultur und will diese zentrale Rolle auch in der digitalen Welt spielen. Mit Inkrafttreten des Gesetzes über die Deutsche Nationalbibliothek (DNBG[2]) vom 22. Juni 2006 (BGBl. I S. 1338) hat die Deutsche Nationalbibliothek auch den Auftrag der Sammlung, Erschließung, Verzeichnung und Archivierung von unkörperlichen Medienwerken (Netzpublikationen) erhalten.

Die Sammelpflicht umfasst sowohl Internetpublikationen mit Entsprechung zum Print-Bereich als auch web-spezifische Medienwerke. Beispiele für Netzpublikationen sind E-Books, elektronische Zeitschriften, Hochschulprüfungsarbeiten, Musikdateien, Digitalisate oder auch Webseiten.

Die Deutsche Nationalbibliothek hat eine Strategie entwickelt, wie dieser Auftrag umfassend erfüllt werden kann, und ein Gefüge aus verschiedenen Komponenten etabliert, die das Vertrauen der Wissensgesellschaft in die zuverlässige Zitierbarkeit und Identifizierung digitaler Objekte stärkt.

Basis für die optimale Erfüllung des erweiterten Sammelauftrags ist die Bereitstellung eines Archivservers zur Ablieferung von Netzpublikationen mit entsprechend automatisierten Geschäftsgängen. Die abgelieferten Netzpublikationen werden dauerhaft archiviert, wobei die Deutsche Nationalbibliothek ständig an der Entwicklung von innovativen Verfahren zur Langzeitarchivierung[3] federführend beteiligt ist. Unter Nutzung von Metadaten wird die automatische

1 Vgl. dazu: nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (Hrsg.) (2008): nestor-Kriterienkatalog vertrauenswürdige digitale Langzeitarchive Version II. nestor-materialien 8. Frankfurt am Main: nestor c/o Deutsche Nationalbibliothek - 40 S. URN: urn:nbn:de:0008-2008021802 oder auch Data Archiving and Networked Services (DANS) (Ed.): Data Seal of Approval – Quality guidelines for digital research data in the Netherlands. http://www.datasealofapproval.org/

2 Vgl. http://www.gesetze-im-internet.de/dnbg/index.html

3 Die Zahl der laufenden und abgeschlossenen Projekte, die der Langzeitarchivierung zuzuordnen sind, ist groß, beispielhaft genannt werden hier: nestor, DP4lib, LuKII, SHAMAN, KEEP, kopal.

Erschließung gewährleistet, die Verzeichnung erfolgt analog zu den gedruckten Publikationen in der Deutschen Nationalbibliografie[4]. Der Einsatz von Persistent Identifiern dient nicht nur der Bereitstellung für die Benutzung, sondern garantiert die langfristige Zitierfähigkeit von Netzpublikationen in einem vertrauenswürdigen Archiv.

Persistent Identifier Strategie der Deutschen Nationalbibliothek

Im Rahmen ihrer Persistent Identifier Strategie hat sich die Deutsche Nationalbibliothek für die systematische Kennzeichnung aller Netzpublikationen, die in den Sammelauf,trag fallen, mit einem URN (Uniform Resource Name[5]) entschieden. Die Entscheidung wurde zugunsten der URN getroffen, weil es sich um ein flexibles, standardisiertes Persistent Identifier System handelt, das von der IETF (Internet Engineering Task Force), einer nicht kommerziellen, offenen Internet-Organisation, verwaltet wird und das speziell auf Nationalbibliotheken zugeschnitten ist und somit beste Voraussetzungen für die optimale Erfüllung des gesetzlichen Sammelauftrags der Deutschen Nationalbibliothek bietet.

Persistenz ist jedoch keine Eigenschaft des URN an sich, sondern erfordert eine Infrastruktur bestehend aus:

- einem Resolvingdienst, der eine verlässliche Service-Infrastruktur zur Administration und Auflösung der Identifier bietet,
- der Langzeitarchivierung, die die langfristige Verfügbarkeit der Ressourcen garantiert,
- einer Policy, die Regeln und Konventionen für die kooperative Nutzung festlegt.

Alle Netzpublikationen erhalten einen URN entweder von der verlegenden Stelle oder von der Deutschen Nationalbibliothek. Zusätzlich fungiert die Deutsche Nationalbibliothek als Zentrale für die Vergabe von Unternamensräumen.

Aufgrund der Erfahrungen, die die Deutsche Nationalbibliothek in den letzten Jahren gesammelt hat, zeichnen sich erste Änderungen in der Vergabepraxis ab: In der Vergangenheit, als es galt, die Idee eines Persistent Identifier Systems einer breiten Öffentlichkeit bekannt zu machen, erhielt jeder Interessierte einen eigenen Unternamensraum. Nachdem jedoch mit zurzeit ca. 6,5 Mio URNs und

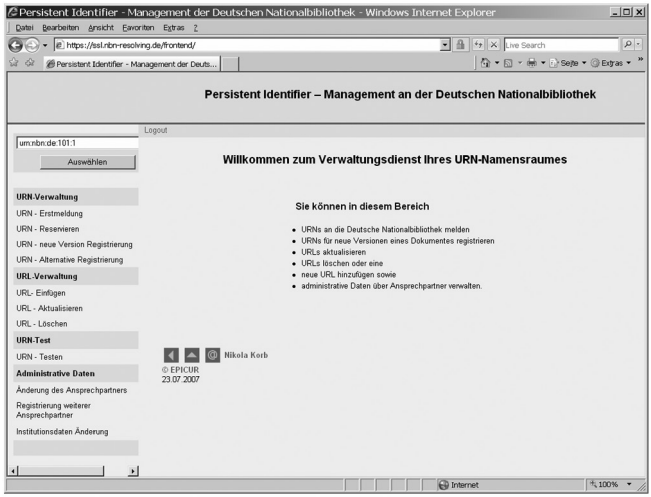
4 Unter http://dnb.dnb.de kann jedermann über öffentliche Netze kostenfrei in der Datenbank der Deutschen Nationalbibliothek recherchieren.

5 Vgl. http://de.wikipedia.org/wiki/Uniform_Resource_Name und auch Neuroth, H./Oßwald, A./Scheffel, R./Strathmann, S. und Jehn, M. (Hrsg.) (2009): nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.0. Kapitel 9.4.1 „Der Uniform Resource Name (URN)“. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20090811490

ca. 420 aktiven Unternamensrauminhabern diese breite Basis besteht, werden zukünftig die Aspekte Vertrauenswürdigkeit, Transparenz und Verbesserung der Servicequalität stärker priorisiert. Dies bedeutet, dass nur noch solche Ablieferer einen eigenen Unternamensraum verwalten können, die die Archivierung in einem vertrauenswürdigen Archiv, die ständige Pflege der URNs und die geeignete personelle Betreuung dauerhaft garantieren können. Dies wird auch in der Policy entsprechend nachvollzogen.

Unternamensrauminhaber, die diese Voraussetzungen erfüllen, legen einen Namensraum fest und lassen diesen bei der Deutschen Nationalbibliothek registrieren. Für ihre eigenen Dokumente vergeben sie URNs und veröffentlichen diese. Spätestens 24 Stunden nach einer Veröffentlichung muss der neue URN im Resolver der Deutschen Nationalbibliothek registriert sein. Besondere Bedeutung hat die Pflege der URLs, die bei jeder Änderung sofort im Resolver aktualisiert werden müssen.

Für die Unternamensrauminhaber steht eine entsprechende Oberfläche zur Administration über die Webseiten der Deutschen Nationalbibliothek zur Verfügung:



Die Auflösung der URNs erfolgt mit einem Resolver[6], der ständig den neuen Anforderungen angepasst wird. Aktuell erfüllt dieser die Aufgabe eines Metaresolvers, der nicht nur die deutschen, österreichischen und schweizerischen URNs auflöst, sondern auch die Weiterleitung zu den URN:NBN-Unternamensräumen

6 Vgl. http://nbn-resolving.org

von Tschechien, Finnland, Ungarn, den Niederlanden, Norwegen und Schweden sowie zu den anderen Persistent Identifier Systemen DOI®, Handle und ARK realisiert.

Beteiligung der Deutschen Nationalbibliothek an europäischen Projekten

Die Deutsche Nationalbibliothek sieht sich als Teil einer heterogenen europäischen und internationalen Persistent Identifier Landschaft, in der nationale Konzepte vielfach nicht interoperabel sind und keine ausreichende Standardisierung existiert. Hier besteht die Notwendigkeit, an der Entwicklung von Standards und einer gemeinsamen Policy mitzuwirken, um eine größere Transparenz für den Nutzer und damit schließlich das notwendige Vertrauen in die Verlässlichkeit von Persistent Identifiern für die Wissenschaft zu stärken.

In diesem Zusammenhang steht die Beteiligung an verschiedenen Projekten auf europäischer Ebene, bei denen die Deutsche Nationalbibliothek als Projektpartner mitgearbeitet hat, zum Beispiel EuropeanaConnect[7] und PersID[8] (Persistent Identifier).

Aus den Ergebnissen dieser europäischen Projekte kann geschlossen werden, dass jetzt und auch zukünftig verschiedene Persistent Identifier Systeme parallel existieren werden. Daher ist eine Kooperation[9] mit den Systemen DOI, Handle, ARK etc. von großer Bedeutung, ebenso wie die bereits begonnene Etablierung eines Meta-Resolvers für alle Persistent Identifier Systeme als Zwischenstufe auf dem Weg zu einer browserbasierten Lösung. Daneben soll die Arbeit an Standards forciert werden,

7 EU-gefördertes Forschungs- und Entwicklungsprojekt, um das europäische Kulturportal Europeana zu einem interoperablen, multilingualen und benutzerorientierten Dienst für alle europäischen Bürgerinnen und Bürger zu machen. Projektpartner waren 30 Institutionen aus ganz Europa, Nationalbibliotheken, Rundfunkarchive, Forschungsinstitute und Wirtschaftsunternehmen.
Nähere Informationen unter: www.europeanaconnect.eu und http://www.dnb.de/DE/Wir/Projekte/Abgeschlossen/europeanaconnect.html

8 Initiative zur Vereinheitlichung und Vernetzung der Persistent-Identifier-Lösungen in Europa. Ziel war eine Förderung der Anwendung von PI-Lösungen, die Entwicklung länder- und institutionenübergreifender abgestimmter und möglichst einheitlicher PI-Strategien, eine Vernetzung der nationalen Resolversysteme zu einem einfach nutzbaren, transparenten und verlässlichen Dienst und die Weiterentwicklung einschlägiger Standards. Das Projekt entstand auf Initiative von Knowledge Exchange, dem Zusammenschluss europäischer Forschungsförderinstitutionen DFG (D), SURF (NL), DANS (NL) und DEFF (DK), Projektteilnehmer waren Institutionen aus den Niederlanden, Italien, Dänemark, Finnland, Schweden und Deutschland.
Nähere Informationen unter: http://www.persid.org und http://www.dnb.de/DE/Wir/Projekte/Abgeschlossen/persId.html

9 Vgl. Ergebnis des Knowledge-Exchange-Workshops „Den Haag Manifesto: Five steps to bringing Persistent Identifiers and Linked Open Data together". 14./15. Juni 2011. Den Haag. http://www.knowledge-exchange.info/Default.aspx?ID=462

so zum Beispiel eine Überarbeitung von RFCs[10] innerhalb der IETF und des DIN-Normentwurfes für vertrauenswürdige Persistent Identifier Systeme, um unterschiedliche Entwicklungen bei den verschiedenen Anwendern zu vermeiden. Angestrebt wird darüber hinaus eine gemeinsame Policy mindestens auf europäischer Ebene.

Aktuelle Entwicklungen des URN-Service in der Deutschen Nationalbibliothek

Aktuell wird in der Deutschen Nationalbibliothek die vorhandene Policy rund um den URN-Service grundlegend überarbeitet, nachdem ausführliche Diskussionen sowohl intern als auch mit Partnerinstitutionen stattgefunden haben. Die neue Policy wird festschreiben, was genau identifiziert wird (welche Art von Dokument, Teile eines Dokuments), welche Voraussetzungen ein Unternamensrauminhaber erfüllen muss und wie die Verantwortlichkeiten verteilt werden. Ein weiterer wichtiger Punkt ist die Festlegung auf Qualitätsstandards.

Auch in technischer Hinsicht entwickelt sich der URN-Service kontinuierlich weiter, zum Beispiel wird eine REST-Schnittstelle[11] für die Pflege von URN-Einträgen implementiert, um qualifizierte Rückmeldungen über den Erfolg oder Misserfolg bei der Registrierung von URNs zu ermöglichen. Außerdem wird der Resolver um eine Linked-Data-Schnittstelle erweitert, um so den geänderten Benutzeranforderungen Rechnung zu tragen.

Auf europäischer Ebene werden die in den abgeschlossenen Projekten benannten Ziele und Aufgaben weiter verfolgt, zum Beispiel in dem neuen Projekt *„URN-Cluster“* mit der schwedischen Nationalbibliothek.

Im URN:NBN-Cluster sollen alle bislang unabhängig voneinander betriebenen nationalen Dienste miteinander verbunden werden. Die Daten werden gegenseitig gespiegelt, statt mehrerer nationaler Resolver wird ein zentraler internationaler Resolver aufgebaut, der URNs aller Ländernamensräume auflösen kann. Dieser Resolving-Dienst wird eine gemeinsam genutzte Software haben, die auf der Grundlage des Resolvers der Deutschen Nationalbibiothek entwickelt wird und redundant bei allen Partnern gleichzeitig parallel betrieben wird. Durch diese Redundanz wird eine 24/7 Verfügbarkeit erreicht.

Auf der technischen Seite wird dieser *„URN-Cluster“* von der Deutschen Nationalbibliothek betreut, deren Resolver über die notwendigen Voraussetzungen verfügt, an die individuellen Bedürfnisse der Partner angepasst zu werden. Am

10 Die Requests for Comments (kurz RFC; zu deutsch Bitte um Kommentare) sind eine Reihe von technischen und organisatorischen Dokumenten des RFC-Editors zum Internet. RFCs behalten auch dann ihren Namen, wenn sie sich durch allgemeine Akzeptanz und Gebrauch zum Standard entwickelt haben. Nähere Informationen unter: http://de.wikipedia.org/wiki/Request_for_Comments

11 Representational State Transfer (REST) bezeichnet ein Programmierparadigma für Webanwendungen

„URN-Cluster“ können sich aber auch solche Partner beteiligen, die über ein leistungsfähiges System verfügen und dies zunächst nicht ändern möchten, allerdings an einer Verbesserung der Verfügbarkeit interessiert sind. Die administrative Seite soll bei den Partnern verbleiben, die sich auf eine gemeinsame Policy einigen, aber grundsätzlich ihre individuelle Entscheidungsfreiheit behalten.

Dieses Projekt ist ein erster Schritt in Richtung des Aufbaus und der Inbetriebnahme eines URN:NBN-Clusters mit mehreren europäischen Partnern, wobei eine spätere Anbindung weiterer Partner an den Cluster mit möglichst geringen Aufwänden verbunden sein soll. Nach Inbetriebnahme des Clusters wird die Deutsche Nationalbibliothek vor allem in beratender Funktion tätig werden. Eine Rolle als technischer Dienstleister ist, je nach Kooperationsvertrag bzw. Geschäftsmodell, aber durchaus denkbar.

Fazit

Digitale Objekte sind bereits fester Bestandteil der Wissensgesellschaft und müssen zuverlässig und dauerhaft die Möglichkeit der Versionierung, Adressierung und Referenzierung bieten. Um dies zu erreichen, ist der Einsatz von Persistent Identifier Systemen unerlässlich, die untrennbar mit einer Langzeitarchivierung in vertrauenswürdigen Archiven verbunden sein müssen.

Es sind weitere Anstrengungen auf diesem Gebiet nötig, um das Vertrauen in die dauerhafte Zitierfähigkeit digitaler Objekte zu stärken und eine größere Transparenz zu schaffen. In der globalen Welt ist dies nur mithilfe von europäischen und internationalen Partnern möglich und erfordert einen ständigen Meinungsaustausch, um auf geänderte Anforderungen reagieren zu können.

Es handelt sich also um ein System, das ständig überdacht, geprüft und modifiziert wird. Auch wenn sich die Deutsche Nationalbibliothek zum jetzigen Zeitpunkt dafür entschieden hat, nur statische Dokumente mit einem URN zu versehen und somit einen Teil der Informationslandschaft unberücksichtigt zu lassen, bedeutet dies nicht, dass die Relevanz dieser Informationen als gering angesehen wird. Vielmehr entsteht daraus die Notwendigkeit einer verteilten nationalen Strategie, die es erforderlich macht, mit wissenschaftlichen Datenzentren, wissenschaftlichen Verlagen und wissenschaftlichen Institutionen im Dialog zu bleiben.

Die Vergabe von DOI-Namen für Sozial- und Wirtschaftsdaten: Serviceleistungen der Registrierungsagentur da|ra

Brigitte Hausstein

Hintergrund

Nicht nur der Zugang zu Literatur, sondern auch allgemein zugängliche und langfristig verfügbare Forschungsdaten sind für eine Wissenschaft, die exzellente Forschungsergebnisse erbringen will, essentiell. Die schnelle Entwicklung der digitalen Technologien und Netzwerke der letzten Jahre hat die Produktion, Verbreitung und Nutzung von Forschungsdaten jedoch radikal verändert. Neue Verfahren und Messinstrumente bringen wachsende, aber auch komplexere Datenmengen und -typen hervor. Entsprechend entwickelte Software-Tools helfen die Fülle der gesammelten Primärdaten zu verwalten, zu interpretieren und in Informations- und Wissenssammlungen zu transformieren. Das wichtigste und allseits präsente Forschungswerkzeug, das Internet, hat die Art und Weise, wie Daten und Informationen ausgetauscht und verfügbar gemacht werden, stark verändert (vgl. Uhlir 2003).

Während für Forschungspublikationen neben den traditionellen Angeboten der freie Zugang (Open Access) immer mehr zur gängigen Praxis wird, sind die Bemühungen hinsichtlich allgemein zugänglicher Datenpublikationen erst am Anfang. Obwohl grundsätzlich die Bereitschaft zur Weitergabe der Primärdaten existiert, scheitert dies oft an den fehlenden Kapazitäten, die für die Aufbereitung und Metadatenbeschreibung notwendig sind. Dies gilt auch für die Sozialwissenschaften, die im Vergleich zu anderen Disziplinen bereits eine ausgeprägte Kultur des *„Data Sharings"* kennen.

Die Verbreitung der Forschungsergebnisse erfolgt fast ausschließlich noch über Publikationen in Fachzeitschriften. Die *„Allianz der deutschen Wissenschaftsorganisationen"* hat jedoch Ende Juni 2010 in den *„Grundsätze(n) zum Umgang mit Forschungsdaten"* eine Regelung für Primärdaten gefordert, um bei Wissenschaftlerinnen und Wissenschaftlern das Bewusstsein für den Handlungsbedarf und für den Nutzen von Primärdaten-Infrastrukturen zu schärfen. Von Seiten der Forschungsfinanzierer wird zunehmend gefordert, nicht nur die Forschungspublikationen, sondern auch die entstandenen Primärdaten im Sinne von Good Scientific Practice öffentlich zugänglich zu machen. Daraus ergibt sich die besondere Bedeutung einer reinen Datenpublikation, mit allen Möglichkeiten der eindeutigen Identifikation und kompakten Zitierung, die für Textpublikationen bereits Standard sind.

Darüber hinaus zeigt sich in allen Wissenschaftsbereichen auch die Dringlichkeit der Kopplung von Datenarchivierung und wissenschaftlicher Literaturpublikation. Die Trennung von Forschungspublikation und zugrundeliegenden Daten erschwert die Evaluation der Publikation und schränkt die Nachvollziehbarkeit der dargestellten Ergebnisse ein. Bislang erfolgt die Verbindung von Daten- und Forschungspublikation nur punktuell. Voraussetzung für die Verbindung von Forschungsprimärdaten und wissenschaftlicher Publikation ist jedoch neben der Langzeitarchivierung der Daten und entsprechender Qualitätssicherung, die Möglichkeit zur Publikation von Daten mit eindeutiger Identifizierbarkeit und Referenzierbarkeit.

Persistent Identifier und Forschungsdatenregistrierung

Ein Weg zur Lösung der geschilderten Problematik ist der Einsatz von speziellen Persistent Identifiern. Deren Funktion entspricht in etwa einer ISBN-Nummer bei gedruckten Werken, die lediglich ein einziges Mal vergeben wird. Hinzu kommt die Unterscheidung zwischen dem Identifier und der Lokation eines Objektes, die es ermöglicht, das Objekt unabhängig von seinem Speicherort zu identifizieren. Dies unterscheidet die Persistent Identifier von einer Universal Ressource Locator (URL). Zur Sicherstellung der eindeutigen Vergabe und der Zuweisung von Kennung und Speicherort bedarf es eines automatisierten Dienstes. Jedem Persistent Identifier werden dabei Adressinformationen, zum Beispiel eine URL zugewiesen. Von zentraler Bedeutung sind hier geeignete organisatorische Maßnahmen, die Verweise auf die tatsächlichen Speicherorte der Ressourcen aktuell halten. Programme können dann über einen sogenannten Resolverdienst den zitierten Persistent Identifier auflösen, so dass ein Zugang zu den zitierten Forschungsdaten möglich wird.

Es existieren mittlerweile für die Identifikation von elektronischen Textpublikationen diverse Systeme von Persistent Identifiern, die technisch die Basis für einen Service auch zur Identifizierung von Daten leisten können: Archival Research Key (ARK), Digital Object Identifier (DOI®), Handle, Library of Congress Control Number (LCCN), Life Science Identifiers (LSID), Persistent URL (PURL), Uniform Resource Name (URN) und weitere. Auf einen gemeinsamen Standard haben sich die verschiedenen Nutzergemeinden jedoch noch nicht geeinigt, da die Systeme im Prinzip gut ineinander überführbar sind. Um die langfristige Eignung zu beurteilen, ist hier weniger die technische als die organisatorische Ausgestaltung relevant.

Das DOI®-Sy stem

Das DOI®-System wurde von der Association of American Publishers entwickelt und wird gegenwärtig von der International DOI Foundation (IDF) verwaltet. Die IDF besteht seit 1998 und unterstützt die Rechteverwaltung für geistiges Eigentum in digitalen Netzwerken, indem sie die Entwicklung und Verbreitung des DOI®-Systems als gemeinsame Infrastruktur für das Content Management fördert. Die IDF ist als not-for-profit Organisation registriert und wird von einem Executive Board, das von den Mitgliedern des IDFs gewählt wird, kontrolliert. Die Mitgliedschaft ist offen für alle Organisationen, die sich mit elektronischem Publizieren und den damit verbundenen Technologien beschäftigen.

Das DOI®-System ist ein verwaltetes System für die persistente Identifikation von Inhalten, die in digitalen Netzwerken angeboten werden. Es kann für die Identifizierung von physikalischen, digitalen oder anderen Objekten benutzt werden. Die Identifikatoren (DOI-Namen) führen direkt zum Speicherort des bezeichneten Objektes. Technisch basiert das DOI-System auf der vom CNRI entwickelten Handle Technology. Es wird ergänzt durch ein Metadatenmodell, um die zum Objekt gehörenden Metadaten mit dem DOI-Namen zu verbinden. Auf der Basis der gemeinsamen Policy und technischen Infrastruktur der IDF wird das DOI®-System durch einen Zusammenschluss von Registrierungsagenturen umgesetzt. Dieses Lizenzmodell wird fälschlicher Weise mit einer kommerziellen Ausrichtung des DOI®-Systems verwechselt. Jede Registrierungsagentur kann jedoch über ihr eigenes Businessmodell für die Vergabe der DOI-Namen entscheiden.

DOI-Namen und Langzeitarchivierung

Das DOI®-System hat gute Aussichten auf Verbreitung und Langlebigkeit. Dies wird nicht zuletzt durch verbindliche Verträge zwischen Registrierungsagentur und Nutzern erreicht, die eine längere Zeitperspektive versprechen. Insbesondere die internationalen wissenschaftlichen Fachverlage setzen fast durchgängig auf die Verwendung von DOI-Namen (vgl. Brase et al. 2009). Das DOI-System überzeugt daher auch für den Einsatz im Wissenschaftsbereich.

Die Verwendung von DOI-Namen im Rahmen der Langzeitarchivierung von digitalen Ressourcen bietet sich aus mehreren Gründen an. Neben der breiten Verwendung garantiert die überwachende Einrichtung der IDF die Einhaltung der Standards und die notwendige Persistenz. Der Identifier selbst kann aber keine dauerhafte Verfügbarkeit sicherstellen; es ist „nur“ eine technische Lösung, die Bestandteil jedes Langzeitarchivierungskonzeptes sein sollte. Langzeitarchivierung impliziert die Sicherung über große Zeiträume. Keines der existierenden Persistent Identifier Systeme besitzt jedoch eine 100-prozentige Gewähr für Dau-

erhaftigkeit. Allerdings ist die Technik des Handle Systems so ausgelegt, dass eine Registrierungsagentur jederzeit komplett selbstständig die Auflösbarkeit ihrer DOI-Namen sicherstellen kann.

Struktur und Resolving eines DOI-Namens

Ein DOI-Name besteht genau wie ein Handle immer aus einem Präfix und einem Suffix, wobei beide durch einen Schrägstrich getrennt werden und das Präfix stets mit „10." beginnt (vgl. Abbildung 1). Das Präfix, das beispielsweise einem bestimmten Datenzentrum zugeordnet ist (in der Abbildung 1: „3478"), ermöglicht die Bildung einer unbegrenzten Anzahl von DOI-Namen, indem auf der Basis des Präfixes und verschiedener Suffixe eine beliebige Reihe von eindeutigen Identifiern gebildet werden können.

Abbildung 1: Struktur eines DOI-Namens

Um einen DOI-Namen zur zugehörigen URL aufzulösen, gibt es verschiedene Möglichkeiten, die alle auf dem zentral betriebenen Handle Server basieren. Zum einen kann er über das vom CNRI angebotene Resolver-plug-in eingegeben und aktiviert werden. Eine andere Möglichkeit ist die Verwendung des Proxy Servers des DOI-Systems (http://dx.doi.org/) bzw. des Handle Systems (http://hdl.handle.net/). Die Eingabe des DOI-Namens zusammen mit der vorangestellten Zeichenkette http://dx.doi.org/ in den Eingabeschlitz jedes beliebigen Browsers führt den Nutzer direkt zum Speicherort des Objektes bzw. zu einer Webseite (landing page), die diese und die Zugangsbedingungen ausführlich beschreibt.

DataCite

Seit 2010 sind das GESIS Leibniz-Institut für Sozialwissenschaften und das ZBW Leibniz-Informationszentrum für Wirtschaftswissenschaften Mitglieder in DataCite. Damit wurde die Voraussetzung (DOI-Vergaberecht) für die Etablierung eines Registrierungsservices für Forschungsdaten auf der Basis von DOI-Namen geschaffen. DataCite ist ein 2009 in London gegründetes internationales Konsortium mit inzwischen sechzehn Mitgliedern aus zehn Ländern, die gemeinsam das Ziel verfolgen, die Akzeptanz von Forschungsdaten als eigenständige, zitierfähige wissenschaftliche Objekte zu fördern.

Während DataCite eine bei der IDF akkreditierte DOI-Registrierungsagentur ist, fungieren GESIS und ZBW als Vollmitglieder in DataCite als DOI-Allocation Agency.

DataCite bietet neben der Mitgliedschaft in der IDF eine international abgestimmte Vorgehensweise bei technischen Lösungen, Standards, Best Practises und Workflows, sowie ein gemeinsames Metadatenschema und einen MetadataStore.

Registrierungsagentur da|ra

GESIS hat im Frühjahr 2010 ein Pilotprojekt unter dem Titel da|ra (Registrierungsagentur für sozialwissenschaftliche Daten) zur Etablierung eines Registrierungssystems für Forschungsdaten der Sozialforschung im deutschsprachigen Raum gestartet. Ziel war es, eine Infrastruktur zu entwickeln, die es ermöglicht, Forschungsdatenbestände mit DOI-Namen zu versehen und sie mit ihren Titeln, Themen, Autoren, Provenienzen, Methoden und Zugangsmöglichkeiten so umfassend wie möglich nachzuweisen sowie findbar und zitierbar zu machen.

In Kooperation mit DataCite wurde dazu mit der technischen Implementierung eines Registrierungstools begonnen und ein spezifisches über das von DataCite hinausgehendes Metadatenmodell entwickelt. Begonnen wurde mit der Registrierung der Forschungsdaten des GESIS-Datenarchivs, um die technischen und organisatorischen Lösungen zu testen.

In einer zweiten Projektphase (bis Ende 2012) wird der Service weiteren Forschungsdatenzentren, Datenservicezentren und institutionalisierten Forschungsprojekten angeboten. Schwerpunkt in dieser Phase ist aber auch die Entwicklung des DOI-Registrierungsservices für Wirtschaftsdaten auf der Basis der vorhandenen technischen Lösung. Dazu wurde die Zusammenarbeit mit der ZBW aufgenommen. Dies lag nahe, da die Nutzergemeinschaften beider Infrastruktureinrichtungen aus eng verwandten Disziplinen stammen.

Die Einbeziehung weiterer Datenrepositorien und Datenkuratoren als Publikationsagenten ist für die Ausbauphase (ab 2013) geplant. In dieser Phase wird der Dienst in Abhängigkeit von der Entwicklung eines Self-Archiving Services bei GESIS auch Einzeldatenproduzenten aus dem sozialwissenschaftlichen Bereich angeboten. Über ein von der DFG gefördertes Projekt werden zusätzlich die Entwicklung von Standards und Guidelines für die DOI-Registrierung sowie die Entwicklung eines erweiterten Nachweissystems realisiert.

Organisation und Verwaltung der Registrierung

Die DOI-Registrierung erfolgt bei da|ra in enger Kooperation mit den datenhaltenden Organisationen, den so genannten Publikationsagenten, die für die Pflege und Speicherung der Forschungsdaten sowie für die Metadatenpflege zuständig sind. Die Datensätze verbleiben bei den Datenzentren, da|ra speichert die Metadaten und macht die registrierten Inhalte über eine Datenbank recherchierbar.

Die Rahmenbedingungen und Voraussetzungen für die DOI-Registrierung sind in einer Policy festgehalten. Darüber hinaus werden der Workflow und sämtliche Verantwortlichkeiten im Registrierungsprozess in einem Service Level Agreement vereinbart. Hierunter fallen Fragen zur Qualitäts- und Persistenzsicherung (Daten und Metadaten), zu Urheberrechten, Versionierungen, zur Verfügbarkeit des Services sowie deren Funktionalitäten. Bei der Festlegung der Details werden die Best Practice Empfehlungen von DataCite berücksichtigt.

Während der Etablierungsphase von da|ra wird der Registrierungsservice kostenneutral angeboten. Nach Abschluss der Einführungsphase werden die Betreiber von da|ra prüfen, ob die Erhebung einer Gebühr erforderlich ist. Ziel ist es, den Service so zu gestalten, dass die Betreiberkosten auch langfristig auf einem niedrigen Niveau gehalten werden können und die Nutzung des Basisservices weiterhin kostenfrei bleibt.

Das Metadatenmodell

Neben dem technischen DOI-Registrierungsservice leistet da|ra auch die Übernahme von Metadatenbeschreibungen in sein Datenbanksystem. Damit wird eine differenzierte Suche aller registrierten Datensätze ermöglicht, eine inhaltliche Beschreibung zur Verfügung gestellt und die Voraussetzung zur einheitlichen Zitation der Daten geschaffen. Unter Berücksichtigung der besonderen Herausforderungen der Dokumentation digitaler Datenobjekte wurde für das Informationssystem ein eigenes Beschreibungsschema (vgl. Hausstein et al. 2011) entwickelt. Dieses basiert auf dem Metadatenschema des GESIS-Datenbestandskatalogs und wurde in Anlehnung an das DataCite-Metadatenschema erweitert. Durch das Mapping mit dem DDI Standard und durch die Verwendung kontrollierter Vokabulare ist die Interoperabilität der erfassten Ressourcen mit anderen Datenbeständen auch im internationalen Rahmen gewährleistet.

Ein wesentliches Element der Datenbeschreibung ist die Möglichkeit der Versionierung, die gerade im Bereich von Primärdaten für die eindeutige Referenzierung eines Datensatzes unerlässlich ist. Da jede Version mit einem eigenen DOI-Namen und eigenen Metadaten versehen ist, wird die eindeutige Bezugnahme auf den einer Analyse zugrundeliegenden Datensatz und damit zum Beispiel auch die Verifizierung von Analyseergebnissen ermöglicht.

Ein Schwerpunkt der Vereinbarung mit dem Publikationsagenten stellt die Sicherstellung der Qualität der Metadaten dar. Bei der Registrierung neu anfallende Metadaten werden durch da|ra zunächst geprüft und bei Bedarf entsprechend dem da|ra-Metadatenschema nachbearbeitet. In der Folge ist der Publikationsagent verantwortlich für die Aktualität der Metadaten, insbesondere die korrekte Auflösung des DOI-Namens, während da|ra für die permanente technische Auflösbarkeit zuständig ist. Hierdurch wird die Persistenz des Angebotes gewährleistet.

Technische Implementation

Die technische Implementation des Registrierungstools erfolgt über eine serviceorientierte Architektur. Zentrale Komponenten, wie beispielsweise das Registrieren eines DOI-Namens oder das Indexieren, wurden in separate Services ausgelagert. Entsprechend verfügt auch das zentrale Informationssystem über Schnittstellen, die als Webservice angesprochen werden können.

