Handbuch Forschungsdatenmanagement

Herausgegeben von
Stephan Büttner, Hans-Christoph Hobohm, Lars Müller

BOCK + HERCHEN Verlag
Bad Honnef
2011

ISBN 978-3-88347-283-6

BOCK+HERCHEN Verlag, Bad Honnef
Printed in Germany

Inhalt

Vorwort

In den allermeisten wissenschaftlichen Vorhaben entstehen digitale Daten als Grundlage für neue Erkenntnisse. Ein großer Teil dieser Daten – manche Schätzungen reichen bis zu 90 % – gehen jedoch in einem relativ kurzen Zeitraum verloren. Sie stehen somit nicht mehr einer weiteren Verwendung und Nachnutzung zur Verfügung. Nicht zuletzt aus dieser Erkenntnis heraus haben seit einigen Jahren das Interesse an und die Bemühungen um ein professionelles Management von Forschungsdaten stark an Bedeutung gewonnen. Im Vordergrund stehen hier nicht das Potenzial und die Leistungsfähigkeit von Soft- und Hardware, sondern vielmehr der Umgang mit digitaler wissenschaftlicher Information im weiteren Sinne. Ziel ist eine sichere Speicherung, eine nachhaltige und langfristige Archivierung und der relativ neue Aspekt einer überregionalen Bereitstellung der Daten. Das Potenzial für eine spätere Nachnutzung dieser Informationen soll erhalten und durch ein professionelles Informationsmanagement ausgebaut werden. Der breite Zugang zu Forschungsdaten, so wird erwartet, erlaubt eine deutliche Verbesserung und neue Perspektiven für das wissenschaftliche Arbeiten. Es wird die Aussicht eröffnet, auf der Grundlage bereits vorhandener Ergebnisse einfacher und schneller zu neuen Erkenntnissen zu gelangen. Diese Nachnutzbarkeit schließt vor allem auch die Möglichkeiten ein, Ergebnisse leicht interdisziplinär in Beziehung setzen zu können.

Zumeist aufgrund ihres besonders kooperativen Charakters nutzen einzelne wissenschaftliche Disziplinen bereits verschiedene Formen des Datenmanagements, in manchen Fällen sogar bereits seit Jahrzehnten. Der Aufbau dieser Strukturen wurde durch die Wissenschaftlerinnen und Wissenschaftler aus den Anforderungen ihres wissenschaftlichen Arbeitens heraus angeregt und in Kombination mit Methoden des Informationsmanagements umgesetzt. Ein systematisches, strukturiertes und vor allem auch nachhaltiges Datenmanagement mit Angeboten für die zahlreichen und diversen Typen wissenschaftlicher Daten in den unterschiedlichen Disziplinen fehlt jedoch. Dieses muss verknüpft sein mit der Implementierung von Methoden zur eindeutigen Identifizierung der Datensätze und zu deren Zitierbarkeit, mit spezifischen Dienstleistungen für die forschende Wissenschaft und der Umsetzung eines freien und fairen Zugangs.

Sowohl auf nationaler als auch auf internationaler Ebene haben verschiedene wissenschaftliche und politische Organisationen auf diesen dringenden Bedarf hingewiesen (OECD, 2007; Allianz, 2008). So wird unter anderem angeregt, Forschungsdaten sollte der Status nationalen Kulturgutes zuerkannt werden. Als wertvolle Wissensressource sollten diese mithin geschützt und erhalten werden, ihre nachhaltige Sicherung und Bereitstellung wird als eine strategisch bedeutsame Aufgabe angesehen. Investitionen in Forschungsinfrastruktur sollten daher zwingend mit der Förderung und Entwicklung einer adäquaten Informationsin-

frastruktur einhergehen, welche die wissenschaftlichen Ergebnisse einer angemessenen Nutzung und Nachnutzung langfristig zur Verfügung stellt.

Mit dem vorliegenden *Handbuch Forschungsdatenmanagement* wird neben diesen grundlegenden Anforderungen und notwendigen Maßnahmen einer der wesentlichen und bisher wenig adressierten Aspekte für die Entwicklung eines erfolgreichen zukünftigen Forschungsdatenmanagements aufgegriffen: die Aus- und Weiterbildung. Nur ein grundsätzliches Verständnis für die Potenziale modernen Informationsmanagements und seine möglichst frühzeitige Umsetzung in wissenschaftlichen Vorhaben werden zu dessen nachhaltigen Etablierung führen. Die langfristige Sicherung und die Bereitstellung zur Nachnutzung, auch im Sinne der „Guten wissenschaftlichen Praxis" (DFG, 1998), müssen zum selbstverständlichen Bestandteil wissenschaftlichen Arbeitens werden und sollte im wissenschaftlichen Fachstudium verankert werden.

Dr. Stefan Winkler-Nees
Deutsche Forschungsgemeinschaft (DFG)
–Wissenschaftliche Literaturversorgungs- und Informationssysteme–
D-53170 Bonn

Literatur

Allianz der deutschen Wissenschaftsorganisationen, 2008. *Schwerpunktinitiative „Digitale Information"*. Online: http://www.allianzinitiative.de/ [Zugriff am 20.07.2011].

DFG (Deutsche Forschungsgemeinschaft), 1998. *DFG Denkschrift: Sicherung Guter Wissenschaftlicher Praxis*. Weinheim: WILEY-VCH. Online: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf [Zugriff am 20.07.2011].

OECD (Organisation for Economic Co-Operation and Development), 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD Publications. Online: http://www.oecd.org/dataoecd/9/61/38500813.pdf [Zugriff am 20.07.2011].

1.0 Einführung

Stephan Büttner, Hans-Christoph Hobohm, Lars Müller
Fachhochschule Potsdam, Fachbereich Informationswissenschaften

Überblick

Forschungsdatenmanagement ist bislang stark disziplinspezifisch geprägt und organisiert. Zu verschieden sind die Anforderungen an Technik, Metadaten und Archivierung, als dass sich einheitliche Modelle und Verfahrensweisen auf allen Wissenschaftsgebieten anwenden ließen. Selten werden die Disziplingrenzen überschritten bei der Betrachtung dieses wichtigen Infrastrukturthemas. Jedes Wissenschaftsgebiet, ja jede Institution geht hier eigene Wege. Nationale und internationale Entwicklungen in Forschungsförderung und Wissenschaftspolitik machen jedoch zunehmend konzertiertes Handeln notwendig. Anlass für den Ruf nach einem geregelten Forschungsdatenmanagement ist aber neben den Fragen der Überprüfbarkeit von Forschungsergebnissen vorwiegend die Hoffnung auf eine erhöhte Verwertung der mit großem Aufwand gewonnenen Daten. Die technischen Möglichkeiten sind vorhanden, nicht nur nie gekannte Mengen an Daten zu produzieren, sondern sie auch in neuer Weise (nach) zu nutzen, neu zu kombinieren und neuen Hypothesen zu unterziehen. So ist z. B. die nicht verklingende Forderung nach interdisziplinärer Forschung mit dem zunehmenden Bedarf verbunden, auf Daten anderer Fächer zuzugreifen, sie miteinander in Beziehung zu setzen und integriert zu analysieren.

Die Wissenschaftsinfrastruktur ist zu einer übergreifenden nationalen, ja oft sogar globalen Aufgabe geworden. So legt die DFG Wert darauf, dass bereits bei der Beantragung von Forschungsprojekten Angaben zum geplanten Datenmanagement und der Nachnutzung von Forschungsdaten gemacht werden. Für Wissenschaftliche Arbeitsgruppen, kleine Institute oder Hochschulen stellt der Aufbau eigener Forschungsdateninfrastrukturen jedoch eine große Herausforderung dar. Zudem sollten Forscherinnen und Forscher möglichst wenig mit administrativen Aufgaben belastet werden. Sie sind deshalb zunehmend angewiesen auf Servicestrukturen für Forschungsdatenmanagement, die langfristige Sicherung und Zugänglichkeit gewährleisten können.

Die angesprochenen Entwicklungen erfordern die Herausbildung von Spezialistinnen und Spezialisten, die Serviceinfrastrukturen für Forschungsdatenmanagement wie auch interdisziplinäre Schnittstellen für alle Phasen des Forschungsprozesses entwickeln und aufrechterhalten können.

Das *Handbuch Forschungsdatenmanagement* ist konzipiert als Leitfaden für das Selbststudium sowie zur Unterstützung der Aus- und Weiterbildung auf dem aktuellen Stand der Diskussion. Sie richtet sich insbesondere an Einsteiger im Forschungsdatenmanagement, aber gleichermaßen auch an wissenschaftliche

Datenkuratoren, IT-Administratoren und Informationswissenschaftler, die ihre Aufgaben im Forschungsdatenmanagement nicht mehr nur einzelfall- oder disziplinorientiert, sondern in Hinblick auf die Arbeit in und an Forschungsdateninfrastrukturen wahrnehmen wollen. Und so war die Aufgabe für die Autorinnen und Autoren in ihrem Kapitel nicht nur den State-of-the-Art darzustellen, sondern das Thema so aufzubereiten, dass z. B. über die Referenzen das weitere Einarbeiten in die Themenfelder erleichtert wird.

Zentrale Aspekte des Forschungsdatenmanagements werden in der vorliegenden Publikation aus informationswissenschaftlicher und anwendungsbezogener Perspektive disziplinübergreifend eingeführt. Wir freuen uns, dass wir zahlreiche ausgewiesene und erfahrene Expertinnen und Experten gewinnen konnten, mit Beiträgen zu ihren Spezialgebieten mitzuwirken. Es ist leider nicht gelungen, für alle Themen, die uns wichtig waren, Autorinnen oder Autoren zu finden. So stellt das Fehlen eines Kapitels zu Qualitätssicherung für Forschungsdaten eine große Leerstelle dar. Auch die Themen Lizenzmodelle für Forschungsdaten und wirtschaftliche Geschäftsmodelle, die beim Austausch von Forschungsdaten zum Tragen kommen, sind von großer Wichtigkeit und fehlen in diesem Band. Wir betrachten es als Herausforderung und Aufgabe, diese Lükken zu schließen. Insofern ist dieses Buch als „Book in progress“ zu verstehen, das in gewissen Abständen aktualisiert wird und werden muss.

Die Publikation gliedert sich in drei Teile. Begonnen wird mit einer allgemeinen Einführung in das Themengebiet. Grundlegende Begriffe werden erläutert, die Entwicklung und Bedeutung des Themas in den Wissenschaften werden dargestellt. Der zweite Teil widmet sich der Praxis des Forschungsdatenmanagements. Hier werden einzelne Problemfelder aufgefächert und praktische Lösungsansätze dargestellt. Im Zentrum stehen praktisch-technische Fragen und Empfehlungen. Im dritten Teil liegt der Schwerpunkt auf strategisch-konzeptionellen Entwicklungen des Forschungsdatenmanagements, die langfristig auf die Entwicklung von Dateninfrastrukturen zielen.

Im ersten Kapitel (1.1) geben die Herausgeber einen Überblick über das Gesamtgebiet im Zusammenhang mit den oft unter dem Schlagwort eScience subsumierten neuen Forschungsstrukturen und gehen zunächst der Frage nach, was unter Forschungsdaten zu verstehen ist. Terminologische Unschärfen und disziplinspezifisch unterschiedliche Verständnisse lassen sich nie ganz vermeiden, die Thematisierung der Unterschiede ist jedoch für die allgemeine Verständigung notwendig. Überlegungen zu den wissenschaftspraktischen Konsequenzen der Entwicklung schließen das Kapitel ab unter Beantwortung der Frage, warum das Handbuch von Informationswissenschaftlern herausgegeben wird.

Im folgenden Kapitel (1.2) führt Stefanie Rümpel in die beiden grundlegenden Modelle des Forschungsdatenmanagements ein: das *Curation Life Cycle* Modell des britischen *Digital Curation Centres* (DCC) das die Prozessschritte des allgemeinen Informationslebenszyklusmodells der Wirtschaft auf digitale

Objekte und speziell auf Daten überträgt. Das Modell des DCC ist weit verbreitet und akzeptiert. Auch das australische *Data Curation Continuum* beschreibt sehr anschaulich die möglichen Erscheinungsweisen von Forschungsdaten und wird deshalb gerne als Basis für weiterführende Erläuterungen genutzt.

Das Autorenteam des Rats für Sozial- und Wirtschaftsdaten (RatSWD), Denis Huschka, Claudia Oellers, Notburga Ott und Gert G. Wagner verdeutlicht basierend auf den Jahrzehnte langen Erfahrungen der Sozialwissenschaften das gesamte Spektrum des Aufbaus von Dateninfrastrukturen (1.3). Das zunehmende Bedürfnis nach „*data sharing*“ übt einen Druck auf alle Beteiligten aus, entsprechende Vereinbarungen und technische Grundlagen zu schaffen, die einen übergreifenden Zugang zu erhobenen Daten gewährleisten (*data access*). Dabei ist zu berücksichtigen, dass „Daten“ disziplinspezifisch völlig unterschiedliche Ausprägungen haben können. Daraus ergeben sich auch sehr unterschiedliche, mehr oder weniger zentralisierte Modelle zur Lösung des Infrastrukturproblems.