Die Metaddaten werden in einer Datenbank gemäß DDI Standard abgespeichert. Dadurch werden ein Im- und Export der Daten im XML-Format auf einfache Weise ermöglicht und ein Austausch mit den Publikationsagenten direkt unterstützt. Für die Funktionalität der Suche wird ein Indexierungsserver eingesetzt (SOLR). Die Oberfläche zum Editieren der Metadaten wird von Projektmitarbeitern oder den Publikationsagenten benutzt, um die Metadaten zu pflegen. Die Edition wird durch die Verwendung von kontrollierten Vokabularen unterstützt, damit möglichst standardisierte Metadaten erstellt werden.

da|ra Services

Die Registrierungsagentur da|ra bietet eine vollständige Infrastruktur für die DOI-Registrierung und die Metadatenverwaltung. Die Betreiber von da|ra sehen sich als Dienstleister für Daten- und Forschungszentren, die ihre Primärdaten mit DOI-Namen registrieren wollen. Dabei kann es sich um Surveydaten, Aggregatdaten, Microdaten, aber auch um qualitative Daten handeln.

Abbildung 2: da|ra Services

Im Rahmen des Registrierungsprozesses erhält jeder Datensatz nach Übermittlung der erforderlichen Metadaten einen eindeutigen DOI-Namen, wobei Granularität und Namensgestaltung vom Publikationsagenten festgelegt werden können. Der DOI-Name kann dann sofort über einen Resolverdienst zur entsprechenden URL aufgelöst werden. Das Serviceangebot von da|ra (vgl. Abbildung 2) umfasst gegenwärtig die DOI-Namensregistrierung, eine DOI-Resolvermöglich-

keit, ein umfangreiches Metadatenbeschreibungsschema, das Metadatenhandling in einer Datenbank sowie verschiedene Möglichkeiten zur Übertragung der Metadaten. Darüber hinaus können interessierte Nutzer über eine einfache und erweiterte Suche in den Metadatenbeständen recherchieren.

Das Angebot zu impact factors and peer review sowie die Verlinkungen von Daten und Publikationen befindet sich noch im Aufbau und wird nach Abschluss des DFG Projektes als Prototyp angeboten.

Gegenwärtig nutzen folgende sozialwissenschaftliche Forschungsdatenzentren und Datenarchive den da|ra Service:

1. Forschungsdatenzentrum des Sozio-oekonomischen Panels (FDZ-SOEP)
2. Forschungsdatenzentrum PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID)
3. Forschungsdatenzentrum Deutscher Alterssurvey (FDZ-DEAS)
4. Projekt Nationales Bildungspanel (NEPS)
5. GESIS Datenarchiv
6. Forschungsdatenzentrum „Internationale Umfrageprogramme“ bei GESIS
7. Forschungsdatenzentrum ALLBUS bei GESIS
8. Forschungsdatenzentrum „Wahlen“ bei GESIS.

Diese haben insgesamt ca. 5200 Studien/Datensätze registriert, wobei der größte Anteil daran noch beim GESIS Datenarchiv liegt. Zusätzlich wurden ca. 2400 Metadatensätze von der iLibrary der OECD, die ihre Bestände über die Partnerschaft mit der DOI-Registrierungsagentur crossref registriert, in das Informationssystem übernommen. Somit stehen mehr als 7500 Metadatensätze für Recherchen im da|ra Informationssystem zur Verfügung.

Ausblick

Im Rahmen der Etablierungsphase des Projektes arbeiten GESIS und ZBW an der Anpassung des Registrierungssystems, um die Veränderungen zu berücksichtigen, die durch die technischen und organisatorischen Neuerungen bei DataCite sowie durch die Erweiterung des Anwendungsbereichs von da|ra entstanden sind. Das betrifft die Überarbeitung des Metadatenschemas, die entsprechende Anpassung der Datenbank, den Anschluss des Registrierungstools an den neugestalteten MetaDataStore von DataCite sowie das Update der da|ra Policy und des Service Level Agreements.

Im Einzelnen bedeutet dies eine Erweiterung des Metadatenschemas, um die Spezifik der Forschungsdaten aus den Wirtschaftswissenschaften abbilden zu können. Gleichzeitig werden die Änderungen, die im Metadatenschema von DataCite vorgenommen wurden, berücksichtigt. Im Dezember 2011 wurde die Version 2.2.1. des da|ra Metadatenschemas verabschiedet. Dieses bildet die Grundlage für die Anpassung des technischen Systems, die noch bis Anfang 2012 andauern wird. In diesem Zusammenhang werden auch die Add-on Services entwickelt. Begonnen wird mit dem Angebot einer facettierten Suche, einem link checker und einem help desk für die Publikationsagenten.

Zukünftig werden GESIS und ZBW die Registrierungsagentur zwar gemeinsam betreiben, die Betreuung der Publikationsagenten wird aber entsprechend der disziplinären Zuordnung jeweils von GESIS oder ZBW erfolgen. Diese und andere daraus resultierenden Veränderungen im organisatorischen Ablauf mussten in der Policy und im Service Level Agreement ihren Niederschlag finden. Die da|ra Policy liegt nun in der Version 2.0 vor.

Das überarbeitete Angebot von da|ra wird ab Anfang 2012 über einen frisch gestalteten Webauftritt unter der neuen Domain www.da-ra.de angeboten. Besuchen Sie uns auf unserer Website und

Make Your Data Citable!

Literatur

Allianz der deutschen Wissenschaftsorganisationen (2010): Grundsätze zum Umgang mit Forschungsdaten. 24. Juni 2010. http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze/

Arbeitsgruppe Fachinformation (2009): Rahmenkonzept für die Fachinformation. Vorlage zur Sitzung des Ausschusses der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder (GWK) am 29.09.2009.

Askitas, N. (2010): What Makes Persistent Identifiers Persistent? Working Paper No. 147. RatSWD Working Paper Series June 2010. http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_147.pdf

Bellini, E./Cirinnà, C. et al. (2008): Persistent Identifiers distributed system for Cultural Heritage digital objects. http://www.bl.uk/ipres2008/presentations_day2/38_Lunghi.pdf

Bleuel, J. (2000): Zitation von Internet-Quellen (Citing of Internet sources). In: Hug, T. (Hrsg.): Wie kommt Wissenschaft zu Wissen? Band 1: Einführung in das wissenschaftliche Arbeiten. Hohengehren: Schneider Verlag.

Blue Ribbon Task Force (2010): Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Brase, J. (2010): DataCite – A global registration agency for research data. Working Paper No 149. RatSWD Working Paper Series July 2010. http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_149.pdf

Brase J. et al. (2009): Approach for a global joint registration agency for research data. Information Services & Use. In: Neuroth H. et al. (Hrsg.): nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.

Brase, J. und Klump, J. (2007): Zitierfähige Datensätze. Primärdaten-Management durch DOIs. In: Wissenschaftskommunikation der Zukunft. 4. Konferenz der Zentralbibliothek. http://dc110dmz.gfz-potsdam.de/contenido/std-doi/upload/pdf/Brase_Wisskomm_2007.pdf

DFG (1998): Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“. Denkschrift der Deutschen Forschungsgemeinschaft. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Diepenbroek, M. und Grobe, H. (2007): PANGAEA® als vernetztes Verlags- und Bibliothekssystem für wissenschaftliche Daten. In: Wissenschaftskommunikation der Zukunft. 4. Konferenz der Zentralbibliothek. http://www.fz-juelich.de/zb/datapool/page/1000/Diepenbroek_Abstract.pdf

Dittert, N./Diepenbroek, M. and Grobe, H. (2002): Data and information management for the CMTT synthesis. Manuskript. http://epic.awi.de/Publications/Dit2002b.pdf

GESIS Report 4/10. Der Aktuelle Informationsdienst für die Sozialwissenschaften. GESIS – Leibniz-Institut für Sozialwissenschaften. Mannheim. September 2010. http://www.gesis.org/fileadmin/upload/institut/presse/gesis_report/gesis_report_10_04.pdf

Hausstein, B. und Zenk-Möltgen, W. (2011): da|ra – Ein Service der GESIS für die Zitation sozialwissenschaftlicher Daten. In: Schomberg, S./Leggewie, C. und Puschmann, C. (Hrsg.): Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. 20./21. September 2011. Köln. Beiträge der Tagung. http://www.hbz-nrw.de/Tagung_Digitale_Wissenschaft.pdf

Hausstein, B./Zenk-Möltgen, W./Wilde, A. und Schleinstein, N. (2011): da|ra Metadatenschema. Version 1.0. GESIS-Working Papers 2011|14. GESIS – Leibniz-Institut für Sozialwissenschaften. DOI:10.4232/10.mdsdoc.1.0

Klump, J./Bertelmann, R. et al. (2006): Data publication in the open access initiative. Data Science Journal 5 (79).

Metadata for the Publication and Citation of Scientific Primary Data (Version 3.0). http://www.icdp-online.org/contenido/std-doi/upload/pdf/STD_metadata_kernel_v3.pdf

Neuroth, H./Oßwald, A./Scheffel, R./Strathmann, S. und Huth, K. (2007): nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3. nestor – Network of Expertise in Long-Term Storage of Digital Resources. Göttingen. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949

Parsons, M.A./Duerr, R. and Minster, J.-B. (2010): Data Citation and Peer Review. Eos, Transactions, American Geophysical Union 91 (34), 24 August 2010, 297-304. http://aurora.gmu.edu/spaceweather/images/2010EO340001.pdf

Paskin, N. (2000): Digital Object Identifier: implementing a standard digital identifier as the key to effective digital rights management. The International DOI Foundation. Kidlington, Oxfordshire, United Kingdom. http://www.doi.org/doi_presentations/aprilpaper.pdf

Reilly, S. (2010): Digital Object Repository Server: A Component of the Digital Object Architecture. D-Lib Magazine 16 (1/2). http://www.dlib.org/dlib/january10/reilly/01reilly.print.html
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. October 2010. http://goo.gl/WrxO
Uhlir, P.F. (2003): Discussion Framework. In: Esanu, J.M. and Uhlir, P.F. (Eds.): The Role of Scientific and Technical Data and Information in the Public Domain. Washington DC: The National Academies Press.
Uhlir, P.F. and Schroder, P. (2008): Chapter 8 – Open Data for Global Science. In: Fitzgerald, B. (Ed.): Legal Framework for e-Research: Realising the Potential. Sydney: Sydney University Press Law Books 39 (2008) 188. http://www.austlii.edu.au/au/journals/SydUPLawBk/2008/39.html

Internet

Allianz der deutschen Wissenschaftsorganisationen
http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze/
California Digital Library: ARK
http://www.cdlib.org/inside/diglib/ark/
Corporation for National Research Initiatives® (CNRI)
http://www.cnri.reston.va.us
crossref
http://www.crossref.org
da|ra
http://www.gesis.org/dara http://www.da-ra.de
DataCite
http://www.datacite.org
Data Documentation Initiative
http://www.ddialliance.org
Digital Object Identifier (DOI®) System
http://www.doi.org
Forschungsdatenzentrum des Sozio-oekonomischen Panels (FDZ-SOEP)
http://www.diw.de/de/diw_02.c.221180.de/fdz_soep.html
Forschungsdatenzentrum Deutscher Alterssurvey (FDZ-DEAS)
http://www.dza.de/informationsdienste/forschungsdatenzentrum-deas.html
Forschungsdatenzentrum PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID)
http://psychdata.zpid.de/

Forschungsdatenzentrum „Internationale Umfrageprogramme“ bei GESIS
http://www.gesis.org/das-institut/kompetenzzentren/fdz-internationale-umfrageprogramme/
Forschungsdatenzentrum ALLBUS bei GESIS
http://www.gesis.org/das-institut/kompetenzzentren/fdz-allbus/
Forschungsdatenzentrum „Wahlen“ bei GESIS
http://www.gesis.org/das-institut/kompetenzzentren/fdz-wahlen/
GESIS – Leibniz Institut für Sozialwissenschaften (Mannheim, Köln, Berlin)
http://www.gesis.org
Homepage der Technischen Informationsbibliothek (TIB) Hannover zu eigener DOI-Registrierungsagentur
http://www.tib-hannover.de/de/die-tib/doi-registrierungsagentur/
Homepage der Technischen Informationsbibliothek (TIB) Hannover zu Projekt CODATA
http://www.tib-hannover.de/de/die-tib/projekte/codata/
Library of Congress: Structure of the LC Control Number
http://www.loc.gov/marc/lccn_structure.html
Life Sciences Identifiers (LSID)
http://lsids.sourceforge.net/
OECD iLibrary
http://www.oecd-ilibrary.org
Persistent Uniform Resource Locator (PURL)
http://purl.oclc.org/docs/index.html
Projekt Nationales Bildungspanel (NEPS)
https://www.neps-data.de/
Publikation und Zitierfähigkeit wissenschaftlicher Primärdaten
http://www.std-doi.de/front_content.php
The Handle System®
http://www.handle.net
Uniform Resource Name (URN) Syntax
http://tools.ietf.org/html/rfc2141
ZBW
http://www.zbw.eu

European Persistent Identifier Consortium - PIDs für die Wissenschaft

Tibor Kálmán, Daniel Kurzawe und Ulrich Schwardmann

Einleitung

Der Umgang mit digitalen Objekten[1] rückt immer stärker in den Fokus der Wissenschaft und dies spiegelt sich auch immer stärker im Forschungsprozess wider. Um die damit einhergehenden Anforderungen an eine ständig wachsende Zahl an digitalen Objekten bewältigen zu können[2], sind nicht nur effiziente Speicherkonzepte notwendig, um Objekte nachhaltig zu lagern, sondern auch Konzepte, um diese Objekte auch zuverlässig zu identifizieren (vgl. Kahn and Wilensky 2006). In diesem Artikel geben wir einen Überblick über den EPIC Verbund, welcher sich mit der langfristigen und nachhaltigen Identifikation von digitalen Objekten beschäftigt.

Im langfristigen Umgang mit Referenzen zu digitalen Objekten ergeben sich spezifische Herausforderungen.[3] Ändert sich der physikalische Ablageort bei einer Umstrukturierung, also bei einem Umzug oder einer Anpassung in der Infrastruktur, ist eine Änderung der Adresse bei allen auf das Objekt verweisenden Referenzen notwendig. Andernfalls kann das Objekt möglicherweise nicht mehr aufgefunden werden. Identifiziert man Objekte mittels eines Uniform Resource Identifiers (URI), wird das Problem deutlich: Ist ein Objekt beispielsweise auf einer Institutswebsite präsentiert und die Struktur der Website ändert sich, ist es nicht mehr unter der ursprünglichen Adresse zu finden. Dazu reicht es aus, wenn sich etwa das Institutskürzel in der Uniform Resource Locator (URL) ändert. Alle Verweise auf die alte Adresse laufen nun ins Leere und müssen abgefangen werden, um den Benutzer zur neuen Position zu verweisen.

Im schlimmsten Fall ist es dem Nutzer, welcher auf das Objekt zugreifen will, nicht möglich nachzuvollziehen, unter welcher Adresse das Objekt zu finden ist oder ob es überhaupt noch existiert.

1 Bei digitalen Objekten handelt es sich in diesem Kontext um jede Art von persistentem und referenzierbarem Datenstream. Dies können Dateicontainer, aber auch andere adressierbare Entitäten, wie etwa Funktionalitäten, welche sich selbst auf Datenstream beziehen, sein. Auch können die Objektbeschreibungen mit inbegriffen sein. Darauf werden wir im Weiteren jedoch nicht dediziert eingehen.

2 Dies fordert etwa die Empfehlung zur guten wissenschaftlichen Praxis der DFG (vgl. Deutsche Forschungsgemeinschaft 1998).

3 Im nestor Handbuch wird ein Überblick über die unterschiedlichsten Herausforderungen und Konzepte gegeben (vgl. Neuroth et al. 2009).

Durch die Verwendung von persistenten Identifikatoren (PIDs)[4] können Objekte langfristig und nachhaltig referenziert werden. Dazu wird ein Vermittler als neutrale Schicht zwischen Objekt und Nutzer eingeschoben und jedem zu identifizierenden Objekt ein eindeutiger PID zugewiesen (siehe Abbildung 1). Dieser PID enthält die aktuelle Referenz zum Objekt und kann je nach Verwendung auch weitere beschreibende Informationen enthalten.[5] PIDs sind hierbei für eine dauerhafte Auflösung konzipiert. Die Struktur stellt sicher, dass es nicht zu Überschneidungen zwischen Identifikatoren kommen kann. Mit Hilfe des Identifikators ist es nun möglich, die aktuell hinterlegte Adresse des Objektes bei dem sogenannten Resolver Dienst abzufragen. Dieser gleicht die übermittelte PID mit einer Datenbank ab und gibt die hinterlegte Adresse zurück.

Um PIDs nachhaltig zu nutzen, ist es notwendig, dass bei einer Änderung im Objektpfad auch die Adresse im PID angepasst wird. Der Vorteil hierbei ist, dass diese Änderung nur noch an der zentralen Stelle innerhalb des PIDs geschehen muss. Selbst in dem zuvor beschriebenen Beispiel, also bei komplexeren Änderungen innerhalb des Data Providers, bleibt die PID Struktur erhalten und es muss nur der Verweis innerhalb der PIDs angepasst werden. Sollten Objekte verworfen werden, kann diese Information ebenfalls in den entsprechenden PIDs hinterlegt werden. Weitere Anfragen auf nicht mehr existente Objekte laufen nun nicht mehr ins Leere, sondern können auf einen entsprechenden Hinweis umgeleitet werden.

Zum einen ist es eine organisatorische Herausforderung, die Identifikatoren langfristig und dezentral aufzulösen und zu verwalten. Dazu ist eine Koordination und Zusammenarbeit von verschiedenen Einrichtungen auf verschiedenen

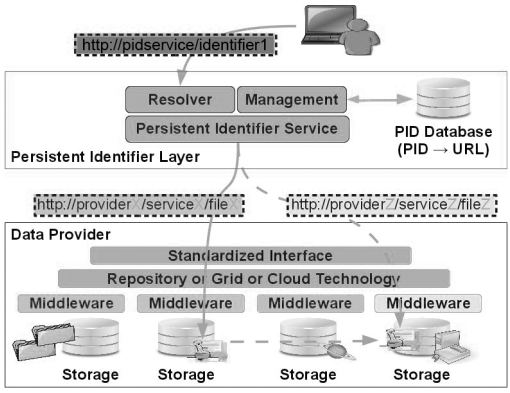


Abbildung 1: Persistent Identifier Service für wissenschaftliche Daten

4 Persistente Identifikatoren oder auch Persistent Identifier werden im Folgenden auch mit PID abgekürzt.

5 Dies kann je nach PID System etwa eine Referenz zu weiteren, zum Objekt gehörenden Metadaten oder die Prüfsumme des Objektes sein.

Ebenen notwendig. Zum anderen ist es aber auch eine große technische Herausforderung, eine solche PID Infrastruktur robust und zuverlässig aufzubauen und langfristig zu betreiben.

Es gibt verschiedene Organisationen und Anbieter, welche PIDs vergeben. Diese unterscheiden sich in Geschäftsmodellen, Policies, Services oder registrierbaren Objekten. Einige Organisationen schränken die Vergabe und Art der Anwendung durch die zugrunde liegenden Geschäftsmodelle ein. Als Beispiel ist das Digital Object Identifier (DOI) System (vgl. Paskin 2010) zu nennen. DOI wird durch die International DOI Foundation (IDF) verwaltet. IDF ist ein Verbund von Registrierungsstellen und finanziert sich durch Mitgliedsbeiträge und Gebühren. Es wird für jeden vergebenen PID eine Gebühr erhoben. Andere Vergabestellen spezialisieren sich auf spezifische Anwendungsbereiche oder geben nur PIDs für persistente Objekte aus. DataCite (vgl. Brase 2009) hat sich auf PIDs für Forschungsdaten spezialisiert. Auch hier werden Gebühren für jeden vergebenen PID verlangt. Entsprechend der Policy von DataCite, müssen die Objekte über eine möglichst lange Zeit vorgehalten werden. Wenn keine längerfristige Aufbewahrungsstrategie vorhanden oder erwünscht ist, können keine PIDs beantragt werden. Alternativ zu diesen oder anderen Organisationen, könnte auch ein eigener Namensraum registriert werden. Doch müsste so auch die notwendige Infrastruktur selber bereitgestellt werden.

Die meisten Organisationen setzten hier auf bewährte PID Technologien (vgl. Broeder et al. 2008). Beispielsweise wird Handle von der Corporation for National Research Initiatives (CNRI)[6] bereitgestellt. Neben Handle gibt es noch weitere verbreitete Technologien, wie etwa Persistent URL (PURL)[7] oder Archival Resource Key (ARK)[8]. Auch Standards zur Adressierung von Objekten, wie etwa Uniform Resource Name (URN)[9], können zur langfristigen Referenzierung von Objekten verwendet werden.[10]

Im Folgenden wird das European Persistent Identifier Consortium (EPIC)[11] vorgestellt. EPIC hat zum Ziel, einen möglichst generischen und kosteneffizienten Dienst zu schaffen. Es gibt keine feste Vorgabe über die Provenienz der Objekte. Auch können PIDs sehr kosteneffizient erzeugt werden und sind so auch für größere Datenbestände oder Daten mit kürzerem Lebenszyklus attraktiv. Im Folgenden wird auf das Konsortium und die organisatorische Struktur von EPIC eingegangen. Darauf aufbauend werden die Infrastruktur, Dienste und Schnittstellen

6 http://www.cnri.reston.va.us
7 http://purl.oclc.org/docs/index.html
8 http://www.cdlib.org/inside/diglib/ark/
9 URN wurde 1992 von der URN-Working Group der Internet Engineering Task Force (IETF) entwickelt.
10 http://www.persistent-identifier.de
11 http://www.pidconsortium.eu

skizziert. Danach werden einige Anwendungsfälle vorgestellt und die Rolle von EPIC als PID-Service Provider aufgezeigt. Abschließend folgt ein Ausblick auf die weiteren Entwicklungen in EPIC.

Das EPIC Konsortium – PIDs für die Wissenschaft

Die langfristige und dezentrale Auflösung und Verwaltung von PIDs ist nicht nur eine technische, sondern ebenso eine organisatorische Herausforderung und erfordert eine enge Koordinierung und Zusammenarbeit von verschiedenen Einrichtungen. In diesem Abschnitt werden Konzepte der organisatorischen Struktur von EPIC und den angegliederten Partnern beschrieben.

EPIC hat zum Ziel, ein dezentrales Konsortium zu gründen, welches die Verwaltung und Bereitstellung von Diensten übernimmt. Um eine längerfristige und nachhaltige Planung zu gewährleisten, sollen die beteiligten Institute durch nationale Programme finanziert werden. Ebenfalls sollte eine langjährige Erfahrung im Betrieb von nachhaltigen Services und von stabilen und hochverfügbaren Diensten vorhanden sein. Um auch Service Level Agreements (SLAs) anbieten zu können, müssen die Diensteanbieter rechtsverbindliche Kooperationen eingehen können. Auch sollte ein enger Kontakt zur Forschung gewährleistet sein, um die notwendige Expertise im Umgang mit Forschungsdaten gewährleisten zu können. In dem Zusammenschluss sollten möglichst viele Benutzergruppen vertreten sein, um ein möglichst generisches Angebot bieten zu können.

Gemeinsam haben die drei Gründungspartner, die Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG, Deutschland), das Stichting Academisch Rekencentrum Amsterdam (SARA, Niederlande) und das IT Center for Science Ltd. (CSC, Finnland), im Herbst 2009 mit einem Memorandum of Understanding den Grundstein für European Persistent Identifier Consortium (EPIC) gelegt. In diesem wurde die Bereitschaft erklärt, gemeinsam einen PID-Service bereitzustellen.

Die GWDG[12] ist eine gemeinsame Einrichtung der Georg-August-Universität Göttingen und der Max-Planck-Gesellschaft. Sie erfüllt die Funktion eines Rechen- und IT-Kompetenzzentrums für die Max-Planck-Gesellschaft und des Hochschulrechenzentrums für die Georg-August-Universität Göttingen. Die GWDG wurde 1970 als gemeinnützige GmbH gegründet. Die GWDG hat zurzeit etwa 25.000 aktive Benutzer. Davon etwa 1000 im Bereich des wissenschaftlichen Rechnens (High Performance Computing). Ebenso ist die GWDG Partner in diversen eScience, Grid und Cloud Projekten. Betreut wird das Rechenzentrum zurzeit von etwa 100 Mitarbeitern.

12 http://www.gwdg.de

Das Amsterdamer Rechenzentrum SARA[13] unterstützt Forscher in den Niederlanden und arbeitet eng mit diversen akademischen Partnern sowie öffentlichen und wirtschaftlichen Einrichtungen zusammen. SARA bietet seit über 40 Jahren IT-Dienstleistungen an.

CSC[14] bietet mit über 180 Mitarbeitern, als Partner der finnischen Forschungsinfrastruktur, hochqualifizierte IT Dienstleistungen und Finnlands leistungsfähigste Supercomputer. Das Rechenzentrum bietet über 3000 Wissenschaftlern Zugang zu diversen Rechenangeboten. CSC wurde 1970 gegründet und wird seit 1993 als gemeinnütziges Unternehmen fortgeführt. Der Standort liegt in Espoo, nahe zum Otanie Campus der Helsinki University.

Der Fokus von EPIC liegt bei den unterschiedlichsten Gruppen innerhalb der europäischen Forschungslandschaft. Dies beinhaltet auch kulturelle Institutionen. Zu der stetig wachsenden Anzahl von Nutzern zählen unter anderem die Max-Planck-Gesellschaft (MPG), das Common Language Resources and Technology Infrastructure Projekt (CLARIN) (vgl. Váradi et al. 2008), das Digital Research Infrastructure for the Arts and Humanities Projekt (DARIAH), TextGrid – Virtuelle Forschungsumgebung für die Geisteswissenschaften, die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) und das Deutsche Klimarechenzentrum (DKRZ)[15]. Im Abschnitt 4 wird an unterschiedlichen Beispielen aufgezeigt, weshalb die von EPIC angebotenen Services in den Projekten verwendet werden.

Schnittstellen zum Erzeugen, Verwalten und Auflösen von PIDs in EPIC

Zum Erzeugen, Verwalten und Auflösen von PIDs werden unterschiedliche Schnittstellen in EPIC angeboten. Für die Auflösung setzt EPIC das weltweit verbreitete Handle System[16] ein. Das Handle System implementiert das Handle System Protokoll[17] und stellt die grundlegenden Funktionen bereit. Diese sind jedoch wenig benutzerfreundlich zu erreichen und stellen nur grundlegende Funktionalitäten zur Verfügung.

Aus diesem Grund wurden, auf dem Handle System aufbauend, weitere Dienste innerhalb von EPIC entwickelt, welche den Umgang mit PIDs erleichtern. Zum Erzeugen und Verwalten von PIDs stehen so etwa ein REST-basiertes Web

13 https://www.sara.nl
14 http://www.csc.fi/english
15 http://www.dkrz.de
16 http://www.handle.net
17 http://www.handle.net/rfc/rfc3652.html

Service Interface sowie verschiedene Weboberflächen zur Verfügung.[18] Im Folgenden werden wir auf die Details der jeweiligen Schnittstellen eingehen.

Auflösung von EPIC PIDs

Zum Auflösen des PIDs benötigt man einen definierten Prozess. Um die Nachhaltigkeit zu gewährleisten, setzt EPIC dazu auf weltweit akzeptierte PID Standards.

Die genaue Spezifikation wird im Handle System Overview[19] und im Handle System Protocol beschrieben. Der Aufbau der Namensräume wird im Handle Namespace and Service Definition[20] definiert und die Handle Namensräume werden durch Prefixes identifiziert. Für jeden Namensraum, somit auch für jeden Prefix, ist ein primärer Server zuständig. Jedem primären Server können mehrere Mirror Server zugewiesen werden. Diese können sowohl für die Lastverteilung als auch beim Ausfall des primären Servers eingesetzt werden. Alle primären und Mirror Server sind in einer zentralen Stelle bei dem globalen Handle System (GHS) registriert.

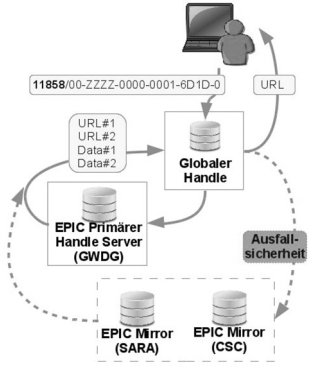


Abbildung 2: Auflösung von Persistent Identifiers innerhalb des European Persistent Identifier Consortiums

Das EPIC Konsortium besitzt einen eigenen Namensraum mit dem Prefix 11858. Alle von EPIC ausgestellten PIDs werden in diesem Namensraum registriert und mit diesem Prefix eingetragen. Dazu betreibt EPIC einen eigenen primären Server und zwei weitere Mirror Server. Der globale Handle Service delegiert alle Anfragen mit dem Prefix 11858 zur Auflösung an die GWDG weiter.

18 http://handle.gwdg.de:8080/pidservice/
19 http://www.handle.net/rfc/rfc3650.html
20 http://www.handle.net/rfc/rfc3651.html

In der Abbildung 2 wird gezeigt, wie ein EPIC PID aufgelöst wird. Eine Benutzeranfrage zur Auflösung eines PIDs wird dem globalen Handle System gestellt. Dieses überprüft, an welchen Namensraum (Prefix) sich die Anfrage wendet. Handelt es sich um den Prefix 11858, wird dieser direkt an das lokale Handle System (LHS) der GWDG geleitet. Dieses sucht in einer Datenbank nach den hinterlegten Daten des Identifikators. Entsprechend der Anfrage, werden die URL oder auch andere Parameter dem globalen Handle Server zurückgegeben. Der globale Handle Server leitet die Antwort an den Benutzer weiter. Der Browser des Benutzers wird angewiesen, eine automatische Weiterleitung auf die zuletzt hinterlegte Adresse des Objektes durchzuführen. Die gesamte Kommunikation beruht auf standardisierten HTTP Methoden, die von den meisten Browsern und HTTP Klienten unterstützt werden. Damit ist die Auflösung eines PIDs (und damit der gesamte PID Layer) für den Benutzer transparent.

Neben dem primären Server kann der globale Handle Server die Mirror Server für die Auflösung involvieren und somit die Last über alle Server eines Prefixes verteilen. Dazu misst der globale Handle Server, wie lange die Auflösung bei einem Server gedauert hat, und berechnet auf dieser Grundlage, bei welchem Server bei der nächsten Anfrage die schnEllste Antwort zu erwarten ist.

Verwaltung von EPIC PIDs

Um die Informationen innerhalb der PIDs konsistent zu halten, müssen diese von Zeit zu Zeit angepasst werden. Hierzu gibt es die Möglichkeit, diese zu aktualisieren bzw. zu verwalten. Da Handle nur eine sehr spezifische und schwer nutzbare Schnittstelle für die Verwaltung zur Verfügung stellt, wurde der PID-Service als eine auf Handle aufbauende Komponente entwickelt. Der Benutzer kann PIDs über eine zentrale Stelle erzeugen. Bei der Registrierung werden die PIDs bei dem primären Server erzeugt und werden dann an die Mirror Server repliziert. Dies ist in Abbildung 3 zu sehen. Bei jeder Änderung des PID Records ist zu gewährleisten, dass alle Änderungen vollständig von den Mirrors registriert werden. Hierbei werden mehrere aufeinanderfolgende Arbeitsschritte durchgeführt. Diese bilden gemeinsam eine Transaktion. Durch die von Handle zur Verfügung gestellten Techniken wird die Transaktionssicherheit des PID-Services gewährleistet.

Als Schnittstelle wird der PID-Service in Form eines RESTful Webservices zur Verfügung gestellt.[21] Es werden Standard HTTP Methoden zur Kommunikation zum Service verwendet. Der Vorteil gegenüber anderen Interfaces liegt in der Verwendung von Standard HTTP Aufrufen wie POST zum Erzeugen, GET zum Anfragen, PUT zum Aktualisieren und DELETE zum Löschen. In dem EPIC PID-Service wurde auf die Implementierung von DELETE verzichtet, da PIDs von der Konzeption her nicht gelöscht werden sollen. Jedoch können sie ablaufen, wer-

21 Unter einem RESTful Interface versteht man eine Methode, welche auf der Arbeit von Fielding (2000) basiert.

den aber in diesem Fall weiterhin vorgehalten. Auch Mechanismen zur Zugriffskontrolle sind Teil des Konzepts. Die Suche und Auflösung der PIDs bleibt frei zugänglich, doch wird für die Erzeugung oder Verwaltung eine Authentifizierung und Autorisierung gegenüber EPIC verlangt. Aufbauend auf dieser Schnittstelle wird auch eine Weboberfläche zur Verwaltung angeboten.

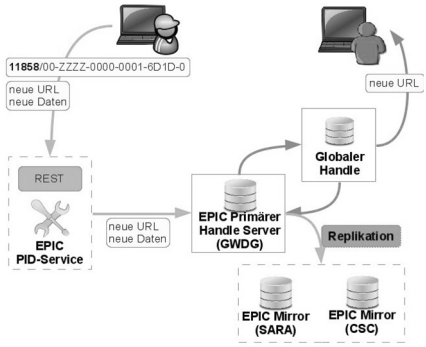


Abbildung 3: Verwaltung von Persistent Identifiers innerhalb des European Persistent Identifier Consortiums

In der Abbildung 3 ist ebenfalls zu sehen, wie ein Benutzer mittels des PID-Services eine PID verwaltet. Hierzu wird als erstes seine Identität gegenüber dem Dienst geprüft und in einem zweiten Schritt werden seine Rechte in einer zentralen Datenbank abgefragt. Um die Identität des Benutzers zu prüfen, werden Benutzername und Passwort per HTTP BASIC Authentifizierung abgefragt. Dabei wird auf eine bewährte und relativ leicht zu realisierende Technik gesetzt. Dies hat den Vorteil, dass keine komplexeren Techniken, wie etwa Grid Zertifikate, notwendig sind. Die Benutzerrechte sind in einer eigenen Datenbank abgelegt. Wird der Name in der Datenbank gefunden, wird mittels einer eindeutigen Identität nach entsprechenden Rechten in der Rechtedatenbank gesucht. In der Datenbank ist auch die Institutszugehörigkeit hinterlegt. Hierdurch werden die Benutzer Gruppen zugeordnet. Dadurch können Gruppenrichtlinien angewandt oder Rechte vererbt werden und die Benutzer können PIDs der entsprechenden Gruppen verwalten. Bei den Gruppen kann es sich auch um virtuelle Entitäten handeln.