Zu Beginn des zweiten, den praktischen Einzelaspekten des Forschungsdatenmangements gewidmeten Teils, beschreiben Heinz Pampel und Roland Bertelmann das Instrument der *Data Policies*, die je nach wissenschaftlichem Kontext empfehlenden oder verpflichtenden Charakter haben können (2.1). Es wird ein Überblick über die Vielfalt der Policies gegeben und die Herausforderungen bei der Umsetzung von Infrastrukturregeln beschrieben, die erfahrungsgemäß gerade im Bereich der Informations- und Wissensverarbeitung besonders schwierig durchzusetzen sind.

Tobias Hillegeist und Gerald Spindler vertiefen in Kapitel 2.2 die schon angeschnittenen rechtlichen Aspekte des Forschungsdatenmanagements. Ein besonderer Schwerpunkt des juristischen Augenmerks auf Forschungsdaten liegt dabei naturgemäß beim Datenschutz, aber auch das Urheberrecht mit der Frage, sind Daten urheberrechtlich geschützt, spielt eine wichtige Rolle. Als besonders beachtenswert erweist sich die Problematik des Schutzes von Digitalen Sammlungen in Datenbanken.

Uwe Jensen beschreibt in Kapitel 2.3 wie Datenmanagementpläne aussehen können und was dabei zu beachten ist. Viele Forschungsförderer (wie z. B. die DFG) verlangen mittlerweile explizite Aussagen zu einem nachhaltigen Datenmanagement. Besondere Probleme sind hierbei die Fragen nach Vertraulichkeit, Qualitätssicherung und Dateiformat. Der langfristige Zugang zu Daten ist trotz langjähriger Erfahrungen der digitalen Langzeitarchivierung auch im Datenbereich noch nicht gelöst.

Eine wichtige Voraussetzung für *data sharing* wie auch für nachhaltige Archivierung ist die Entwicklung und Vereinbarung allgemein gültiger Metadatenstandards zur Beschreibung und Identifikation von Daten. Uwe Jensen, Alexia Katsanidou und Wolfgang Zenk-Möltgen geben in Kapitel 2.4 dazu einen pro-

funden Überblick und erläutern den Stand der Metadatenentwicklung im Hinblick auf Forschungsdaten.

Die grundlegenden strategischen Konzepte von Kapitel 1.2 aufgreifend erläutern Andreas Aschenbrenner und Heike Neuroth in Kapitel 2.5 die verschiedenen Formen und Ausprägungen von Repositorien, in denen nicht nur wie bisher Textdokumente, sondern auch Daten gesammelt und zur Nutzung zur Verfügung gestellt werden. Ein Schwerpunkt dieses Kapitels ist auch die Darstellung der unterschiedlichen technologischen Konzepte von Repositoriensystemen.

Jens Klump geht anschließend (2.6) zusammenfassend auf Konzepte und Methoden der Langzeitarchivierung von Daten ein. Hier gibt es zwar schon eine Reihe von Werkzeugen, Probleme bereitet aber immer noch ihre Implementierung in der Wissenschaftspraxis. An diesem Beispiel wird deutlich, dass ein wirklich tiefgreifender kultureller Wandel erforderlich ist, um die technischen Potentiale der Datenmengen zu nutzen.

Matthias Razum schildert in Kapitel 2.7 Systeme und Systemarchitekturen für das Datenmanagement. Viele in den vorhergehenden Kapiteln angeschnittenen Teilaspekte des Themenfeldes kann man hier in ihrer technischen Umsetzung beschrieben sehen: funktionale Anforderungen, die Berücksichtigung des Lebenszyklus, Fragen des Zugangs und der Vertraulichkeit über Authentifizierung und ein Ausblick auf die Grid-Technologie als moderner technischer Basis.

Zum Abschluss des Überblicks über die konkreten Teilaspekte des Forschungsdatenmanagements stellen Bettina Berendt, Joaquin Vanschoren und Bo Gao die unter den Bedingungen der eScience sich ergebenden neuen Möglichkeiten der Datenvisualisierung und Datenanalyse dar (2.8). Visualisierung als Dimensionalitätsreduktion dient nunmehr nicht nur der reinen Dokumentation und Ergebnisdarstellung, sondern wird zum Instrument der kollaborativen Datenexploration, das besondere Anforderungen an das Design der Datenstrukturen stellt.

Der dritte Teil versucht einen Ausblick auf zukünftige Erfordernisse im Zusammenhang mit dem Forschungsdatenmanagement. Michael Lautenschläger greift die schon in den Anfängen befindliche Entwicklung zu regelrechten *Data Curation Services* auf. Ziel ist es, die archivierten Daten auch nach vielen Jahren für wissenschaftliche, interdisziplinäre Nachnutzung bereitzustellen. Fachspezifisch oder zentral wird sich zunehmend die Notwendigkeit ergeben, Institutionen übergreifende Dienste anzubieten, die die Wissenschaftler von den technischen und infrastrukturellen Aufgaben der Datenhaltung und -pflege entlastet (3.1).

In vielen Disziplinen gibt es bisher keine etablierte Kultur der Veröffentlichung wissenschaftlicher Daten. Sünje Dallmeier-Tiessen stellt an Hand konkreter Beispiele die Entwicklung von Publikationsmodellen für Datensätze dar (3.2). Die immer häufiger zu hörende Forderung, auch die Erhebung und Publi-

kation von Daten soll zur wissenschaftlichen Reputation einzelner Wissenschaftler beitragen können, wird bereits von verschiedenen Initiativen realisiert.

Ein ebenfalls aktuell sehr dynamisches Entwicklungsgebiet beschreiben Günther Neher und Bernd Ritschel im Kapitel 3.3 über die semantische Vernetzung von Forschungsdaten mittels Ontologien. Besonders die *Linked Open Data* Bewegung hat hierbei schon beachtliche Erfolge der Vernetzung und Weiterverwertung von Daten erzielt, und es wird ersichtlich, welche Potenziale systematische Anwendungen von Ontologiesprachen auch im Zusammenhang mit Forschungsdaten bergen.

Erich Weichselgartner, Armin Günther und Ina Dehnhard behandeln den immer noch brisanten Aspekt der nachhaltigen Archivierung von Forschungsdaten (3.4). Trotz aller Erfolge im allgemeinen Verständnis der Notwendigkeit digitaler Langzeitarchivierung, zeigt das Kapitel, wo die Probleme liegen, welche Selektionskriterien angemessen sind, welche Herausforderungen zu bewältigen sind und welche Trends sich abzeichnen.

Zum Abschluss beschreiben Stephan Büttner, Stefanie Rümpel und Hans-Christoph Hobohm in Kapitel 3.5 die Rollen von Informationswissenschaftlern in diesem neuen Handlungsfeld. Kompetenzen und Tätigkeitsfelder werden diskutiert. Es wird aufgezeigt, dass es sich um die Ausdifferenzierung neuer Berufsbilder und gleichzeitig um die teilweise Neuausrichtung alter Berufe handelt. Neben dem *data curator* wird auch ein *data librarian* seinen Platz finden. Die Entwicklung für neue Aus- und insbesondere Weiterbildungsangebote sowie die Anpassung der Curricula existierender Studienangebote ist in vollem Gange.

Danksagung

Unser Dank gilt in erster Linie den Autorinnen und Autoren. Erst durch ihre Beiträge konnte die Idee, ein Handbuch zum Thema Forschungsdatenmanagement herauszugeben, in die Realität umgesetzt werden. Trotz einiger terminlicher und sonstiger Widrigkeiten brachten alle Verständnis für die teilweise doch kurzfristigen Nachfragen während der Bearbeitung auf. Wir freuen uns sehr über das Einverständnis aller Beteiligten, dass das Handbuch neben der Printversion des Verlags zugleich auch unter einer Creative-Commons-Lizenz online erscheinen kann. Hier gilt unser Dank dem Verlag für die Nachsicht bei den produktionstechnischen Deadlines und insbesondere Herrn Andreas Bock und Herrn Bernhard Wambach, die unsere Sonderwünsche geduldig ertrugen.

Bei der redaktionellen Bearbeitung haben uns die studentischen Mitarbeiterinnen Carolin Jäckel und Julia Lindner unschätzbare Hilfe geleistet. Für weitere infrastrukturelle Unterstützung und manche Motivation danken wir Christoph Höwekamp und Judith Pfeffing. Das Grundkonzept des Bandes hätte ohne die Diskussionen mit Claudia Oellers, Heinz Pampel, Thomas Wetzel und vielen

anderen sicher anders ausgesehen. Für die Einstellung der Beiträge auf dem edoc-Server der Fachhochschule Potsdam als Open-Access Dokumente danken wir Frau Ingrun Griesa von der Hochschulbibliothek. Und last but not least ist es der Verdienst von Stefan Winkler-Nees, uns bei der Annäherung an das Thema immer wieder ermuntert zu haben, auch als kleinerer Player im Feld der *Big Sciences* mitzumischen.

Die Herausgeber und Autoren hoffen, dass dieses Handbuch eine Hilfe sowohl für den Einstieg in die Hintergründe und zugrundeliegenden Theorien und Ansätze bietet und vielleicht die eine oder andere Anregung bei der Umsetzung konkreter Praxisprojekte beim Umgang mit der ständig steigenden Zahl von Forschungsdaten liefert.

1.1 Research Data Management

Stephan Büttner, Hans-Christoph Hobohm, Lars Müller
Fachhochschule Potsdam, Fachbereich Informationswissenschaften

1.1.1 Forschungsdatenmanagement: Infrastruktur für eScience

Forschungsdaten haben durch die Entwicklung von Computern und Speichermedien in einem so großen Umfang zugenommen, dass schon vor Jahren von der „Datenflut“ (Hey & Trefethen, 2003) gesprochen wurde. Automatische Messungen erzeugen neue Datensätze im Sekundentakt, elektronische Speicherung und Verknüpfung von Forschungsdaten ermöglichen, immer größere Datenmengen mit statistischen Verfahren und Visualisierungswerkzeugen auszuwerten. Bislang analoge Medien wie Texte und Fotografien werden in digitale Daten umgewandelt und automatisch verarbeitet. Es entsteht damit ein wachsender Fundus an Daten, der mit herkömmlichen wissenschaftlichen Verfahren nicht mehr erschöpfend genutzt werden kann. Die globale und interdisziplinäre Verknüpfung und Auswertung dieser Datenmengen eröffnet völlig neue Dimensionen wissenschaftlicher Erkenntnis. Damit das in den Daten implizit enthaltene Wissen zu Tage gefördert werden kann, müssen neue Verfahren und Instrumente entwickelt werden. (vgl. President’s Information Technology Advisory Committee, 2005, S. 56f.).

Gut veranschaulichen lässt sich das Potenzial dieser Entwicklung anhand „Virtueller Observatorien“: astronomische Messdaten aus unterschiedlichsten Quellen werden hierfür unter einer Oberfläche zusammengeführt und dienen für sich als Datenbasis weiterer Forschung. Anstelle des Blicks durch ein Teleskop in den Himmel tritt der Blick „durch“ Analyse- und Visualisierungstools in die Daten (Bell, Hey & Szalay, 2009). Dies ist eine Entwicklung, die außer in der Astronomie auch in anderen, den besonders datenintensiven Wissenschaften, wie bspw. der Genomanalyse oder der Erforschung des Klimawandels, aber auch in den Sozial- und Wirtschaftswissenschaften zu beobachten ist. Von Fachwissenschaftlerinnen und -wissenschaftlern werden Forschungsdaten in so großem Umfang erhoben, dass deren Verwaltung, Auswertung und Weiterverarbeitung nur noch durch intensiven Einsatz von Computertechnologie zu leisten ist. In Verbindung mit Informationstechnologien werden Daten somit gleichzeitig Grundlage und Ergebnis wissenschaftlicher Erkenntnisprozesse. Die Daten werden zu Daten aus anderen Quellen in Beziehung gesetzt und bilden eine neue Datenbasis für weitere Berechnungen. Dies geschieht häufig in einer technischen Umgebung, die vernetztes und kooperatives Forschen unterstützt. Die inzwischen gängige Bezeichnung dafür ist eScience (*enhanced science*).

> „e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it.“ (John Taylor, Director General of Research Councils, Office of Science and Technology, zitiert nach: „National e-Science Centre definition of e-Science“, 2006)[1]

Die Umwälzungen durch den Einzug der eSciences in den datenintensiven Wissenschaften hielt der amerikansche, bei Microsoft arbeitende Computerwissenschaftler Jim Gray für so gravierend, dass er von der Entstehung eines vierten Forschungsparadigmas sprach. Nach der rein empirisch ausgerichteten und beobachtenden Wissenschaft (1) und der auf Theorie und Modellentwicklung basierten Wissenschaft (2) entstand durch die informationstechnologischen Möglichkeiten eine Wissenschaft, die komplexe Phänome in Simulationen durchrechnete (3). Mittlerweile entwickeln sich in der „Daten getriebenen“ (*data driven*) Wissenschaft (4) gänzlich neue Wege zu wissenschaftlicher Erkenntnis auf der Grundlage der Exploration von massenhaft vorhandenen oder erhobenen Daten (Hey, Tansley & Tolle 2009, S. xvii f.)