Zuverlässigkeit und Nachhaltigkeit des EPIC PID-Services

Dadurch, dass der EPIC PID-Service auf dem Handle System aufsetzt, kann es auf eine stabile und bewährte Basis zurückgreifen. Viele weitere Organisationen setzen auf das Handle System.[22] So wird auch eine nachhaltige Systempflege

22 http://www.handle.net/apps.html

garantiert. Die grundlegende Technologie des Handle Systems existiert seit über zwanzig Jahren. Aufgrund des verteilten Systemdesigns skaliert die Auflösung der EPIC PIDs sehr gut.

Der globale Handle Server wird weltweit repliziert. Der globale Mirror Server für Europa wird bei der GWDG betrieben. Damit ist EPIC größtenteils unabhängig von Handle. Wie anfangs in Abschnitt 2 beschrieben, ist die organisatorische Struktur von EPIC auf eine nachhaltige Planung und Finanzierung ausgelegt.

Syntax von EPIC PIDs

Um den Aufbau der EPIC PIDs zu beschreiben, betrachten wir zunächst folgendes Beispiel: Ein Objekt ist mit dem PID 11858/00-ZZZZ-0000-0001-6D1D-0 bei EPIC registriert. Dieser PID folgt nun folgender Struktur: ‚PREFIX/FLAG-INST-NUM1-NUM2-NUM3-C'. Der Prefix identifiziert den Namensraum, welcher bei EPIC standardmäßig 11858 ist. Das zweite Attribut ‚FLAG' wird zurzeit noch nicht verwendet. Das dritte Feld ‚INST' wird zur Unterteilung des Namensraums genutzt. So können etwa Institute ihren eigenen Teilbereich im Namespace erhalten. Dieser Bereich wird mit vier alphanumerischen Großbuchstaben kodiert und zeigt an, welchem Institut die PIDs zugeordnet sind. Die Felder ‚NUM1' bis ‚NUM3' werden vom PID-Service als fortlaufende Nummer vergeben und bestehen aus 12 Hexadezimalzeichen. Das letzte Feld ‚C' ist für eine Prüfsumme vorgesehen, um die Integrität des Identifikators sicherzustellen.

Anwendungsfälle von EPIC Persistent Identifiers

Die internationale Kooperation der drei Service Provider in EPIC verlief bisher zielführend. Die in Kapitel zwei beschriebene organisatorische Struktur ermöglicht eine sehr enge und miteinander abgestimmte Zusammenarbeit. So werden alle Entscheidungen gemeinsam getragen. Auch bei neu beantragten Forschungsprojekten, wie beispielsweise CLARIN[23] und DARIAH[24], TextGRID[25] oder EUDAT[26] wurde EPIC direkt involviert. Die unterschiedlichen Gruppen haben die Möglichkeit, bei jährlichen Nutzertreffen direkt an EPIC mitzuwirken.

Die im Abschnitt 3 beschriebenen EPIC Dienste werden in mehreren Projekten produktiv genutzt. Im Folgenden werden mit ausgewählten Beispielen einige Anwendungsfälle beschrieben. Diese Beispiele decken ein breites Spektrum von unterschiedlichen Szenarien ab.

23 http://www.clarin.eu

24 Sowohl bei den deutschen Teilprojekten als auch auf europäischer Ebene.

25 http://www.textgrid.de

26 http://www.eudat.eu

Zu den Kernaufgaben der GWDG zählt auch der Bereich der Langzeitarchivierung. Dadurch wird Forschern ein nachhaltiger Zugang zu Forschungsdaten ermöglicht. PIDs werden in mehreren Bereichen zur Referenzierung und Identifikation der archivierten Objekte verwendet. Jedes Objekt, welches archiviert wird, enthält zu Beginn einen PID, welcher von da an das Objekt identifiziert. In diesem Konzept hat der PID eine zentrale Rolle: Der PID enthält die Adresse, unter der das Objekt erreichbar ist. Zusätzlich wird auch die Prüfsumme hinterlegt, um die Integrität des Objektes zu validieren. Mit diesen beiden Angaben wird der PID an mehreren Stellen im Arbeitsablauf verwendet, um es in verschiedenen Layern zu identifizieren.

In weiteren Forschungsprojekten, wie etwa TextGRID oder CLARIN, werden PIDs nicht nur zur Referenzierung von Objekten, sondern auch zur Referenzierung von Objektbereichen verwendet. Beispiele sind etwa die Referenzierung von Textstellen in Dokumenten oder der Verweis auf genaue Bereiche innerhalb von digitalen Medien. Um dies zu ermöglichen, wurde in EPIC die Unterstützung von Fragment Identifier eingeführt.

In anderen Anwendungsfällen, etwa wie bei der Max Planck Digital Library (MPDL)[27] werden zusätzliche Metadaten direkt im PID benötigt. Dazu werden Dublin Core Metadaten[28] direkt in den PID aufgenommen. Weitere Metadaten können auch als zusätzliche Referenz im PID hinterlegt werden. EPIC empfiehlt hierbei, den PID möglichst frei von semantischen Informationen zu halten und die Metadaten nur als Referenz zu hinterlegen.

Das Deutsche Klimarechenzentrum nutzt PIDs zur Verwaltung von Klimadaten. Diese werden zum Aufbau von Kollektionen benutzt. Da Daten in vielen Projekten, wie etwa in dem Coupled Model Intercomparison Project (CMIP5)[29], in unterschiedlichen Systemen weiter verwendet werden, können PIDs als systemübergreifende Lösung zur Identifizierung genutzt werden. Dadurch vermeidet man Inkonsistenzen, welche sich bei dem Zusammenspiel von unterschiedlichen Systemen ergeben können. Im Collaborative Climate Community Data and Processing Grid (C3Grid)[30] nutzt das DKRZ PIDs, um Daten und Metadaten in einem Objekt zusammenzuführen und so dauerhaft zu verknüpfen. Auch in EUDAT werden EPIC PIDs auf ähnliche Weise verwendet.

Innerhalb des Infrastrukturprojekts DARIAH werden PIDs ebenfalls an verschiedensten Stellen eingesetzt. Es werden Objekte im Kontext der Langzeitarchivierung, aber auch in anderen Diensten, wie in den DARIAH Entwicklungen Collection Registry oder Schema Registry, durch EPIC PIDs identifiziert.[31]

27 http://www.mpdl.mpg.de

28 Das Dublin Core Metadatenschema ist eine Metadatenkonvention für ein Minimalset an Metadaten.

29 http://cmip-pcmdi.llnl.gov/cmip5/

30 https://verc.enes.org/c3web

31 In Tonne et al. (2012) wird eine Data Federation für Geisteswissenschaftler im Detail beschrieben.

Ein Vorteil der EPIC PIDs ist es, dass auch Anwendungsfälle ermöglicht werden, bei denen die Identifikatoren länger existieren können als die Daten selbst. Sollten die Daten obsolet werden, kann dies im PID hinterlegt werden.

Ein Beispiel hierfür ist die Windkanal Strömungsforschung im Max-Planck-Institut für Dynamik und Selbstorganisation. Dort werden eine große Anzahl von experimentellen Daten erzeugt. Erst im späteren Verlauf stellt sich heraus, welche Daten bestehen bleiben. Doch werden bereits zur Erzeugung PIDs zur Referenzierung benötigt. Aus diesem Grund ist es wichtig, dass Daten auch nach der Referenzierung entfernt werden können.

Die Beispiele zeigen, dass die konsortielle Struktur und aufgebaute technische Infrastruktur produktiv und effektiv funktionieren und nachhaltig einen PID-Service für die Wissenschaft bereitstellen können.

Zusammenfassung und Ausblicke

PIDs werden in den verschiedensten Arbeitsabläufen innerhalb der Forschung und Wissenschaft benötigt und sollten Teil der Strategie zur Langzeitarchivierung und des nachhaltigen Umgangs mit Forschungsdaten sein. Mit EPIC ist ein Konsortium gegeben, welches eine Antwort auf die organisatorischen Herausforderungen bietet. Dabei stellt EPIC den europäischen Forschern einen hochverfügbaren PID-Service und eine Infrastruktur zur Auflösung von PIDs bereit. Durch den PID-Service wird auch ein Dienst, welcher einen neuen REST-basierten Zugang zu PIDs bietet, zur Verfügung gestellt.

Die technische Infrastruktur und der PID-Service werden aktiv gepflegt, um technischen Entwicklungen standzuhalten und weitere, von den Anwendern gewünschte Funktionen zu integrieren. Hierzu werden zurzeit weitere Funktionen geplant, wie etwa die Batch Verarbeitung von PIDs, eine föderierte Authentifizierungs- und Autorisierungsstruktur und die Verwaltung von replizierten Objekten, die durch mehrere Adressen referenziert werden.

Um die Handle Infrastruktur auf internationaler Ebene unabhängiger zu gestalten, wird das Handle System fortan von International Telecommunication Union Telecommunication Standardization Sector (ITU-T)[32] übernommen. EPIC wird sich hier aktiv einbringen und strebt eine feste Position in den aufzubauenden Gremien an. Im Hinblick auf mögliche Erweiterungen des EPIC Konsortiums, werden die organisatorischen Konzepte evaluiert und entsprechend angepasst, um auch den Anforderungen eines wachsenden Konsortiums gerecht zu werden.

32 http://www.itu.int/ITU-T/

Literatur

Brase, J. (2009): DataCite – A Global Registration Agency for Research Data. In: Proceedings of the 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology (COINFO ,09). Washington: IEEE Computer Society. DOI: 10.1109/COINFO.2009.66

Broeder, D./Dreyer, M./Kemps-Snijders, M./Witt, A./Kupietz, M. and Wittenburg, P. (2008): Deliverable D2.2: Persistent and Unique Identifiers (CLARIN-2008-2). http://www.clarin.eu/files/wg2-2-pid-doc-v4.pdf

Deutsche Forschungsgemeinschaft (1998): Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft" – Vorschläge zur Sicherung guter wissenschaftlicher Praxis. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf ISBN: 3-527-27212-7

Fielding, R.T. (2000): Architectural Styles and the Design of Network-based Software Architectures. Doctoral Dissertation: University of California. ISBN:0-599-87118-0

Kahn, R. and Wilensky, R. (2006): A Framework for Distributed Digital Object Services. International Journal on Digital Libraries 6. DOI: 10.1007/s00799-005-0128-x

Neuroth, H./Oßwald, A./Scheffel, R./Strathmann, S. und Huth, K. (2010): nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.0. 2009. http://nestor.sub.uni-goettingen.de/handbuch/ ISBN: 978-3-940317-48-3

Paskin, N. (2010): Digital Object Identifier (DOI) System. In: Encyclopedia of Library and Information Sciences, 3rd Edition, 1586-1592. ISBN: 978-0-8493-9712-7

Tonne, D./Stotzka, R./Jejkal, T./Hartmann, V./Pasic, H./Rapp, A./Vanscheidt, P./Neumair, B./Streit, A./Garcia, A./Kurzawe, D./Kalman, T./Bribian, B.S. and Rybicki, J. (2012): A Federated Data Zone for the Arts and Humanities. In: Proceedings of the 20th International Euromicro Conference on Parallel, Distributed, and Network-Based Processing, 189-207.

Váradi, T./Krauwer, S./Wittenburg, P./Wynne, M. and Koskenniemi, K. (2008): CLARIN: Common Language Resources and Technology Infrastructure. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco. ISBN: 2-9517408-4-0

Internet

Archival Resource Key (ARK) h
ttp://www.cdlib.org/inside/diglib/ark/
Collaborative Climate Community Data and Processing Grid (C3Grid)
https://verc.enes.org/c3web
Common Language Resources and Technology Infrastructure Projekt (CLARIN)
http://www.clarin.eu/
Corporation for National Research Initiatives (CNRI)
http://www.cnri.reston.va.us/
Coupled Model Intercomparison Project Phase 5 (CMIP5)
http://cmip-pcmdi.llnl.gov/cmip5/
DataCite
http://www.datacite.org/
Deutsches Klimarechenzentrum (DKRZ)
http://www.dkrz.de/
Digital Object Identifier (DOI) System
http://www.doi.org/
Digital Research Infrastructure for the Arts and Humanities Projekt (DARIAH)
http://www.de.dariah.eu/
European Data Infrastructure (EUDAT)
http://www.eudat.eu/
European Persistent Identifier Consortium (EPIC)
http://www.pidconsortium.eu/
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
http://www.gwdg.de/
The Handle System
http://www.handle.net/
The Handle System, Current Applications
http://www.handle.net/apps.html
Handle System Namespace and Service Definition (RFC 3651)
http://www.handle.net/rfc/rfc3651.html
Handle System Overview (RFC 3650)
http://www.handle.net/rfc/rfc3650.html
Handle System Protocol (ver 2.1) Specification (RFC 3652)
http://www.handle.net/rfc/rfc3652.html
International DOI Foundation
http://www.doi.org

International Telecommunication Union Telecommunication Standardization Sector
http://www.itu.int/ITU-T/
IT Center for Science Ltd. (CSC)
http://www.csc.fi/english/
Max Planck Digital Library (MPDL)
http://www.mpdl.mpg.de/
Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)
http://www.sub.uni-goettingen.de/
Persistent Uniform Resource Locator (PURL)
http://purl.oclc.org/docs/index.html
PID-Service der GWDG
http://handle.gwdg.de:8080/pidservice/
Stichting Academisch Rekencentrum Amsterdam (SARA)
http://www.sara.nl/
TextGrid – Virtuelle Forschungsumgebung für die Geisteswissenschaften
http://www.textgrid.de/
Uniform Resource Identifier (RFC 3896)
http://www.ietf.org/rfc/rfc3986.txt/
Uniform Resource Name (URN) Syntax
http://tools.ietf.org/html/rfc2141

C LANGZEITARCHIVIERUNG IN DER PRAXIS

C

Forschungsdaten in den Geowissenschaften

Jens Klump

Einleitung

Forschungsdaten aus den Geowissenschaften sind so vielfältig wie das Forschungsgebiet selbst. Die Geowissenschaften betrachten die feste Erde und die Prozesse an ihrer Oberfläche. Diese Vorgänge stehen gleichzeitig auch in Wechselwirkung mit Vorgängen in der Biosphäre, Atmosphäre, Hydrosphäre und Kryosphäre. Die Erforschung des gesamten Planeten, seiner Entstehung, seines Trabanten und seiner Nachbarn erstreckt sich über ein räumlich und zeitlich sehr ausgedehntes Forschungsgebiet. Ein Forschungsgebiet mit solch einem „universellen" Anspruch, wie es hier der Fall ist, weist viele Überschneidungen mit anderen Disziplinen auf. Ein Kapitel über Forschungsdaten in den Geowissenschaften kann daher nur exemplarisch einige Bereiche beleuchten.

Forschung in den Geowissenschaften ist in vielen Fällen dadurch gekennzeichnet, dass die untersuchten Phänomene episodischer Natur sind. Das Phänomen, zum Beispiel ein Erdbeben, ist nur in dem einen Moment messbar und dann nie wieder. Oder das „Experiment" hat in der Natur bereits stattgefunden und muss von den Forschern im Gelände gefunden, untersucht und dort Proben genommen werden. Diese Orte sind oft in schwer zugänglichen Regionen der Erde und müssen durch Expeditionen oder wissenschaftliche Tiefbohrungen erschlossen werden. Der große Aufwand, der notwendig ist, um die benötigten Daten zu gewinnen, führt dazu, dass auch diese Daten sehr wertvoll sind, selbst wenn sie prinzipiell ein weiteres Mal gewonnen werden könnten.

Aus diesem Grund besteht für viele Daten ein großes Interesse an einer späteren Nutzung (Pfeiffenberger 2007). Die hohen Kosten der Datenerhebung, zum Beispiel in der Geophysik, machen manchen älteren Datenbestand für die Neuprozessierung interessant. Dies ist allerdings nicht allgemein gültig. Insbesondere im Bereich geochemischer Analytik sind die Fortschritte so immens, dass hier ältere Datenbestände nur noch wenig nachgenutzt werden. Dieses Muster findet sich auch in anderen Disziplinen wieder (Severiens und Hilf 2006). Ältere, bisher unerschlossene Datenbestände werden nur in seltenen Fällen für die Nachnutzung erschlossen, da der hiermit verbundene Aufwand sehr hoch ist. Der Wiederbeschaffungswert der Daten lässt sich im akademischen Bereich nur grob schätzen, er dürfte jedoch bei Großforschungseinrichtungen im Bereich einiger Millionen Euro pro Jahr und Institution liegen. Allein der Wiederbeschaffungs-

wert von Erkundungsdaten in der Rohstoffindustrie wird auf einige Milliarden Euro pro Firmenarchiv geschätzt (Hawtin und Lecore 2011).

Abgesehen von proprietären Daten, zum Beispiel aus der Erkundung von Rohstofflagerstätten, unterliegen Forschungsdaten in den Geowissenschaften nur in seltenen Fällen Zugangs- oder Nutzungsbeschränkungen. Sollten Beschränkungen vorliegen, so sind diese meist durch die Nutzungsvereinbarungen mit den Datenproduzenten bestimmt, nicht durch gesetzliche Vorgaben. In den meisten Verbundprojekten werden inzwischen unter den Projektpartnern Vereinbarungen über den Umgang mit Daten getroffen. Dabei wird anerkannt, dass es unter Forschern einen starken *sense of ownership* in Bezug auf Forschungsdaten gibt, auch wenn dieser urheberrechtlich strittig ist (Tenopir et al. 2011). Aus diesem Grund wird den Forschern meist eine Frist von bis zu zwei Jahren nach Projektende für ausschließliche Nutzung der Daten eingeräumt. Einige Projekte geben ihre Daten bereits schon zur Laufzeit des Projekts zur Nutzung durch Dritte frei (Pfeiffenberger und Klump 2006). Unabhängig von der unklaren Urheberrechtssituation für Forschungsdaten besteht die Möglichkeit, diese Daten mit einer Lizenz zu versehen (Ball 2011). Seitens der Institutionen wird in den Geowissenschaften der offene Zugang zu wissenschaftlichem Wissen im Sinne der *„Berliner Erklärung"* (Open Access) (Berlin Declaration 2003) nachdrücklich unterstützt (Pampel 2009). Ein erwünschter Nebeneffekt ist auch, dass Veröffentlichungen, deren Daten zugänglich sind, signifikant häufiger zitiert werden als Veröffentlichungen, bei denen dies nicht der Fall ist (Sears 2011).

Es sollte noch erwähnt werden, dass es auch technische Gründe gibt, aus denen der direkte Zugriff auf die Daten gesperrt wird. Bei sehr großen Datensätzen – mehrere Gigabyte und größer – kann der Vertrieb aus technischen Gründen nicht unmittelbar durch einen Zugriff des Nutzers auf die Daten über das Internet erfolgen.

Typen der Datenherkunft

Grundsätzlich lassen sich in den Geowissenschaften drei Typen von Datenproduktion unterscheiden, die in ihren Datenvolumina und -strukturen stark voneinander abweichen:

- Daten aus Sensorsystemen, Dateninfrastrukturen und Großinstrumenten mit automatisierter Prozessierung
- Daten aus numerischer Modellierung
- individuell hergestellte Datensätze aus Labordaten, Felderhebungen und Literaturrecherche.

Im Bereich der Großgeräte und Sensorsysteme fallen zum Teil Datenmengen von mehreren Terabyte pro Jahr an. Die Daten werden in automatisierten Abläufen weiterverarbeitet und bereitgestellt. Die Strukturen dieser Datenbestände sind in sich meist homogen mit standardisierten Daten- und Metadatenformaten. Im Bereich der numerischen Modellierung sind die Datenstrukturen und -volumina ähnlich wie im vorgenannten Bereich, jedoch werden die Arbeitsabläufe zurzeit erst ungenügend durch automatisierte Workflows unterstützt. Im Bereich der Großgeräte und Sensorsysteme liegen Forschungsdaten und Metadaten im Allgemeinen in standardisierten Formaten vor. Die semantisch homogenen Datenstrukturen im Bereich der Großgeräte, Sensorsysteme und Modellierung begünstigen die Verwendung von standardisierten Formaten.

Der Umgang mit Daten aus Großgeräten, Sensorsystemen und numerischen Modellen ähnelt damit anderen Bereichen des Big Data Science, wie zum Beispiel der Teilchenphysik. Wie in anderen Bereichen des Big Data Science wächst die Kapazität zur Erzeugung neuer Daten schneller als die Möglichkeit, diese längerfristig zu speichern. Aus diesem Grund sind hier auch die ersten Ansätze für die Entwicklung von Regelungen und technischen Verfahren zur Langzeiterhaltung dieser Datenbestände zu beobachten.

Im Bereich der individuell hergestellten Forschungsdaten fallen nur vergleichsweise geringe Datenmengen an, dafür sind die Stückkosten zur Herstellung der Datensätze sehr hoch. Die Datenstrukturen, Metadaten und Arbeitsabläufe orientieren sich an den individuellen Anforderungen der Projekte. Standardisierte Datenformate finden daher kaum Anwendung, da die heterogenen Projekte untereinander semantisch inhomogene Strukturen bedingen. Die erwarteten Steigerungsraten sind geringer als die Zunahme der Kapazität der Speichermedien.

Betrachtet man die Kosten pro Einheit, die bei der Erstellung von Datensätzen mit geringen Volumen anfallen, also gewissermaßen die „Lohnstückkosten", so sind diese Kosten enorm hoch. Hier finden sich, neben manuell erhobenen Daten, auch die Ergebnisse von Datensynthesen wieder, weshalb diesem sogenannten *Long-tail* der Forschungsdaten eine besondere Bedeutung beigemessen wird. Diese Daten und ihre semantische Heterogenität zu erfassen erfordert neue Werkzeuge, die sich möglichst nahtlos in die Arbeitsabläufe der Forschung einfügen, so dass die Überführung der Daten in stabile und nachnutzbare Formen die Wissenschaftler nicht mit zusätzlichen Aufgaben belastet (Feijen 2011; Klump und Ulbricht 2011; Razum et al. 2009).

Datenmanagementpläne sind bisher wenig verbreitet. In Projekten mit großen Datenmengen ist die Notwendigkeit eines Datenmanagementplans evident und wird daher bereits in der Antragsphase berücksichtigt. Auch in großen Verbundprojekten gibt es ein systematisches Datenmanagement, oft flankiert von

einer Vereinbarung zwischen den Projektteilnehmern über den Umgang mit im Projekt gewonnenen Daten. In der Mehrzahl der Projekte mit kleinen Datenmengen dagegen ist systematisches Datenmanagement bisher noch nicht verbreitet.

Erst wenn die Daten in eine archivierbare Form gebracht und mit Metadaten versehen wurden, wird eine langfristige Archivierung der Daten möglich. Dieser Schritt, die Überführung von Daten in eine archivierbare Form, ist der aufwendigste, teuerste und riskanteste im gesamten Lebenszyklus von Forschungsdaten (Beagrie et al. 2010; Digital Preservation Testbed 2005). Auf der Seite der Datenzentren besteht das Problem, dass diese meist immer noch als „Silo" angelegt sind, d.h. der Inhalt der Systeme ist nicht über automatisierte Verfahren zugänglich und auch die Überführung von Daten in diese Systeme erfordert immer noch viele manuelle Eingriffe. Dies ist insbesondere bei der Überführung von Forschungsdaten in ein Datenarchiv problematisch, weil Medienbrüche stets eine Hürde im Lebenszyklus der Daten darstellen, an denen das Risiko besonders hoch ist, dass die Kette der Bearbeitungsschritte im Datenlebenszyklus abreißt.

Nachnutzung und Datenveröffentlichung

Der große Aufwand zur Gewinnung der Daten und die oft episodische Natur der beobachteten Phänomene machen die gewonnenen Daten für die gemeinsame Nutzung in Forschungsverbünden und für eine spätere Nachnutzung wertvoll. Schon sehr früh wurde erkannt, dass Strukturen für die Bereitstellung und Langzeiterhaltung der Daten notwendig sind. Bereits für das Internationale Geophysikalische Jahr (IGY, 1957 bis 1958) wurden die World Data Center (WDC) eingerichtet, um die Daten des IGY bereitzustellen und zu erhalten (Dittert et al. 2001; Pfeiffenberger 2007).

Um Forschungsdaten aus den Geowissenschaften nachnutzbar zu machen, wurden seit den 1990er Jahren, zusätzlich zum eher als Archiv konzipierten WDC-System, zentrale disziplinäre Datenportale aufgebaut. Eine wichtige Rolle in den beiden bereits skizzierten Komplexen der WDC und der Meeresgeologie spielt PANGAEA/WDC-MARE, das Mitte der 1990er Jahre in einer Zusammenarbeit zwischen dem Alfred-Wegener-Institut für Polar- und Meeresforschung und dem Fachbereich Geologie der Universität Bremen aufgebaut wurde (Diepenbroek et al. 2002). Seit 2001 hat der offen zugängliche Teil des PANGAEA den Status eines ICSU WDC als World Data Center for Marine and Environmental Sciences (WDC-MARE).

Auch andere deutsche geowissenschaftliche Großforschungsreinrichtungen, wie zum Beispiel das Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum (GFZ) und das Helmholtz-Zentrum für Ozeanforschung GEOMAR betreiben Datenarchive und -portale und bieten diese Forschern aus anderen Institutionen

zur Nutzung an. Zwischen den deutschen ICSU WDC sowie dem GFZ und dem GEOMAR besteht eine enge Zusammenarbeit bei der Entwicklung von Werkzeugen und Diensten für Erfassung, Archivierung, Nachweis und Vertrieb von Forschungsdaten (Lautenschlager et al. 2005). Dieser Verbund von Datenzentren hat das Potenzial als künftiger nationaler Dienstleister für die Langzeitarchivierung geowissenschaftlicher Forschungsdaten.

In den letzten Jahren wurde das System der ICSU WDC modernisiert und in das World Data System (WDS) überführt. Ein wichtiger Aspekt des WDS ist die Einführung einer Akkreditierung der beteiligten Datenzentren anhand eines Kriterienkatalogs für die Vertrauenswürdigkeit für digitale Archive (Dittert et al. 2007). Vergleichbare Kriterienkataloge werden auch in anderen Bereichen, wie zum Beispiel der Erdbeobachtung angewendet (Albani et al. 2010). Auch wenn die Ausprägungen dieser Kriterienkataloge im Detail voneinander abweichen, so bauen sie alle auf den selben Prinzipien auf (Klump 2011).

Im Jahr 2000 wurde im Rahmen des deutschen Beitrags zu CODATA ein Projekt initiiert, das ein Konzept erarbeiten sollte, wie wissenschaftliche Daten publiziert und damit zitierbar gemacht werden können. Im Rahmen des DFG-Projekts *„Publikation und Zitierbarkeit wissenschaftlicher Primärdaten“* (STD-DOI) wurden ein Konzept und eine technische Infrastruktur aufgebaut, um Daten mittels Digital Object Identifier (DOI) eindeutig identifizierbar und damit auch zitierbar zu machen (Brase und Klump 2007). Aus diesem Projekt ist 2009 DataCite e.V. hervorgegangen, das als Verbund von Großbibliotheken die Strukturen für die Veröffentlichung und Zitierbarkeit von Forschungsdaten betreibt und weiterentwickelt.

Mit der Entwicklung eines Systems für die Publikation und Zitierbarkeit von Forschungsdaten stellte sich die Frage, wie die Qualität der veröffentlichten Daten geprüft werden kann. Inzwischen haben einige Fachzeitschriften Kriterien für die Bewertung der inhaltlichen Qualität von Daten und für Peer-Review Verfahren formuliert (Dallmeier-Thiessen et al. 2010; Pfeiffenberger und Carlson 2011). Für einzelne Fälle, in denen einheitliche Konzepte und standardisierte Datenformate existieren, wurden bereits Verfahren zur Prüfung der Konsistenz und Struktur der Daten entwickelt (DIN 2007; IUCr 2008).

Wie bereits erwähnt, werden Datenveröffentlichungen zunehmend auch in Bibliothekskatalogen nachgewiesen. Zusätzlich entwickelt sich auch eine direkte Zusammenarbeit zwischen Datenzentren und Verlagen. So wird in ScienceDirect (Elsevier) angezeigt, ob in PANGAEA/WDC-MARE Daten zu dem in ScienceDirect angezeigten Artikel vorgehalten werden. Zusätzlich werden in einer eingebetteten Landkarte die Orte angezeigt, an denen die beschriebenen Proben gewonnen wurden (Abbildung 1).

Zwischen geowissenschaftlichen Datenzentren, Herausgebern von Fachzeitschriften und Verlagen finden regelmäßig Treffen statt, um sich über die Anfor-

derungen und Vorgehensweise abzustimmen. Auf gemeinsamen Veranstaltungen auf internationalen Konferenzen werden die Konzepte und Angebote zur Veröffentlichung von Forschungsdaten und deren Verknüpfung mit wissenschaftlichen Veröffentlichungen den Fachwissenschaftlern vorgestellt.

Persistente Identifikatoren werden jedoch nicht nur für Daten eingesetzt. Auch Probenstücke werden in Sammlungen der Institute, Museen und staatlichen geologischen Dienste archiviert. In diesen Sammlungen werden die originalen Proben als Referenzmaterialien für mögliche weitere Untersuchungen aufbewahrt. Auf Grund des schnellen methodischen Fortschritts werden insbesondere in der Geochemie an Stelle der Daten oft die Probenstücke selber nachgenutzt, um sie mit neuen oder verbesserten Methoden erneut zu bearbeiten. Ähnlich wie bei Daten wurde auch bei Probenstücken festgestellt, dass eine eindeutige Identifizierbarkeit der Stücke notwendig ist. An einzelnen besonders wertvollen Probenstücken werden über Jahre hinweg immer neue Analysen gemacht und veröffentlicht. Bisher gab es in den meisten Fällen keine international gültige Namenskonvention. Zudem wurde in vielen Projekten eine bereits verabschiedete Namenskonvention später nicht durchgehend eingehalten. Dies führte teilweise zu einer Verwirrung bei den Probenbezeichnungen, so dass sich Daten, die in der Literatur zu bestimmten Stücken veröffentlicht wurden, nicht mehr eindeutig zu Proben zuordnen und damit auch nicht in größer angelegten Studien integrieren lassen.

Um die Anforderung einer eindeutigen Benennung von Proben zu lösen, wurde vorgeschlagen, eine International Geo Sample Number (IGSN) einzuführen (Lehnert und Klump 2008). Auf dem American Geophysical Union Fall Meeting

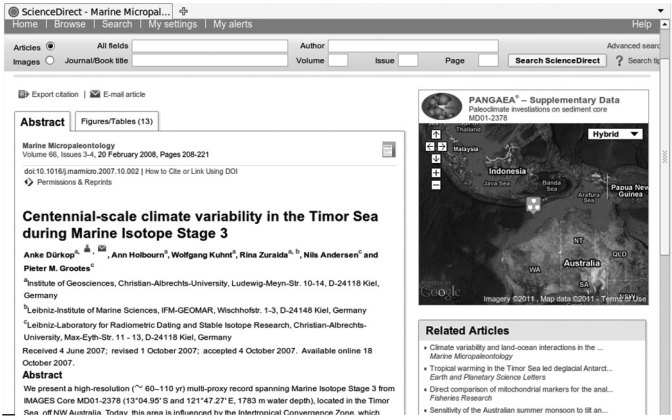


Abbildung 1: Datenveröffentlichungen in WDC-MARE/PANGAEA werden in ScienceDirect (Elsevier) zusammen mit dem jeweiligen Artikel angezeigt. Die eingeblendete Landkarte zeigt den Ort der Probennahme und verknüpft den Artikel mit der Datenveröffentlichung.

2011 in San Francisco, Kalifornien, wurde im Dezember 2011 ein Trägerverein für die Einführung der IGSN nach dem Vorbild des DataCite e.V. gegründet; die zur Registrierung und globalen Auflösung von IGSN notwendige technische Infrastruktur ist inzwischen in Betrieb gegangen.

Die schnelle Entwicklung der Rechentechnik, Betriebssysteme und Software führt zu neuen Herausforderungen an die Reproduzierbarkeit von Forschungsergebnissen. Schon eine Änderung der Rundungsalgorithmen beim Wechsel von einer Rechenplattform zu einer anderen führt zum Teil zu signifikant abweichenden Ergebnissen (Ince et al. 2012; Peng 2011). Es entwickeln sich daher Bestrebungen, auch für Software Verfahren zur eindeutigen Identifikation und Zitierbarkeit zu entwickeln.

Ausblick

Die zunehmende Vernetzung von interdisziplinären Arbeitsgruppen erfordert eine intensivere Vernetzung von Daten und Forschungsdaten-Infrastrukturen. Der wissenschaftliche Erkenntnisgewinn wird auch in den Geowissenschaften zunehmend durch Forschungsansätze ergänzt, die sich auf die Analyse bereits vorhandener Daten stützen. Als Ergänzung zu Empirie, Theorie und Simulation wird diese Vorgehensweise *Data intensive science* oder auch das „*Vierte Paradigma*" der Forschung genannt (Hey et al. 2009; McNally et al. 2012).

Gemessen an anderen Disziplinen ist der Umgang mit Forschungsdaten und deren Langzeitarchivierung in Teilen der Geowissenschaften bereits weit entwickelt. Und dennoch wird auch hier erst ein kleiner Teil der Daten in langfristige Strukturen überführt. Denn auch wenn sich die Publikation von Daten allmählich als anerkanntes Verfahren durchsetzt, so scheuen noch viele Wissenschaftler den vermuteten Aufwand, Daten für eine Langzeitarchivierung, und gegebenenfalls Veröffentlichung, aufzubereiten. Die bereits vollzogenen Änderungen in der Förderpolitik der DFG und der Europäischen Kommission haben mit dazu beigetragen, ein Umdenken über den Wert von Daten einzuleiten. Dennoch wird die Verpflichtung zum Datenmanagement deutlich als Last wahrgenommen (Feijen 2011).