Voraussetzung für eine solche, paradigmatisch verstandene „*data-driven-science*“ ist ein systematisches Management, im Sinne von: Infrastruktur, Regelwerken und – was häufig vergessen wird – *zusätzlichen* personellen Ressourcen (vgl. Bell, Hey & Szalay, 2009, und: Lynch, 2009, S. 181) Die in einem System integrierte Verarbeitung und Darstellung heterogener Datenmengen aus ganz unterschiedlichen Quellen, erfasst mit verschiedenen Instrumenten, erfordert einen langen Prozess an Transformation, Speicherung und Übermittlung. Dieser Prozess muss bewusst und nachvollziehbar gestaltet werden, damit die erzeugten Daten ihre wissenschaftliche Aussagekraft behalten und für die Auswertung zugänglich bleiben. Das ist die Aufgabe des Forschungsdatenmanagements. Es findet in dem Bewusstsein statt, dass lokale Lösungen Bestandteil einer übergreifenden Forschungsdateninfrastruktur sein müssen.

Das Forschungsdatenmanagement muss so gestaltet werden, dass Datenzugriff und -auswertung unabhängig vom Datenerzeuger möglich wird und bleibt. Neben der technischen Speicherung und Lesbarkeit der Forschungsdaten müssen ausreichend Informationen zu ihrer Interpretation in Metadaten überliefert werden. Damit die Forschungen „in den Daten“ nachprüfbar bleiben und auch über mehrere Stufen veränderte Datensätze als Basis weiterer Forschungen dienen können, muss die Zuverlässigkeit der Daten sichergestellt und ihr Verarbeitungsprozess nachvollziehbar dokumentiert werden, um die Nachweiskette zu sichern. Institutionen und Personen, die Verantwortung für das Forschungsdatenmanagement übernehmen, müssen bei den Fachwissenschaftlerinnen und -wissenschaftlern großes Vertrauen hinsichtlich der Archivierung und Rechte-

1. National e-Science Centre definition of e-Science. Retrieved from http://www.nesc.ac.uk/nesc/define.html [Zugriff am 19.08.2011].

verwaltung für die anvertrauten Forschungsdaten genießen (vgl. PARSE. Insight, 2009, S. 7).

Aus der Perspektive des *Information Professionals* ist eScience eine Infrastrukturaufgabe. Ziel der Infrastrukturanstrengungen für Forschungsdaten muss sein, gleitende, nahtlose Übergänge zu schaffen im Zusammenspiel von Institutionen, Disziplinen und technischen Systemen. In der Praxis ist das Forschungsdatenmanagement bislang überwiegend in den Disziplinen verankert. Das interdisziplinär geprägte Forschungsfeld Klimawandel macht deutlich, wie wichtig perspektivisch die integrierte Analyse von Daten unterschiedlichster Fachwissenschaften ist. Es bedarf übergreifender Anstrengungen, die bestehenden disziplinspezifischen Lösungen zu einer wirklichen Infrastruktur zu vernetzen und diese nachhaltig aufrecht zu erhalten (Müller, 2011).

1.1.2 Was sind Forschungsdaten?

Die verschiedenen Wissenschaften sprechen von Messdaten, Rohdaten, empirischen Daten, Quelldaten, Forschungsrohdaten etc., verwenden viele unterschiedliche Begriffe und haben auch unterschiedliche Konzepte von dem, was für sie ein „Datum“ ist. Mittlerweile kristallisiert sich heraus, dass im deutschen Sprachgebrauch der Begriff „Forschungsdaten“ bevorzugt wird. Damit gibt es aber noch keine wirkliche Einigung darüber, was darunter verstandenen wird. Manche Wissenschaften sehen in dem ‘Gegebenen’ (lat. *datum*) eher das Dokument für etwas zu beschreibendes oder als Fakt aus der Realität zu erklärendes. Das vorherrschende Modell wissenschaftlichen Arbeitens bleibt in allen Gray’schen Paradigmen ein eher induktives: geprägt vom kritischen Rationalismus wird zunächst die Untersuchungshypothese generiert, um diese an der Realität zu überprüfen. Dazu sind „Spuren“, Repräsentationen der Realität notwendig, an denen die eigenen Ideen abgearbeitet werden können. Jede Wissenschaft hat eine solche Basis: die Naturwissenschaften die Beobachtungen und Messungen mit Geräten oder die Erhebung von „Proben“, die Sozialwissenschaften die Befragungen von Personen oder die Beobachtung von sozialen Situationen und die Geisteswissenschaften die Analyse von kulturellen Artefakten oder die experimentelle Produktion von solchen. In manchen Situationen werden explizit „Daten“ unter bestimmten Fragestellungen erhoben oder gar generiert, manche andere Wissenschaften finden solche Spuren der Realität vor. Manche Daten werden aus rein beobachtender Perspektive benutzt, wie z. B. die Verwendung von prozessproduzierten Daten (etwa aus Verwaltungshandeln), manche Daten werden „nachgenutzt“, nachdem sie zu anderen analytischen Zwecken als denen der wissenschaftliche Fragestellung erhoben wurden (etwa klinische Daten) und schließlich werden auch Daten aus vorhandenen, unter Forschungsgesichtspunkten erhobene Daten in neuen Kontexten nachgenutzt (sog. Sekundäranalyse).

Üblicherweise werden die Daten für Forschungsfragen unter gezielten Fragestellungen (Hypothesen) ausgesucht. In einzelnen Disziplinen wie der Medizin unterliegen sogar die Arbeitshypothesen einem *Reviewing*-Verfahren und es wird von der *Community* streng geprüft, ob der Forschungsprozess auch die ursprüngliche Fragestellung unverändert verfolgt hat. Die neuen, schlichtweg vorhandenen Datenmengen, die solcher Hypothesen bezogenen Analyse nicht unterlegen haben, bieten sich regelrecht an, rein induktiv - in dem Fall ohne Hypothese – exploriert zu werden. Die Eingrenzung oder Auswahl von zu analysierenden Daten wird dabei dann von dem Datenanalyseansatz vorgegeben. Je nach Instrument etwa des *data mining* oder der visuellen Datenexploration (vgl. Kap. 2.8) und seiner dahinter liegenden Algorithmen „entstehen““ neue Ideen oder gar Forschungshypothesen. Das neue Paradigma ermöglicht somit grundsätzlich neue wissenschaftliche Vorgehensweisen (Kell & Oliver, 2004[2]).

Bei der Erhebung oder Selektion von genuinen Forschungsdaten kommen jedoch ebenfalls schon Perspektive verkürzende Mechanismen ins Spiel, wie der Einsatz bestimmter Messmethoden, die bewusste Ausblendung von Begleitfaktoren des untersuchten Objektes: auch wenn in großen Datenmengen Hypothese freie Restmengen übrigbleiben, sind diese doch meist (noch) mit einer bestimmten Intention erfasst worden. Beide Perspektiven, die des Originalansatzes und die der Sekundäranalyse, ohne Interferenzen weiter zu verwerten, ist eine der Herausforderungen des Forschungsdatenmanagements. Für die Weiterverwertbarkeit großer Datenmengen in der eScience gilt es schon im Forschungsdesign auch an mögliche Synergieeffekte zu denken. Denkbar ist jetzt auch das reine „*Screening*“ von Daten aus der Realität ohne weiterreichende Analysefragestellung mit dem Ziel möglichst großer Vollständigkeit – je nach Datenart und Objekt wären die technischen Speicherkapazitäten sogar vorhanden für dieses „interessenlose Wohlgefallen“ des technischen Beobachtens der Welt – und tatsächlich hat die Datenanalyse oft auch ästhetische Aspekte in ihren Ergebnissen.

Auch die weitere Verwaltung ‘gewonnener’ Daten birgt Gefahren des Verlustes an Beziehung vom Referenten (zeichentheoretisch gesprochen) – zumindest kann sich durch Speicherung, Codierung, Datenbankmanagementsystem oder Abfragemechanismen ein Rauschen einstellen, dass gerade in der Weiterverwertung von Daten problematisch sein könnte, da es nicht intentional im Forschungsansatz (der Hypothese) kontrolliert wird. Auch hier liegt eine der wesentlichen Herausforderungen des Forschungsdatenmanagements. Insgesamt wird deutlich, dass zur Nutzung der neuen Möglichkeiten der eScience auch neue Aufgaben und Rollen für alle Beteiligten entstanden sind. Um Daten sinnvoll weiterverwenden zu können, damit diese auch neue wissenschaftliche Erkenntnisse generieren können, die die Welt korrekt abbilden, muss diesen

2. Dank an Erich Weichselgartner für den Hinweis auf diese Literaturstelle!

Aufmerksamkeit und Pflege zuteilwerden; sie sollen eben nicht nach der ersten Falsifizierung einer Hypothese in der Schublade (resp. Festplatte) verschwinden. Interessanterweise spricht die junge Disziplin des Forschungsdatenmanagements hierbei tatsächlich von der Datenkuratierung und kreiert damit die Rolle des *data curators* (vgl. Kap. 3.5). Die amerikanische Vereinigung der Forschungsbibliotheken wählte gar in einem Strategiepapier für die *National Science Foundation* den Begriff des „*stewardship*" (ARL, 2006). Ganz wie die Pflege von Objekten des kulturellen Erbes z. B. in Museen dreht es sich um die Bewahrung potenzieller wissenschaftlicher Erkenntnis als „Erbe" vorhergehender Forschungen.

Durch den Bezug auf die wissenschaftliche Wiederverwertung von Daten lässt sich auch ableiten, in welchem Maße Datenkuratierung tatsächlich relevant ist. Die wissenschaftliche Datenkuratierung dient der Erzeugung von neuem wissenschaftlichen Wissen und lässt sich damit als neues Aufgabenfeld abgrenzen von IT-Sicherheit, IT-Infrastruktur, Management von Daten in laufenden Verwaltungsprozessen, ja sogar der Fragestellungen der Langzeiterhaltung und Archivierung digitaler Daten und Objekte. Die Beiträge im vorliegenden Band – aber auch viele internationale Diskussionen um die neuen Anforderungen, die das Forschungsdatenmanagement an die Wissenschaft stellt – gehen davon aus, dass es sich um eine regelrechte Infrastrukturaufgabe handelt, die weder rein technisch gelöst, noch den Fachwissenschaftlern alleine überlassen werden kann. Es liegt vielfach auf der Hand, dass es sich hierbei um eine Aufgabe für die Informationswissenschaften handelt. Eines der grundlegenden informationswissenschaftlichen Theoreme beinhaltet auch schon das Konzept der Daten als eine Basis dessen, worum sich Informationsspezialisten kümmern. Unter eher zeichentheoretischen Überlegungen spricht die Informationswissenschaft von der „DIKW Hierarchie", die beschreibt, dass Informationen (I) auf den Daten der emprischen Ebene (D=*data*) aufbauen und in Wissen (K=*knowledge*) bzw. Weisheit (W=*wisdom*) münden können (Hobohm, 2010). Um *knowledge management* hat sich die Informationswissenschaft in den letzten Jahren intensiv gekümmert, die Frage nach der Weisheit wird letztlich wohl eher den Philosophen überlassen bleiben, aber die *Basis* von Informationsarbeit ist *Daten*management. Dies wurde bisher eher unter dem Aspekt der technischen *Daten*verarbeitung behandelt, wenn z. B. bibliothekarische oder dokumentarische *Daten*banken konzipiert und gepflegt werden. Der Bereich des Forschungsdatenmanagements zeigt aber, dass auch hier (dank der enormen informationstechnischen Entwicklung) unter anderen, nicht technischen Aspekten neue informationswissenschaftliche Arbeitsfelder entstanden sind.

1.1.3 Entwicklungen in der Wissenschaftspraxis

Die Praxis, Forschungsdaten zu publizieren und damit für andere nachprüfbar und -nutzbar zu machen, hat eine Tradition, die bis ins vorletzte Jahrhundert zurückreicht (Weichselgartner et al., Kap. 3.4 in diesem Band). Jahrzehnte lang dominierte aber die Ablage in lokalen Systemen, die Datenspeicherung und -weitergabe wurde auf der Ebene von Arbeitsgruppen oder Instituten praktiziert. Dies ist bis heute in den „*small Sciences*“ gängige Praxis (vgl. Müller, 2011, S. 133f.).