Hemmnisse bei der Umsetzung einer Strategie zur Langzeitarchivierung von Forschungsdaten sind meist die fehlenden organisatorischen und technischen Strukturen. Es fehlen Ansprechpartner in den Institutionen und Werkzeuge, die den datenkuratorischen Prozess unterstützen. Insbesondere für das Datenmanagement in Projekten müssen weiter Werkzeuge und Konzepte entwickelt werden, die eine nahtlose Integration der datenkuratorischen Aufgaben in die Arbeitsabläufe der Forschung ermöglichen.

In ihren aktuellen Verwendungsrichtlinien verlangt die DFG von ihren Antragstellern einen Datenmanagementplan.

> *„Wenn aus Projektmitteln systematisch (Mess-)Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch – sofern vorhanden – die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien." (DFG 2010)*

Um diesen Anspruch zu erfüllen, müssen in den nächsten Jahren auch von den Datenarchiven neue Angebote entwickelt werden. Diese müssen von Geschäftsmodellen flankiert sein, die es erlauben, diese Dienstleistungen mit den Projekten abzurechnen. Aktuell fällt es den Datenzentren noch schwer, die Kosten der Langzeitarchivierung von Forschungsdaten zu beziffern. Bei allen Fortschritten, die in den vergangenen Jahren zu verzeichnen waren, müssen integriertes Datenmanagement und die Langzeitarchivierung von Forschungsdaten erst noch Bestandteil des wissenschaftlichen Alltags und des wissenschaftlichen Wertesystems werden. Wir arbeiten daran.

Literatur

Albani, M./Beruti, V./Duplaa, M./Giguere, C./Velarde, C./Mikusch, E./ Serra, M./Klump, J. and Schroeder, M. (2010): Long term preservation of earth observation space data – European LTDP Common Guidelines (Version 1.1). Frascati, Italien: European Space Agency. Ground Segment Coordination Body. http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf

Ball, A. (2011): How to License Research Data. JISC How-to Guides. Edinburgh, Großbritannien: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/license-research-data

Beagrie, N./Lavoie, B.F. and Woollard, M. (2010): Keeping research data safe 2. Bristol, Großbritannien: Joint Information Systems Committee (JISC). http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx

Berlin Declaration (2003): Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/

Brase, J. und Klump, J. (2007): Zitierfähige Datensätze: Primärdaten-Management durch DOIs. In: WissKom 2007: Wissenschaftskommunikation der Zukunft. 4. Konferenz der Zentralbibliothek. Forschungszentrum Jülich. 6.-8. November 2007. Bd. 18. Herausgegeben von R. Ball. Jülich: Forschungszentrum Jülich, 159-167. http://edoc.gfz-potsdam.de/gfz/10493

Dallmeier-Thiessen, S./Pfeiffenberger, H./Puschmann, C. and Stein, D. (2010): Peer reviewed data publication in earth system sciences. In: Towards Open Access Scholarship: Selected Papers from the Berlin 6 Conference. Düsseldorf: düsseldorf university press, 77-84. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hbz:061-20100722-142254-7

DFG (2010): Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für die Antragstellung und ergänzenden Leitfäden für die Antragstellung für Projekte mit Verwertungspotenzial, für die Antragstellung für Projekte im Rahmen einer Kooperation mit Entwicklungsländern. Bonn: Deutsche Forschungsgemeinschaft (DFG). http://www.dfg.de/download/formulare/1_02/1_02.pdf

Diepenbroek, M./Grobe, H./Reinke, M./Schindler, U./Schlitzer R./Sieger, R. and Wefer, G. (2002): PANGAEA – an information system for environmental sciences. Computers & Geosciences 28 (10), 1201-1210. doi:10.1016/S0098-3004(02)00039-0.

Digital Preservation Testbed (2005): Costs of Digital Preservation. From digital volatility to digital permanence. Den Haag, The Netherlands: Nationaal Archief of the Netherlands. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf

DIN (2007): Qualitätsmodell für die Beschreibung von Geodaten. PAS 1071:2004-10. Berlin: Deutsches Institut für Normung.

Dittert, N./Diepenbroek, M. and Grobe, H. (2001): Scientific data must be made available to all. Nature 414 (6862), 393. doi:10.1038/35106716.

Dittert, N./Diepenbroek, M. and Grobe, H. (2007): Toward a Networked Publication and Library System for Scientific Data. EOS, Transactions, American Geophysical Union 88 (48), 525.

Feijen, M. (2011): What researchers want. Utrecht, The Netherlands: SURFfoundation. http://www.surffoundation.nl/en/publicaties/Pages/Whatresearcherswant.aspx

Hawtin, S. and Lecore, D. (2011): The business value case for data management – a study. London, United Kingdom: Common Data Access Ltd.
http://www.oilandgasuk.co.uk/datamanagementvaluestudy/

Hey, T./Tansley, S. and Tolle, K. (eds.) (2009): The Fourth Paradigm: Data-Intensive Scientific Discovery. 1.1 ed. Redmond, WA: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Ince, D.C./Hatton, L. and Graham-Cumming, J. (2012): The case for open computer programs. Nature 482 (7386), 485-488. doi:10.1038/nature10836.

IUCr (2008): Notes for authors 2008. Acta Crystallographica Section C 64 (1), e2-e9. doi:10.1107/S0108270107067698.

Klump, J. (2011): Criteria for the Trustworthiness of Data Centres. D-Lib Magazine 17 (1/2). doi:10.1045/january2011-klump.

Klump, J. and Ulbricht, D. (2011): PanMetaDocs – A tool for collecting and managing the long tail of „small science data". EOS, Transactions, American Geophysical Union 92 (53, Fall Meet. Suppl.), IN23C-1461.

Lautenschlager, M./Diepenbroek, M./Grobe, H./Klump, J. and Paliouras, E. (2005): World Data Center Cluster „Earth System Research" – An Approach for a Common Data Infrastructure in Geosciences. EOS, Transactions, American Geophysical Union 86 (52, Fall Meet. Suppl.), IN43C-02.

Lehnert, K. and Klump, J. (2008): Facilitating Research in Mantle Petrology with Geoinformatics. 9th International Kimberlite Conference 9IKC. Frankfurt (M). http://www.cosis.net/abstracts/9IKC/00250/9IKC-A-00250-1.pdf

McNally, R./Mackenzie, A./Hui, A. and Tomomitsu, J. (2012): Understanding the 'Intensive' in 'Data Intensive Research': Data Flows in Next Generation Sequencing and Environmental Networked Sensors. IJDC 7 (1), 81-94. doi:10.2218/ijdc.v7i1.216.

Pampel, H. (2009): Bericht über den Workshop „Offener Zugang zu Daten – eine Herausforderung" im Rahmen der Open-Access-Tage 2008 am 10.10.2008 in Berlin. LIBREAS. Library Ideas 14 (1). http://www.libreas.eu/ausgabe14/017pam.htm

Peng, R.D. (2011): Reproducible Research in Computational Science. Science 334 (6060), 1226-1227. doi:10.1126/science.1213847.

Pfeiffenberger, H. (2007): Offener Zugang zu wissenschaftlichen Primärdaten. Zeitschrift für Bibliothekswesen und Bibliographie 54 (4-5), 207-210.

Pfeiffenberger, H. and Carlson, D. (2011): „Earth System Science Data" (ESSD) – A Peer Reviewed Journal for Publication of Data. D-lib 17 (1-2). doi:10.1045/january2011-pfeiffenberger.

Pfeiffenberger, H. und Klump, J. (2006): Offener Zugang zu Daten – Quantensprung in der Kooperation. Wissenschaftsmanagement, (Special 1), 12-13.

Razum, M./Schwichtenberg, F./Wagner, S. and Hoppe, M. (2009): eSciDoc Infrastructure: A Fedora-Based e-Research Framework. In: Research and Advanced Technology for Digital Libraries. Bd. 5714. Heidelberg: Springer Verlag, 227-238. http://dx.doi.org/10.1007/978-3-642-04346-8_23

Sears, J.R. (2011): Data Sharing Effect on Article Citation Rate in Paleoceanography. EOS, Transactions, American Geophysical Union 92 (53, Fall Meet. Supp.), IN53B-1628.

Severiens, T. und Hilf, E.R. (2006): Langzeitarchivierung von Rohdaten. nestor-Materialien 6. Frankfurt (M): nestor - Kompetenznetzwerk Langzeitarchivierung. http://nbn-resolving.de/urn:nbn:de:0008-20051114018

Tenopir, C./Allard, S./Douglass, K./Aydinoglu, A.U./Wu, L./ Read, E./ Manoff, M. and Frame, M. (2011): Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6 (6), e21101. doi:10.1371/journal.pone.0021101.

Langzeitarchivierung am Deutschen Klimarechenzentrum

Hans Luthardt

Einführung

Die Aktivitäten der Klimaforschung in Deutschland haben in den letzten Jahren stark zugenommen und die Entwicklung der Klimamodelle hin zu Erdsystemmodellen (ESM) vorangetrieben. Dank der wachsenden verfügbaren Rechenkapazitäten konnten die Modelle in ihrer Komplexität und Auflösung weiterentwickelt werden und die Klimaprojektionen für verschiedene Szenarien für das nächste Jahrhundert durchgeführt und verbessert werden. Dieses Wachstum an Rechenleistung führt unmittelbar zu einem entsprechenden Anwachsen des erzeugten Datenvolumens. Die wissenschaftliche Auswertung dieser Daten und die Forderung der Geldgeber, diese Daten, insbesondere wenn sie Grundlage von Veröffentlichungen sind, mindestens 10 Jahre aufzubewahren[1], macht im Bereich der Erdsystemforschung ein Langzeitarchivierungs-Angebot für große Datenvolumina erforderlich.

Genau hier liegt eine Aufgabe des Deutschen Klimarechenzentrums (DKRZ)[2] Als überregionale Serviceeinrichtung betreibt es ein Rechenzentrum für die Durchführung von Klimasimulationen. Ferner hält das DKRZ alle für die Verarbeitung und Auswertung einschlägiger Daten notwendigen technischen Einrichtungen vor, pflegt und entwickelt allgemein für die Klimaforschung relevante Anwender-Software, berät und unterstützt seine Nutzer in Fragen der Datenver-

1 Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten (Stand 26.06.2008):
„Forschungsprimärdaten bilden einen wertvollen Fundus an Informationen, die mit hohem finanziellem Aufwand erhoben werden. Je nach Fachgebiet und Methode sind sie replizierbar oder basieren auf nicht wiederholbaren Beobachtungen oder Messungen. In jedem Fall sollten die erhobenen Daten nach Abschluss der Forschungen öffentlich zugänglich und frei verfügbar sein. Dieses ist die wesentliche Voraussetzung dafür, dass Daten im Rahmen neuer Fragestellungen wieder genutzt werden können sowie dafür, dass im Falle von Zweifeln an der Publikation die Daten für die Überprüfung der publizierten Ergebnisse herangezogen werden können."
1997 veröffentlichte die DFG „Vorschläge zur Sicherung guter wissenschaftlicher Praxis" mit 16 Empfehlungen. Die Empfehlung 7 lautet
„Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden." (DFG (2009): Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf)
Die Forderung, Daten verfügbar zu halten, steht für die DFG somit schon seit über 10 Jahren im Raum.
2 http://www.dkrz.de

arbeitung und beteiligt sich an nationalen und internationalen Projekten und Kooperationen mit dem Ziel der Verbesserung der Infrastruktur für die Klimamodellierung. Außerdem bietet das DKRZ einen Langzeitarchivierungsservice für seine Nutzer, aber auch für externe Projekte an.

DKRZ Infrastruktur

Um den Anforderungen der Klima- und Erdsystemforschung in Bezug auf die Modellrechnungen mit Erdsystemmodellen gerecht werden zu können, ist die Infrastruktur des DKRZ mit einem Höchstleistungsrechner und einem Hochleistung-Speichersystem ausgestattet. Es ermöglicht die Bearbeitung komplexer Erdsystemmodelle sowie die Speicherung und das Prozessieren umfangreicher Datenvolumina. Entscheidend für den Bedarf an Rechenleistung und Datenspeicherung sind dabei folgende Faktoren, die kontinuierlich wachsende Anforderungen zur Folge haben:

- Komplexität der Modelle
- Räumliche/zeitliche Auflösung
- Ensemble Rechnungen
- Lange Simulationszeiten
- Anzahl der Parameter/Komplexität der Physik im Modell

Dies führt neben dem wachsenden Bedarf an Rechenkapazität zu einer entsprechenden Nachfrage an Speicherplatz, der entsprechend steigt. Dazu kommen Beobachtungs-Datensätze über klimarelevante Zeiträume, die der Forschung zugänglich gemacht werden sollen. Auch diese können große Volumina umfassen (zum Beispiel Satellitenprodukte).

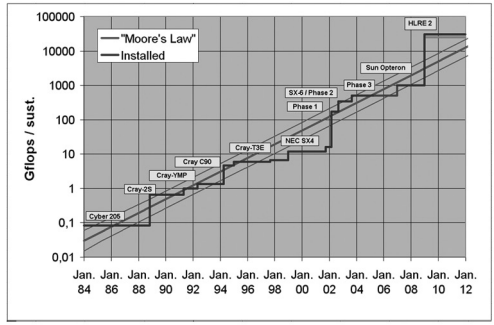


Abbildung 1: Entwicklung der Rechenleistung am DKRZ

Rechenleistung

Die Rechenleistung am DKRZ ist in den letzten dreißig Jahren beständig gewachsen.

Die Rechenkapazität, die den Nutzern des DKRZ zur Verfügung gestellt wird, basiert im Wesentlichen auf einem Höchstleistungsrechner, einem System von Rechnern für das Prozessieren der Daten und einem Graphiksystem zur Visualisierung. Die technischen Daten des aktuellen Höchstleistungsrechners (IBM-Power 6) sind nachfolgend zusammengestellt.

Rechnerhardware:

- 158 Teraflops (158 * 1012 Gleitkommaoperationen / Sekunde)
- 264 IBM Power6-Rechnerknoten
- 16 Dual-Core-Prozessoren pro Knoten (insgesamt 8.448 Kerne)
- Mehr als 20 Terabyte Hauptspeicher
- 7 PetaByte Festplattenspeicher (7 * 1015 Byte)
- Infiniband-Netzwerk mit 7,6 TeraByte/s aggregierter Übertragungsrate

Abbildung 2: Speichersysteme des DKRZ

Datenmanagement

Für die Archivierung der Daten stehen am DKRZ die in Abbildung 2 dargestellten Speichersysteme zur Verfügung.

Die Speicher-Infrastruktur (ohne Plattenspeicher) umfasst gegenwärtig:

- 7 automatische Sun StorageTek SL8500-Bandbibliotheken
- 8 Roboter je Bibliothek
- mehr als 67.000 Stellplätze für Bänder mit Gesamtkapazität von ca. 100 Petabyte
- 88 Bandlaufwerke
- bidirektionale Bandbreite von 5 Gigabyte/s

Die Abteilung Datenmanagement bietet dazu umfangreiche Unterstützung für alle Stadien im Lebenszyklus von Klimadaten[3]

3 http://www.dkrz.de/daten?set_language=de

- Laufzeitumgebung zur Durchführung aufwendiger Modellrechnungen (Integrated Model and Data Infrastructure)
- Unterstützung bei der Durchführung von Rechnungen, die von großen Teilen der Klimaforschungsgemeinschaft konzipiert wurde und genutzt werden soll („Community-Rechnungen")
- Unterstützung virtueller Forschungsumgebungen
- Langzeitarchivierung von Klimadaten (World Data Center Climate, WDC-Climate)
- Unterstützung des redaktionellen Prozesses und Qualitätskontrolle zur Publikation von Klimadaten (DataCite Primärdatenpublikation)[4]
- Erstellung der Metadaten als Voraussetzung der Datenspeicherung in der CERA (Climate and Environmental Archiving and Retrieval) Datenbank
- Internetbasierter Zugriff, Datensuche und interdisziplinärer Datenzugriff (CERA Datenbank)
- Erzeugung von Antriebsdaten für die Modellläufe

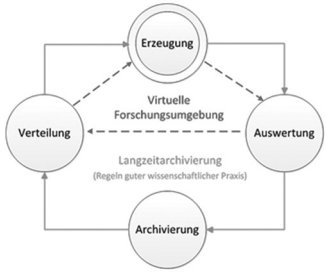


Abbildung 3: DATA Life Cycle für Klima-Modelldaten

Dabei ist von einer Entwicklung des zu speichernden Datenvolumens ausgegangen worden, die sowohl die Entwicklung der verfügbaren Rechenkapazität als auch die zu erwartenden Forschungsaktivitäten berücksichtigt.
Den Verlauf der Speicherung von Daten am DKRZ (gesamtes Speichervolumen) in den letzten Jahren und die abgeschätzte Entwicklung für die nächsten Jahre zeigt Abbildung 4.

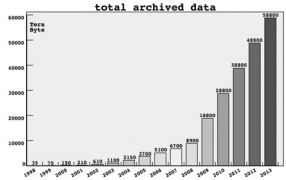


Abbildung 4: Erwartete Entwicklung der Datenspeicherung bis 2013

4 http://www.dkrz.de/daten/Datenpublikation?set_language=de

World Data Centre for Climate (WDC-Climate)

Die überwiegende Anzahl der Datensätze, die vom DKRZ langzeitarchiviert sind, werden im Rahmen des World Data Centre for Climate angeboten und verteilt. Dies bedeutet unter anderem, dass diese Daten frei zugänglich sind und den Regeln der ICSU-WDS (International Council for Science – World Data System)[5] unterliegen.

Für das vom DKRZ betriebene WDC-Climate gilt:

- Start: Eingerichtet im Januar 2003
- Betrieben von: Model and Data (M&D/MPIMET) und DKRZ, ab 2010 nur DKRZ
- Mission: Daten für Klima- und Erdsystemforschung werden gesammelt, gespeichert und verteilt
- ICSU Politik: Langzeitarchivierung (10 Jahre +) und unbegrenzter Zugang für akademische Nutzung, Teile der Daten sind generell zugänglich
- Beschränkung: Nur Klimadatenprodukte in der CERA DB, keine Speicherung von Rohdaten
- Metadaten: Daten werden nur zusammen mit Metadaten gespeichert, basierend auf dem CERA2 Metadaten-Modell
- PID: Möglichkeit einer Veröffentlichung für Primärdaten (DOI)
- Inhalt: Schwerpunkt sind Daten von Klimamodellen und zugehörigen Beobachtungen
- Kooperationen: Mit thematisch ähnlichen WDCs wie WDC-MARE (Bremen) und WDC-RSAT (Oberpfaffenhofen)
- Zugang: Webbasiert (CERA WEB-Portal)
- Personal: 5 Personen
- Datenbestand: Ca. 440 Terabyte (Ende 2011)
- Datenvolumen: 1 PetaByte/Jahr erwartete Zuwachsrate (ab 2011)

Metadaten	
Anzahl der Projekte	*80*
Anzahl der Experimente	*1.438*
Anzahl der Gruppen	*230*
Anzahl der Datensätze	*187.802*
Anzahl der ‚Additional Infos'	*237*

5 http://www.icsu-wds.org

Daten	
Größe der Datenbank in TByte	*434*
Anzahl der Container	*183.038*
Anzahl der Blobs (kleinste Dateneinheit)	*8.586.769.505*

Tabelle 1: Status der CERA Datenbank

Das WDC-Climate ist im Web[6] erreichbar.

Die bisher gespeicherten Klimadaten haben dabei unterschiedliche Strukturen, je nachdem ob sie aus Modellrechnungen (Klimaszenarien, Reanalysen) oder von Beobachtungssystemen (Boden-beobachtungen, Flugzeugmessungen oder Satellitendaten) stammen. Neben unterschiedlichen Strukturen liegen auch verschiedene Dateiformate (zum Beispiel GRIB, netCDF, ASCII) vor.

Der überwiegende Teil der Datensätze stammt aus Modellläufen mit numerischen Klimamodellen, enthält aber auch zahlreiche andere Daten, die für die Erdsystemforschung relevant sind und die öffentlich zugänglich sind. Beispiele hierfür sind:

- Ergebnisse von globalen und regionalen Klimamodellen verschiedener Zentren für Klimamodellierung (CCCma, CCSR/NIES, CSIRO, GFDL, HADLEY, MPIfM, NCAR) basierend auf IPCC-Emissionsszenarien
- Daten aus wissenschaftlichen Projekten HOAPS (Satellitendaten), CARIBIC (Daten der zivilen Luftfahrt), GOP, COPS
- Modellhafte Beobachtungen
- Reanalysedaten

CERA Datenmodell

Entscheidende Voraussetzung für die semantische Suche nach und das Arbeiten mit den in der CERA-Datenbank abgelegten Daten ist deren umfangreiche Beschreibung. Diese wird für das WDC-Climate in einer Datenbank durch das Metadatenmodell CERA2 (Climate and Environmental Retrieval and Archiving)[7,8] sichergestellt. Dessen Struktur ist in Abbildung 5 schematisch dargestellt und besteht aus verschiedenen Blöcken, die die Beschreibung der Daten in vielen Aspekten ermöglicht.

Die modulare Struktur des Datenmodells erlaubt eine einfache Integration neuer Beschreibungsblöcke und unkomplizierte Anpassung an neue Anforderun-

6 http://www.dkrz-de/data-en/wdcc

7 Lautenschlager, M./Toussaint, F./Thiemann, H. and Reinke, M. (1998): The Cera-2 Data Model. http://mms.dkrz.de//pdf/reports/ReportNo.15.pdf

8 http://www.dkrz.de/daten-en/cera/data_model

gen. Das CERA Datenmodell wurde 1998/1999 entwickelt und ist seitdem ohne substantielle Änderungen in Betrieb.

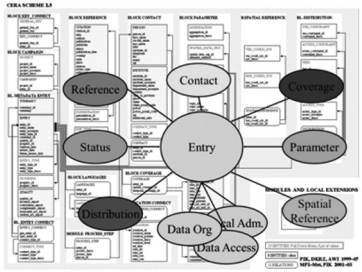


Abbildung 5: CERA2 Metadaten-Modell

Die Version CERA-2 ist auf einer SQL-basierten Datenbank implementiert, die über eine webbasierte Benutzerschnittstelle (Abbildung 7) erlaubt, Suchanfragen durchzuführen und sich Daten ausliefern zu lassen, wobei ein Teil der Daten in die Datenbank integriert ist.

Aufgrund des großen Datenvolumens, der unterschiedlichen Datentypen und der wissenschaftlichen Weiterverwertung dieser Daten, werden hohe Anforderungen an die WDC-Climate Datenbank gestellt:

- Hochleistungsrechner, Speicherung und Visualisierungssysteme, die für Klimaforschung optimiert sind
- Parallelisierung und Optimierung von Klimamodellen und Arbeitsabläufen
- Effizientes Management von größten Datenmengen
- 3D-Visualisierung für die Kommunikation von Forschungsergebnissen sowie
- Unterstützung für aktuelle Projekte der Klimaforschung

Schnittstellen zur Datenbank

Die Benutzerschnittstelle zur CERA-2 Datenbank ist schematisch in Abbildung 6 dargestellt. Das CERA-2 Web Portal[9] bietet eine komfortable Schnittstelle zur WDC-Climate Datenbank CERA und Zugriff auf die Daten.

Das Web Gateway (Abbildung 7)[10] erlaubt die Suche nach verschiedenen Kriterien wie Projekten, Experimenten, Variablen, Schlüsselworten und anderem. Ein Herunterladen der gefundenen Datensätze oder Teilen von ihnen wird nach einer erfolgreichen Suche angeboten, sofern ein Benutzeraccount existiert und die Daten freigegeben sind.

9 http://cera-www.dkrz.de/WDCC/ui/Index.jsp
10 http://cera-www.dkrz.de/WDCC/ui/Index.jsp

Es besteht auch die Möglichkeit, Datensätze über ein Java-basiertes Kommandozeilen-Tool (jblob) auf den lokalen Rechner zu transferieren. Dies ermöglicht die Einbindung des Downloads in Skripte.

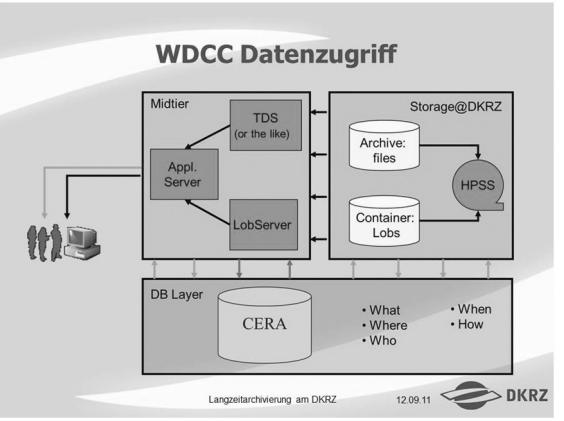


Abbildung 6: Schematische Darstellung des webbasierten Zugriffs auf die WDC-Climate Datenbank CERA

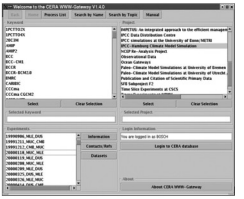


Abbildung 7:CERA-2 Web Portal zur Suche und zum Download

Langzeitarchivierungsservice

Das DKRZ bietet einen Service zur Langzeitarchivierung (LZA)[11] von Daten an, die relevant für Klima- und Erdsystemforschung sind. Er umfasst die Archivierung (mit Metadaten) und eine Verteilung der archivierten Daten über eine webbasierte Schnittstelle. Der Zeithorizont der Archivierung beträgt 10 Jahre und länger. Die Daten werden im High Performance Storage System (HPSS) gespeichert und mit den Metadaten verknüpft, die in der CERA Datenbank des WDC-Climate abgelegt sind. Eine doppelte Sicherung ist realisiert.

11 http://www.dkrz.de/daten/langzeitarchivierung

Die Nutzung der CERA Datenbank erlaubt auch eine Zugriffskontrolle, die eine zeitlich beschränkte Einschränkung des Zugriffsrechts auf bestimmte Nutzergruppen ermöglicht, beispielsweise, um die Daten zunächst im Rahmen eines Projektes wissenschaftlich zu verwerten.

Workflow

Der LZA-Prozess am DKRZ läuft wie in der schematischen Darstellung in Abbildung 8 gezeigt ab. Hierbei hat der Datenprovider verschiedene Schritte durchzuführen, bei denen das DKRZ Datenmanagement beratend und unterstützend beteiligt ist. Dazu stehen auch die vom DKRZ verfügbar gemachten Tools zur Verfügung. Die eingekreisten Ziffern kennzeichnen Schritte, die in der Kostenberechnung für die Archivierung DKRZ-ferner Daten berücksichtigt werden.

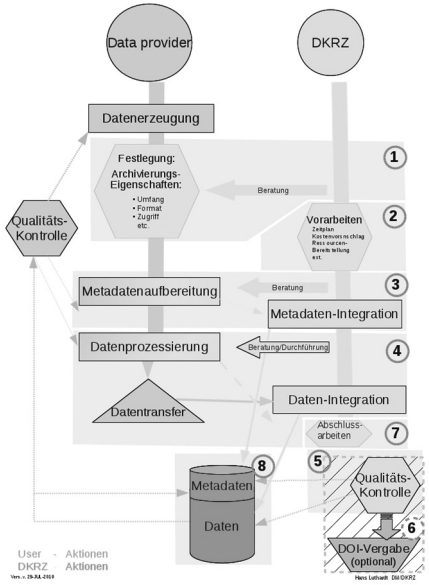


Abbildung 8: LZA Workflow

Datenstrukturen

Die Struktur der in der Datenbank zu speichernden geowissenschaftlichen Daten ist unterschiedlich. Gegenwärtig besteht das überwiegende Volumen der Daten aus Resultaten von Rechnungen mit numerischen Modellen, die auf dreidimensionalen Gittern (Abbildung 9) vorliegen. Daneben sind auch beobachtungsbasierte Datensätze mit anderen Strukturen vorhanden.

Die Speicherung der Daten erfolgt in einer Weise, die an der erwarteten Nutzung der Daten und an einer Minimierung der erforderlichen Transfervolumina orientiert ist.

Dies führt dazu, dass diese dreidimensionalen Daten meist nach Schichten und Variablen (Abbildung 10) aufgeteilt und als Zeitserien zweidimensionaler Felder in der Datenbank abgelegt werden.

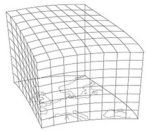


Abbildung 9: Schematische Darstellung eines Modellgitters (Beispiel)

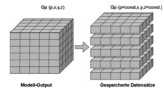


Abbildung10: Aufbereitung der Datensätze (Model-Output)

Werkzeuge

Graphische Oberfläche zur Erzeugung der LZA-Metadaten

Das DKRZ stellt für die Nutzer seines LZA-Services eine webbasierte, graphische Oberfläche (GUI) zur Verfügung (Abbildung 11)[12], mit der die erforderlichen Metadaten erzeugt werden können. Dieser Satz von Metadaten umfasst nur einen kleinen Teil des im CERA2-Metadatenmodell enthaltenen Umfangs. Es steht den Nutzern aber frei, die Möglichkeiten weitergehend (unter Verwendung von XML-Templates) auszuschöpfen.

Abbildung 11: Webbasiertes GUI zur Metadatengenerierung

12 http://cera-www.dkrz.de/LTA_metadata

cdo post processing Paket
Das am Max-Planck-Institut für Meteorologie entwickelte Paket *„climate data operators"* (cdo)[13] erlaubt eine komfortable Bearbeitung von Klima-(Modell-) Daten in den Standardformaten (zum Beispiel GRIB oder netCDF). Es umfasst nicht nur Formatkonvertierungen, sondern auch statistische und arithmetische Operatoren zur Auswertung der Datensätze.

CERA Web-Interface (CERA Portal)
Die Datenbank ist im Web (siehe auch Abbildung 7)[14] sichtbar und erlaubt die Sichtung der Metadaten einschließlich der Suche nach verschiedenen Einträgen (Projekt, Experiment, Variable, Variable aus experimentspezifischen Codelisten). Die Datenbank liefert die Zeiger (Pointer) zu den Dateien des Langzeitarchives, die den Metadaten zugeordnet sind, oder den direkten Zugriff auf Daten, die im WDC-Climate archiviert sind. Dazu ist ein Nutzeraccount erforderlich, der auf Antrag kostenfrei per E-Mail zu erhalten ist.
Die Nutzer-Authentifizierung erlaubt neben dem Führen einer Nutzungsstatistik auch die Information der Nutzer zu datensatzspezifischen Problemen und die Information der Datenlieferanten zur Nutzung ihrer Datensätze.

Command-line Tool jblob
Mit dem Java-basierten Kommando jblob[15] lassen sich Datensätze auch über die Kommandozeile herunterladen und damit der Download in Skripte integrieren. Für die Verwendung von jblob werden die Authentifizierung sowie der Name des herunterzuladenden Datensatzes benötigt. Auch eine Begrenzung der Daten (Bereich von Datensatzelementen/Records), zum Beispiel auf Zeitperioden, ist möglich.

Qualitätssicherung
Für die wissenschaftliche Qualität der in der Langzeitarchivierung abgelegten Daten, die auch Voraussetzung für die Vergabe eines DOI (siehe Kapitel Datenpublikation) ist, ist der Datenlieferant zuständig.

Das DKRZ führt ergänzend eine technische Qualitätskontrolle durch, die eine korrekte Speicherung der Daten im Langzeitarchiv sicherstellt und die Konsistenz von Daten und Metadaten gewährleistet.
„Bit-stream Preservation" als Sicherstellung der Unversehrtheit der Daten ist dabei Teil der technischen Qualitätskontrolle.

13 https://code.zmaw.de/projects/cdo
14 http://cera-www.dkrz.de/WDCC/ui/Index.jsp
15 http://cera-www.dkrz.de/CERA/jblob/

Die Qualitätkontrolle, die durch den Datenerzeuger durchgeführt werden muss, betrifft die wissenschaftliche Korrektheit der gespeicherten Daten, die Prüfung der Richtigkeit und Vollständigkeit der Metadaten und die Überprüfung der Konsistenz zwischen Daten und Metadaten im Langzeitarchiv.

Das Datenmanagement des DKRZ überprüft die Korrektheit des am DKRZ gegebenenfalls durchgeführten Postprocessings und der Speicherung der Daten im Langzeitarchiv sowie die Konsistenz und Vollständigkeit der gespeicherten Datensätze. Zudem werden die Zugriffs- und Download-Mechanismen sichergestellt.

Datenpublikation

Das DKRZ bietet im Zusammenhang mit der Datenspeicherung an, auf die Daten einen 'Persistent Identifier' zu vergeben und die Daten im Rahmen von DataCite in Bibliothekskatalogen zu registrieren und zitierfähig zu machen. Hierzu werden die Daten mit einem Digital Object Identifier (DOI) versehen, der von der TIB in Hannover verwaltet wird.

Die Auflösung des DOI erfolgt über einen globalen Handle-Server.

Voraussetzung dafür ist, dass die Daten und Metadaten qualitätsgeprüft sind und nicht mehr geändert werden (dürfen). Das WDC-Climate ist einer der DataCite[16, 17] Partner zur Sicherstellung der Langzeitarchivierung.

Kosten

Die Nutzerklientel des DKRZ setzt sich zusammen aus:

- Nutzern aus den Instituten der Gesellschafter
- Nutzern, die aus dem Kontingent des BMBF Ressourcen zugewiesen bekommen haben
- Externe Nutzer, die gegen Gebühren DKRZ-LZA-Ressourcen nutzen

Projekte, die bereits Computerressourcen des DKRZ nutzen, haben in der Regel auch (auf Antrag) Kontingente für die Langzeitarchivierung zugewiesen bekommen.

Für externe Nutzer und Projekte fallen dagegen Kosten für die Langzeitarchivierung an, da ihnen noch keine DKRZ Ressourcen zugewiesen wurden.