Die Notwendigkeit des systematischen Forschungsdatenmanagements offenbarte sich zuerst in den datenintensiven Wissenschaften. Stark automatisierte Messtechniken und Verfeinerung von Messgeräten beschleunigen das gewaltige Anwachsen der Datenmengen. Mit dem stetigen Anstieg der zu speichernden und zu verarbeitenden Datenmengen wurde es erforderlich, Informationssysteme zu entwickeln, die diese Daten langfristig speichern und für weitere Nutzung zugänglich halten. Insbesondere für Messdaten, die nur einmalig erhoben werden können (z. B. Wetterdaten), besteht langfristiger Bedarf, wiederholt darauf zuzugreifen (vgl. Klump 2009, S. 111). Forschungsdatenmanagement ist zudem auch aus wirtschaftlichen Überlegungen erforderlich. Im Vergleich zu Kosten und Aufwand, die besonders in Großforschungsprojekten für Datenerhebungen aufgebracht werden müssen (man denke z. B. an wissenschaftliche Bohrungen, Polarexpeditionen oder Satellitenmissionen), sind die Kosten für das Forschungsdatenmanagement gering. Zentrale Entwicklungen zu Datenmanagement und -publikation stammen aus dem Umfeld der datenintensiven Wissenschaften, weil sie aus ihrer Praxis heraus schon lange gezwungen sind, den dynamischen Lebenszyklus von Forschungsdaten handhabbar zu machen. Das Kernproblem bestand dabei von Anfang an weniger in der technischen Speicherung der Daten, als in der langfristigen Interpretierbarkeit der Datensets. „The logical availability of the data is rapidly degrading due to a lack of structural knowledge about the data and a lack of information describing the data (metadata).“ (Diepenbroek et al., 1998, S. 655). In den vergangenen zehn Jahren wurden große Fortschritte in Entwicklung und Anwendung bspw. von digitaler Langzeitarchivierung und Grid-Technologien erzielt, die Abstimmung der Technik mit den institutionellen und sozialen Bedingungen in den Wissenschaften ist jedoch eine bleibende Herausforderung (Klump, 2009, S. 109).

Eine große Herausforderung ist die rasant steigende Menge der Daten. Erfolgte früher die Datenerhebung manuell durch den Wissenschaftler oder Laboranten, werden die Daten heute automatisch erzeugt bzw. erfasst. Dabei werden auch Daten erfasst, die früher vernachlässigt wurden. Das führte dann oft zu Problemen mit der Reproduzierbarkeit, wenn Messungen ohne identische, weil nicht bekannte, Rahmenbedingungen wiederholt wurden. Durch die zunehmende Menge an Daten und der wachsenden Leistungsfähigkeit der Informati-

onstechnologie hat sich auch die Geschwindigkeit der Datenverarbeitung wesentlich beschleunigt. Erinnert sei in diesem Zusammenhang an das Apollo-Programm der NASA. Die Auswertung der gewonnenen Daten kam u.a. wegen nicht ausreichender informationstechnologischer Verarbeitungskapazitäten nie zum Abschluss.

Doch es gibt auch z. T. durchaus berechtigte Vorbehalte, Hemmschwellen beim Umgang mit Daten bei den einzelnen Akteuren. „Den Wissenschaftlern fehlen Anreize, um z. B. Daten als Publikation zu werten. Es gibt Angst vor „Missinterpretation" der Daten durch Dritte." (Büttner & Rümpel, 2011, S. 107). Dies ist auch der Grund für die Dreiteilung im bekannten „*Data Curation Continuum*" Modell (Treloar & Harboe-Ree, 2008, S. 6 sowie Kap. 1.2 Rümpel „Der Lebenszyklus von Forschungsdaten" in diesem Band). Die Daten der sog. „*Private Research Domain*", der Entstehungsphase haben einen hohen Zugriffsschutz. Eine Weitergabe der Daten an Dritte ist unerwünscht, da Gefahr besteht, das andere Wissenschaftler damit eine Publikation machen, ohne die oft langwierigen Messungen durchgeführt zu haben. Erst in der nächsten Phase, der „*Shared Research Domain*" erfolgt eine, wenn auch beschränkte Weitergabe an kooperierende Wissenschaftler.

> „Zudem ist eine Bereitschaft zur Beschreibung der Daten mit Metadaten kaum ausgeprägt. Es fehlen Vorgaben (*Policies*) von Seiten der Verwaltung, den Forschungsförderern und den Verlagen, zum Umgang mit Forschungsdaten. Zunehmend greift die Erkenntnis, dass auch die Informationstechnologie nicht Probleme löst, sondern ein Tool zur Problemlösung ist." (Büttner & Rümpel, 2011, S. 107).

Ein z. Z. sich stark entwickelndes Gebiet sind die Verknüpfung der Daten, das sog. „*Linked Open Data* Konzept". Dabei werden die Datenbestände einer Domäne mit Datenbeständen anderer Domänen wechselseitig verknüpft. Dies eröffnet völlig neue Möglichkeiten. Daten des Meteorologischen Dienstes könnten z. B. wertvolle Hinweise für die Luftfahrt enthalten, Daten aus Untersuchungen zum Klimawandel bieten reichlich Hinweise für Leistungsdaten von Nutztieren etc. Dazu ist dann eine aktive Kooperationen zwischen den Einrichtungen unentbehrlich. Diesen Themen wird im Kapitel 3.3 „Semantische Vernetzung von Forschungsdaten, Ontologien und Ontologiesprachen" Neher, Ritschel nachgegangen.

Neben diesen technischen Entwicklungen und Konzepten kommt es aber vielmehr darauf an, ein Bewusstsein, eine *Awareness*, für Forschungsdaten zu entwickeln und zu verbreiten. Dabei geht es z. B. um die Einführung von Forschungsdatenpolicies auf Instituts- wie auch Domänenebene. Sie sind Voraussetzung die o.a. Hemmschwellen zu überschreiten. Diese *Policies* müssen auf die Förderung einer Publikationskultur für Forschungsdaten zielen (Müller, 2011).

Ähnliche Diskussionen zur rasanten technologischen Entwicklung, zu Fragen der Beschreibung der Daten mit Metadaten, der Weitergabe von Daten etc. wurden vor Jahren zum elektronischen Publizieren geführt. Das Ergebnis ist bekannt – es hat sich durchgesetzt.

1.1.4 Notwendigkeit der Weiterentwicklung des Forschungsdatenmanagements

Nachdem renommierte Wissenschaftler wie Herbert Van de Sompel, Carl Lagoze und andere in vielbeachteten Beiträgen (z. B. 2004) grundsätzlich neue Strukturen für die wissenschaftliche Arbeit gefordert und deren Möglichkeit aufgezeigt hatten, entstand vor allem im anglomerikanischen Bereich die Erkenntnis der Notwendigkeit wissenschaftspolitsch steuernden Handels („..z. B. NSF 2005, 2007; ARL, 2006; Lyon, 2007; Treloar & Harboe-Ree, 2008)“. In Deutschland wurde dies aufgegriffen von den großen Wissenschaftsorganisationen, die sich 2008 zur „Schwerpunktinitiative Digitale Information“ zusammenschlossen (Allianz, 2008). Auch die Deutsche Forschungsgemeinschaft wies zunehmend auf die Notwendigkeit des Forschungsdatenmanagements hin (vgl. Vorwort) und in der von der Wissenschaftsgemeinschaft Gottfried Leibniz (WGL) 2010 ins Leben gerufenen und von der GWK beauftragten Kommission „Zukunft der Informationsinfrastruktur (KII)“ erhielt besonders der Aspekt Forschungsdaten einen großen Stellenwert. Der im Frühjahr 2011vorgelegte Gesamtbericht (WGL, 2011) fasst eine Reihe von Forderungen zusammen, die davon ausgehen, dass es sich bei Forschungsdaten zum größten Teil um „nationales Kulturgut“ handelt, das „im Sinne einer hoheitlichen Aufgabe dauerhaft gesichert und [...] bereitgestellt werden“ muss (WGL 2011, S. B109). Hierbei wird betont, dass in Deutschland auf diesem Feld besonderer Nachholbedarf besteht.

> „Die Erwartungen der Wissenschaftlerinnen und Wissenschaftler als Nutzer im Themenfeld Forschungsdaten betreffen v. a. den Aufbau und die Verstetigung von Disziplin-getriebenen, nutzerorientierten Infrastrukturen. Dabei sind vier Handlungsfelder zu berücksichtigen: nachhaltige Sicherung, Erschließung/Bereitstellung, Nachnutzung und langfristige Bewahrung von Forschungsdaten. Um der Aufgabe gerecht zu werden, ist eine nationale Allianz zwischen allen Akteuren nötig: Die Wissenschaftler als Datenproduzenten und Datennutzer, die Hochschulen und Forschungseinrichtungen, die Forschungsförderer, die Datenzentren und Infrastruktureinrichtungen sowie Bund und Länder müssen kooperieren, um gemeinsam den gewaltigen Herausforderungen zu begegnen und Deutschland anschlussfähig für den europäischen und internationalen Wettbewerb zu halten.“ (ebd.)

Neben grundlegenden Forderungen nach Aufbau einer Infrastruktur (samt Regelwerk, s.o.) wird besonders auf die Notwendigkeit der „Verankerung von einschlägigen Ausbildungsangeboten für Wissenschaftler (Schlüsselkompetenz) und für Daten-Kuratoren (Lehre)“ (ebd.) hingewiesen. Der vorliegende Band will zu diesen beiden Aspekten schon jetzt beitragen.

Die Gemeinsame Wissenschaftskonferenz hat den von der Kommission vorgelegten Bericht begrüßt und den Wissenschaftsrat aufgefordert, in seinem an mehrere Empfehlungen zur Forschungsinfrastruktur in Deutschland im Frühjahr 2011 anknüpfendes Gesamtkonzept der Wissenschaftsinfrastruktur in Deutschland aufzugreifen. Entsprechende Stellungnahmen sind frühestens 2012 zu erwarten. Ob bei der Geschwindigkeit der Entwicklung der Cyberinfrastruktur (NSF, 2007) so viel Zeit ist, wird sich herausstellen.

1.1.5 Auswirkung der Datenorientierung auf Informationswissenschaften und Information Professionals

Datenmanagement als vorzugsweise technische Disziplin zu verstehen greift viel zu kurz. Gleiches gilt für die lange vorherrschende Auffassung, die Informationstechnologie löse Probleme. Vielmehr greift zunehmend die Erkenntnis dass Informationstechnologie ein Tool zur Problemlösung ist. Aktuelle Ergebnisse von Forschungsprojekten (Wibaklidama 2010, CopaL 2011) sowie Graduierungsarbeiten an der FH Potsdam (Rümpel, 2010) zeigen anschaulich, dass Datenmanagement eine neue Ausprägung des Wissensmanagements ist. Informations- und Wissensmanagement ist jedoch ein originäres Thema der Informationswissenschaften. Auch beim Datenmanagement geht es um klassische informationswissenschaftliche Themen, wie Bewertung und Einordnung in Kontexte, um Metadaten, um Ontologien. Als essentiell haben sich Metadaten und der Bezug zu den datenerhebenden Experten ergeben – eine Erkenntnis, die schon lange aus dem Wissensmanagement bekannt ist.

Die Informationsinfrastruktur ist traditionell am Ende des geistigen Schaffensprozesse verortet. Sie konzentriert sich auf die Ergebnisse der Wissenschaftler, die Dokumente, die Publikationen, erschließen diese und stellen sie der wissenschaftlichen Community über Datenbanken oder Bibliotheksdienste zur Verfügung. Forschungsdaten standen bisher nicht im Fokus der Informationsinfrastruktur. Es war eher gängige Praxis, dass diese Daten oft für immer verloren gingen, z. B. wenn das Forschungsthema oder Forschungsprojekt beendet wurde. In der Wissenschaft nehmen virtuelle Forschungsumgebungen (eng.: *Virtual Research Environments* VRE) in den letzten Jahren ständig zu. „Durch E-Science wächst die Erkenntnis, dass der Wert der Forschung insbesondere in den Daten steckt und sich daher das Arbeitsspektrum auf das Primärobjekt, die Forschungsdaten, erweitern muss.“ (Büttner & Rümpel, 2011). Die VRE begleiten den Wissenschaftler im Forschungsprozess von der Ideengenerierung über

die experimentelle Datenerhebung, der Aggregation der Daten, dem Austausch mit der *scientific community*, bis zur Publikation. Für die notwendigerweise eingebundene Informationsinfrastruktur ändert sich das Aufgabengebiet gravierend. Vom Dienstleister für Endprodukte zum Mitgestalter des Forschungsprozesses. Wenn anfangs vom 4. Forschungsparadigma gesprochen wurde, tut sich für die Informationsinfrastruktur hier ebenfalls ein Paradigmenwechsel auf.

Es wird daher wichtig, zu beobachten, ob dies Auswirkungen auf die Nachfrage von Infrastrukturspezialisten haben wird. Wird es konkret Konsequenzen für Ausbildung von Informationswissenschaftlern haben? Gibt es ggf. demnächst Studiengänge die z. B. „Data Librarian" ausbilden? Diesen Themen wird im Kapitel 3.5 „Informationswissenschaftler im Forschungsdatenmanagement" von Büttner, Rümpel & Hobohm nachgegangen.

Literaturhinweise

Allianz der deutschen Wissenschaftsorganisationen, 2008. *Pressemitteilung: Wissenschaftsorganisationen starten Schwerpunktinitiative zur „Digitalen Information"* Online: http://www.allianzinitiative.de/de/aktuelles_und_presse/12062008/ [Zugriff am 18.08.2011] (vgl. http://www.allianz-initiative.de/de/ [Zugriff am 18.08.2011]).