Als Kostenfaktoren werden hierbei berücksichtigt:

- Personalkosten für Arbeiten im Zusammenhang mit dem Einfüllen von Daten und Metadaten
- Kosten für die Archivierungsmedien (mehrere Generationen)
- Umlage der Betriebskosten des DKRZ
- Pflege der Daten während des Archivierungszeitraumes

16 dkrz.de/daten-en/Datapublication/konzept-datacite-tib-metadatakernel?set_language=en
17 http://www.datacite.org

Stand des Archives

Ende 2011 umfasst der in der Datenbank des WDC-Climate gespeicherte Datenbestand ca. 440 TeraByte (siehe auch Abbildung 12 und Tabelle 1)[18].

Der Umfang der LZA-Speicherung in Files außerhalb des WDC-Climate ist dagegen noch gering, da dieser Service erst seit Anfang des Jahres 2011 in der beschriebenen Form angeboten wird.

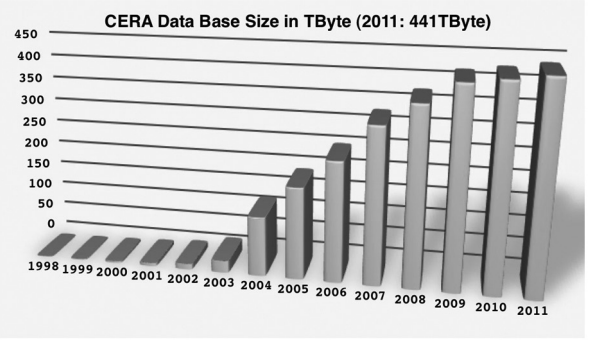


Abbildung 12: Entwicklung des Speichervolumens in der CERA Datenbank

Der aktuelle Bestand der CERA (WDC-Climate) Datenbank ist in Tabelle 1 zusammengefasst.

Ausblick

Die Größe der Datensätze, die in der Erdsystemforschung erzeugt, bearbeitet und gespeichert werden, wird unter anderem mit steigender, zur Verfügung stehender Rechnerleistung weiter wachsen. Einen Überblick über typische Dateigrößen bei Klimasimulationsrechnungen gibt folgende Übersicht:

Horizontalauflösung des Klimamodells
T42: 128 * 64 = 8192 Punkte pro Globalfeld T106: 160 * 320 = 51200 Punkte pro Globalfeld
Erforderliche Speichereinheiten (GRIB Format)
Horizontalfeld (Zugriffseinheit) • 17.1 kB (T42) • 100.1 kB (T106)

18 http://www.dkrz.de/daten-en/wdcc/statistics

Unix Filegröße für monatsweise akkumulierte Ergebnisse mit 6 Std. Speicherintervall und 300 2d Variablen (Physikalische Einheit):
- 616 MB (T42)
- 3500 MB (T106)

240 Jahre Modellintegration (Logische Einheit)
- 1.7 TB (T42)
- 10 TB (T106)

Diese wachsende Rechenkapazität wird unter anderem dazu verwendet werden,
- die Gitterauflösung zu verfeinern,
- die physikalischen Prozesse in den Modellen zu verbessern und neue Prozesse einzubinden,
- mehr Klimasubsysteme zu berücksichtigen,
- mehr Ensemble-Rechnungen durchzuführen.

Damit wird der Bedarf an Speicherkapazität in diesem Forschungsbereich weiter deutlich zunehmen und der Bedarf an performanten Speichersystemen mit effizienten Such- und Zugriffsverfahren auch am DKRZ weiter an Bedeutung gewinnen.

Zusammenfassung

Der aktuelle Langzeitarchivierungsservice des DKRZ für wissenschaftliche Daten baut auf der bereits seit längerer Zeit bewährten Hard- und Software-Infrastruktur des WDC-Climate auf und erlaubt damit komfortable Speicher- und Zugriffsmechanismen. Der Service ist allerdings noch in der Anlaufphase. Die dabei gesammelten Erfahrungen, speziell für die Archivierung externer Daten, werden sich in einer Anpassung der Abläufe und gegebenenfalls auch in der Kostenstruktur widerspiegeln.

Die bisherigen Erfahrungen zeigen jedoch, dass das gewählte Konzept praktikabel ist und zunehmend von internen Nutzern des DKRZ und externen Projekten – auch aus dem EU-Bereich – genutzt wird.

Das GESIS Datenarchiv für Sozialwissenschaften

Reiner Mauer

Das GESIS Datenarchiv für Sozialwissenschaften (DAS)

> *„Der eigentliche Sinn jedweder langfristigen Archivierung liegt darin, dass einmal erfasste Forschungsdaten erneut genutzt werden können. Die Effekte solcher Nachnutzung werden umso größer sein, wenn Forschungsdaten in dezidierte, in einer Community weithin bekannte und akzeptierte Repositorien eingepflegt und von dort abgerufen werden können". (Kleiner 2012: 10)*

Das Datenarchiv für Sozialwissenschaften stellt Forschungsdaten, vorwiegend aus nationalen und international-vergleichenden Umfragen für die Sekundärnutzung bereit. Die Studien werden gemäß klar definierten methodisch-technischen Anforderungen akquiriert und sodann bedarfsorientiert aufbereitet, dokumentiert, langfristig gesichert und der wissenschaftlich interessierten Öffentlichkeit zugänglich gemacht.

Die Sozialwissenschaften verfügen über eine vergleichsweise lange Tradition in der Archivierung und Nachnutzung von Forschungsdaten. So werden zur Beschreibung und Analyse sozialer Sachverhalte oder zur Entwicklung und Überprüfung sozialwissenschaftlicher Theorien häufig Forschungsdaten herangezogen, die nicht speziell zur Beantwortung der jeweiligen Forschungsfrage erhoben wurden. Oder es werden Daten verwendet, die von vornherein beispielsweise als Mehrthemenbefragung darauf ausgerichtet sind, Analysepotentiale für unterschiedliche Forschungsfragen bereitzustellen. Diese in den Sozialwissenschaften als Sekundäranalyse bezeichnete Nachnutzung von Daten durch nicht unmittelbar an der Datenerhebung beteiligte Dritte (oder auch durch den Primärforscher zur Analyse neuer Forschungsfragen) wurde nicht zuletzt erst durch den Aufbau entsprechender Infrastrukturen ermöglicht und befruchtete diese wiederum. Dies geschah sowohl durch das Angebot auf ein reiches Datenangebot zugreifen zu können als auch durch die gezielte Förderung des data sharing und der Sekundäranalyse als wertvoller Forschungsmethode:

> *„ ... high priority was assigned to convincing the community of empirical social scientists that a data archive was not a repository for 'used data' but a research facility. This meant the promotion of secondary*

analysis as an approach that could lead to findings as interesting as those of primary research." (Scheuch 1990: 96).

Der Aufbau entsprechender Datenserviceinfrastrukturen begann bereits kurz nach dem Zweiten Weltkrieg mit der Gründung des Roper Centers, welches dann allerdings auch erst im Jahr 1957 für eine weitere Öffentlichkeit zugänglich wurde (Scheuch 2003: 386) und entwickelte eine gewisse Dynamik in den 1960er Jahren. Beginnend mit der Etablierung des Zentralarchivs in Köln, folgte eine Reihe von weiteren Gründungen von sozialwissenschaftlichen Datenarchiven in den Folgejahren (wie beispielsweise dem ICPSR in Ann Arbor, dem Steinmetz Archiv in den Niederlanden und dem UK Data Archive in Colchester). Ein weiterer Schub an Gründungen war dann in den 1990er Jahren zu verzeichnen, insbesondere in den Staaten Mittel- und Osteuropas. Allein in Europa sind gegenwärtig 21 sozialwissenschaftliche Datenarchive im Council of European Social Science Data Archives (CESSDA[1]) organisiert. Dieses 1976 gegründete Kooperationsnetzwerk nationaler Archive entwickelt sich gegenwärtig im Rahmen des ESFRI-Prozesses zu einem sogenannten European Research Infrastructure Consortium (ERIC) und damit zu einer Körperschaft europäischen Rechts, die dann auch dauerhaft koordinierende Funktionen für die Datenlandschaft in Europa übernehmen kann (Quandt und Mauer 2012: 66).

Die in der jüngeren Vergangenheit enorm gewachsene Aufmerksamkeit hinsichtlich der Bedeutung von Forschungsdaten hat auch in den Sozialwissenschaften zu einer erneuten Dynamik beim Aufbau von Forschungsdateninfrastrukturen geführt. In Deutschland sind dabei insbesondere die Aktivitäten des Rates für Sozial- und Wirtschaftsdaten (RatSWD[2]) und die in seinem Umfeld entstandenen Forschungsdatenzentren (FDZ) zu nennen.

Institutionelle Verankerung

1960 als Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln gegründet, ist das Archiv heute eine von fünf wissenschaftlichen Abteilungen von GESIS – Leibniz-Institut für Sozialwissenschaften, die mit ihrem forschungsbasierten Service- und Produktangebot den Forschungsprozess der empirischen Sozialforschung in seiner gesamten Breite abdecken. Dabei dient der Forschungsdatenzyklus als Leitbild zur Strukturierung und Verknüpfung der Angebote. GESIS ist Mitglied der Leibniz-Gemeinschaft und erbringt mit seinen über 250 Mitarbeitern an drei Standorten (Mannheim, Köln, Berlin) grundlegende, überregional und international bedeutsame forschungsbasierte Dienstleistungen.

1 http://www.cessda.org
2 http://www.ratswd.de

Leitgedanken bei der Gründung

Die Untersuchung des sozialen Wandels, also beispielsweise von politischem, technologischem oder ökonomischem Wandel oder von auch Veränderungen im Wertesystem, ist eines der klassischen Themen sozialwissenschaftlicher Forschung. Sollen derartige Veränderungen im Zeitverlauf empirisch untersucht werden, besteht die Notwendigkeit auf bereits erhobene Daten zurückzugreifen, da es nicht immer oder nur bedingt möglich ist, entsprechende Daten retrospektiv zu erheben, die beispielsweise Auskunft über Einstellungen oder Verhalten in der Vergangenheit geben können. Die Archivierung sozialwissenschaftlicher Daten und die damit häufig erst mögliche Nachnutzung, *„... extend[s] any particular survey from a mere snapshot of reality into a continuous observation."* (Scheuch 1990: 96).

Neben diesem zentralen methodischen Motiv spielten forschungsökonomische Argumente eine gleichsam bedeutende Rolle bei der Gründung. Insbesondere die Beobachtung, dass mit vergleichsweise hohem finanziellem und zeitlichem Aufwand erhobene Daten in der Regel nur sehr begrenzt in einem einzigen Forschungskontext analysiert wurden, häufig aber auch über darüber hinausgehende, bisher nicht ausgeschöpfte Analysepotentiale verfügen. Gleichfalls sollte die Möglichkeit, auf bereits erhobene Daten zurückzugreifen, zur Vermeidung unnötiger Doppelungen beitragen und Forschende in stärkerem Ausmaß dazu anregen, neue Erhebungen auf der Basis bestehender Arbeiten aufzubauen, um somit auch eine größere Vergleichbarkeit von Forschungsergebnissen zu erreichen. Ganz allgemein zielte die Gründung des Datenarchivs darauf ab, den Zugang zu Forschungsdaten – gerade auch für Studierende und Nachwuchswissenschaftler – zu erleichtern und auf diesem Wege die empirische Sozialforschung und im Speziellen Sekundäranalysen sowie Methodenforschungen zu unterstützen.

Aufgaben

Zu den zentralen Aufgaben des GESIS Datenarchivs zählt insbesondere die Langzeitarchivierung von Forschungsdaten, also die langfristige Sicherstellung von Interpretierbarkeit und Verfügbarkeit der archivierten Bestände und die Bereitstellung dieser Daten. Dies geschieht mit dem Ziel, Sekundäranalysen ganz allgemein und im Besonderen Raum- und Zeitvergleiche zu ermöglichen. Des Weiteren leistet die Arbeit des Archivs einen wichtigen Beitrag zur Erhöhung von Transparenz und Überprüfbarkeit wissenschaftlicher Forschung, indem es Replikationen und Re-Analysen ermöglicht. [3]

3 Wenngleich, wie Wagner und Huschka (2012) sicher zu Recht anmerken, die Überprüfung von Forschungsergebnissen auch in den Sozial- und Wirtschaftswissenschaften nicht gerade eine vorherrschende Praxis ist und die bloße Verfügbarkeit von Forschungsdaten allein nicht ausreicht, um Replikationen zu fördern. Jedoch gilt andersherum natürlich auch, dass ohne die Zugriffsmöglichkeit auf die fraglichen Daten Replikationen erst gar nicht möglich sind.

Eine weitere wichtige Aufgabe des GESIS Datenarchivs besteht darin, den Zugang zu internationalen Forschungsdaten zu erleichtern. Diese geschieht einerseits durch die Archivierung ausländischer Datenbestände[4] und hierbei insbesondere durch die Betreuung großer interkulturell und intertemporal vergleichender Umfrageprogramme, wie etwa die Eurobarometer Studien der Europäischen Kommission oder die Daten der Europäischen Wertestudie (EVS)[5]. Zum anderen aber auch durch die Vermittlung von Daten, die bei anderen Archiven gehalten werden. Als Mitglied in internationalen Archivnetzwerken wie IFDO (International Federation of Data Organizations for Social Science[6]) und dem europäischen Archivverbund CESSDA (Council of European Social Science Data Archives[7]) sowie als nationaler Repräsentant des ICPSR (Inter-university Consortium for Political and Social Research[8]) vermittelt das GESIS Datenarchiv für deutsche Forscher den Zugang zu sozialwissenschaftlichen Forschungsdaten weltweit.

Ferner unterstützt das Datenarchiv Primärforscher bei der Sicherung, Dokumentation, Aufwertung und Bereitstellung ihrer Daten. Sei es durch die Beratung in allen Fragen des Datenmanagements, aber auch durch praktische Unterstützung der Arbeit oder die Bereitstellung von Tools.

Datenbestand

Der Datenbestand umfasst gegenwärtig knapp 5.100 öffentlich zugängliche Studien. Dabei handelt es sich zum größten Teil um Mikrodaten der Umfrageforschung sowie um historische Zeitreihen und Aggregatdaten. Thematisch liegt der Fokus auf Studien zu sozial- und politikwissenschaftlichen Fragestellungen. Ein wichtiger und intensiv betreuter Schwerpunkt sind dabei Daten der interkulturell vergleichenden Sozialforschung, die eine Vielzahl von Ländern und überwiegend lange Zeiträume abdecken – wie etwa das International Social Survey Programme (ISSP), das jährlich eine gemeinsame Umfrage zu sozialwissenschaftlich relevanten Themen durchführt und seit seiner Gründung 1984 auf 48 Mitgliedsländer im Jahre 2011 angewachsen ist[9]. Auch kontinuierliche nationale Erhebungsprogramme, wie beispielsweise die seit 1980 von GESIS bzw.

4 So hält das Datenarchiv beispielsweise eine größere Sammlung von Daten aus Mittel- und Osteuropa oder die Daten des Japanese General Social Survey.

5 http://www.europeanvaluesstudy.eu

6 http://www.ifdo.org

7 http://www.cessda.org

8 www.icpsr.umich.edu

9 http://www.issp.org bzw. http://www.gesis.org/issp/ . Weitere wichtige von GESIS (teilweise in Kooperationen) betreute internationale Umfrageprogramme sind etwa die Eurobarometer Studien der Europäischen Kommission, die European Values Study (EVS) oder auch die Comparative Study of Electoral Systems (CSES), die in rund 40 Ländern politische Einstellungen und Wahlverhalten untersucht. Eine Übersicht und weiterführende Hinweise finden sich unter http://www.gesis.org/das-institut/kompetenzzentren/fdz-internationale-umfrageprogramme/

seinen Vorgängerinstituten durchgeführte Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS)[10] oder die erstmals anlässlich der Bundestagswahl 2009 durchgeführte German Longitudinal Election Study (GLES)[11] werden aufwendig betreut. Darüber hinaus hält das GESIS Datenarchiv umfangreiche Bestände zu politischen Einstellungen und Verhalten, Werten, Jugend, Gesundheit, Mediennutzung und vielen weiteren sozialwissenschaftlich relevanten Themenbereichen vor.

Datenarchivierung@DAS: Konzeptionell

Auf der Grundlage des 1999 veröffentlichten OAIS Referenzmodells entwickelte sich ein zunehmendes Bewusstsein und mittlerweile weithin geteiltes Verständnis für die Notwendigkeit von archivarischen Konzepten, die darauf abzielen, den langfristigen Erhalt von und den Zugang zu digitalen Informationen sicherzustellen. Mit dem mittlerweile als ISO-Standard etablierten Referenzmodell wurde sowohl eine Nomenklatur als auch ein Rahmen bereitgestellt, der es ermöglicht, unterschiedlichste (digitale) Archive und deren Archivierungskonzepte in einer einheitlichen Form zu beschreiben und darauf aufbauend ihre Funktionsfähigkeit und auch Vertrauenswürdigkeit beurteilen zu können.

Entsprechend dem damaligen technologischen Stand waren die ersten Jahre der Datenarchivierung in Köln verständlicherweise nicht ‚digital' geprägt, sondern wurden von der Notwendigkeit bestimmt, auf Lochkarten vorgehaltene Forschungsdaten zu verarbeiten. Die Sicherung der Forschungsdaten erfolgte durch Herstellung und Aufbewahrung von Duplikaten dieser Lochkarten – also auf einem Speichermedium aus Papier – und die Datenverarbeitung beschränkte sich weitgehend auf das Auszählen und Sortieren der Karten mittels sogenannter Fachzählsortiermaschinen. Dies änderte sich jedoch relativ bald durch den Einsatz von Großrechnern (zunächst extern und dann mit eigenem Rechenzentrum) und entwickelte große Dynamik mit dem Aufkommen von Mikrocomputern, so dass man spätestens seit den siebziger Jahren des letzten Jahrhunderts das Datenarchiv auch als digitales Langzeitarchiv bezeichnen kann.

Der sich am OAIS orientierende Kriterienkatalog vertrauenswürdige digitale Langzeitarchive von nestor (2006: 2) definiert ein digitales Langzeitarchiv als eine *„Organisation [...], die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit digitaler Objekte sowie für ihre Interpretierbarkeit*

10 http://www.gesis.org/allbus

11 http://www.gles.eu/ bzw. http://www.gesis.org/wahlen/gles/. Darüber hinaus betreut GESIS auch die vom ZDF beauftragte und der Forschungsgruppe Wahlen seit 1977 monatlich durchgeführten Politbarometer sowie die im Auftrag der ARD erhobenen DeutschlandTrends. Siehe hierzu: http://www.gesis.org/wahlen/

zum Zwecke der Nutzung durch eine bestimmte Zielgruppe übernommen hat." Die beiden wesentlichen Bestandteile dieser Definition, nämlich die Übernahme der Verantwortung für die Langzeitarchivierung von digitalen Objekten und die Verfügbarkeit dieser Objekte – im Fall des Datenarchivs von Forschungsdaten – sowie die Zielgruppenorientierung (Sozialwissenschaften) sind explizit in der Satzung von GESIS[12] verankert. So bestimmt § 2, mit welchen konkreten Aufgaben der Vereinszweck – nämlich die *„Förderung der sozialwissenschaftlichen Forschung"*[13] – erfüllt wird. Dazu zählen unter anderem die *„c) Archivierung, Dokumentation und Langzeitsicherung sozialwissenschaftlicher Daten, einschließlich ihrer Erschließung sowie qualitativ hochwertigen Aufbereitung besonders relevanter Daten für Sekundäranalysen, ..."*[14] sowie die *„e) Schaffung eines benutzerfreundlichen und hochqualitativen Zugangs zu allen für die empirische Sozialforschung relevanten Informationen und Daten ..."*[15].

Um eine möglichst breite und optimale Nachnutzung zu ermöglichen, steht das aktive Management der Daten und insbesondere die Schaffung von Mehrwert durch Aufbereitung, Dokumentation und die Verknüpfung von Daten seit den Gründungstagen im Mittelpunkt der Arbeit des Datenarchivs. Dieses Verständnis von Datenarchivierung bei GESIS lässt sich in aktueller Terminologie am besten mit dem Begriff ‚data curation' beschreiben:

> *"Digital curation is all about maintaining and adding value to a trusted body of digital information for future and current use; specifically, the active management and appraisal of data over the entire life cycle. [...] builds upon the underlying concepts of digital preservation whilst emphasising opportunities for added value and knowledge through annotation and continuing resource management [...]." (JISC 2006: 1)*

Diese Schwerpunktsetzung erfolgte bereits bei Gründung in bewusster Abgrenzung zu den Erfahrungen, die Scheuch – einer der beiden Gründungsväter des Archivs – beim Besuch des Roper Centers in den fünfziger Jahren machte. Heute eines der bedeutendsten sozialwissenschaftlichen Archive, war es damals nicht zuletzt aufgrund der fehlenden finanziellen Ausstattung mehr oder weniger eine reine Lagerstätte für Lochkarten (Scheuch 1990: 95).

12 GESIS – Leibniz-Institut für Sozialwissenschaften e.V., Satzung vom 19.06.2010. http://www.gesis.org/das-institut/der-verein/satzung/ [21.05.2012]

13 Ebd., § 2 (1), Satz 1

14 Ebd., § 2 (2) (c)

15 Ebd., § 2 (2) (e)

Datenarchivierung@DAS: Funktional

Die zuvor erfolgte konzeptionelle Einordnung des Datenarchivs soll im Folgenden um eine Beschreibung der konkreten Arbeitsweisen des Archivs erweitert werden (die aus Platzgründen naturgemäß nur einen groben Überblick über die Abläufe geben kann). Am Anfang der Archivierungstätigkeit steht die Akquisition von Studien, also die Entscheidung darüber, welche Daten überhaupt in das Archiv aufgenommen werden sollen. Daran schließt sich die konkrete Übernahme der Forschungsdaten in das Archiv an (Ingest). In der Folge durchlaufen die Studien verschiedene Aufbereitungs- und Dokumentationsschritte, wobei es einen gemeinsamen Standard gibt, den alle Studien durchlaufen, und darüber hinausgehende aufwendigere Prozesse, die nur für ausgewählte bzw. besonders betreute Studien durchgeführt werden. Im Anschluss an diese werden Daten und Dokumentation an den Datenservice sowie an die Langzeitarchivierung übergeben.[16]

Akquisition

Forschungsdaten – im Regelfall Daten der empirischen Umfrageforschung – gelangen auf unterschiedlichen Wegen ins Archiv: Sie werden entweder gezielt akquiriert oder aber dem Archiv von Primärforschern zur Archivierung angeboten bzw. erfolgt die Übergabe regelmäßig im Rahmen längerfristiger Kooperationen. Grundsätzlich werden maschinenlesbare Datensätze aus allen Bereichen der Sozialwissenschaften archiviert, sofern (1) die Studien Aussagen über die deutsche Bevölkerung oder über Teile von ihr erlauben, (2) an der Untersuchung deutsche Forscher beteiligt waren, unabhängig davon, ob sie sich auf Deutschland bezieht oder nicht, und (3) die Studie ganz allgemein für die sozialwissenschaftliche Forschung von Interesse sein könnte. Gezielt akquiriert werden dagegen meist Daten zu bestimmten Forschungsgebieten (beispielsweise der international vergleichenden Sozialforschung), besonders prominente und bedeutende Erhebungsprojekte oder solche, die bereits im Archiv vorhandene Kollektionen vervollständigen. Neben den zuvor genannten Kriterien, müssen die aufzunehmenden Studien bestimmten formalen und technischen Anforderungen entsprechen. Insbesondere müssen neben den Datensätzen selbst, auch alle für eine Sekundärnutzung notwendigen Materialien vorhanden sein und an das Archiv übergeben werden (Erhebungs- bzw. Messinstrumente, Codepläne, Methodenberichte etc.). Diese den Datensatz begleitenden Materialien und Dokumente bilden auch die Grundlage für die vom Datenarchiv im Rahmen der Archivierung erzeugten standardisierten Metadaten.

16 Für den ebenfalls von GESIS im Rahmen von da|ra angebotenen Datenregistrierungsservice sei auf den Beitrag von Brigitte Hausstein in diesem Band verwiesen.

Um eine Archivierung und damit insbesondere eine Nachnutzung durch Dritte optimal vorzubereiten, ist es hilfreich, diese möglichst frühzeitig einzuplanen. Idealerweise beginnt dies bereits bei der Vorbereitung einer Datenerhebung. Insofern ist es auch das Bestreben des Datenarchivs, möglichst früh im Lebenszyklus einer Studie anzusetzen. Auch wenn sich langsam das Wissen um die Vorteile einer proaktiven Planung verbreitet – nicht zuletzt durch die Diskussion um die Notwendigkeit von Datenmanagementplänen –, so stellt sie doch derzeit (noch) nicht den Regelfall in der Archivierungspraxis dar. Vielmehr werden Studien häufig erst lange nach der Erhebung bzw. nach dem Projektende zur Archivierung angeboten bzw. für diese freigegeben. Dies verursacht je nach Zustand von Daten und Dokumentation einen teils erheblichen Mehraufwand.

Ingest – Aufnahme ins Archiv

Vor der eigentlichen Aufnahme einer Studie ins Archiv, wird diese zunächst gemeinsam mit dem Datengeber vorbereitet. Dazu werden Absprachen über Umfang, Formate, Aufbereitungs- und Dokumentationsziele getroffen sowie geklärt, unter welchen Bedingungen die Daten weitergegeben werden. Diese Absprachen werden schließlich in Form einer Archivierungsvereinbarung festgehalten, in der auch die für eine Archivierung und Weitergabe notwendige Übertragung von Nutzungsrechten an das Archiv festgehalten wird.

Entsprechen die aufzunehmenden Studien den oben genannten Anforderungen und liegt eine entsprechende Archivierungsvereinbarung vor, werden sie beim Dateneingang einer standardisierten Eingangskontrolle unterzogen. Dabei besteht das Submission Information Package bzw. Übergabeinformationspaket (SIP)[17] in der Regel aus einem oder mehreren Datensätzen, den dazugehörigen Messinstrumenten (Fragebogen), sowie weiteren Materialien mit Informationen zum Forschungsprozess und Publikationen. Diese Eingangskontrolle umfasst dabei unter anderem folgende Schritte:

- Technische Kontrollen: Formate, Lesbarkeit, Virenfreiheit etc.
- Vollständigkeit und Nutzbarkeit: Sind Daten, Messinstrumente und Dokumente vollständig übergeben worden und beziehen sie sich aufeinander? Insbesondere: Passen Erhebungsinstrument und Daten zusammen?
- Konsistenz: Gibt es Werte außerhalb des zulässigen Bereichs (wild codes)? Sind fehlende Werte definiert und wenn ja, wie? Stimmen die Daten mit der im Fragenbogen vorgegebenen Filterführung überein (question routing)? Gibt es widersprüchliche Merkmalskombinationen?

17 „Ein Informationspaket, das vom Produzenten an das OAIS geliefert wird, um es zur Konstruktion oder zur Aktualisierung eines oder mehrerer AIPs und/oder den dazugehörigen Erschließungsinformationen zu benutzen“ (nestor 2012: 15).

- Datenschutz: Da es sich bei den archivierten Studien vorrangig um Mikrodaten handelt, werden sie darüber hinaus auch auf datenschutzrechtlich relevante Aspekte untersucht.

Bei der Eingangskontrolle festgestellte Inkonsistenzen oder sonstige Fehler werden dokumentiert und, sofern möglich, nach Rücksprache mit den Datengebern korrigiert.

In Abhängigkeit vom Ergebnis der Eingangskontrolle und einem zuvor festgelegten Aufbereitungs- und Dokumentationsziel folgen der Eingangskontrolle weitere Arbeitsschritte. So werden auf Datenträgern gelieferte Daten und Materialien von den Original-Medien getrennt. Digitale Objekte werden in definierte Langfristsicherungsformate überführt, Materialien im Papierformat werden digitalisiert und die Originale in einem speziell dafür vorgesehenen Archivraum abgelegt. Sodann werden beschreibende Metadaten in Form einer sogenannten Studienbeschreibung angelegt (siehe unten) sowie weitere interne (administrative und technische) Metadaten in einer zentralen Datenbank erzeugt. Ein Teil der eingehenden Studien wird über die Standardarchivierung hinaus einer umfassenderen Datenaufbereitung und -dokumentation zugeführt (siehe unten). Die Ergebnisse dieser Aufbereitungs- und Dokumentationsarbeiten werden nach Abschluss wiederum in die Langfristsicherung übergeben. Schlussendlich werden die für den Datenservice bestimmten Objekte in ein oder mehrere Dissemination Information Packages oder Auslieferungsinformationspakete (DIP)[18] überführt. Dabei erhalten alle Datensätze eine standardisierte Versionsinformation und werden mit einem persistenten Identifikator in Form eines DOI versehen. Die dafür notwendige Registrierung erfolgt bei der von GESIS betriebenen Registrierungsagentur da|ra.

Datenaufbereitung / -dokumentation

Ein wesentlicher Teil der Ressourcen des Archivs wird eingesetzt, um Forschungsdaten so aufzubereiten und zu dokumentieren, dass sie durch Dritte – also nicht unmittelbar an der Datenerhebung beteiligte – genutzt werden können. Der Umfang dieser Arbeiten ist einerseits abhängig vom Ausgangszustand der übernommenen Daten und Materialien und andererseits vom konkreten Aufbereitungs- und Publikationsziel der jeweiligen Studie. Prozesse, die alle Studien durchlaufen, sind aufbauend auf der oben beschriebenen Eingangskontrolle einfache Aufbereitungen (beispielsweise zur Fehlerkorrektur oder zur Herstellung eines vollständig gelabelten Analysedatensatzes), Versionierung und die Zuweisung eines entsprechenden DOI-Namens, sowie die Erstellung einer Studienbe-

18 „Ein Informationspaket, abgeleitet aus einem oder mehreren AIPs, das der Endnutzer als Antwort auf eine Anfrage an das OAIS erhält." (nestor 2012: 9)

schreibung, die in standardisierter Form (kompatibel zur Metadatenspezifikation der internationalen Data Documentation Initiative, DDI) inhaltliche, methodische und technische Charakteristika einer Studie spezifiziert.[19] Sie ermöglicht externen Nutzern das Auffinden der Studien sowie die Beurteilung der Relevanz für die eigene Forschung. Diese Studienbeschreibungen werden sowohl in deutscher als auch in englischer Sprache in einer Datenbank gepflegt, die online recherchierbar Nutzern zur Verfügung gestellt wird (siehe unten zum Datenbestandskatalog).

Adding value

Um eine möglichst breite und optimale Nachnutzung zu ermöglichen, werden für die Forschung besonders relevante Daten mit teils erheblichem Aufwand mit Mehrwert versehen. Diese Studien durchlaufen zusätzliche teils sehr intensive Aufbereitungs- und Dokumentationsschritte. Dabei handelt es sich hauptsächlich um solche Studien, die sich für die interkulturell vergleichende Forschung eignen und/oder kontinuierlich durchgeführt werden und sich somit grundsätzlich in Zeitreihen einordnen lassen. Für diese meist großen Umfrageprogramme bietet das GESIS Datenarchiv eine auf die jeweiligen Bedarfe maßgeschneiderte Unterstützung im Datenmanagement. Häufig ist GESIS selbst in verschiedenen Funktionen an der Datenerhebung dieser Programme beteiligt[20] oder führt diese, wie im Fall der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften selbst durch. Service und Forschung zu diesen langjährigen GESIS-Aktivitäten wurden mittlerweile in Forschungsdatenzentren gebündelt, die sich an den Kriterien des Rates für Sozial- und Wirtschaftsdaten (RatSWD) orientieren.[21] Diese bieten Forschern nicht nur hochwertig aufbereitete und dokumentierte Daten an, sondern bedienen auch den besonderen Bedarf an Beratung und Wissensvermittlung, der mit der Nutzung derartiger Daten einhergeht.

Die derart betreuten Daten werden je nach Zielvorgabe weitergehend standardisiert und aufbereitet, harmonisiert und zu komplexen zeit- und/oder ländervergleichenden Studien integriert. Die Dokumentation der Studien wird je nach Zielvorgabe erweitert um Metadaten sowohl auf Studien- als auch auf Datensatzebene (Standardisierung der Datensatzstruktur, Variablennamen/-label, etc.). Darüber hinaus erfolgt eine strukturierte Dokumentation des Fragebogens, ergänzt um Anmerkungen zur Datenqualität auf Variablenebene und weiteren Kontextinformationen. Dieser umfangreiche Bestand an strukturierten Metadaten wird dann einerseits über verschiedene (Meta)Datenportale und Recherche-

19 Zum Metadatenschema der Studienbeschreibungen im Datenbestandskatalog siehe Zenk-Möltgen und Habbel (2012).

20 So ist GESIS beispielsweise am International Social Survey Programme nicht nur für die Datenarchivierung zuständig, sondern – vertreten durch die Abteilung Dauerbeobachtung der Gesellschaft – auch für die Durchführung der deutschen Teilerhebung und übt darüber hinaus weitere internationale Funktionen insbesondere im Methodenbereich aus.

21 http://www.gesis.org/das-institut/kompetenzzentren/

systeme, wie etwa dem Datenportal ZACAT, zur Verfügung gestellt und dient andererseits aber auch dazu, begleitende Dokumentationen etwa in der Form von Codebüchern, Variablenreports oder Methodenberichten zu erstellen. Diese sind insbesondere bei komplexeren Studien oder Studienkollektionen neben den Daten selbst und den dazugehörenden Messinstrumenten häufig Voraussetzung für eine sekundäranalytische Nutzung durch Dritte.