ARL (Association of Research Libraries), 2006: To Stand the Test of Time. Long-term Stewardship of Digital Data Sets in Science and Engineering, A Report to the National Science Foundation. Arlington. Online: http://www.arl.org/bm~doc/digdatarpt.pdf [Zugriff am 18.08.2011].

Bell, G., Hey, T., & Szalay, A. 2009. Computer Science: Beyond the Data Deluge. *Science*, *323*(5919), S. 1297–1298.

Büttner, S. & Rümpel, S., 2011. Bibliotheken und Bibliothekare im Datenmanagement. In: S. Schomburg, C. Leggewie, H. Lobin, & C. Puschmann, Hrsg. 2011. *Konferenz: Digitale Wissenschaft. Stand der Entwicklung digital vernetzter Forschung in Deutschland.* Köln, Deutschland 20.-21. Sept. 2010. 2. Aufl. Köln, S.107–114. Online: http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf [Zugriff am 18.08.2011].

CopaL (Communities of Practice für den Wissens- und Technologietransfer agrarwissenschaftlicher Institute der Leibniz-Gemeinschaft), 2011. *Forschungsprojekt*. vgl. http://iw.fh-potsdam.de/4725.html [Zugriff am 18.08.2011].

Diepenbroek, M. et al., 1998. PANGAEA information system for glaciological data management. *Annals of Glaciology*, (27), S. 655–660. Online: Retrieved from http://hdl.handle.net/10013/epic.11222.d001 [Zugriff am 18.08.2011].

Hey, A. J. G. & Trefethen, A. E., 2003. *The Data Deluge: An e-Science Perspective*.Wiley and Sons. Online: http://eprints.ecs.soton.ac.uk/7648/ [Zugriff am 18.08.2011].

Hey, T. Tansley, T. & Tolle, K., 2009. Jim Gray on eScience: A Transformed Scientific Method. In: A. Hey St. Tansley & K.M. Tolle, 2009. *The Fourth Paradigm Data-Intensive Scientific Discovery.* Redmond, Wash.: Microsoft Research, S. xix-xxxiii. Online: http://research.microsoft.com/en-us/collaboration/fourthparadigm/ [Zugriff 18.08.2011].

Hobohm, H. C., 2010. „DIKW Hierarchie“. In: K. Umlauf & S. Gradmann, Hrsg. 2010. *Lexikon der Bibliotheks- und Informationswissenschaft.* Stuttgart: Hiersemann, S. 222f.

Kell, D.B. & Oliver, S.G., 2004.: Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* 26(1), S. 99–105.

Klump, J., 2009. *Digitale Forschungsdaten*. Online: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_292.pdf [Zugriff am 18.08.2011].

Lynch, C., 2009. Jim Gray's Fourth Paradigm and the Construction of the Scientific Record. In: S. Tansley & K. Tolle, Hrsg. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond, Wash.: Microsoft Research, S. 177–183. Online: http://research.microsoft.com/en-us/collaboration/fourthparadigm/ [Zugriff am 18.08.2011]

Lyon, L. 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships.* Consultancy Report. Online: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf [Zugriff am 18.08.2011].

Müller, L., 2011. An der Schwelle zum Vierten Paradigma: Datenmanagement in der Klimaplattform. In S. Schomburg, C. Leggewie, H. Lobin, & C. Puschmann, Hrsg. 2011 *Konferenz: Digitale Wissenschaft. Stand der*

Entwicklung digital vernetzter Forschung in Deutschland. Köln, Deutschland 20.-21. Sept. 2010. 2. Aufl. Köln, S.131–137. Online: http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf [Zugriff am 18.08.2011].

NSF (National Science Foundation)/ National Science Board, 2005. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century.* Arlington. Online: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf [Zugriff am 17.08.2011]

NSF (National Science Foundation), 2007. *Cyberinfrastructure. Vision for 21st Century Discovery* (NSF 07–28). Arlington. „Online:“ http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf [Zugriff am 18.08.2011].

PARSE.Insight, 2009. *Road Map*. Deliverable D2.1. Online: http://www.parse-insight.eu/downloads/PARSE-Insight_D2-1_DraftRoadmap_v1-1_final.pdf [Zugriff am 18.08.2011].

President's Information Technology Advisory Committee, 2005. *Computational Science: Ensuring America's Competitiveness.* Online: http://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf [Zugriff am 18.08.2011].

Rümpel, S., 2010. *Data Librarianship – Anforderungen an Bibliothekare im Forschungsdatenmanagement.* Diplomarbeit, Fachhochschule Potsdam.

Treloar, A. & Harboe-Ree, C., 2008. Data management and the curation continuum. How the Monash experience is informing repository relationships. *VALA2008 14th Biennial Conference*. Melbourne, Australien 5.-7. Feb. 2008. Online: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf [Zugriff 18.08.2011]

Van de Sompel, H. Payette, S. Erickson, J. Lagoze, C. & Warner, S., 2004. Rethinking scholarly communication. Building the system that scholars deserve. *D-Lib Magazine,* 10(9). doi:10.1045/september2004-vandesompel.

„Wibaklidama“, 2010. *Wissenbasiertes Klimadatenmanagement.* Online: http://wibaklidama.fh-potsdam.de/ [Zugriff am 18.08.2011].

WGL (Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz e.V./ Leibniz Gemeinschaft), 2011. Kommission Zukunft der Informationsstruktur: *Gesamtkonzept für die Informationsinfrastruktur in Deutschland.* Empfehlungen im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder, April 2011. Online via: http://www.wgl.de/?nid=infrastr&nidap=&print=0 [Zugriff am 14.08.2011].

1.2 Der Lebenszyklus von Forschungsdaten

Stefanie Rümpel
Fachhochschule Düsseldorf

Für Wissenschaftler sind Veröffentlichungen unentbehrlich und werden als „Währung“ angesehen. Die Forschungsdaten, auf denen die Publikation basiert, sind aber i.d.R. nicht enthalten.

> „Mit beginnender Analyse und Interpretation von Daten, werden unter Umständen nicht mehr alle Details eines Rohdatensatzes transportiert. […] [Beispielsweise] werden zusammengeführte Einzelmessungen unter Umständen nur noch als Mittelwert dargestellt, obwohl ursprünglich eine ganze Reihe von Forschungsdatensätzen erzeugt wurde […].“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 26–27)

Doch gerade die Daten sind deutlich interessanter und relevanter für weitere Forschungsprozesse, um einen Mehrwert zu erreichen (Sietmann, 2009, S. 154).

> „ […] data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself.“ (Gold, 2007)

Mit anderen Worten: Forschungsdaten nur als Grundlage für eine Publikation zu verwenden, missachtet deren Wert. Gegenwärtig wird im Forschungsprozess meist darauf verzichtet, auf bereits erhobene und gespeicherte Daten zurückzugreifen. Eher werden kostenintensive Messwiederholungen in Kauf genommen.

Die Technische Informationsbibliothek (TIB) formulierte, bezogen auf das Fachgebiet Chemie, die Missstände des Forschungsdatenmanagements deutlich.

> „Der bisherige Umgang mit Forschungsdaten in der Chemie beinhaltet keine allgemein anerkannten Standards hinsichtlich einer Nachnutzbarkeit oder langfristigen Verfügbarkeit. Überwiegend existiert keine Qualitätssicherung, keine gesicherte Langzeitarchivierung, kein gesicherter Nachweis sowie keine Erschließung der Forschungsdaten und somit keine Datensicherheit.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 5)

Diese Vorgehensweise ist vorherrschend, da Daten aus vielen Forschungsprozessen i.d.R. noch gar nicht dauerhaft gespeichert oder aufbereitet werden. Gründe sind sowohl auf Seiten der Wissenschaftler als auch der Institutionen zu finden, beispielsweise Unwissenheit über persistente und qualitative Verwaltung der Daten, Hemmnisse bezüglich der Datenspeicherung oder fehlende Transparenz der gespeicherten Daten in Repositorien.

Um die Distanz von Wissenschaftlern und allen Involvierten im Forschungsprozess gegenüber der Aufbereitung von Daten zu mindern, erscheint es wesentlich, das Bewusstsein der Wissenschaftler für die Notwendigkeit einer Nachvollziehbarkeit der Forschung zu fördern. Dabei muss beachtet werden, dass die dauerhafte Speicherung, Pflege und Bereitstellung von Forschungsdaten einen erheblichen Arbeitsaufwand erfordert.

Die Stimmen nach einem verantwortungsvollen und organisierten Umgang mit Forschungsdaten werden immer lauter:

> „Einhergehend mit der Bearbeitung von Forschungsdaten steigt die Gefahr von Fehlern und Fehlinterpretationen. Umso komplexer die Experimente, Datenstrukturen und Fragestellungen, desto relevanter wird die Verfügbarkeit von ursprünglichen Forschungsdaten, um Ergebnisse kritisch zu evaluieren. Daher ist der öffentliche Zugang zu den Forschungsdaten im wissenschaftlichen Erkenntnisgewinn eminent.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 27)

Der gegenwärtige Lebenszyklus von Forschungsdaten sieht jedoch anders aus.

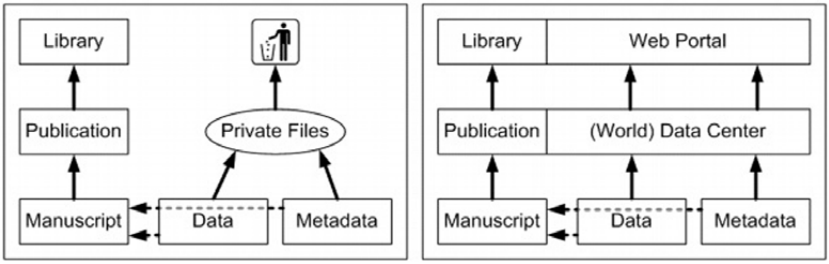

Abb. 1: *(links) Schematische Darstellung des wissenschaftlichen Informationsflusses in der Forschung (= bekannter Weg), (rechts) Veränderter Umgang mit Daten* (Klump et al., 2006, S. 80 nach Helly, Staudigel & Koppers, 2003, S. 2)

> „[...] Forschungsdaten sind die Grundlage jeglicher wissenschaftlicher Arbeit. Ausgehend vom Experiment durchlaufen Forschungsdaten viele, dem Wissenschaftler bekannte Stadien, die letztendlich als Erkenntnisgewinn in einer wissenschaftlichen Publikation münden. Danach verliert sich der bis dahin so klare Weg der Forschungsdaten, was deren Dokumentation, langfristige Speicherung oder Nachnutzbarkeit für andere Wissenschaftler betrifft.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 7)

Der Anspruch ist, die Daten aus den „Papierkörben“ der Forscher heraus in das Licht der Öffentlichkeit zu bringen. Sicher, die Speicherung von vielen einzelnen Daten ist arbeitsintensiver als die Speicherung einer einzelnen Text-Publi-

kation. Sie besitzen außerdem eine enorme Heterogenität und Komplexität, wodurch sie zu „[…] eine[r] wertvolle[n], jedoch schwierig zu handhabende[n] Ressource […]“ (NESTOR, 2009, S. 1) werden. Erforderlich ist also eine konsequent qualitative und persistente Verwaltung.

Hilfe bei der komplexen Verwaltung von Forschungsdaten gibt deren Lebenszyklus. Dieses wird im Folgenden mit Hilfe von zwei Modellen gezeigt. Beide beschreiben die verschiedenen Lebensphasen von Forschungsdaten, betrachten dies jedoch aus verschiedenen Blickwinkeln. Im ersten Modell werden die theoretischen Anforderungen an den Umgang mit Daten aufgeführt, im anderen die notwendigen technischen Bedingungen im Laufe des Lebenszyklus benannt.

1.2.1 Curation-Lifecycle-Model

Um einen Mehrwert von Forschungsdaten zu erhalten, ist eine adäquate Verwaltung notwendig. Ihr Lebenszyklus erstreckt sich über verschiedene Phasen, die von der Entstehung in wissenschaftlichen Arbeitsprozessen bis zur nachnutzbaren Archivierung reichen. Die Anforderungen an das Management von Forschungsdaten gehen weit über die Langzeitarchivierung hinaus (NESTOR, 2009, S. 1–2). Alle Tätigkeiten des Forschungsdatenmanagements werden durch das „*Curation Lifecycle Model*“, erstellt vom *Digital Curation Centre* (DCC), identifiziert (DCC, 2010).

> „It is important to note that the model is an ideal. In reality, users of the model may enter at any stage of the lifecycle depending on their current area of need.“ (DCC, 2010)

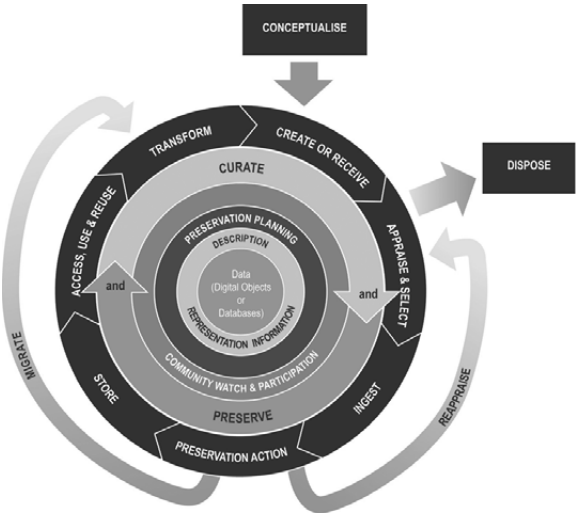

Abb. 2: *Curation Lifecycle Model* (DCC, 2010)

Die Abbildung zeigt ein sich aus mehreren Ebenen zusammensetzendes Kreismodell. Mittig sind die Tätigkeiten angeordnet, welche Daten während des gesamten Lebenszyklus begleiten: „*Data Preservation*“ (Datenerhaltung) und „*Data Curation*“ (Datenpflege). Beide ergänzen sich und bilden die Kernprozesse der *Digital Curation*. Diese Arbeiten müssen im gesamten Lebenszyklus von Forschungsdaten erfolgen. *Preservation* bezieht sich auf die Bewahrung der Daten im Sinne der digitalen Langzeitarchivierung. Um Daten nutzbar zu gestalten und zu behalten, wird eine Pflege notwendig, subsumiert unter dem Begriff „*Data Curation*“.

Die sequenziellen Tätigkeiten sind im äußeren Kreis dargestellt. Mit der *Konzeption* des Forschungsvorhabens erfolgt der Einstieg in den Kreislauf. Bereits vor der eigentlichen Forschungstätigkeit sind Überlegungen bezüglich der anfallenden Daten in dem Modell integriert. Die Deutsche Forschungsgemeinschaft (DFG) fordert beispielsweise seit kurzem die Berücksichtigung des Forschungsdatenmanagements bereits bei Beantragung von Forschungsvorhaben. Vor dem Start der Forschung müssen nun alle relevanten Fragen bezüglich des Umgangs mit Forschungsdaten beantwortet werden (Winkler-Nees, 2010, S. 23).

Der nächste Schritt der Datenverwaltung ist die *Datenerstellung* und die *Datenübernahme*. Der Punkt „*Datenübernahme*“ macht klar, dass es sich um einen Zyklus handelt bzw. handeln kann. In dieser „Lebensphase“ kann nach erhobenen Forschungsdaten recherchiert und diese im eigenen Forschungsprozess übernommen werden. Die Daten sind aber nur dann wiederverwendbar, wenn deren Anreicherung mit Informationen so umfassend ist, dass sie transparent werden. Wo, wann und wie sind die Daten erhoben worden? In vielen wissenschaftlichen Fächern ist es gegenwärtig jedoch nicht möglich nach passenden Forschungsdaten zu recherchieren, weil es keine ausreichenden Übersichten über die vorliegenden Daten gibt. Der Schwerpunkt bei der Verwaltung liegt gegenwärtig noch auf der Verwaltung von neu erhobenen Daten.

Nicht alle Daten, die erhoben wurden, müssen gespeichert werden.

> „Derzeit werden in den meisten Institutionen alle Primärdaten so lange gespeichert, bis diese irgendwann schleichend verloren gehen.“ (Severiens & Hilf, 2006, S. 29)

Es muss eine *Bewertung* erfolgen, welche Daten speicherwürdig sind. Daran schließt sich die *Selektion* jener Forschungsdaten an, die letztendlich gespeichert werden. Daten, die bei der Prüfung nicht als speicherwürdig erachtet wurden, können im Sinne der Richtlinien bzw. rechtlichen Anforderungen aussortiert werden. Die DCC bezeichnet diesen Vorgang als „*Dispose*“.

> „[…] im Laborbetrieb [wird] eine große Menge an Forschungsdaten produziert, die eher in den Bereich der Qualitätskontrolle von laufenden Prozessen fallen und nicht relevant für Publikationen sind. Für solche

> Forschungsdaten ist eine Speicherung in institutionellen Repositorien vorstellbar. Erst bei der Zusammenfassung von wissenschaftlichen Ergebnissen und deren Aufbereitung für eine Veröffentlichung werden Forschungsdatensätze für die Untermauerung wissenschaftlicher Erkenntnisse und Thesen herangezogen. Solche Forschungsdaten sind von Relevanz für die langfristige Speicherung und öffentliche Zugänglichkeit.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 30)

Mit der Speicherung werden Maßnahmen zur *Preservation* notwendig.

> „Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.“ (DCC, 2010)

Nach der Durchführung der *Preservation* schließt sich die „Langzeitspeicherung“ an. Dabei muss der *Zugriff* auf die Daten bzw. auch die *Benutzung* oder die sich daraus resultierende *Wiederverwendung* gewährleistet sein.

Die Komplexität des Datenmanagements wird offensichtlich. Gegenwärtig existieren Bereiche der Wissenschaft, in denen die Datenverwaltung schon recht gut funktioniert. In anderen wiederum trifft man beim Thema „Datenverwaltung“ auf Diskrepanzen als auch auf Vorbehalte.

Auch im folgenden zweiten Modell werden diese Komplexität sowie die Notwendigkeit von Fachpersonal für die Umsetzung des Forschungsdatenmanagements deutlich.

1.2.2 Data Curation Continuum

Treloar und Harboe-Ree (2008) veranschaulichten in ihrem Model „*Data Curation Continuum*“, das an der *Monash University* in Australien entwickelt wurde, die unterschiedlichen Phasen im Lebenszyklus von Forschungsdaten. Der Forschungsprozess wurde in diesem Modell in drei Domänen unterteilt. Es illustriert, dass jeder Bereich unterschiedliche, teilweise gegensätzliche Ansprüche besitzt. Teils werden sogar verschiedene Technologien erforderlich.

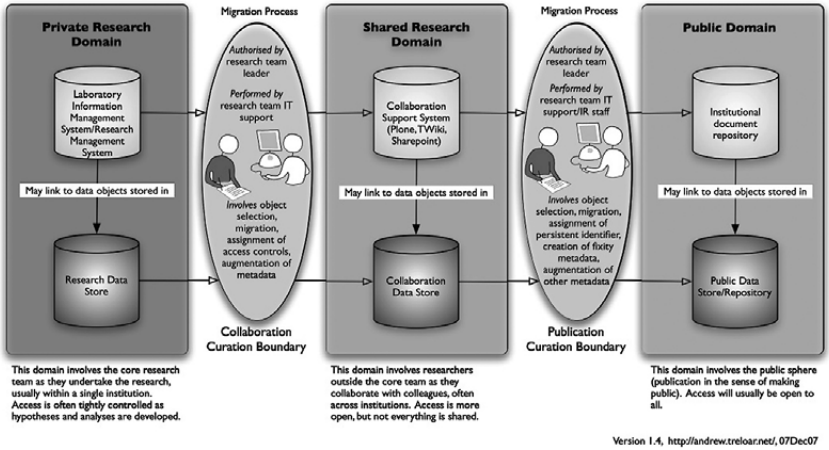

Abb. 3: *Data Curation Continuum* (Treloar & Harboe-Ree, 2008, S. 6)

Daten werden in einer Forschergruppe, der sog. „*Private Research Domain*“ erzeugt. Für die Arbeit in dieser Phase werden „*research management*“ Systeme benötigt, die einen Überblick über die gesamten Datenbestände geben. Ab der Entstehungsphase müssen Sie mit einem hohen Zugriffsschutz und Metadaten versehen werden. Metadaten ergeben sich einerseits durch die gerätespezifische Generierung, andererseits vergibt der Wissenschaftler zusätzlich Metadaten, um eine persönliche Verwaltung zu erhalten (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 32). Die Datenspeicherung erfolgt in einem „*Research Data Store*“. Sind die Forscher bereit, Teilergebnisse ihrer Forschung anderen für Analysen zugänglich zu machen, erfolgt eine Migration in die sog. „*Shared Research Domain*“. Dies ergibt sich beispielsweise, wenn dem Vorgesetzen oder kooperierenden Wissenschaftlern die bisherige Forschung präsentiert wird (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 32). Für diesen Austausch, auch bezeichnet als „*Data Sharing*“, werden Systeme notwendig, die eine kollektive Arbeitsweise unterstützten, wie Plone oder TWiki. Die Datenobjekte selbst befinden sich in Repositorien. Somit kann eine starke Strukturierung der Datensammlungen erfolgen und ausgefeilte Zugriffsrechte formuliert werden. Mit dem Abschluss der Forschungstätigkeiten erfolgt die Migration zur sog. „*Public Domain*“. Die fertigen Forschungsergebnisse (beispielsweise Publikationen) werden in die institutionellen Repositorien migriert, ein bekannter Prozess. Zusätzlich muss eine Verlinkung auf die mit *Digital Object Identifiers* (DOI) und Metadaten versehene Datenobjekte erfolgen, die sich in einem öffentlichen Forschungsdaten-*Repository* befinden. (Treloar & Harboe-Ree, 2008, S. 5–7)

Elementar ist die Anreicherung der Daten mit vollständigen Metadaten, damit eine Recherche, Identifizierung und Wiederverwendbarkeit eindeutig gewährleistet wird (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 33).

Es ist möglich, die Gesamtheit des Forschungsprozesses durch die Nutzung eines *Repository* umzusetzen. Gegenwärtig gibt es jedoch unterschiedliche Anforderungen in den verschiedenen wissenschaftlichen Domänen, wodurch die technische Umsetzung in der Praxis komplex ist (Treloar & Harboe-Ree, 2008, S. 7). Bisher werden grundsätzlich institutionelle Repositorien verwendet, die für die Verwaltung und den Nachweis von Dokumenten konzipiert wurden. Diese ebenfalls für das Forschungsdatenmanagements zu verwenden, ist wegen ihrer Inflexibilität schwierig. Für Speicherung und *Curation* von Forschungsdaten muss eine Lösung existieren, die beispielsweise eine variable Vergabe von Metadaten erlaubt, wie es von der jeweiligen Disziplin gefordert wird, um zukünftig die Daten auch nachnutzen zu können. Die meisten Softwarelösungen für Repositorien (OPUS, *eprints* etc.) unterstützen dies noch nicht. Fedora macht hier eine Ausnahme. Mit dem *Open Source Projekt **F**lexible **E**xtensible **D**igital **O**bject **R**epository **A**rchitecture Commons* (FEDORA), entwickelt an der *Cornell University*, steht ein fertiges System zur Verfügung, das beliebige digitale Objekte (Daten, Textdateien, Metadaten, Bilder, Videos, Webseiten etc.) verwalten kann. (siehe Beitrag von Aschenbrenner & Neuroth, Kap. 2.5)

Neben institutionellen und disziplinären Repositorien werden Forschungsdatenspeicher notwendig, die allesamt jedoch miteinander verknüpft arbeiten sollten.

1.2.3 Fazit

Die Darlegung des Lebenszyklus von Forschungsdaten anhand der beiden Modelle verdeutlicht einerseits die Komplexität andererseits aber auch die theoretische Machbarkeit der Speicherung von Forschungsdaten. Wie bereits geschildert, existieren Einrichtungen, in denen das Forschungsdatenmanagement erfolgreich praktiziert wird. Beispielsweise wird Pangaea vom Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) gemeinsam mit dem Zentrum für Marine Umweltwissenschaften (MARUM) gehostet.

> „The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. The system guarantees long-term availability of its content through a commitment of the operating institutions.“ (AWI & Center for Marine Environmental Sciences)

Die Diskussionen zu „Forschungsdaten“ und insbesondere deren Management wird weiter dadurch erschwert, da es eine hohe Domänenspezifik gibt, die eine Übertragung auf andere Wissenschaftsdisziplinen nicht per se zulässt.

> „Vielmehr muss es Ziel sein, in Zusammenarbeit mit den Fachgesellschaften disziplinspezifische Ansätze zu entwickeln, die dann prototypisch realisiert werden können. Dabei wird es Disziplinen geben, die wie die Geowissenschaften eine zentrale Datenzentrenstruktur benötigen, aber auch Disziplinen, die unter Verwendung allgemeingültiger Standards individuelle Lösungen in Form von verteilten Repositorien betreiben.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 14)

Neben der Entwicklung der notwendigen Techniken und Systeme zur Datenspeicherung gibt es noch einige grundsätzliche Fragen, deren Beantwortung derzeitig noch nicht erfolgte. Sietmann zählt dazu die folgenden:

> „So wirft die Transformation der gesamten Prozesskette von der Erzeugung über die Speicherung bis zur Bewahrung und Pflege von Forschungsdaten, die sich in dem Begriff „*Open Data*“ verdichtet, Fragen über Fragen auf. Wer standardisiert die Metadaten? Wer setzt die „*Data Policies*“? Wie erzeugt man die Anreize, dass Forscher ihre Daten und Programme verfügbar machen? Wer trägt die Aufwendungen, dass sie verfügbar bleiben?“ (Sietmann, 2009, S. 160)

Sicherlich könnten durch den Austausch von Erfahrungen, die Einen von den Anderen lernen. Ein Problem, das alle Disziplinen und Beteiligte betrifft, ist die Frage nach der personellen Umsetzung des Forschungsdatenmanagements. Wissenschaftler werden i.d.R. die Daten nicht verwalten. Wer dabei welche Unterstützung gibt, bzw. ob eine mehr oder weniger vollständige Ausgliederung dieser Arbeit möglich ist, bleibt Gegenstand der Diskussion.