Access / Datenbereitstellung

Ein zentraler Bestandteil der Archivierungstätigkeit ist die Bereitstellung der Forschungsdaten für die Nachnutzung, sei es für Replikationszwecke oder aber – und das gilt für die überwältigende Anzahl der Fälle – um neue Forschungsfragen mithilfe dieser Daten zu beantworten. Mit mehr als 30.000 Datenweitergaben in 2011 und der Bearbeitung von mehr als tausend Anfragen rund um das Datenangebot pro Jahr wird der Service des GESIS Datenarchivs intensiv genutzt – weit überwiegend für die akademische Forschung und Lehre. Dabei beschränkt sich die Nutzung nicht auf Deutschland, sondern je nach Angebot werden die Dienstleistungen zwischen 30 % und 70 % von Nutzern außerhalb Deutschlands in Anspruch genommen. Die Bereitstellung der archivierten Daten und Dokumente wird durch vom Datengeber vergebene Zugangskategorien geregelt.[22] Rund Dreiviertel der archivierten Studien ist dabei frei zugänglich für die akademische Forschung und Lehre, für ein Viertel der Studien ist vor einer Weitergabe eine schriftliche Genehmigung durch den Datengeber einzuholen. Daten und zugehörige Materialien, wie beispielsweise Fragebögen oder Variablenreports, werden entweder kostenfrei zum Download in verschiedenen Portalen und Systemen angeboten oder können über den Datenservice bestellt werden. Für die individuelle Bereitstellung von Daten wird dann allerdings eine Bereitstellungsgebühr in Rechnung gestellt.

Einen zentralen Zugangspunkt zu den Beständen des Archivs bietet der Datenbestandskatalog, der zu allen archivierten Studien ausführliche Informationen in der Form von Studienbeschreibungen bereitstellt.[23] Neben bibliographischen Angaben, werden insbesondere Inhalte und methodische Aspekte beschrieben, wie etwa die Stichprobenziehung oder die in der Studie eingesetzten Erhebungsverfahren. Darüber hinaus finden sich auch Hinweise auf Veröffentlichungen und die Versionsgeschichte, sowie aktuell bekannte Fehler in den Daten. Neben der Möglichkeit einfache oder auch verknüpfte Suchen durchzuführen, bietet der Datenbestandskatalog die Möglichkeit Daten und Dokumente dieser Studien direkt herunterzuladen oder über ein Warenkorbsystem zu bestellen.

22 http://www.gesis.org/unser-angebot/daten-analysieren/datenservice/benutzungsordnung/#3_Zugangskategorien

23 http://www.gesis.org/dbk

Neben dem zentralen Zugang über den Datenbestandskatalog existieren weitere, spezialisierte Angebote, um beispielsweise erweiterte Dokumentation oder zusätzliche Funktionen für die Nutzer anbieten zu können. So bietet das Datenportal ZACAT[24] einen direkten Zugang zu gegenwärtig rund 600 Studien. Der Schwerpunkt des Angebotes liegt dabei auf international vergleichenden Studien sowie auf Daten der Wahlforschung, unter anderem Eurobarometer, ISSP, European Values Study, ALLBUS, Politbarometer und Wahlstudien (Deutschland, Osteuropa). Zu jeder Studie wird eine ausführliche Dokumentation sowohl auf Studien- als auch auf Variablenebene, inklusive der vollständigen Frage- und Antworttexte des Fragebogens, angeboten. Für einige Studien sind Frage- und Antworttexte auch mehrsprachig dokumentiert. So sind beispielsweise für die vierte Welle der European Value Study alle in den 49 Teilnahmeländern eingesetzten originalsprachlichen Fragebögen dokumentiert. Neben der Suche in den strukturierten Metadaten auf Studienebene können mithilfe der erweiterten Suche auch die kompletten Frage- und Antworttexte des Fragebogens sowie Variablen- und Wertelabel durchsucht werden und somit aus dem Bestand von weit über 200.000 Variablen einzelne Indikatoren gezielt identifiziert werden. Darüber hinaus ermöglicht das System (einfache) Analysen wie Häufigkeitsauszählungen, Kreuztabellen, Regressionen und deren graphische Darstellung (Balken- und Kuchendiagramme). Die Analysen können auch unter Verwendung von Gewichtungsvariablen durchgeführt werden. Die Datensätze selbst sind in verschiedenen gängigen Formaten herunterladbar (unter anderem SPSS, SAS, Stata, csv). Für den Download können auch subsets von Datensätzen gebildet werden (Auswahl von Variablen oder Fällen). Der Zugang ist gebührenfrei, Analyse und Download erfordern eine Registrierung.

Studien mit Zeitreihen bzw. mit zeitreihenfähigen Daten zur historischen Demografie, zur empirischen Sozial- und Wirtschaftsgeschichte sowie zur Historischen Statistik Deutschlands werden mit der Datenbank HISTAT[25] bereitgestellt. Die Datenbank ist thematisch untergliedert (zum Beispiel Erwerbstätigkeit, Bevölkerung, Bildung und Wissenschaft etc.) und bietet derzeit Zugang zu rund 260.000 Zeitreihen. Neben diesen Zeitreihen selbst, enthält die Datenbank ausführliche Studienbeschreibungen, die Auskunft über inhaltliche und methodische Aspekte der Studien geben sowie die verwendeten Quellen beschreiben. Neben verschiedenen Recherchemöglichkeiten erlaubt das auf Zeitreihen spezialisierte System auch den Download der Studien bzw. Zeitreihen im Excel- oder CSV-Format.

24 http://zacat.gesis.org

25 http://www.gesis.org/unser-angebot/daten-analysieren/daten-historische-studien/zeitreihen/

Mit SIMon[26], dem Social Indicators Monitor, stellt GESIS ein weiteres Informationssystem mit Zeitreihendaten zur Verfügung. Angeboten werden deutsche und europäische Sozialindikatoren, mit deren Hilfe sozialstruktureller Wandel und die Entwicklung der Lebensbedingungen und Lebensqualität der Bevölkerung beobachtet und analysiert werden kann. SIMon erlaubt neben einer auf verschiedenen Wegen unterstützten Auswahl und Recherche von Indikatoren, einfache Möglichkeiten der Tabellenmanipulation und Datenanalyse sowie die grafische Darstellung der Indikatoren, etwa als Liniengrafik, Histogramm oder Boxplot oder auf Landkarten. Ebenso wird der Export von Tabellen und Grafiken unterstützt, so dass diese in anderen Anwendungen weiterverarbeitet werden können.
Über die oben beschriebenen Portale und Systeme hinaus, bietet das Web-Angebot der GESIS eine Fülle an weiteren Informationen rund um das Datenangebot, insbesondere zu den besonders betreuten Umfrageprogrammen bzw. einzelnen Studien in diesen Serien.[27]

Langzeitarchivierung

Seit seiner Gründung im Jahr 1960 ist die langfristige Sicherung von Daten und zugehörigen Materialien eine der Kernfunktionen des Datenarchivs. Dies ist notwendig, da die durch digitale Objekte repräsentierten Informationen durch Einbußen in ihrer Integrität, Authentizität und Vertraulichkeit sowie den gänzlichen Verlust der Verfügbarkeit und Nutzbarkeit bedroht sind (Dobratz und Schoger 2010: 94). Auch wenn die grundsätzlichen Arbeitsabläufe lange vor der Etablierung des OAIS-Standards (ISO 14721:2003) entwickelt wurden, bildet dieser doch einen relevanten Bezugsrahmen und liefert eine Nomenklatur für die Darstellung. Wie eingangs erwähnt, lag und liegt der Fokus der Arbeit des Datenarchivs darauf, eine möglichst breite Nachnutzung der archivierten Forschungsdaten zu ermöglichen. Vor diesem Hintergrund ergab sich mehr oder weniger zwangsläufig auch die Notwendigkeit, Maßnahmen zum Erhalt des Datenbestandes durchzuführen. Maßnahmen, die heutzutage unter dem Begriff „Langzeitarchivierung“ diskutiert werden.

Nach Abschluss der oben unter Ingest und Aufbereitung/Dokumentation beschriebenen Prozesse erfolgt die Archivierung im engeren Sinne durch die Erzeugung des Archival Information Packages oder Archivinformationspaketes (AIP)[28] und dessen Überführung in den zentralen Archivspeicher. Die Zusammensetzung des AIP variiert dabei je nach Aufbereitungs-, Dokumentations- und Publikationsziel einer Studie. Es enthält unter anderem:

26 http://www.gesis.org/unser-angebot/daten-analysieren/soziale-indikatoren/simon-social-indicators-monitor/

27 http://www.gesis.org/unser-angebot/daten-analysieren/umfragedaten/

28 „Ein Informationspaket, bestehend aus der Inhaltsinformation und den dazugehörigen Erhaltungsmetadaten, das innerhalb eines OAIS aufbewahrt wird.“ (nestor 2012: 8)

- vom Datengeber übernommene Original-Objekte (zum Beispiel Datensätze, Dokumente) bzw. Kopien davon (SIP)
- im Archiv aufbereitete Datensatzversionen
- Aufbereitungssyntax, die den Bezug zwischen Original- und Archivversion herstellen, sowie weitere Dokumentation der Aufbereitung
- im Archiv erzeugte Metadaten auf Variablen- und Studienebene (im Archiv erstellte DDI-Instanzen) zur Beschreibung der Daten sowie weitere technische und administrative Metadaten.

Zusätzlich zu dem oben beschriebenen AIP werden die im Verlauf des Ingestprozesses ebenfalls erzeugten Dissemination Information Packages im Studienarchiv abgelegt, also die Objekte des AIP, die Nutzern über verschiedene Services in geeigneten Formaten zur Verfügung gestellt werden (OAIS: Access Funktion).

Organisation des Archivspeichers
Der Archivspeicher ist in Form einer dateibasierten Verzeichnisstruktur organisiert. Alle zu einer Studie gehörenden digitalen Objekte (AIP und DIP) werden dort so abgelegt, dass sie den Archiv-Lebenszyklus einer Studie reflektieren. Die Dateien werden nach bestimmten Regeln in definierte Verzeichnisse abgelegt und nach einem einheitlichen Schema benannt (die Originaldateien behalten ihren ursprünglichen Namen). Durch dieses Vorgehen werden die technischen und logischen Beziehungen der Objekte zueinander ausgedrückt.

Das Studienarchiv unterliegt restriktiven Zugriffsrechten. Ein erweiterter Kreis der Archivmitarbeiter hat zwar Leserechte, jedoch besitzen nur zwei Personen das Recht, Dateien hinzuzufügen, zu löschen oder zu ändern. An verschiedenen Punkten im Workflow werden digitale Objekte erzeugt, die der Langfristsicherung zugeführt werden müssen. Der Austausch erfolgt über ein spezielles Verzeichnis außerhalb des digitalen Archivs. Der für den Archivspeicher zuständige Mitarbeiter übernimmt von dort alle anfallenden Objekte, überprüft sie auf die Einhaltung der Langfristsicherungsstandards (Formate, Namenskonventionen etc.), passt sie ggf. an und überführt sie dann an die entsprechende Stelle im Archivverzeichnis. Durch diesen Ablauf wird sichergestellt, dass nur autorisierte Personen langfristsicherungsfähige Objekte in das digitale Archiv überführen bzw. aus dem Archiv entfernen können.

Physische Speicherung und Erhaltung der Archivdaten
Neben den zuvor beschriebenen organisatorischen und technischen Maßnahmen und dem Vorhalten einer modernen und leistungsfähigen IT-Infrastruktur setzt das GESIS Datenarchiv weitere Maßnahmen ein, die zur Sicherstellung des physischen Erhalts von archivierten Objekten geeignet sind und im Desasterfall eine schnelle Wiederherstellung ermöglichen. Dazu gehören die üblichen Maßnah-

men und Vorkehrungen, die Teil eines aktuellen IT-Sicherheitskonzeptes sein sollten. Dies reicht von grundlegenden Maßnahmen zum physischen Schutz vor Ort (wie dauerhaft verschlossene Rechenzentren, die nur für Berechtigte zugänglich sind, über Rauch- und Wassermeldeanlage, Temperaturüberwachung, unterbrechungsfreie Stromversorgung etc.), räumlich getrennter und redundanter Datenhaltung (onsite, off site, offline), Diversifizierung der eingesetzten Speichertechnik sowie regelmäßige Medienmigration.

Sicherstellung der langfristigen Nutzbarkeit und Interpretierbarkeit

Diese technischen Maßnahmen alleine reichen allerdings nicht aus, um die Nutzbarkeit der Datenbestände langfristig zu erhalten, da sie primär auf die physische Sicherung der digitalen Objekte in ihrem jeweiligen Zustand ausgerichtet sind (Bitstream-Preservation). Die weitaus größere Herausforderung für das Langzeitarchiv ist es aber, die Daten, Metadaten und Dokumente so zu erhalten, dass deren Lesbarkeit bzw. Interpretierbarkeit auch in der Zukunft garantiert werden kann. Insbesondere die Bedrohung der digitalen Bestände durch technischen Fortschritt bzw. dessen Kehrseite (Hardware und Speichermedien, die nicht mehr unterstützt oder produziert werden, sowie veraltete Betriebssysteme, Applikationen und Dateiformate) ist eine Konstante in der über fünfzigjährigen Archivarbeit. Der Erhalt der Interpretierbarkeit der archivierten Studien wird in der Hauptsache durch Migrationsstrategien erreicht. Bei Bedarf wurden in der Vergangenheit auch Emulationen bzw. Virtualisierungen als mittelfristige Lösung eingesetzt, bis die Bestände komplett migriert werden konnten.[29]

Die langfristige Nutzbarkeit des Archivbestandes wird insbesondere durch folgende Maßnahmen gesichert:

1. Die technische Entwicklung der im Archiv eingesetzten Speichertechnik, Medien, Software und insbesondere der damit verbundenen Dateiformate wird kontinuierlich verfolgt.
2. Bei der Aufnahme von Studien ins Langzeitarchiv werden Daten und Dokumente in definierte und standardisierte Formate überführt. Diese Maßnahme führt zu einer Reduktion der im Archiv vorhandenen Formate, was einerseits das Verfolgen der technischen Entwicklungen

29 Als zentrale Produktions- und Archivierungsplattform (Aufbereitung, Dokumentation, Archivierung) des Datenarchivs diente bis Ende der 90er Jahre eine IBM Mainframe. Nach und nach wurde die Datenaufbereitung und -dokumentation sowie ein Großteil des Archivbestandes in zeitgemäßere Arbeitsumgebungen migriert. Da der Betrieb des Großrechners hohe Wartungs- und Unterhaltungskosten verursachte, wurde die Hardware abgeschafft und der Betrieb in eine virtuelle Umgebung verlagert. Die Migration der noch auf Mainframe-Medien (Magnetspulen und -kassetten, Festplatten) befindlichen Archivbestände konnte dann mit einem deutlich niedrigeren Zeitdruck realisiert werden.

erleichtert und andererseits im Fall einer Migration, den dafür notwendigen Ressourcenbedarf deutlich verringert.

3. Wenn Bestände von technischen Entwicklungen bedroht sind oder sie in der vorliegenden Form nicht mehr zeitgemäß verarbeitet werden können, werden Strategien zur Migration entwickelt und umgesetzt. Je nach Notwendigkeit setzen dabei die konkreten Maßnahmen auf unterschiedlichen Ebenen an: Reine Datenträgermigrationen (sprich, die Kopie von einem Medium auf ein anderes) finden im Prinzip kontinuierlich statt. Formatmigrationen, da ungleich aufwendiger, werden nur dann vorgenommen, wenn die Gefahr besteht, dass die archivierten Objekte aufgrund des Formates in Zukunft nicht mehr nutzbar sind oder wenn mit der Migration so große Vorteile für die Nutzung oder die Archivarbeit einhergehen, dass der Aufwand zu rechtfertigen ist.[30] Dabei wird sichergestellt, dass die wesentlichen Inhalte der Objekte nicht verändert werden, und durch entsprechende Verfahren und Dokumentation der Prozesse, die Authentizität der migrierten Objekte erhalten bzw. wiederhergestellt.

Ausblick

In der letzten Dekade hat sich die (sozialwissenschaftliche) Datenlandschaft sehr dynamisch entwickelt. Eine Vielzahl neuer Datenserviceeinrichtungen sind entstanden, wie etwa die im Umfeld des RatSWD angesiedelten Forschungsdatenzentren[31], wichtige, auf Dauer angelegte Erhebungsprogramme wurden etabliert, die rund um diese komplexen Studien eigene Daten(zugangs)infrastrukturen aufgebaut haben (zum Beispiel NEPS[32], SHARE[33]) und immer mehr – auch kleinere – Forschungsprojekte stellen ihre Forschungsdaten selbstständig für eine Nachnutzung zur Verfügung. Darüber hinaus verändern sich auch die Daten selbst: Komplexere Forschungsdesigns, neue Formen von Daten und insbesondere auch neue (technische) Möglichkeiten der Verbindung verschiedener Datenquellen und -formen – auch über Disziplingrenzen hinweg – werden zunehmend wichtiger. Insgesamt hat das Thema Forschungsdaten enorm an Bedeutung

30 Eines der größten Migrationsprojekte der letzten Jahre betraf den umfangreichen Bestand an sogenannten Codebüchern. Neben der Dokumentation auf Studienebene beinhalten diese eine vollständige Dokumentation der Variablen des Datensatzes (inklusive der vollständigen Frage- und Antworttexte). In einem mehrjährigen Projekt wurde dabei die in den Codebüchern enthaltene Dokumentation in den XML-basierten Metadatenstandard der Data Documentation Initiative (DDI, www.ddialliance.org) und somit in ein auch für die Langfristsicherung geeignetes Format überführt.

31 http://www.ratswd.de/dat/fdz.php

32 https://www.neps-data.de

33 http://www.share-project.org

gewonnen und damit verbundene Fragen beispielsweise der Zugänglichkeit, des Datenschutzes und auch der Langzeitarchivierung rücken in den Fokus der Aufmerksamkeit. Gleichzeitig wachsen die Anforderungen an Forschende bzw. Forschungsprojekte bezüglich des Umgangs mit den von ihnen erzeugten Daten. Das Erstellen von Datenmanagementplänen bzw. ganz allgemein, die Adressierung von Fragen des zukünftigen Umgangs mit generierten Forschungsdaten wird mehr und mehr Bestandteil der regulären Projektplanung und spielt zunehmend auch bei der Beantragung von Forschungsgeldern eine wichtige Rolle.[34]

Diese Entwicklungen greift das Datenarchiv bereits seit einiger Zeit in unterschiedlichen Projekten und Initiativen auf. Insbesondere mit Blick auf die zunehmend verteilte Dateninfrastruktur und die damit auch tendenziell einhergehende Unübersichtlichkeit aus Nutzersicht hat GESIS seine Aktivitäten im Bereich strukturierender und übergreifender Dienste deutlich verstärkt. So schafft die von GESIS gegründete und mittlerweile gemeinsam mit der Zentralbibliothek Wirtschaftswissenschaften (ZBW) betriebene Registrierungsagentur für Sozial- und Wirtschaftsdaten (da|ra) die Voraussetzungen für eine dauerhafte Identifizierung, Sicherung und Lokalisierung von Forschungsdaten. Damit wird nicht nur die Grundlage für eine zuverlässige und dauerhafte Zitierbarkeit von Daten geschaffen und Datenproduzenten auch ein Reputationsgewinn aus solchen Zitationen ermöglicht, sondern es gehen damit auch deutlich verbesserte Möglichkeiten einher, Publikationen und zugrundeliegende Daten effizienter zu verbinden. Desweiteren arbeitet GESIS im Kontext von da|ra seit Ende 2011 am Aufbau eines umfassenden Datennachweissystems, das existierende Datenbestände auch über die bei da|ra registrierten hinaus beschreibt, auf die entsprechenden Datenquellen verweist und Hinweise über die Zugänglichkeit gibt. Dieser zentrale Nachweis sozialwissenschaftlicher Datenbestände, der in einer späteren Ausbaustufe auch noch um eine Selbstmeldekomponente erweitert wird, soll dazu beitragen, die an vielen verschiedenen Stellen erzeugten und zunehmend auch zur Nachnutzung bereitstehenden Datenbestände für die Wissenschaft sichtbar und damit auch nutzbar zu machen.

Mit dem kürzlich gegründeten 'Archiving and Data Management Training and Information Center'[35] fördert GESIS systematisch das Bewusstsein für die Wichtigkeit guten Datenmanagements und der Bedeutung einer professionellen Archivierung und Pflege (data curation) über den gesamten Lebenszyklus von Daten. Entsprechende Kenntnisse werden durch Trainingskurse, gezielte Beratung und die Bereitstellung von Informationsangeboten vermittelt. Die Angebote des Training Centers richten sich dabei sowohl an Wissenschaftler und Projekte,

34 Für eine ausführliche Darstellung zur Bedeutung von Datenmanagementplänen und ihrem Aufbau siehe Jensen (2011).

35 http://www.gesis.org/en/archive-and-data-management-training-and-information-centre/training-center-home/

die Daten generieren bzw. planen dies künftig zu tun, als auch an Experten aus Datenserviceeinrichtungen, Archiven oder Bibliotheken, um diese im Hinblick auf ihre Tätigkeit weiter zu professionalisieren.

Angesichts der eingangs dargestellten Entwicklungen bedarf aber auch die eigentliche Archivierungstätigkeit des GESIS Datenarchivs im engeren Sinne einer stetigen Anpassung und Weiterentwicklung. Das Datenarchiv bietet mittels eines relativ komplexen Workflows ein weites Spektrum häufig sehr eng miteinander verbundener Dienstleistungen an. Um aber auch solchen Projekten oder Institutionen bedarfsgerechte Angebote machen zu können, die nur einzelne oder mehrere Komponenten aus dem Dienstleistungsportfolio in Anspruch nehmen wollen, beispielsweise Langzeitarchivierungsdienste, sollen zukünftig verstärkt separate, aber standardisierte Dienstleistungen angeboten werden, die dann (weitgehend frei) kombinierbar sein sollen. In diesem Zusammenhang sollen insbesondere Hintergrunddienste für andere Datenzentren im Bereich der Langzeitarchivierung angeboten werden. Darüber hinaus wird das Dienstleistungsangebot des Archivs Anfang 2013 um einen neuen Service erweitert, nämlich die vom Nutzer (weitgehend) eigenständig durchgeführte Publikation und Distribution von Daten mittels eines Daten-Repositoriums (DATORIUM). Mit diesem Angebot reagiert GESIS auf den Bedarf von Datenproduzenten und Forschern nach schnellen und flexiblen Publikations- und Distributionswegen, die es ihnen insbesondere erlauben, Forschungsergebnisse mit anderen Wissenschaftler/inne/n zu teilen und diesen somit auch Sichtbarkeit zu geben. Auch im Hinblick auf den Zugang zu datenschutzrechtlich sensibleren Varianten der bei GESIS gehaltenen Daten wird es in 2013 eine Erweiterung des Services geben. Bisher stellt das Datenarchiv Daten in der Regel als vollständig oder faktisch anonymisierte Public bzw. Scientific Use Files zur Verfügung. Die dafür notwendigen Anonymisierungsmaßnahmen sind in Einzelfällen nahezu unmöglich (zum Beispiel bei Elitestudien) oder können auch teilweise mit einer erheblichen Einschränkung des Analysepotentials für bestimmte Auswertungszwecke einhergehen (zum Beispiel durch die Löschung ganzer Variablen oder identifizierbarer Fälle oder durch Vergröberung oder Zufallsüberlagerung bestimmter zur Re-Identifikation geeigneter Merkmale wie Berufsklassen oder kleinräumige regionale Identifikatoren). Um zukünftig auch die Zugänglichkeit solcher Daten für wissenschaftliche Forschungszwecke systematisch zu verbessern, baut GESIS derzeit ein Secure Data Center auf, das in einer ersten Ausbaustufe für schutzwürdige Datenbestände Vor-Ort-Zugänge an den GESIS Standorten ermöglicht. In einer weiteren Ausbaustufe soll dieser Zugang dann auch um die Möglichkeit eines Fernzugriffes erweitert werden.

Eine der größten Herausforderungen für die Zukunft geht für das Datenarchiv jedoch sicher mit dem Aufkommen neuer Datenformen und -typen einher, deren Archivierung teilweise völlig neue Anforderungen an das Datenarchiv stellen würde.

Literatur

Dobratz, S. und Schoger, A. (2010): Grundkonzepte der Vertrauenswürdigkeit. In: Neuroth, H./Oßwald, A./Scheffel, R./Strathmann, S. und Huth, K. (Hrsg): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3 – 2010, Kapitel 5:2 (ohne Seitenzahl).

Jensen, U. (2011): Datenmanagementpläne. In: Büttner, S./Hobohm, H.-C. und Müller, L. (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock + Herchen, 71-82.

JISC (Joint Information Systems Committee) (2006): Digital Preservation. Continued access to authentic digital assets. Briefing paper 20. November 2006, 1. http://www.jisc.ac.uk/media/documents/publications/digitalpreservationbp.pdf [21.05.2012]

Kleiner, M. (2012): Vorwort. In: Neuroth, H./Strathmann S./Oßwald. A./Scheffel, R./Klump, J. und Ludwig, J. (Hrsg.): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Boizenburg: vwh, 9-13.

nestor-Arbeitsgruppe OAIS-Übersetzung / Terminologie (Hrsg.) (2012): Referenzmodell für ein Offenes Archiv-Informations-System – Deutsche Übersetzung. nestor-Materialien 16. Frankfurt am Main: nestor. urn:nbn:de: 0008-2012051101.

nestor-Arbeitsgruppe Vertrauenswürdige Archive – Zertifizierung (Hrsg.) (2006): Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. Version 1 (Entwurf zur Öffentlichen Kommentierung). nestor-Materialien 8. Frankfurt am Main: nestor. urn:nbn:de:0008-2006060710.

Quandt, M. und Mauer, R. (2012): Sozialwissenschaften. In: Neuroth, H./Strathmann, S./Oßwald. A./Scheffel, R./Klump, J. und Ludwig, J. (Hrsg.): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Boizenburg: vwh, 61-81.

Scheuch, E.K. (1990): From a data archive to an infrastructure for the social sciences. International Social Science Journal (123), 93-111.

Scheuch, E.K. (2003): History and visions in the development of data services for the social sciences. International Social Science Journal (177), 385-399.

Wagner, G. und Huschka, D. (2012): Datenverfügbarkeit reicht nicht, um Replikationsstudien zur Routine zu machen. Working Paper No. 194. RatSWD Working Paper Series Januar 2012. Berlin: RatSWD.

Zenk-Möltgen, W. und Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. GESIS Technical Reports 2012-01.

Forschungsdatenmanagement in den Wirtschaftswissenschaften – Ausgewählte Dienste und Projekte der Deutschen Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft (ZBW)

Olaf Siegert, Ralf Toepfer und Sven Vlaeminck

Einleitung

In einer zunehmend digitalen und vernetzten Forschungslandschaft reicht es für wissenschaftliche Bibliotheken nicht mehr aus, der Scientific Community klassische Dienstleistung wie den (lokalen) Zugang zu Büchern und Fachzeitschriften anzubieten. Vielmehr ist es notwendig, die Wissenschaft als Infrastrukturanbieter im gesamten Publikations- und Forschungsprozess zu unterstützen. So weist unter anderem der Wissenschaftsrat in seinen übergreifenden Empfehlungen zu Informationsinfrastrukturen zu Recht darauf hin, dass neben den traditionellen Aufgaben wissenschaftlicher Bibliotheken *„...neue Funktionen wie der Betrieb von potenziell global nutzbaren Repositorien zur Sicherung eines langfristigen Zugangs zu digitalen Medien oder Forschungsdaten [an Bedeutung gewinnen].“*[1]

Auch der Rat für Sozial- und Wirtschaftsdaten (RatSWD) hält fest, dass in den aktuellen Diskussionen zur Neuausrichtung der Informationsinfrastruktur wissenschaftliche Bibliotheken im Allgemeinen und zentrale Fachbibliotheken im Besonderen eine wichtige Rolle bei der Dokumentation sowie der Bereitstellung von Forschungsdaten spielen könnten.[2]

Die Rolle der wissenschaftlichen Bibliotheken im Zusammenhang mit Forschungsdaten ist aktuell noch nicht hinreichend spezifiziert, kann jedoch vielfältige Formen annehmen[3]. Wie die im Folgenden beschriebenen Dienste und Projekte der ZBW andeuten, können Informationsinfrastruktureinrichtungen aufgrund ihrer hohen Metadatenkompetenz, ihrer großen Erfahrung im Bereich

1 Wissenschaftsrat (2011): Übergreifende Empfehlungen zu Informationsinfrastrukturen. Drs. 10466-11 vom 28.01.2011. Berlin, 14. http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf

2 Vgl. German Data Forum (RatSWD) (ed.) (2010): Building on Progress: Expanding the Research Infrastructure for the Social, Economic, and Behavorial Sciences. Vol. 1. Opladen & Farrington Hills, MI: Budrich UniPress Ltd., 39.
http://www.budrich-unipress.de/media/products/0611760001296569161.pdf

3 Vgl. erhellend hierzu Feijen, M. (2011): What researchers want. www.surf.nl/nl/publicaties/Documents/What_researchers_want.pdf

der Archivierung sowie nicht zuletzt aufgrund der Dauerhaftigkeit ihrer Aufgaben eine wichtige Rolle beim weiteren Ausbau einer Forschungsdateninfrastruktur einnehmen. Erfolgskritisch ist dabei, dass dies nur gemeinsam mit den Stakeholdern, insbesondere der wirtschaftswissenschaftlichen Forschung, erfolgreich gestaltet werden kann. Die Aktivitäten finden dabei nicht im luftleeren Raum statt, sondern fügen sich in die existierende sozial- und wirtschaftswissenschaftliche Forschungsdateninfrastruktur ein, deren Dach in der Bundesrepublik Deutschland der RatSWD bildet.

Data-Sharing in den Sozial- und Wirtschaftswissenschaften

Den Hintergrund der von der ZBW in Kooperation mit anderen Partnern beschriebenen Dienste und Projekte bildet der weitere Auf- und Ausbau einer Kultur des Teilens von Forschungsdaten (data sharing) in den Sozial- und Wirtschaftswissenschaften, wie sie von Huschka et al. (2011)[4] beschrieben wurde. Beim Teilen von Forschungsdaten geht es einerseits um die Nachnutzung bereits erhobener Daten und andererseits um die Nachprüfbarkeit von Forschungsergebnissen. Eine gute Dokumentation bildet dabei die Voraussetzung dafür, dass Forschungsdaten im Rahmen einer Sekundärnutzung sinnvoll verwendet werden können. Eine Publikationskultur mit detaillierten Beschreibungen der Forschungsdaten, die in der Praxis über vermehrte Zitationen belohnt wird, hat sich in den Wirtschafts- und Sozialwissenschaften (anders als in Teilen der Naturwissenschaften[5]) allerdings aus verschiedenen Gründen bislang noch nicht durchgesetzt:

Erstens wurde in den Sozial- und Wirtschaftswissenschaften seit dem 19. Jahrhundert im Wesentlichen mit amtlichen Statistikdaten gerechnet, für die es keine persönliche Autorenschaft gibt[6]. Fehlende Autorenschaft wurde auch zum Standard für von Sozial- und Wirtschaftswissenschaftlern selbst erhobenen und aufbereiteten Forschungsdaten (zum Beispiel bei Surveys). Vor diesem Hintergrund fehlen bis heute die Anreize für Forscher/innen, sich mit einer umfangreichen Datenaufbereitung für Dritte zu beschäftigen. Eine nutzerfreundliche und

4 Vgl. Huschka, D./Oellers, C./Ott, N. und Wagner, G.G. (2011): Datenmanagement und Data Sharing: Erfahrungen in den Sozial- und Wirtschaftswissenschaften. In: Büttner, S./Hobohm, H.-C. und Müller, L. (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock+Herchen Verlag, 35-48. http://www.forschungsdatenmanagement.de/; Wiederabdruck in diesem Band.

5 Vgl. Piwowar, H.A./Day, R.S. and Fridsma, D.B. (2007): Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2 (3), e308. doi:10.1371/journal.pone.0000308
Zudem existieren zum Beispiel in den Geowissenschaften inzwischen schon eigene Fachzeitschriften, wo Forscher/innen ihre Daten als Zeitschriftenartikel veröffentlichen können, wie zum Beispiel „Earth System Science Data“ (http://www.earth-system-science-data.net). Darüber hinaus werden Journals bezüglich Datensätzen weiterentwickelt, wie zum Beispiel im Fall der International Journal of Epidemiology (http://ije.oxfordjournals.org/), bei der in der Rubrik “Cohort Profile“ Datensätze dokumentiert und zitierbar gemacht werden.

6 Typisch ist hier zumeist die Angabe: „Quelle: Statistisches Jahrbuch“ oder „Statistisches Bundesamt“.

damit replikationsfreundliche Aufbereitung von Forschungsdaten ist zudem sehr zeitaufwendig, weil die meist sehr komplexen Berechnungsschritte entsprechend dokumentiert werden müssen. Für diese Arbeit gibt es bislang im Wissenschaftssystem, das immer stärker durch Zeitschriftenaufsätze dominiert wird, keine angemessene Anerkennung.[7]

Durch die Forderung eines freien Zugangs zu den verwendeten Forschungsdaten als Bedingung für die Publikation ihrer Arbeiten entsteht (zweitens) für Wissenschaftler/innen ein Moral Hazard Problem[8]. Denn die bereits unter großem Arbeitsaufwand aufbereiteten Daten werden auch einer wissenschaftlichen Community zur Verfügung gestellt, die sie nutzen kann, obwohl von Nutzerseite kein Beitrag zur Datenaufbereitung geleistet wurde. Die relevanten referierten Zeitschriften veröffentlichen zudem bislang keine zitierfähigen Beschreibungen und Dokumentationen von Daten, wodurch die Zitation für die geleistete Datenaufbereitung als wesentliches Belohnungsinstrument für publizierende Forscher/innen entfällt. Dies kann zu einer Schieflage in der Forschung führen, da Wissenschaftler/innen, die neue Daten generieren und für andere aufbereiten, dafür keine Reputation erlangen und es somit schwerer haben, Universitätskarrieren zu verfolgen. Zudem befürchten viele Forscher/innen einen Missbrauch der Daten durch Dritte, zum Beispiel durch falsche Interpretation oder durch Nutzung der Daten ohne korrekte Zitation der Urheberin bzw. des Urhebers.[9] Schließlich ist drittens die Rechtslage bei der Weitergabe von Datensätzen in vielen Fällen nicht ausreichend geklärt, was ebenfalls zu einer großen Zurückhaltung im Bereich „Data Sharing“ führt.

Damit ist der Hintergrund skizziert, vor dem die nachfolgend beschriebenen Projekte und Dienste der ZBW zu sehen sind.

Open-Access-Journal „Economics“

Erste Erfahrungen im Umgang mit wirtschaftswissenschaftlichen Forschungsdaten hat die ZBW beim Aufbau des Open-Access-Journal *„Economics“* erworben, das gemeinsam mit dem Institut für Weltwirtschaft (IfW) aufgebaut wurde und seit 2007 online ist. Da unter den Einreichungen auch viele empirische Beiträge sind, stellte sich die Frage, wie mit den Berechnungen und Rohdaten umzugehen

7 Bislang wird eher vereinzelt, aber doch von prominenter Seite, beklagt, dass die Arbeit an Forschungsdaten und deren Zur-Verfügung-Stellung von den Belohnungssystemen der Forschungsgemeinschaften nicht angemessen gewürdigt werden (vgl. Lane, J. (2010): Let's make science metrics more scientific. Nature 464, 488-489. Vgl. auch dies. (2009): Assessing the Impact of Science Funding. Science 324 (5932), 1273-1275).

8 Kirchgässner, G. (2000): Homo oeconomicus. Das ökonomische Modell individuellen Verhaltens und seine Anwendung in den Wirtschafts- und Sozialwissenschaften. 2. Auflage. Tübingen: Mohr Siebeck.

9 Diese Aussagen wurden durch eine Umfrage unter Wirtschaftsforschern im Rahmen des EU-Projekts “Economists Online” bestätigt: Vgl.: http://www.neeoproject.eu/NEEO_UserStudy_1.pdf

ist bzw. wie sie in den Publikationsprozess einzubeziehen sind. Ähnlich wie mehrere renommierte US-Fachzeitschriften wurde bei Economics eine sogenannte *„Data Availability Policy"* verabschiedet und in Kraft gesetzt. Diese besagt, dass Autoren empirischer Untersuchungen nach Akzeptanz ihres Beitrags alle notwendigen Daten, Programme und Beschreibungen/Dokumentationen an das Journal schicken, die für eine Replikation der Berechnungen nötig sind[10]. Ergänzend dazu wurde für das Journal ein Datenarchiv[11] aufgebaut, bei dem die Quellen mit bibliographischen Metadaten beschrieben werden (vgl. Abbildung 1).
Außerdem wird für jeden Datensatz ein Persistent Identifier (in Form eines sog. „handle") vergeben, um eine eigenständige Zitation unabhängig vom Zeitschriftenaufsatz zu ermöglichen. Die inhaltliche Steuerung des gesamten Journalprozesses inklusive Datenarchiv erfolgt durch den Herausgeber Institut für Weltwirtschaft, während die ZBW ihr bibliothekarisches und technisches Know-how einbringt.

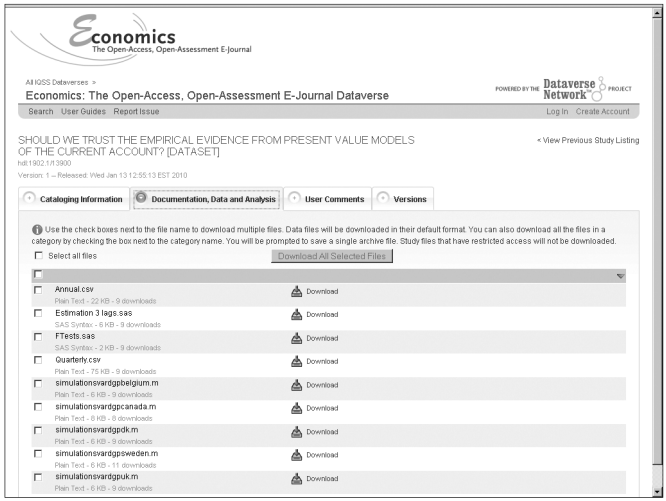


Abbildung 1: Einzelelemente eines Datensatzes im Datenarchiv von „Economics"

10 http://www.economics-ejournal.org/submission/data-availability-policy
11 http://www.economics-ejournal.org/special-areas/data-sets

Economists Online

Ein Beispiel für die Integration eines Datenarchivs in ein Publikationsportal stellt Economists Online[12] dar. Dabei wurde im Rahmen eines EU-Projekts ein Open-Access-Portal aufgebaut, das im Wesentlichen auf den institutionellen Repositorien von 20 europäischen Universitäten und Forschungseinrichtungen basiert und das die kompletten Publikationslisten von etwa 1000 Wirtschaftsforschern nachweist – nach Möglichkeit inklusive Volltext im Open Access, was bei etwa 40% der Veröffentlichungen der Fall ist (ca. 35.000 Volltexte). Die Autoren wurden zudem gebeten, bei empirischen Arbeiten ihre verwendeten Daten, Programme und Dokumentationen an die Repositorien zu liefern, damit diese zur Replikation und für Re-Analysen ebenfalls im Open Access zur Verfügung stehen. Die gelieferten Datensätze werden in einem Datenarchiv[13] inklusive Metadatenbeschreibung und Persistent Identifier bereitgestellt (vgl. Abbildung 2).

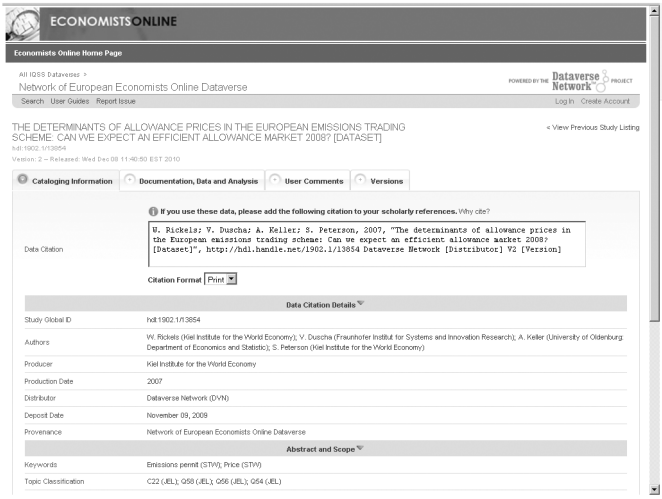


Abbildung 2: Nachweis eines Datensatzes zu einer Publikation in Economists Online

Allerdings sind nur relativ wenig Forschende der Bitte nachgekommen, ihre Datensätze an das Repository zu senden. So konnten „nur" 97 statt der angestrebten 160 Datensätze eingeworben werden. Die Hintergründe für die Zurückhaltung beim Teilen von Forschungsdaten werden im Projektabschlussbericht[14] wie folgt zusammengefasst:

12 http://www.economistsonline.org
13 http://dvn.iq.harvard.edu/dvn/dv/NEEO
14 http://itswww.uvt.nl/its/voorlichting/PDF/NEEO/D1.7-NEEO-Final-Report-2010.pdf

- *„Acquiring datasets from economists is difficult. Many academics are reluctant, to give away their data. Or they have legal (not their data) or ethical (data related to persons) reasons for not disclosing their data.*
- *Many economists who are willing to share their data have already data in the public domain.*
- *Data is not enough, to understand the data we need documentation, supplementary material and the „codes".“*[15]

Wie bereits angesprochen liegt ein weiteres Problem zudem darin, dass Forschende kaum Anreize zur Publikation von Daten und weiteren Materialien haben, da das Wissenschaftssystem solche Tätigkeiten nicht hinreichend „entlohnt".

EDaWaX

Die bei Economists Online gemachten Erfahrungen bildeten den Ausgangspunkt für die Überlegungen des von der DFG geförderten Projekts *„European Data Watch Extended“* (EDaWaX). So wird insbesondere der Aspekt, dass die Bereitstellung der „reinen" Daten, ohne weitere ergänzende Informationen (Dokumentation, Codes, etc.) für Replikationsanalysen nicht ausreicht, adressiert. Das auf 24 Monate angesetzte Projekt wird von der ZBW in Kooperation mit dem RatSWD[16] und dem Institut für Innovationsforschung, Technologiemanagement und Entrepreneurship an der LMU München[17] durchgeführt. Ziel ist es, im Rahmen eines ganzheitlichen Ansatzes ein publikationsbezogenes Datenarchiv am Beispiel der Fachzeitschrift *„Schmollers Jahrbuch/Journal of Applied Social Science Studies“*[18] zu entwickeln.

Um dies zu erreichen, ist eine umfassende Analyse bereits bestehender Lösungen und Rahmenbedingungen für die Implementierung eines solchen Datenarchivs notwendig. Diese Analysen erfolgen in der ersten Phase des Projekts. Dabei werden zunächst auf Basis einer fachwissenschaftlich fundierten Analyse die Anreizprobleme untersucht, die bislang verhindern, dass Daten für Replikationsanalysen in adäquater Form bereitgestellt und genutzt werden.

Parallel dazu werden bereits existierende Lösungen im Kontext von Datenarchiven und wirtschaftswissenschaftlichen Fachzeitschriften sowie die rechtlichen Rahmenbedingungen im Hinblick auf die Eignung für die Anforderungen aus EDaWaX[19] untersucht.

15 Ebd., 29.

16 http://www.ratswd.de

17 http://www.inno-tec.bwl.uni-muenchen.de/personen/professoren/harhoff/index.html

18 http://schmollersjahrbuch.diw.de/schmollersjahrbuch/

19 Die Arbeiten und Ergebnisse des Projekts werden laufend im Blog http://www.edawax.de kommuniziert.

Erste Ergebnisse dieser Untersuchungsschritte lassen sich wie folgt zusammenfassen:

- In einem definierten Sample von 141 nationalen und internationalen wirtschaftswissenschaftlichen Fachzeitschriften konnten 29 Zeitschriften mit Data Availability Policy aufgefunden werden, die sich hinsichtlich ihrer Qualität allerdings erheblich unterscheiden.
- Auch die Bereitstellung und Speicherung der Datensätze in den bestehenden Datenarchiven von Fachzeitschriften wurde untersucht. Hier zeigte sich, dass die allermeisten Daten als zip-Files unter Supplementary Materials zum Download angeboten wurden, und mithin den Anforderungen an Replikationsanalysen in den seltensten Fällen genügen.
- Die Datenarchive werden höchst unterschiedlich gepflegt. Im Rahmen einer Zufallsstichprobe wurde ermittelt, dass durchschnittlich nur 29,3% aller Artikel der 29 Zeitschriften mit Data Availability Policy über angehängte Datensätze verfügen.
- Die eingesetzte Infrastruktur ist zudem nicht persistent und erlaubt keine standardisierte Zitation der Urheber. Die Daten werden nicht nachgewiesen und können dementsprechend auch nur schwer aufgefunden werden. Auch der Aspekt der Langzeitarchivierung wird nicht adressiert. Es zeigt sich deutlich, dass der Aufbau weiter reichender infrastruktureller Lösungen notwendig ist.
- Da bei Forschungsdatenzentren die Datenbereitstellung und das Data Management integrale Aspekte ihrer Arbeit sind, ist eine Kooperation mit diesen sinnvoll. Erste Ergebnisse einer Untersuchung von 46 nationalen und internationalen Forschungsdatenzentren und potentiellen Data Hosts zeigen jedoch, dass diese oftmals keine externen Daten annehmen. Wenn Daten angenommen werden, dann meist nur für spezifische Fachdisziplinen.

Auf Basis der Analyseergebnisse wird in einem zweiten Schritt ein Metadatenschema für die Beschreibung und Auszeichnung der Daten entwickelt bzw. existierende Metadatenschemata (zum Beispiel das da|ra Metadatenschema[20]) für die Zwecke von EDaWaX angepasst.

Die Erkenntnisse münden schließlich im dritten Schritt in die Pilotanwendung eines innovativen publikationsbezogenen Datenarchivs, das in Kooperation mit der renommierten Fachzeitschrift *„Schmollers Jahrbuch/Journal of Applied Social Science Studies“* aufgebaut wird. Projektergebnis wird somit unter anderem ein publikationsbezogenes Datenarchiv sein, das die veröffentlichten Textpublikationen und die dafür verwendeten Forschungsdaten sowie ergänzenden Dokumentationen in einem zusammenhängenden Kontext präsentiert und damit für Dritte nachvollziehbar macht.

20 http://www.gesis.org/dara/home/technische-informationen/dara-metadatenschema

da|ra

Die Problematik der Zitation von Forschungsdaten wird im Kontext von da|ra adressiert. da|ra[21] (Datenregistrierungsagentur) ist ein Service für Datenproduzenten und datenhaltende Organisationen zur Registrierung sozial- und wirtschaftswissenschaftlicher Forschungsdaten. Die Idee dabei ist, dass im Web zugängliche Forschungsdatensätze dauerhafte stabile Internetadressen (sog. Persistent Identifier) erhalten. Dies ermöglicht eine leichtere Zitierbarkeit der Datensätze und damit eine höhere eigene Sichtbarkeit. Den gleichen Weg sind vor einigen Jahren die großen Fachzeitschriften gegangen, die für ihre Online-Ausgaben ebenfalls Persistent Identifier in Form von DOIs (Digital Object Identifier) verwenden.

da|ra richtet sich vor allem an Datenarchive, Forschungsdatenzentren und Servicedatenzentren. Diese können damit zum Beispiel Umfragedaten, Aggregatdaten, Mikrodaten oder Daten aus Quellenstudien registrieren und mit DOIs versehen lassen. Der Service wurde von GESIS aufgebaut und wird im Kontext von DataCite[22] gemeinsam mit der ZBW betrieben, um Datenproduzenten im Bereich Wirtschafts- und Sozialwissenschaften einen Service aus einer Hand bieten zu können[23]. da|ra ist ein gutes Beispiel dafür, wie Forschungsdatenzentren und Bibliotheken im Kontext des Managements von Forschungsdaten kooperieren können. Die Forschungscommunity wird in Form eines wissenschaftlichen Beirats eingebunden, der die strategische Weiterentwicklung des Service steuert.

Fazit

Vor dem Hintergrund zunehmender digitaler Verfügbarkeit von Informationen werden neue Anforderungen aus der Wissenschaft an die Forschungsbibliotheken herangetragen. Über die „klassischen" Dienste rund um die Bereitstellung von Fachinformation für die Endnutzung hinaus, gilt es, Services für den gesamten Forschungs- und Publikationsprozess anzubieten. Mit ihren Grundkompetenzen und Erfahrungen in unter anderem den Bereichen Metadaten, Langzeitarchivierung, Dokumentation und Nutzerservices können Bibliotheken zu einem wichtigen Akteur beim weiteren Auf- und Ausbau der Forschungsdateninfrastruktur werden. Die oben skizzierten Ansätze aus der Praxis der ZBW deuten an, wo Schwerpunkte bibliothekarischer Arbeit im Forschungsdatenmanagement liegen könnten; nämlich im Metadatenmanagement, in der Verfügbarmachung von Forschungsdaten sowie in der Organisation und Vergabe persistenter Iden-

21 http://www.gesis.org/dara
22 http://datacite.org
23 http://www.zbw.eu/presse/pressemitteilungen/2010_12_01.htm

tifikatoren für Forschungsdaten. Die Rolle der wissenschaftlichen Bibliotheken im Zusammenhang mit Forschungsdaten ist aktuell noch nicht hinreichend ausdiskutiert. Der Dialog zwischen den Forschenden, den Forschungsdatenzentren, Datenarchiven, Verlagen, Forschungsförderern und Bibliotheken muss weiter intensiviert werden, um zu einer optimalen Rollenverteilung beim Ausbau der Forschungsdateninfrastruktur zu kommen.

Stärkung der Forschungskooperation und des Datenmanagements in der Psychologie mit PsychData

Erich Weichselgartner, Armin Günther und Ina Dehnhard

Entstehung von Forschungsdatenzentren

Verschiedene technische und gesellschaftliche Entwicklungen haben in den letzten Jahren dazu geführt, dass Forschungsdaten eine immer größere Beachtung finden und von manchen sogar als das „neue Gold" bezeichnet werden. Durch die Möglichkeit, gigantische Datenmengen zu erheben und zu verarbeiten, tritt die „Data Driven Science" als neues Wissenschaftsparadigma auf den Plan (Murray-Rust 2007). Zahlreiche wissenschaftspolitische Erklärungen fordern den freien Zugang zu Forschungsdaten (unter anderem ICSU Principles for Dissemination of Scientific Data 2002; Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen 2003; OECD Declaration on Access to Research Data from Public Funding 2004). Die bedeutenden Forschungsförderer wünschen die maximale Ausschöpfung der von ihnen vergebenen Mittel und fordern (NIH 2003; Wellcome Trust 2010) oder empfehlen (DFG 2009) die Weitergabe der Forschungsdaten.

Bei aller Zustimmung zu diesen politisch und ökonomisch motivierten Forderungen darf nicht übersehen werden, dass die Archivierung zwar für alle wissenschaftlichen Disziplinen einen Gewinn bringen kann (zum Beispiel zur Verhinderung wissenschaftlichen Fehlverhaltens), aber erst die rege Nachnutzung der archivierten Daten die volle Rendite erbringt. Ob und wie sich die Daten anderer für die eigene Forschung nutzen lassen, hängt jedoch stark von der wissenschaftlichen Disziplin ab. Während es in den Sozial- und Wirtschaftswissenschaften eine lange Tradition der Sekundärnutzung von Daten gibt, wird deren Nutzen in der Psychologie von vielen immer noch in Frage gestellt. Breckler (2009) fasst die Lage in der Psychologie prägnant zusammen:

> *„The data culture of psychology is one of limited sharing, and then only among a select few with demonstrated competence and legitimate need." (S. 41)*

Gegen diese weltweit übliche Praxis hat PsychData, das 2002 am Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) gegründete Forschungsdatenzentrum für die Psychologie, einen schweren Stand, obwohl es zumindest ein nationaler Versuch ist, einen Kulturwandel im Fach einzuleiten. Da die meisten Forschungsdatenzentren disziplinär organisiert sind, um den jeweiligen Anforderungen unterschiedlicher wissenschaftlicher Disziplinen entsprechen zu können, zählt PsychData zu einem von zahlreichen disziplinspezifischen Forschungsdatenzentren, die beispielsweise beim World Data System (38 Mitglieder) oder beim Rat für Sozial- und Wirtschaftsdaten (20 Einrichtungen der Dateninfrastruktur) nachgewiesen werden. Was war die Motivation für die Gründung von PsychData und wie sieht die Archivierungspraxis aus?

Entstehung von PsychData

PsychData wurde auf Anregung von forschungsaktiven Fachvertretern geschaffen, die aufwendig erhobene Längsschnittdaten der Nachwelt erhalten wollten und die ihre Daten in existierende Archive nicht fachgerecht einspielen konnten, da diese an die Verhältnisse in der Psychologie nicht angepasst waren. Die Unterstützung des Faches für das in der Gründungsphase befindliche Archiv erläuterte der Präsident der Deutschen Gesellschaft für Psychologie (DGPs) in seinem Bericht zur Lage der Psychologie (Silbereisen 2003):

> *„Will man die Forschungskooperation stärken, dann bedarf es besonderer Anreize. Dies wiederum hat viel zu tun mit der Infrastruktur, etwa hinsichtlich der Ausstattung mit Geräten oder der Vorhaltung von Daten und methodischer Expertise. Gerade bei letzterem Gesichtspunkt ist die Psychologie in Deutschland schon gegenüber den Sozialwissenschaften benachteiligt – bis heute fehlt uns eine Einrichtung, welche die Daten wichtiger psychologischer Untersuchungen zumindest für ein Fachpublikum zur Sekundäranalyse bereit hält. Von einer Datenbank der Psychologie im Sinne der Vorhaltung von Originalerhebungen erwarte ich mir nicht nur eine bessere Verwertung vorhandener Befunde, sondern vor allem auch einen Anstoß, solche Untersuchungen mit Langzeiteffekt überhaupt zu beginnen." (S. 3)*

Die Arbeiten am Forschungsdatenzentrum PsychData begannen mit Unterstützung der Deutschen Forschungsgemeinschaft (DFG) im Jahre 2002 (Montada und Weichselgartner 2002; Krampen und Weichselgartner 2005). Bereits im Jahr darauf wurde mit der Archivierung der ersten Datensätze begonnen. Das Projekt ist am Leibniz-Zentrum für Psychologische Information und Dokumentation

(ZPID) in Trier angesiedelt, einem 1972 gegründeten, überregionalen Fachinformationszentrum für die Psychologie in den deutschsprachigen Ländern. Diese institutionelle Verankerung von PsychData stellt den Fortbestand des Archivs langfristig sicher, ein Aspekt, der bei einem auf möglichst dauerhafte Sicherung von Daten ausgerichteten Forschungsdatenzentrum nicht unerheblich ist. Ein weiterer Vorteil dieser institutionellen Verankerung besteht darin, dass PsychData mit anderen Produkten der Informationsinfrastruktureinrichtung ZPID, wie zum Beispiel PSYNDEX, einer Referenzdatenbank für psychologische Literatur und Testverfahren, oder PsychAuthors, einer Autoren-Datenbank, unmittelbar vernetzt werden kann. Verglichen mit anderen nationalen oder gar internationalen Datenzentren ist die Ausstattung von PsychData mit Personal- und Sachmitteln sicherlich als bescheiden zu betrachten. Dennoch konnte seit der Einrichtung von PsychData, nicht zuletzt aufgrund der günstigen institutionellen Rahmenbedingungen, eine Infrastruktur entwickelt werden, die den Vergleich mit großen Forschungsdatenzentren nicht scheuen muss und durchaus internationale Beachtung gefunden hat (Sablonnière et al. 2012; Ruusalepp 2008).

Ziele und Inhalte von PsychData

Entsprechend dem Serviceauftrag seiner Trägereinrichtung ZPID richtet sich das Angebot von PsychData zunächst an die deutschsprachige Psychologie im internationalen Kontext. Doch selbst bei einer Fokussierung auf deutschsprachige Länder war von vorneherein klar, dass schon aufgrund des Umfangs der wissenschaftlichen Datenproduktion eine umfassende Archivierung psychologischer Forschungsdaten nicht Ziel des Datenzentrums PsychData sein konnte. PsychData hat es sich vielmehr zur Aufgabe gemacht, besonders bedeutsame Datenbestände aus der psychologischen Forschung zu sichern (vgl. Weichselgartner et al. 2011: 195). Konkret gelten als Selektionskriterien für PsychData (vgl. Weichselgartner 2008):

- Daten aus psychologischen Längsschnittstudien, die sich aufgrund ihrer naturgemäß langwierigen Erhebungsdauer (einige Jahre bis Jahrzehnte), meist aufwendigen Anlage und historischen Bedeutung unmittelbar für eine Archivierung empfehlen.
- Daten aus repräsentativen Querschnittstudien, bei denen aufgrund ihres Umfangs oder des Aufwands bei der Datenerhebung eine Neuerhebung vielfach nicht in Frage kommt.
- Daten zur Verteilung von Testwerten psychologischer Testverfahren in repräsentativen Stichproben. Diese ‚Normdaten' stellen (auch in ihrer zeitlichen Staffelung) wichtige Bezugsgrößen bei der Anwendung der entsprechenden Testverfahren etwa in der Psychodiagnostik dar.

- Daten zur Verbreitung und zum Verlauf von psychischen und psychosomatischen Erkrankungen. Diese Daten sind für die allgemeine, historische oder kulturvergleichende epidemiologische Forschung bedeutsam.
- Daten aus Studien unter historisch einmaligen Rahmenbedingungen. Diese Daten sind prinzipiell nicht replizierbar und erhalten ihre Relevanz nicht zuletzt auch unter dem Gesichtspunkt der Bewahrung des kulturellen Erbes.

Aufgrund der vielfach konstatierten Zurückhaltung vieler Wissenschaftler bei der Veröffentlichung ihrer Daten (Wolins 1962; Wicherts et al. 2006; Botella und Ortego 2010), dienen diese (prinzipiell offenen und erweiterbaren) Selektionskriterien derzeit allerdings nicht als Ausschlusskriterien für die Aufnahme von Forschungsdaten. Es sind vielmehr Selektionskriterien, an denen sich PsychData in erster Linie bei der aktiven Akquise neuer Datenbestände orientiert. Daneben akzeptiert PsychData aber auch Dateneinreichungen aus allen Feldern der Psychologie, die nicht in eine der genannten Kategorien fallen, wenn sie das Minimalkriterium erfüllen, dass die Daten in mindestens einer qualitätsgeprüften Veröffentlichung verwendet worden sind.

Der Datenbestand des Forschungsdatenzentrums PsychData wächst seit seiner Einrichtung langsam, aber kontinuierlich. Derzeit umfasst der Archivbestand von PsychData ca. 30 Millionen Datenpunkte aus 37 Studien mit insgesamt 61 Forschungsdatensätzen. Daneben bestehen Backfiles für ca. 300 Forschungsdatensätze aus groß angelegten Längsschnittstudien, die im Rahmen des Möglichen nach und nach erfasst werden. Im Folgenden werden die Komponenten und Prozesse von PsychData, sowie die Integration mit weiteren Informationsangeboten genauer dargestellt. Dabei sollen besonders die beiden Prinzipien deutlich werden, die für das Konzept von PsychData konstitutiv sind: Das ist zum einen die Fachorientierung, d.h. die Orientierung an den spezifischen Bedürfnissen des Fachs Psychologie, und zum anderen die Qualitätsorientierung, d.h. der Anspruch, Datenarchivierung für Datengeber und Datennehmer auf einem hohen Qualitätsniveau anzubieten.

Strukturen

Die in einem psychologischen Forschungsprojekt generierten Daten werden in PsychData zusammen mit Metadaten sowie ergänzenden Dateien zu einem sogenannten „Datensatz“ zusammengefasst. (Der Ausdruck „Datensatz“ ist hier also nicht in seinem engen datenbanktechnischen Sinne zu verstehen.) Diese drei Komponenten eines PsychData-Datensatzes – Forschungsdaten, Metadaten und ergänzende Dateien – sollen nun kurz beschrieben werden.

Forschungsdaten

Nicht nur zwischen, sondern ebenso innerhalb wissenschaftlicher Disziplinen trifft man auf „Forschungsdaten" in einer Vielzahl unterschiedlicher Erscheinungsformen und Formate. Das gilt auch für die Psychologie: Die Videoaufzeichnung einer Mutter-Kind-Interaktion, ein Transskript der Dialoge zwischen den Interaktionspartnern, das vom Kind gemalte Bild, die EKG-Aufzeichnung der Herzfrequenz bei der Mutter oder ihre Antworten in einem Fragebogen zum Erziehungsstil – all dies kann in einem gewissen Sinne als „Forschungsdatum" verstanden und zur weiteren wissenschaftlichen Auswertung und Nutzung archiviert werden.

In PsychData wird der Begriff des Forschungsdatums in einem engeren Sinne verwendet: Als Forschungsdaten archiviert werden im Wesentlichen quantitative Daten, die sich aus psychologischen Messungen ergeben. Die verwendeten Messverfahren können dabei sehr unterschiedlicher Natur sein: Fragebögen, psychologische Tests, physiologische Messinstrumente usw. Entscheidend ist, dass psychologisch relevante empirische Strukturen (empirische Relationen) in numerische Strukturen (numerische Relationen) abgebildet werden. Diese Messwerte werden in PsychData als Forschungsdaten archiviert. „Qualitative Daten" wie Texte, Bilder, audiovisuelle Aufzeichnungen etc. werden nicht als Forschungsdaten erfasst, sie können allenfalls (in gewissen Grenzen) als „Stimulusmaterial" bei der Dokumentation der Datenerhebungsverfahren mit archiviert werden.

Damit folgt PsychData der überwiegenden Praxis in der psychologischen Forschung, psychologische Strukturen und Prozesse mit Hilfe quantitativer Verfahren abzubilden und zu analysieren. Für psychologische Forschungsansätze, die einer qualitativen Methodologie verpflichtet sind und die ein anderes Verständnis von Forschungsdaten und Empirie haben, befindet sich an der Universität Bremen ein eigens auf diesen Datentypus spezialisiertes Datenzentrum (Archiv für Lebenslaufforschung (ALLF[1])) im Aufbau, das bereits über einen Grundbestand an qualitativen Interviewdaten verfügt.

Die quantitativen psychologischen Forschungsdaten werden in PsychData in standardisierter Form archiviert und zur weiteren Nutzung zur Verfügung gestellt. Formal betrachtet müssen sich die Forschungsdaten als Datentabellen oder -matrizen darstellen lassen, bei denen jede Zelle den Messwert für eine Variable (= Spalte) bei einer Untersuchungseinheit (= Zeile; in der Regel entspricht die Untersuchungseinheit einer Person, möglicherweise aber auch einem Personenpaar oder einer Familie etc.) enthält. Komplexere zum Beispiel geschachtelte Datenstrukturen sind nicht möglich – allenfalls können in den Metadaten strukturelle Abhängigkeiten zwischen Variablen beschrieben werden. Technisch gesehen müssen sich die Datenmatrizen entsprechend als CSV-Dateien (mit den Variablennamen in der Kopfzeile) ablegen lassen. Auch bei dem Archivierungs-

1 http://www.lebenslaufarchiv.uni-bremen.de/index.php?site=about&lang=de

format der Forschungsdaten handelt es sich um das universell lesbare Textformat ASCII. Zusätzlich erfolgt die Speicherung der Daten in einer entsprechend strukturierten (eine Variable – ein Datensatzfeld) MySQL-Datenbank.

In dieser Standardisierung der Forschungsdaten sowohl in ihrer logischen Struktur als auch ihrem Speicherformat unterscheidet sich PsychData von Archiven, die Forschungsdaten in heterogenen, von den Datengebern selbst gewählten Datenstrukturen und Dateiformaten archivieren (wie zum Beispiel Dryad[2]). Diese Standardisierung ermöglicht beispielsweise Qualitätskontrollen (zum Beispiel Abgleich von Kodebüchern mit Daten) oder die Entwicklung von Suchprozeduren auf Variablenebene, die bei nichtstandardisierter Archivierung der Forschungsdaten nicht möglich wären. Zusätzlich wird durch die Möglichkeit, sämtliche Daten in nichtproprietären Textformaten auszugeben, die kurz- und langfristige Zugänglichkeit der Daten verbessert.

Auf der anderen Seite resultiert aus dieser Standardisierung ein nicht unerheblicher Aufwand bei der Archivierung der Daten. Diese werden von den Datengebern in der Regel in proprietären Dateiformaten geliefert (insbesondere als sav-Dateien, dem Format für Datendateien der in der Psychologie weit verbreiteten Statistiksoftware SPSS) und müssen erst in das PsychData kompatible Format konvertiert werden. Aufgrund von Problemen mit fehlenden Werten, speziellen Datenformaten usw. ist in der Regel eine zusätzliche, mehr oder weniger aufwendige Nachbearbeitung erforderlich.

Metadaten

Das von PsychData verwendete Metadatenset orientiert sich an den internationalen Metadatenstandards Dublin Core Metadata Initiative[3] und Data Documentation Initiative DDI 2.0[4]. Durch diese Verwendung einschlägiger Metadatenstandards wird die Interoperabilität sowie Durchsuchbarkeit der Inhalte gewährleistet. Während es sich beim Dublin Core Schema um ein Metadatenset handelt, das auf verschiedenste Web-Ressourcen angewendet werden kann, wurde DDI für die Beschreibung sozial- und wirtschaftswissenschaftlicher Datensätze entwickelt.

Zu jedem Datensatz in PsychData werden neben den eigentlichen Forschungsdaten (den „Messwerten“) Metadaten auf Variablen-, Studien- und Dateiebene erfasst.

Metadaten auf Variablenebene

Die Metadaten auf Variablenebene werden in sogenannten „Kodebüchern“ zusammengefasst. Jeder in einer Datenmatrix enthaltenen Variablen entspricht genau ein Kodebucheintrag. In diesem Kodebucheintrag wird in knapper Form

2 http://www.datadryad.org
3 http://dublincore.org
4 http://www.ddialliance.org

dokumentiert, was mit der jeweiligen Variablen erfasst oder gemessen wurde und was die in der Datenmatrix enthaltenen Zahlen und Zeichenketten bedeuten. Erst auf Basis des Kodebuchs kann ein Datennutzer beispielsweise erkennen, dass die Zahlen in Spalte 7 einer Datenmatrix den in Millisekunden gemessenen Reaktionszeiten der getesteten Versuchspersonen bei Aufgabe 5 im Gedächtnistest X entsprechen. Das Kodebuch dient also der Dekodierung der in den Datenmatrizen enthaltenen Zahlen und Zeichenketten: Ohne das zugehörige Kodebuch blieben diese bedeutungslose, uninterpretierbare Zeichen. Eine detaillierte Dokumentation der teilweise komplexen Datenerhebungsverfahren (zum Beispiel bei Leistungstests oder psychophysiologischen Messungen) kann das Kodebuch allerdings nicht leisten. Es benennt aber in der Regel das verwendete Verfahren, so dass der Datennutzer die entsprechenden Dokumentationen zu diesen Datenerhebungsverfahren identifizieren und zu Rate ziehen kann. Wenn der Datengeber zugestimmt hat, werden außerdem ergänzende Dateien, die dem Verständnis der Forschungsdaten dienen, ebenfalls dem Datennutzer bereitgestellt.

Wie schon bei den Forschungsdaten, so ist auch das Format der variablenbezogenen Metadaten standardisiert: Die Syntax eines PsychData-Kodebuchs ist fest vorgegeben. Damit sind auch die Informationen, die auf dieser Meta-Ebene zu einer Variablen gegeben werden können, festgelegt und begrenzt. Gespeichert wird das Kodebuch ebenso wie die Datenmatrizen in zweifacher Form: Zum einen werden alle variablenbezogenen Metadaten in einer MySQL-Datenbank erfasst, zum anderen werden Textdateien generiert, so dass der Zugriff auf ein PsychData-Kodebuch auch ohne Spezialsoftware möglich ist.