Literaturhinweise

AWI (Alfred Wegener Institute for Polar and Marine Research) & Center for Marine Environmental Sciences. *PANGAEA.* Data Publisher for Earth & Environmental Science. Online: http://www.pangaea.de/about/ [Zugriff am 17.07.2011].

DCC (Digital Curation Centre), 2010. *DCC Curation Lifecycle Model*. Online: http://www.dcc.ac.uk/resources/curation-lifecycle-model [Zugriff am 01.06.2011].

Gold, A., 2007. Cyberinfrastructure, Data, and Libraries, Part I. A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine*, 13(9/10). Online: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html [Zugriff am 25.04.2011].

Helly, J. Staudigel, H. & Koppers, A, 2003. ‚Scalable models of data sharing in Earth sciences'. *Geochemistry, geophysics, geosystems.* G 3, an electronic journal of the earth sciences, 4(1), S. 1–14. Online: http:// www.beamreach.org/research/data_sharing_model_GC2002.pdf [Zugriff am 17.07.2011].

Klump, J. et. al., 2006. Data Publication in the Open Access Initiative. *Data Science Journal*, 5(15 June 2006), S. 79–83. Online: http:// www.mad.zmaw.de/fileadmin/extern/Publications/datapublication.pdf [Zugriff am 17.07.2011].

NESTOR – Kompetenznetzwerk Langzeitarchivierung, 2009. *Digitale Forschungsdaten bewahren und nutzen – für die Wissenschaft und für die Zukunft*. NESTOR Arbeitsgruppe Grid /e-sience und Langzeitarchivierung. (NESTOR-Bericht) Online: http://nbn-resolving.de/nbn:de:0008-2009071031. [Zugriff am 17.07.2011].

Severiens, T. & Hilf, E.R., 2006. *Langzeitarchivierung von Rohdaten – Studie zum Stand vorhandener Forschungsdaten und Rohdaten aus wissenschaftlichen Tätigkeiten: Erfordernisse und Eignung zur Archivierung bzw. Zurverfügungstellung in Deutschland (Primärdaten)*. Online: http:// nbn-resolving.de/urn:nbn:de:0008-20051114018.

Sietmann, R., 2009. Rip. Mix. Publish. Der Wissenschaft steht ein radikaler Wandel im Umgang mit Forschungsdaten bevor. *c't*, (14), S. 154–161.

TIB (Technische Informationsbibliothek) Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010. *Konzeptstudie „Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie“*. Online:

http://www.tib-hannover.de/fileadmin/projekte/primaer-chemie/Konzeptstudie_Forschungsdaten_Chemie.pdf [Zugriff am 25.04. 2011].

Treloar, A. & Harboe-Ree, C., 2008. *Data management and the curation continuum. How the Monash experience is informing repository relationship*. Online: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf [Zugriff am 15.05.2011].

Winkler-Nees, S., 2010. *Der Umgang mit Forschungsdaten in Wissenschaft und Lehre*. Bad Honnef. Online: http://www.dfg.de/download/pdf/dfg_magazin/wissenschaftliche_karriere/heisenberg_treffen_2010/forschungsdaten.pdf [Zugriff am 01.06.2011].

1.3 Datenmanagement und Data Sharing: Erfahrungen in den Sozial- und Wirtschaftswissenschaften

Denis Huschka [1], Claudia Oellers [1], Notburga Ott [1, 2]
Gert G. Wagner [1, 3, 4]
[1] Rat für Sozial- und Wirtschaftsdaten
[2] Ruhr-Universität Bochum
[3] Deutsches Institut für Wirtschaftsforschung (DIW Berlin)
[4] Technische Universität Berlin

Die Menge der für Forschungszwecke zur Verfügung stehenden Daten vergrößert sich beständig (King, 2011). Jedoch werden unter 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 Phänotypen von Forschungsdaten erfordern spezifische Datenmanagementstrategien. Oft beschreiben die Daten die Ausprägung einer Eigenschaft eines Individuums oder einer Organisation, wie z. B. 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 & 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 (z. B. dem GESIS Datenarchiv für Sozialwissenschaften – vormals Zentralarchiv für empirische Sozial-

forschung an der Universität Köln) sind alle vom Rat für Sozial- und Wirtschaftsdaten akkreditierten Forschungsdatenzentren (FDZ) und Datenservicezentren (DSZ) Teil dieser Forschungsinfrastruktur. Die FDZ und DSZ 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 sharings* 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 (z. B. 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 die Beschreibung der Inhalte, Qualität, Analysepotenziale, Aussagekraft und insbesondere über Verknüpfungsmöglichkeiten zwischen Datensätzen. Es reicht also nicht jeden einzelnen Datensatz verfügbar zu machen. Für eine breite Nutzung in der Wissenschaft ist ein „Informationsportal“ notwendig, in welchem ein an einem bestimmten Thema interessierter Forscher alle erforderlichen Informationen über 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.

1.3.1 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, da die mehrfache Nutzung der Daten diese nicht zerstört (wie das z. B. 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 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 Rat für Sozial- und Wirtschaftsdaten (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 Rückkopplungs-

prozesse, 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 und 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 i. d. R. nicht im Literaturverzeichnis von Veröffentlichungen zitiert und entsprechend erntet 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 sharings* 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 v.a. 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 Wirtschaftsforschung (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, das allen Datenschutz- 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 *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.

1.3.2 Data Access

Sozial-, verhaltens- und wirtschaftswissenschaftliche Daten weisen oft Charakteristika auf, die 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[1] 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, die durch angepasste technische und infrastrukturelle Lösungen, d. h. durch ein intelligentes Datenmanagement, überwunden werden können. 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 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

1. http://www.forschungsdatenzentrum.de/campus-file.asp [Zugriff am 10.08.2011].

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.

1.3.3 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 Dokumenta-

tionen, den sog. 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 (vgl. Kap. 3.1). 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 u. U. unwiederbringlich zu vernichten. Die Langzeitarchivierung ist eine in seiner 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. In 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.

1.3.4 Modell II: Zugang zu den Daten und Serviceleistungen durch Forschungsdatenzentren[2]

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 (z. B. statistische Ämter) und/oder besonders komplizierte Datensatzstrukturen anbieten (z. B. prospektive Längsschnitterhebungen) und deshalb eine enge Verbindung zwischen Datenproduzent und Datennutzer wünschenswert ist. Auch die Einhaltung des Datenschutzes kann in „eigenen“ FDZ 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

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

bezüglich der zum großen Teil personenbeziehbaren Daten nicht ohne weiteres möglich. In der Zwischenzeit sind neben den genannten sechs Datenzentren eine ganze Reihe weiterer Forschungsdatenzentren hinzugekommen, die über den Rat für Sozial- und Wirtschaftsdaten akkreditiert und organisiert werden (http://www.ratswd.de/dat/fdz.php). Derzeit (Stand Sommer 2011) gibt es 19 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 war, 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 FDZ betreibenden Datenproduzenten in der Regel die besten Experten im Umgang mit den eigenen Daten sind. Ein zentraler Aspekt der Akkreditierungsrichtlinien des RatSWD für FDZ und DSZ 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 u.a. in der Festlegung gemeinsamer Kriterien und Standards als Antwort auf gemeinsame rechtliche und organisatorische Voraussetzungen, welche das Modell Datenarchiv ausschließen, ihren Ausdruck. Auch die Weiterentwicklung von Verfahren des *onsite* 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 Messmechanis-

men, 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 v.a. 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 – u. U. 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.

Herausforderungen: Im Feld der FDZ 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.

1.3.5 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 sollte, 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, wird ihre Qualität und Verfügbarkeit steigen.

Literaturhinweise

Bender, S. Himmelreicher, R. Zühlke, S. & Zwick, M., 2009. *Improvement of Access to Data from the Official Statistics.* (RatSWD Working Paper Nr. 118) Online: http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_118.pdf [Zugriff am 09.08.2011].

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

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

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

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

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

Habich, R. Himmelreicher, R. K. & Huschka, D., 2010. *Zur Entwicklung der Dateninfrastruktur in Deutschlan*d. (RatSWD Working Paper Nr. 157) Online: http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_157.pdf [Zugriff am 09.08.2011].

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

King, G., 2011. Ensuring the Data Rich Future of the Social Sciences. *Science*, 331, S. 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.* Online: http://www.leibniz-gemeinschaft.de/?nid=infrastr [Zugriff am 09.08.2011].

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

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

Solga, H. & 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).* (RatSWD Working Paper Nr. 1) Online: http://www.ratswd.de/download/RatSWD_WP_2007/RatSWD_WP_01.pdf [Zugriff am 09.08.2011].

Weichselgartner, E., 2011. *Disziplinenspezifische Aspekte des Archivierens von Forschungsdaten am Beispiel der Psychologie*. (RatSWD Working Paper Nr. 179) Online: http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_179.pdf [Zugriff am 12.07.2011].

Winkler-Nees, S., 2011. *Anforderungen an wissenschaftliche Informationsinfrastrukturen.* (RatSWD Working Paper Nr. 180) Online: http://www.ratswd.de/download/RatSWD_WP_2011/RatSWD_WP_180.pdf [Zugriff am 09.08.2011].

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

2.1 „*Data Policies*“ im Spannungsfeld zwischen Empfehlung und Verpflichtung

Heinz Pampel [1], Roland Bertelmann [2]
[1] Helmholtz-Gemeinschaft, Helmholtz Open Access Projekt
[2] Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum GFZ, Bibliothek des Wissenschaftsparks Albert Einstein

Abstract

Unter Beachtung disziplinärer Anforderungen beginnen Akteure aus Wissenschaft, Wissenschaftsmanagement und Infrastruktureinrichtungen Aussagen zum Umgang mit Forschungsdaten zu tätigen. Je nach Akteur und Zielgruppe variieren diese Aussagen, die häufig unter dem Begriff Policy gefasst werden. Der Beitrag gibt einen Überblick über die Vielfalt der Policies und beschreibt die Herausforderungen bei der Umsetzung dieser empfehlenden oder verpflichtenden Aussagen.

2.1.1 Einführung

Die Diskussion um einen zeitgemäßen Umgang mit wissenschaftlichen Daten[1] hat in den letzten Jahren deutlich an Dynamik gewonnen. Fachgesellschaften, Förderorganisationen, wissenschaftliche Infrastruktureinrichtungen und Verlage beginnen sich den Herausforderungen rund um die dauerhafte Zugänglichkeit wissenschaftlicher Daten anzunehmen und Konsequenzen für das eigene Handeln zu formulieren, die häufig in so genannten *Data Policies* beschrieben werden.

Bei Forschungsdaten kann es sich z. B. um grafische, numerische oder auch textuelle Objekte handeln, die in ihren Ausprägungen je nach Disziplin variieren. Aufgrund der Heterogenität der Daten müssen Aussagen, die unter dem Terminus *Data Policy* gefasst werden, nach Fach und nach Akteur differenziert betrachtet werden.

Unter Berücksichtigung der übergreifenden Diskussion zum Umgang mit wissenschaftlichen Daten[2] sollen insbesondere Aussagen zur Zugänglichkeit der Daten betrachtet werden. Die Zugänglichkeit ist die zentrale Vorrausetzung für eine mögliche Nachprüfbarkeit und Nachnutzbarkeit der Daten. Während die Nachprüfbarkeit die inhaltliche und formale Qualitätssicherung der Daten

1. Im deutschen Sprachraum werden diese Daten, die im Englischen auch unter den Begriffen *Scientific Data* oder *Research Data* gefasst werden, häufig mit den Termini Forschungsdaten und wissenschaftliche Daten beschrieben.
2. Siehe dazu Abschnitt 2. Interdisziplinäre *Policies*.

fokussiert, kann im Rahmen der Nachnutzung eine Weiterbearbeitung der Daten, auch in anderen Kontexten, vorgenommen werden.

In der folgenden Betrachtung wird zwischen interdisziplinären, disziplinären und institutionellen Positionen unterschieden. Darüber hinaus wird auf die Positionen von Zeitschriften im Rahmen ihrer *Editorial-Policies* eingegangen.

2.1.2 Interdisziplinäre *Policies*

Vor dem Hintergrund eines international beachteten Falls wissenschaftlichen Fehlverhaltens in Deutschland[3] verabschiedete das Präsidium der Deutschen Forschungsgemeinschaft (DFG) 1997 „Vorschläge zur Sicherung guter wissenschaftlicher Praxis“. Die Vorschläge wurden auf Basis bestehender Regelungen im Ausland formuliert und sind an wissenschaftliche Institutionen und deren Mitarbeiter adressiert. Ihrem Selbstverständnis nach definieren sich die Vorschläge nicht als „detailliertes Regelsystem“. Anliegen des Papiers ist es wissenschaftlichen Einrichtungen „einen Rahmen für eigene Überlegungen“ zu geben (DFG, 1998, S. 6). In den Vorschlägen werden sechzehn Empfehlungen beschrieben. In der Empfehlung 7 „Datenhaltung“ wird der Umgang mit wissenschaftlichen Daten aufgegriffen: „Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden.“ (DFG, 1998, S. 12f). Diese Empfehlung betont die Notwendigkeit einer Nachprüfbarkeit der Daten. Sie trifft keine Aussage zur Zugänglichkeit und einer darauf ggf. folgenden Nachnutzung (Klump, 2010).