Die Vor- und Nachteile dieser Standardisierung auf Kodebuchebene entsprechen den Vor- und Nachteilen standardisierter Datenmatrizen. Sie dient insbesondere der Qualitätssicherung. Nur auf diese Weise kann beispielsweise durch automatisierte Prüfalgorithmen sichergestellt werden, dass jede Variable in einer Datenmatrix auch durch einen Kodebucheintrag dokumentiert wird, gültige und fehlende Werte deklariert sind usw. Auf der anderen Seite ist die Erstellung der Kodebücher in der Regel zeitaufwendig, da sie sich nur in Ausnahmefällen automatisch aus den Originaldateien der Datengeber generieren lassen.

Metadaten auf Studienebene

Bestandteil jedes PsychData-Datensatzes ist eine detaillierte, wiederum standardisierte Dokumentation der wissenschaftlichen Untersuchung, in deren Rahmen die archivierten Daten erhoben wurden. Diese studienbezogenen Metadaten dienen der Erfassung all jener Informationen, die über die Variablenbeschreibungen hinaus notwendig sind, um die Forschungsdaten langfristig interpretieren zu können. Konkret umfassen sie bibliografische Angaben, den wissenschaftlichen Kontext der Studie und naturgemäß in besonderem Maße Informationen zu den Rahmen- und Durchführungsbedingungen der eigentlichen Datenerhebung.

Angereichert werden sie durch die Vergabe von psychologiespezifischen Klassifikationen und Schlagwörtern gemäß den PSYNDEX-Terms (ZPID 2011), wodurch die Absuchbarkeit der Metadaten verbessert wird.

Besonders bei der Auswahl der in PsychData erfassten Metadaten auf Studienebene wurden fachspezifische Aspekte berücksichtigt (Weichselgartner 2008). Ziel war es, die Informationen zu erfassen, die aus fachlich-psychologischer Sicht für die Evaluation der Datenqualität besonders wichtig sind und die es potenziellen Datennehmern ermöglichen, die Relevanz dieser Daten im Rahmen eigener Problemstellungen einzuschätzen.

So sind genaue Angaben zur Stichprobenziehung, der Probandenrekrutierung, der Stichprobengröße und dem Rücklauf (bzw. Ausfall) aus fachlicher Sicht oftmals wichtige Basisinformationen zur Qualitäts- und Relevanzbeurteilung psychologischer Daten, die entsprechend in den studienbezogenen Metadaten von PsychData auch erfasst werden (siehe Tabelle 1). Die Bedeutung und Gewichtung dieser Kriterien hängt wiederum vom Grundtyp der Datenerhebung ab. So kommt der Stichprobengröße und Aspekten der Repräsentativität beispielsweise in einem experimentellen Untersuchungsdesign eine andere, weniger gewichtige Bedeutung zu als in einer Erhebung, bei der Normdaten zu einem Testverfahren erhoben werden sollen. Entsprechend werden in den studienbezogenen Metadaten auch drei Grundtypen der Datenerhebung unterschieden: *„Experimentaldaten"*, bei denen die Forschungsdaten im Rahmen eines experimentellen oder quasiexperimentellen Untersuchungsdesigns erfasst wurden, *„Befragungsdaten"*, bei dem die Daten in einem nichtexperimentellen Setting durch die Befragung von Personen hinsichtlich ihres eigenen oder fremden Erlebens und Verhaltens gewonnen wurden, und *„Testdaten"*, die vor allem der Entwicklung oder Weiterentwicklung eines psychologischen Testverfahrens dienen.

Stichprobe	*Quotenstichprobe*
Probandenrekrutierung	*Die Versuchspersonen wurden über die Schulen und Klassenleherer rekrutiert und im Klassenverband befragt.*
Stichprobengröße	*1434 Probanden*
Rücklauf/Ausfall	*Von den 811 in Welle 1 befragten Jugendlichen der Kohorte A (11,5 jährige) konnten in den Folgewellen erreicht werden: 704 Welle 2; 634 in Welle 3; 566 in Welle 4; 566 in Welle 5; 484 in Welle 6; 408 in Welle 7.* *Von den 623 in Welle 1 befragten Jugendlichen der Kohorte B (14,5 jährige) konnten in den Folgewellen erreicht werden: 543 in Welle 2; 439 in Welle 3; 352 in Welle 4.*

Tabelle 1: Auszug aus PsychData-Metadaten auf Studienebene: Angaben zur Erhebungsmethode (Silbereisen und Eyferth 2004)

Neben diesen fachspezifischen Aspekten orientieren sich die von PsychData verwendeten Metadaten auf Studienebene zusätzlich an den internationalen Metadatenstandards Dublin Core Metadata Initiative und Data Documentation Initiative (DDI 2.0). Während es sich beim Dublin Core Schema um ein Metadatenset handelt, dass auf verschiedenste, fachlich nicht weiter eingegrenzte Web-Ressourcen angewendet werden kann, wurde DDI speziell für die Beschreibung sozialwissenschaftlicher Datensätze entwickelt. Derzeit wird DDI von PsychData zwar noch nicht als Ausgabeformat genutzt, doch sind entsprechende Entwicklungen geplant, um auf diese Weise die Interoperabilität und Durchsuchbarkeit der PsychData-Inhalte weiter zu verbessern. Bereits jetzt ist durch die durchgängige Standardisierung der Metadaten auch auf Studienebene eine strukturierte Suche über die verschiedenen PsychData-Datensätze (bzw. Studien) hinweg zum Beispiel nach Schlagwörtern möglich.

Da nicht prinzipiell ausgeschlossen werden kann, dass über die standardisierten Metadaten hinaus noch zusätzliche Informationen zu einem vertieften Verständnis der Daten benötigt werden, stellen Literaturangaben ebenfalls einen Bestandteil der Metadaten auf Studienebene dar. Dabei wird sowohl Literatur angegeben, die direkt auf den Datensatz bezogen ist, als auch solche, die einen weiteren Einblick in das zugehörige Forschungsgebiet geben kann.

Metadaten auf Dateiebene

Zu allen Dateien, die einem PsychData-Datensatz zugerechnet werden, werden ebenfalls Metadaten erfasst, die eine kurze formale (zum Beispiel Dateiname) und inhaltliche Beschreibung dieser Dateien umfassen. Zu diesen Dateien gehören die von den Datengebern übermittelten Originaldateien mit den Forschungsdaten und gegebenenfalls weitere Dateien mit ergänzendem Material wie Fragebögen, Instruktionsmaterial, Steuerprogramme, etc. in unterschiedlichen Dateiformaten sowie die standardisierten PsychData-Datenmatrizen und PsychData-Kodebücher als einfache Textdateien. In der Praxis werden meist nur letztere auch an Datennutzer weitergegeben; die Erlaubnis zur Weitergabe anderer zusätzlicher Dateien wird vom Datengeber bei der Dateneingabe ins Archiv festgelegt.

Ergänzende Dateien

Forschungsdaten und sämtliche Metadaten auf Variablen-, Studien- sowie Dateiebene werden in PsychData in standardisierter Form in einer MySQL-Datenbank erfasst. Zusätzlich können noch ergänzende Dateien als dritte Komponente eines PsychData-Datensatzes archiviert werden. Diese ergänzenden Dateien enthalten in nicht standardisierter Form Informationen und Unterlagen, die für das Verständnis und die Dokumentation eines Datensatzes bedeutsam sind wie Stimulusmaterialien, Fragebögen, Kodieranweisungen und ähnliches. Diese

Inhalte selbst werden nicht in der MySQL-Datenbank erfasst, sondern es erfolgt lediglich eine summarische Inhaltsangabe in den dateibezogenen Metadaten (siehe oben).

Datenübergabe und -archivierung

Bei einer Übergabe von Forschungsdaten an das Datenarchiv PsychData stellen Rechtssicherheit, Aufbereitung der Forschungsdaten nach hohen Qualititätsstandards und die Zitierbarkeit als Grundlage für eine wissenschaftliche Anerkennung der Datenproduktion wichtige Prinzipien dar. Um Rechtssicherheit herzustellen, wird zunächst ein Datengebervertrag zwischen dem Wissenschaftler und PsychData geschlossen. Darin erklärt der Datengeber unter anderem, dass er mit einer Weitergabe der Forschungsdaten für wissenschaftliche Zwecke (Forschung und Lehre) durch PsychData einverstanden ist. Zur Archivierung werden vom Datengeber neben den Forschungsdaten selbst alle Materialien benötigt, die zur Erstellung der Metadaten auf Variablen- und Studienebene notwendig sind. Im Fall der Metadaten auf Variablenebene kann dies ein bereits vorhandenes Kodebuch sein, der Originalfragebogen oder eine vollständig dokumentierte Datendatei der jeweiligen verwendeten Statistiksoftware. Zur Übergabe der Metadaten auf Studienebene stellt PsychData ein Dokumentationsformular bereit, das vom Datengeber auszufüllen ist. Aus den gelieferten Informationen wird durch verschiedene Überarbeitungs-, Umwandlungs- und Prüfschritte ein konsistenter PsychData-Datensatz erstellt. Ein wichtiger Bearbeitungsschritt stellt dabei die Überprüfung der Anonymisierung der Forschungsdaten dar. Da in psychologischen Untersuchungen häufig sensible Daten erhoben werden, spielen Belange des Datenschutzes eine wichtige Rolle bei der Dateneingabe, -aufbereitung und -weitergabe. Im Regelfall werden bereits anonymisierte Forschungsdaten an PsychData übergeben.

Web-basiertes Dokumentationstool

Die gesamte Bearbeitungsprozedur bindet in hohem Maß Zeit- und Arbeitsressourcen, garantiert allerdings auch ein hohes Qualitätsniveau der archivierten Forschungsdaten und zugehöriger Metadaten. Um diesen Zeit- und Arbeitsaufwand zu reduzieren, gleichzeitig aber hohen Qualitätsansprüchen gerecht zu werden, wird von PsychData seit 2010 ein eigens entwickeltes, web-basiertes Dokumentationstool zur Datenübergabe bereitgestellt. Mit diesem können sowohl die Metadaten auf Studienebene erstellt, als auch Kodebücher und Forschungsdaten entweder direkt eingegeben oder hochgeladen werden. Sowohl bei Kodebüchern als auch bei Forschungsdaten erfolgen automatische Validitäts- und

Konsistenzkontrollen, so dass erste Prüfschritte bereits durchlaufen wurden und eine weitere Bearbeitung im Archiv wesentlich zügiger von statten gehen kann.

Falls während der Überarbeitung durch das Archiv Fragen oder Unklarheiten auftauchen, werden diese direkt mit dem Datengeber geklärt. Auch die Bereitstellung der Forschungsdaten erfolgt erst nach einer finalen Revision des PsychData-Datensatzes durch den Datengeber. Erst dann werden die studienbeschreibenden Metadaten auf der Psychdata-Homepage veröffentlicht, wo sie entweder thematisch (Browsen) oder nach Schlagwörtern durchsucht werden können. Die Forschungsdaten und die Metadaten auf Variablenebene sowie vorhandene zusätzliche Materialien können nur nach Abschluss eines Datennehmervertrags von PsychData bezogen werden. Im Gegensatz zu den Metadaten auf Studienebene sind die Kodebücher nicht direkt über die Homepage verfügbar. Dieses Vorgehen ist darin begründet, dass in der Psychologie häufig publizierte Testverfahren angewendet werden, die kommerziell über spezielle Testverlage vertrieben werden. Um den Bedenken der Testverlage (und auch der Testautoren) bezüglich eines möglichen Missbrauchs der Fragebogenitems entgegenzuwirken, können die Kodebücher nicht frei im Internet eingesehen werden (vgl. Fahrenberg 2012), wie es beispielsweise in den Sozialwissenschaften üblich ist.

Die technische Archivierung der Forschungsdaten und der Metadaten erfolgt durch quelloffene Software (unter anderem Unix, MySQL, Apache, PHP), das Zusammenspiel von mehreren räumlich getrennten Servern und definierte Backups auf magnetischen und optischen Medien. Integrität und Schutz vor Manipulation werden durch Prüfsummen und eine abgestufte Zugriffskontrolle erreicht.

Alle bei PsychData archivierten Forschungsdatensätze erhalten ab dem Zeitpunkt der Bereitstellung einen Digital Object Identifier (DOI), durch den eine permanente Verfügbarkeit garantiert ist. Dies entspricht zum einen dem Anspruch einer langfristigen Verfügbarkeit von Forschungsdaten, zum anderen ermöglicht es die Zitierbarkeit von Forschungsdaten. Diese Zitierfähigkeit wird als wichtiger Schritt gesehen, damit der Bereitstellung von Daten eine größere wissenschaftliche Anerkennung zu Teil werden kann (Lautenschlager und Sens 2003). Zusätzliche Vorteile entstehen, weil von den DOI-Registrierungsagenturen Metadaten-Suchmaschinen entwickelt werden, die das schnelle Auffinden von Forschungsdaten ermöglichen. Die Vergabe der DOIs für PsychData-Datensätze erfolgt über da|ra[5], einem DOI-Registrierungsservice des Leibniz-Instituts für Sozialwissenschaften GESIS in Kooperation mit DataCite[6].

5 http://www.gesis.org/dara
6 http://www.datacite.org

Datenabruf

Da es sich bei den in PsychData archivierten Forschungsdaten um (wenn auch faktisch anonymisierte) personenbezogene Daten handelt, werden sie ausschließlich für die wissenschaftliche Forschung und Lehre zur Nachnutzung bereitgestellt (sogenannte *„scientific use files"*). Um Forschungsdaten zu beziehen, muss ein Datennehmervertrag mit PsychData geschlossen werden, in welchem sich der Nutzer verpflichtet, keine Deanonymisierungsversuche zu unternehmen, die Daten nicht kommerziell zu nutzen und – im Falle einer Veröffentlichung – den Datensatz zu zitieren. Nach postalischer Zusendung des Datennehmervertrags erhält der Nutzer die Forschungsdaten zusammen mit den variablenbeschreibenden Metadaten (Kodebüchern) auf einer revisionssicheren CD-ROM zugeschickt.

Visibilität

Damit Forschungsdaten nachgenutzt werden können, müssen sie auffindbar sein. Durch die Standardisierung der studienbeschreibenden Metadaten, die zusätzlich psychologische Schlagwörter und Klassifikationen beinhalten, können PsychData-Datensätze mit weit verbreiteten Internet-Suchmaschinen wie Google, aber auch mit spezialisierten Suchmaschinen wie PsychSpider[7] recherchiert werden. PsychSpider ist eine Psychologie-Suchmaschine, die Webseiten indiziert, deren Inhalte sich mit Psychologie und psychologischen Themen auseinandersetzen. Sie gehört ebenfalls zu den vom ZPID entwickelten disziplinspezifischen Informationsprodukten. Zusätzliche Recherchemöglichkeiten werden durch Metadaten-Suchmaschinen realisiert, die von den DOI-Registrierungsagenturen da|ra und DataCite entwickelt werden.

Durch die Vernetzung mit weiteren Informationsprodukten des ZPID wird die Visibilität von Forschungsdaten ebenfalls erhöht. Traditionell werden in der Psychologie wie in den meisten anderen Wissenschaften auch Forschungsergebnisse publiziert und diese Publikationen von der Forschungsgemeinschaft rezipiert. Wissenschaftler sollten daher bereits bei der Literaturrecherche Hinweise zum Vorhandensein von Forschungsdaten erhalten, die den publizierten Ergebnissen zugrunde liegen. In der psychologischen Referenzdatenbank PSYNDEX des ZPID enthält jeder Literaturnachweis eine Information, falls zugehörige Forschungsdaten in PsychData verfügbar sind. Durch eine direkte Verlinkung kann der Forscher dann die studienbeschreibenden Metadateninformationen aufrufen (Abbildung 1). Umgekehrt besteht auch eine Verknüpfung von den PsychData-Metadaten zu PSYNDEX-Nachweisen. Über die in den Metadaten enthaltenen

7 http://www.psychspider.de

Literaturangaben kann sich der Forscher direkt die vorhandenen PSYNDEX-Informationen zu der gesuchten Literatur anzeigen lassen.

Um die Bereitstellung von Forschungsdaten als wichtigen Aspekt wissenschaftlichen Arbeitens hervorzuheben, wurde auch eine Vernetzung mit der Autorendatenbank PsychAuthors realisiert. Die Datenbank PsychAuthors ist eine Art „Who is who“ der deutschsprachigen Psychologie. Jeder Psychologe, der wissenschaftlich publiziert, kann bei PsychAuthors ein Autorenprofil anlegen lassen. Zentrales Element dieses Profils ist dabei die vollständige Publikationsliste; zusätzlich enthalten sind Angaben zum aktuellen Dienstort, beruflichen Werdegang, Forschungs- und Lehrinteressen und weitere Funktionen im Wissenschaftsbetrieb. Hat ein Forscher Datensätze über PsychData bereitgestellt, wird dies in seinem Autorenprofil vermerkt und eine direkte Verlinkung zu den studienbeschreibenden Metadaten der Forschungsdaten eingefügt.

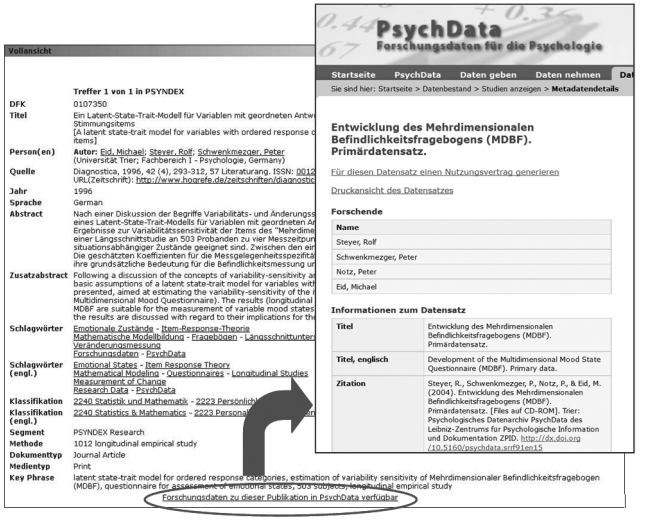


Abbildung 1: Verlinkung von PSYNDEX mit PsychData

Dienstleistungen

PsychData entwickelt zusätzliche Dienstleistungen, um die Open Data Bewegung in der Psychologie zu unterstützen. Zahlreiche Datenzentren, manchmal aber auch einzelne Forscher oder Institute stellen Forschungsdaten bereit, die für psychologische Analysen eine wertvolle Datenbasis darstellen. Deswegen bietet das ZPID eine spezielle Recherchemöglichkeit nach diesen Datensätzen mit der

Psychologie-Suchmaschine PsychSpider an. Eine weitere Dienstleistung besteht in der Bereitstellung des web-basierten Dokumentationstools, das von Wissenschaftlern für eigene Dokumentationszwecke und Data-Sharing-Aktivitäten genutzt werden kann (Dehnhard und Weiland 2011).

Recherchemöglichkeit nach psychologischen Forschungsdaten
Für die bessere Auffindbarkeit von Datensätzen im Bereich der Psychologie wurde in der Suchmaschine PsychSpider eine Kollektion „Forschungsdaten" aufgebaut. Diese Kollektion ermöglicht es, Forschungsdatensätze im Internet zu recherchieren, die psychologische Fragestellungen behandeln oder für psychologische Forschungsfragen interessante Informationen liefern können. Zum kontinuierlichen Ausbau dieser Kollektion werden die Angebote anderer Datenzentren und Datenanbieter hinsichtlich psychologierelevanter Forschungsdatensätze durchsucht und anschließend die zugehörigen Metadateninformationen indiziert.

Unterstützung des Datenmanagements
Wesentliche Voraussetzung für die Nachnutzbarkeit von Forschungsdaten ist eine vollständige Datendokumentation, da nur durch sie die langfristige Interpretierbarkeit der Daten gewährleistet bleibt. In der Forschungspraxis erfolgt die Dokumentation leider häufig nachlässig. Als Gründe können die fehlende Verbreitung und Kenntnis von Dokumentationsstandards sowie der zusätzlich erforderliche Zeit- und Arbeitsaufwand genannt werden (Wicherts 2006; Postle 2002). Außerdem unterstützen gängige Statistik-Software-Programme den Forscher meist wenig bei der Dokumentationserstellung (Freedland und Carney 1992). Um Forscher bei der Datendokumentation so früh wie möglich im Forschungsprozess zu unterstützen, stellt PsychData ein web-basiertes Dokumentationstool bereit, das für eigene Dokumentations- oder Data-Sharing-Aktivitäten genutzt werden kann. Eine anschließende Datenübergabe an das PsychData-Archiv ist möglich, jedoch nicht obligatorische Bedingung für die Nutzung des Dokumentationstools.

Wissenschaftler erhalten mit dem Dokumentationstool das erforderliche Handwerkszeug, um ihre Forschungsdaten nach bewährten Standards zu dokumentieren und durch automatische Validitäts- und Konsistenzkontrollen die Qualität der Daten sicherzustellen. Ein wesentlicher Vorteil durch die Verwendung des Tools besteht in der Ablage von Forschungsdaten und kompletter Dokumentation an nur einem Speicherort. Dies ist in der sonst gängigen Forschungspraxis eher nicht der Fall, was langfristig zu einem Verlust der Dokumentation und damit zusammenhängend der Interpretierbarkeit der Daten führen kann (Freedland und Carney 1992). Für die Benutzung des Tools ist lediglich eine Registrierung erforderlich. Durch ein Rechte-Management-System wird Data-Sharing ermöglicht, indem jeder Nutzer anderen Nutzern Zugriff auf die eigenen Forschungsdaten gewähren kann. Die Art des Zugriffs ist abstufbar: Von Lese-

rechten über Schreibrechte bis hin zum „grant"-Recht, d.h. dem Recht, ebenfalls Rechte vergeben zu können. Natürlich kann auch eine Übergabe der Daten an das PsychData-Archiv erfolgen, wobei durch die Benutzung des Dokumentationstools der Workflow der Dateneingabe wesentlich vereinfacht ist, da die Daten bereits in einer standardisierten und kontrollierten Form vorliegen.

Die Einsatzbereiche des Tools liegen neben eigenen Dokumentationszwecken in der Unterstützung von vernetztem Arbeiten in Forschungsgruppen oder -instituten. Außerdem kann es ein effizientes Werkzeug für die Aus- und Weiterbildung in der Psychologie darstellen: Bei der Betreuung von Diplomarbeiten und Dissertationen oder auch in der Lehre können durch seine Verwendung Prinzipien des „Best Practice" der Datendokumentation vermittelt werden. Durch den Einblick in die Forschungsdaten selbst sind zudem verbesserte Betreuungsmöglichkeiten gegeben. Allerdings sind die Funktionalitäten des Dokumentationstools noch eingeschränkt, sollen aber in weiteren Entwicklungsschritten verbessert werden. So ist es bisher gemäß den Archivierungsformaten in PsychData nur möglich, ASCII-Dateien hochzuladen oder zu exportieren. Direkte Upload- und Export-Funktionen von Dateiformaten der gängigen Statistikprogramme SPSS oder STATA sind noch nicht gegeben. Auch die automatische Erstellung von Kodebüchern aus gut dokumentierten Datendateien stellt eine Entwicklungsherausforderung dar.

Ausblick

Zwar hat PsychData für ein Archiv noch eine relativ kurze Bestandsdauer zu verzeichnen, trotzdem kann konstatiert werden, dass bisher gute Erfahrungen mit der vorhandenen Archivierungspraxis gemacht wurden. In einigen Bereichen – wie zum Beispiel der Versionierung von Datensätzen – fehlen noch Erfahrungswerte, entsprechende Vorgehensweisen und Routinen befinden sich in Entwicklung. Aufgrund der sehr begrenzten Ressourcen von PsychData, aber auch als Dienstleistungsangebot für die Wissenschaft, spielt die Automatisierung von Arbeitsprozessen und damit verbunden die Vereinfachung der internen und externen Workflows eine wichtige Rolle. Wesentlicher Gesichtspunkt bleibt die Aufrechterhaltung eines hohen Qualitätsniveaus von Forschungsdaten und Forschungsmetadaten. Dies wird auch zukünftig einen wichtigen Aspekt innerhalb der technischen und strukturellen Fortentwicklung ausmachen.

Eine grundlegende Änderung der nationalen Forschungskultur in der Psychologie in Richtung einer stärkeren Würdigung von Datenbereitstellung und -nachnutzung wird noch einige Zeit brauchen. Allerdings lässt sich im Vergleich zur Gründungszeit des Archivs vor allem bezüglich der Datennachnutzung schon ein positiver Trend feststellen. Angestrebtes Ziel für die psychologische Forschungspraxis ist das „data lifecycle management" durch PsychData.

Literatur

Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen (2003). http://oa.mpg.de/files/2010/04/Berliner_Erklaerung_dt_Version_07-2006.pdf [31.05.2012]

Botella Ausina, J. y Ortego Maté, C. (2010): Compartir datos: Hacia una investigación más sostenible. Psicotema 22, 263-269.

Breckler, S.J. (2009): Dealing with data. Monitor on Psychology 40 (2), 41.

de la Sablonnière, R./Auger, E./Sabourin, M. and Newton, G. (2012): Facilitating Data Sharing in the Behavioural Sciences. Data Science Journal 11, DS29-DS43.

Dehnhard, I. und Weiland, P. (2011): Toolbasierte Datendokumentation in der Psychologie. In: Griesbaum, J./Mandl, T. und Womser-Hacker, C. (Hrsg.): Information und Wissen: global, sozial und frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft. Boizenburg: Hülsbusch, 74-84.

DFG (Deutsche Forschungsgemeinschaft) (2009): Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten. Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme/ Unterausschuss für Informationsmanagement. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf [24.05.2012]

Fahrenberg, J. (2009): Open Access – nur Texte oder auch Primärdaten? RatSWD Working Paper Series Nr. 200. Berlin: Rat für Sozial- und Wirtschaftsdaten. http://www.ratswd.de/download/RatSWD_WP_2012/RatSWD_WP_200.pdf [21.11.2012]

Freedland, K.E. and Carney, R.M. (1992): Data Management and Accountability in Behavioral and Biomedical Research. American Psychologist 47 (5), 640-645.

ICSU (International Council for Science) (2002): Principles for Dissemination of Scientific Data. http://www.codata.org/codata/data_access/principles.html [31.05.2012]

Krampen, G. und Weichselgartner, E. (2005): Dokumentation und Archivierung von Rohdatensätzen aus der psychologischen Forschung. DFG Sachbeihilfe GZ: 554 922 (1) Uni Trier BIB44 TRuv 01-02.

Lautenschlager, M. und Sens, I. (2003): Konzept zur Zitierfähigkeit wissenschaftlicher Primärdaten. Information Wissenschaft & Praxis 54, 463-466.

Montada, L. und Weichselgartner, E. (2002): Dokumentation und Archivierung von Rohdatensätzen aus der psychologischen Forschung. DFG Sachbeihilfe GZ: 554 922 (1) Uni Trier BIB44 TRuv 01-01.

Murray-Rust, P. (2007): Data Driven Science – A Scientist's View. In: NSF/JISC 2007 Digital Repositories Workshop. http://www.sis.pitt.edu/~repwkshop/papers/murray.html [24.05.2012]

NIH (US National Institutes of Health) (2003): NIH Data Sharing Policy. February 2003. http://grants.nih.gov/grants/policy/data_sharing/ [24.05.2012]

OECD (Organisation for Economic Co-operation and development) (2004): Declaration on Access to Research Data from Public Funding. http://www.oecd.org/document/0,3746,en_21571361_44315115_25998799_1_1_1_1,00.html [31.05.2012]

Postle, B.R./Shapiro, L.A. and Biesanz, J.C. (2002): On Having One´s Data Shared. Journal of Cognitive Neuroscience 14, 838-840.

Ruusalepp, R. (2008): Infrastructure, Planning and Data Curation. A comparative study of international approaches to enabling the sharing of research data. Version 1.6. 30. November 2008. http://www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf [31.05.2012]

Silbereisen, R.K. (2003): Zur Lage der Psychologie – neue Herausforderungen für Internationalität und Interdisziplinarität. Psychologische Rundschau 54, 2-11.

Silbereisen, R.K. und Eyferth, K. (2004): Berliner Jugendlängsschnitt „Jugendentwicklung und Drogen". Primärdaten der ersten Erhebungswelle (Jugendlichenstichprobe) 1982. (Version 1) [Files auf CD-ROM]. Trier: Psychologisches Datenarchiv PsychData des Leibniz-Zentrums für Psychologische Information und Dokumentation ZPID. doi: 10.5160/psychdata.rems82be29

Weichselgartner, E. (2008): Fünf Jahre Primärdatenarchivierung in der Psychologie: Ein Erfahrungsbericht. In: Ockenfeld, M. (Hrsg.): Verfügbarkeit von Information. Frankfurt a. M.: DGI, 259-267.

Weichselgartner, E./Günther, A. und Dehnhard, I. (2011): Archivierung von Forschungsdaten. In: Büttner, S./Hobohm, H.-C. und Müller, L. (Hrsg.): Handbuch Forschungsdatenmanagement. Bad Honnef: Bock + Herchen Verlag, 191-202.

Wellcome Trust (2010): Policy on data management and sharing. http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm [24.05.2012]

Wicherts, J.M./Borsboom, D./Kats, J. and Molenaar, D. (2006): The Poor Availability of Psychological Research Data for Reanalysis. American Psychologist 61, 726-728.
Wolins, L. (1962): Responsibility for raw data. American Psychologist 17, 657-658.
ZPID (Zentrum für Psychologische Information & Dokumentation) (2011): PSYNDEX Terms:Deskriptoren/Subject Terms zur Datenbank PSYNDEX (Lit & AV, Tests). Trier: ZPID.

Verzeichnis der Autorinnen und Autoren

Reinhard Altenhöner
Deutsche Nationalbibliothek
Leiter der Abteilung Informationstechnik

Ina Dehnhard
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)
Arbeitsbereich psychologisches Datenarchiv PsychData

Prof. Dr. Armin Günther
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)
Arbeitsbereich PsychOpen Europäische Open Access; psychologisches Datenarchiv PsychData

Brigitte Hausstein
GESIS - Leibniz-Institut für Sozialwissenschaften
Leiterin der da|ra - Registrierungsagentur für Sozial- und Wirtschaftsdaten

Stefan Hein
Deutsche Nationalbibliothek
Informatiker, Softwareentwickler

Nicole von der Hude
Deutsche Nationalbibliothek
Bereich Digitale Dienste

Denis Huschka
Geschäftsstelle des Rates für Sozial- und Wirtschaftsdaten (RatSWD)
Geschäftsführer

Tibor Kálmán
Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG)
Arbeitsgruppe eScience

Dr. Christian Keitel
Landesarchiv Baden-Württemberg
Abteilung Fachprogramme und Bildungsarbeit, stellvertretender Abteilungsleiter und Referatsleiter Überlieferungsbildung

Dr. Jens Klump
Deutsches GeoForschungsZentrum (GFZ) am Helmholtz-Zentrum Potsdam

Daniel Kurzawe
Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG)
Arbeitsgruppe eScience

Dr. Hans Luthardt
Deutsches Klimarechenzentrum/Datenmanagement (DKRZ)
Abteilung Datenmanagement

Reiner Mauer
GESIS - Leibniz-Institut für Sozialwissenschaften
Datenarchiv für Sozialwissenschaften, Teamleiter Akquisition, Sicherung Datenbereitstellung

Claudia Oellers
Geschäftsstelle des Rates für Sozial-und Wirtschaftsdaten (RatSWD)
wissenschaftliche Mitarbeiterin

Prof. Dr. Notburga Ott
Ruhr-Universität Bochum
Professorin für Sozialpolitik und Institutionenökonomik
Rat für Sozial- und Wirtschaftsdaten (RatSWD), Stellvertretende Vorsitzende

Sabine Schrimpf
Deutsche Nationalbibliothek
wissenschaftliche Mitarbeiterin Informationstechnik

Natascha Schumann
GESIS – Leibniz-Institut für Sozialwissenschaften
Bereich Langzeitarchivierung, wissenschaftliche Mitarbeiterin
nestor Geschäftsstelle an der Deutschen Nationalbibliothek (von 2008 bis 2012)

Dr. Ulrich Schwardmann
Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG)
stellvertretender Leiter der Arbeitsgruppe eScience

Olaf Siegert
ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften | Leibniz-Informationszentrum Wirtschaft
Leiter der Abteilung Elektronisches Publizieren

Ralf Toepfer
ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften | Leibniz-Informationszentrum Wirtschaft
Stellvertretender Leiter der Abteilung Elektronisches Publizieren

Sven Vlaeminck
ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften | Leibniz-Informationszentrum Wirtschaft
Projektmanager European Data Watch Extended (EDaWaX)

Prof. Dr. Gert G. Wagner
Deutsches Institut für Wirtschaftsforschung (DIW)
Vorstandsvorsitzender
Rat für Sozial- und Wirtschaftsdaten (RatSWD), Vorsitzender

PD Dr. Erich Weichselgartner
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)
stellvertretender wissenschaftlicher Leiter und Leiter der Bereiche IT und Entwicklung

Wolfgang Zenk-Möltgen
GESIS - Leibniz Institut für Sozialwissenschaften
Abteilung: Datenarchiv für Sozialwissenschaften, Teamleiter Archivinstrumente und Prozesse

Die Langzeitarchivierung von Forschungsdaten ist eine Voraussetzung für gute wissenschaftliche Praxis. Sie weist drei zentrale Bereiche auf: die Dokumentation der Forschungsdaten, deren langfristige Aufbewahrung sowie die Bereitstellung eines Zugangs zu den Daten. Ohne diese infrastrukturellen wie organisatorischen Voraussetzungen sind die Daten für die wissenschaftliche Sekundärnutzung, also für die Überprüfung von Ergebnissen und auch für die Beantwortung neuer Forschungsfragen nur eingeschränkt verfügbar.

Das vorliegende Buch gibt einen Überblick über bestehende Standards und liefert einen Beitrag zur Diskussion über Voraussetzungen zur Archivierung von Datenbeständen. Es ist somit gleichermaßen für Infrastruktureinrichtungen, Fachbibliotheken, Archive, Wissenschaftler und alle, die im weitesten Sinne mit der Verfügbarmachung von Forschungsdaten betraut sind, lesenswert.