Diese Empfehlungen müssen bei der Inanspruchnahme von Mitteln der DFG eingehalten werden. Darüber hinaus forderte die DFG Mittelempfänger seit 1998 auf, an ihrer Einrichtung entsprechend den Empfehlungen eigene Regeln zur Sicherung einer guten wissenschaftlichen Praxis zu etablieren.[4]

Auch in anderen Ländern wurden ähnliche Regelungen verabschiedet, die sich nicht auf eine wissenschaftliche Disziplin oder eine Institution beschränken. In Dänemark wurde bereits 1992 mit dem *Danish Committee on Scientific Dishonesty* (DCSD) ein nationales Gremium zur Sicherung von Standards einer guten wissenschaftlichen Praxis geschaffen. Auch in den dänischen Regelungen wird der Umgang mit wissenschaftlichen Daten berücksichtigt. In der aktuellsten Version aus dem Jahre 2009 wird die Zugänglichkeit wissenschaftlicher Daten thematisiert: „Upon publishing of the results from a project, the institution ought to make data available to any scientist with relevant interest in and

3. Siehe dazu die Pressemitteilung der „*Task Force F. H.*“ (DFG, 2000).
4. Siehe dazu Abschnitt 4. Institutionelle *Policies*.

assumptions for using them conditional upon the approval of the authorities (e.g. *The Danish Data Protection Agency*).“ (DCSD, 2009).

Mit den technologischen Entwicklungen rund um das Internet eröffnen sich den Disziplinen neue Möglichkeiten im Umgang mit Information und Wissen, die unter dem von Jim Gray beschriebenen „*fourth paradigm*“ der wissenschaftlichen Arbeit gefasst werden können (Hey et al. 2009). Vor diesem Hintergrund unterzeichneten führende Wissenschaftsorganisationen 2003 die „Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen“. Dieses Dokument bildet einen Grundpfeiler vielfältiger Aktivitäten, die unter dem Begriff *Open Access* gefasst werden. Die Unterzeichner erklären ihr Bestreben neben klassischen Textpublikationen auch „raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material“ im Internet frei zugänglich und nachnutzbar zu machen (Berliner Erklärung, 2003).

Für große Aufmerksamkeit sorgten die „*Principles and Guidelines for Access to Research Data from Public Funding*“, die die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) 2007 veröffentlichte. Ziel des Papiers, das eine Steigerung des gesellschaftlichen Nutzens durch frei zugängliche Forschungsdaten fordert, ist es u.a. eine „culture of openness and sharing of research data among the public research communities“ (OECD, 2007, S. 11) in den Mitgliedstaaten der OECD zu fördern.

Vor dem Hintergrund einer breiten Diskussion in den wissenschaftlichen Disziplinen[5] verankerten die *European Science Foundation* (ESF) und die *European Heads of Research Councils* (EUROHORCs) die Forderung nach einem offenen Zugang zu qualitätsgesicherten Forschungsdaten in ihrer gemeinsamen Vision des europäischen Forschungsraums (ESF & EUROHORCs, 2008) und der darauf aufbauenden Strategie (ESF & EUROHORCs, 2009).

In Deutschland wurde die Diskussion 2008 im Rahmen der Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen aufgegriffen und 2010 in „Grundsätze zum Umgang mit Forschungsdaten“ gebündelt (Allianz 2008, 2010). Diese Grundsätze tragen der disziplinspezifischen Abhängigkeit der Arbeit mit Forschungsdaten Rechnung und formulieren übergreifende und elementare Aspekte eines zeitgemäßen Umgangs mit wissenschaftlichen Daten. Neben Aussagen zu rechtlichen Rahmenbedingungen wird beispielsweise die Notwendigkeit einer professionellen Anerkennung des „*data sharings*“ thematisiert. Dabei unterstützen die Partnerorganisationen der Allianz „die langfristige Sicherung und den grundsätzlich offenen Zugang zu Daten aus öffentlich geförderter Forschung“ (Allianz, 2010).

5. Siehe dazu z. B. die Schwerpunktausgaben „*Big Data*“ und „*Sharing Data*“ der Nature (2008, 2009a) sowie die Special *Online Collection* „Dealing with Data“ der Science (2010).

Die beschriebenen Papiere geben den Rahmen für empfehlende oder verpflichtende Aussagen, die auf Disziplinen und Institutionen wirken und sich teilweise in den *Editorial Policies* wissenschaftlicher Zeitschriften widerspiegeln.

2.1.3 Disziplinäre *Policies*

Über die rahmengebenden interdisziplinären Positionspapiere hinaus gibt es insbesondere in den Geo-, Lebens-, und Sozialwissenschaften vielfältige *Policies* bezüglich des Umgangs mit wissenschaftlichen Daten. Solche disziplinären Spezifikationen sind nötig, da wissenschaftliche Daten heterogen sind und ihr Umgang durch die fachlichen Wissenschaftskulturen geprägt ist.

In den Lebenswissenschaften wirken insbesondere die „Gute klinische Praxis" und die „Grundsätze der Guten Laborpraxis" auf den Umgang mit Daten. Beide Grundsätze, die auf Ebene der OECD formalisiert wurden, sind in Deutschland gesetzlich verankert.[6]

Ein international beachtetes Beispiel für eine Forschungsdaten-*Policy* sind die *Bermuda Principles*, die 1996 im Rahmen des *Human Genome Project* formuliert wurden. In ihnen heißt es: „All human genomic sequence data generated by centers funded for large-scale human sequencing should be freely available and in the public domain to encourage research and development and to maximize the benefit to society." (Smith & Carrano, 1996). Mit den *Bermuda Principles* hat sich eine wissenschaftliche Community in Abstimmung mit Förderorganisationen selbstverpflichtende Regelungen geschaffen. Die Diskussion über den Umgang mit biologischen Daten hält bis heute an. Die lebhafte Debatte in der Community macht deutlich, dass eine *Policy* kein statisches Dokument ist, sondern weiterentwickelt werden sollte und begleitender Maßnahmen bedarf.[7] Weiter machen die *Bermuda Principles* die Bedeutung disziplinärer *Policies* deutlich: Eine Veröffentlichung wissenschaftlicher Daten vor der eigentlichen Interpretation im Rahmen einer Textpublikation, wie in den *Bermuda Principles* vorgesehen ist, ist in vielen Disziplinen undenkbar. So gehen die Sozialwissenschaften einen anderen Weg. In der „*Data Management Policy*" des *International Social Science Council* (ISSC) aus dem Jahr 1994 wird die Bedeutung einer zeitverzögerten Zugänglichkeit der Daten betont: „While the rights of scholars to publish from their data must be protected, a guideline period of one year from data collection to availability will in most cases give such protection

6. Die „Gute klinische Praxis" ist im Arzneimittelgesetz verankert, die „Grundsätze der Guten Laborpraxis" im Chemikaliengesetz.

7. Siehe dazu z. B. das 2009 verabschiedete „*Toronto Statement*" (2009) und die Diskussion um die „*Rome Agenda*" (Schofield et al., 2009).

and good cause should be given for embargo beyond one year; a shorter period of exclusive rights would be commendable.“ (Ferris, o. J.)

So muss, ausgehend von der disziplinübergreifenden Notwendigkeit der guten wissenschaftlichen Praxis, die Forderung nach der offenen Zugänglichkeit der Daten an die Bedürfnisse der jeweiligen Disziplinen angepasst werden.

Die lange Tradition der Sozialwissenschaften im „*sharing*“ von Daten wird durch ihre Infrastruktureinrichtungen gestützt; so gibt es beispielsweise im Rahmen der Dachorganisation der sozialwissenschaftlichen Datenarchive, dem *Council of European Social Science Data Archives* (CESSDA) ein Übereinkommen zur Zusammenarbeit der europäischen Datenarchive (CESSDA, 2011). Das Beispiel zeigt, dass *Policies* nicht nur auf Wissenschaftler, sondern auch auf disziplinäre Infrastruktureinrichtungen wirken.

Eine besondere Herausforderung stellt der nachhaltige Umgang mit Forschungsdaten in wissenschaftlichen Großprojekten dar. Je mehr Personen und Institutionen an einem Projekt beteiligt sind, desto notwendiger ist es, sich auf übergreifende Standards im Umgang mit erhobenen Daten zu einigen. Als Beispiel mag hier das „Internationales Polarjahr 2007–2008“ dienen. In dem Großprojekt wurde eine *Data Policy* verabschiedet, die für die beteiligten Partner bindend ist. Diese baut auf übergreifenden Positionspapieren des *International Council for Science* (ICSU) und der *World Meteorological Organisation* (WMO) auf. In der Einführung wird der Fokus der *Policy* wie folgt beschrieben: „This policy aims to provide a framework for these data to be handled in a consistent manner, and to strike a balance between the rights of investigators, the rights of indigenous peoples, and the need for widespread access through the free and unrestricted sharing and exchange of both data and metadata.“ Das Papier trifft, ausgehend von einer Definition der betroffenen Daten, Aussagen zu folgenden Themen: Zugänglichkeit und Austausch sowie Publikation und Erhaltung der im Projekt erhobenen Daten. (IPY, 2008).

Bemerkenswert ist, dass z. B. in den Geowissenschaften die Notwendigkeit von Regelungen zum Umgang mit Forschungsdaten bei der strategischen Ausrichtung des Faches berücksichtigt wird. So wird das Thema in der 2010 von der Senatskommission für Geowissenschaftliche Gemeinschaftsforschung der DFG veröffentlichten Strategieschrift „Dynamische Erde – Zukunftsaufgaben der Geowissenschaften“ aufgegriffen. Dort heißt es u. a.: „Förderrichtlinien sollten in Zukunft nicht nur zu einem Abschlussbericht verpflichten, sondern auch dazu, die gewonnenen Daten zu publizieren. Bei den Fördereinrichtungen und den Datenzentren müssen entsprechende Kontroll- und Abnahmemechanismen etabliert werden. Dazu gehören auch Mechanismen, um datenproduzierende Projekte zu managen.“ (Wefer et al., 2010, S. 327).

Während die beschriebenen Beispiele zeigen, dass sich einige Disziplinen des Themas annehmen, sind andere Fachgebiete noch sehr zurückhaltend.

2.1.4 Institutionelle *Policies*

Zur erfolgreichen Umsetzung von disziplinären *Policies* nehmen sich neben Förderorganisationen auch wissenschaftliche Institutionen des Themenfeldes an.

In Deutschland müssen wissenschaftliche Einrichtungen, die DFG-Mittel in Anspruch nehmen, entsprechend den „Vorschlägen zur Sicherung guter wissenschaftlicher Praxis“ eigene Regeln etablieren, die auch die bereits genannte Empfehlung 7 „Datenhaltung“ berücksichtigen.[8] Der Wortlaut dieser Empfehlung wurde beispielsweise vom Senat der Max-Planck-Gesellschaft (MPG) übernommen und erweitert. So heißt es in den Regeln der MPG: „Es muss entweder vom Institut oder zentral sichergestellt werden, dass Daten zumindest für diesen Zeitraum lesbar verfügbar bleiben. Für berechtigte Interessenten muss der Zugang zu den Daten gewährleistet sein.“ (MPG, 2009). An einigen Institutionen wird darüber hinaus auch die Erhaltung der Gerätschaften, die zur Erhebung der Daten verwendet werden, angeregt, so z. B. an der Universität Siegen (2001): „Wann immer möglich, sollen Präparate und Geräte, mit denen Primärdaten erzielt wurden, für denselben Zeitraum aufbewahrt werden.“

Auch in anderen Ländern sind solche Regeln verankert. Nach dem „*Australian Code for the Responsible Conduct of Research*“ muss jede australische Wissenschaftseinrichtung eine *Policy* zum Umgang mit wissenschaftlichen Daten verankern (NHMRC, 2007). So findet sich in den „*Research Data Management Guidelines*“ der australischen *Charles Darwin University* (2010) der folgende Absatz: „In general, research data must be retained for a minimum of 5 years from the date of any publication which is based upon that data or the end of the project, if the research has not been published. Retaining data for 15 years or more may be necessary for clinical trials. When considering the length of time that data should be retained, disciplinary practice, professional standards, contractual obligations and relevant legislation, codes and guidelines must also be taken into account.“ (CDU, 2010).

Über die Verankerung an wissenschaftlichen Institutionen hinaus, kommt der Einführung von Empfehlungen und Verpflichtungen auf Seiten der Forschungsförderer eine bedeutende Rolle zu. Über ihre Verwendungsrichtlinien haben diese die Möglichkeit den Umgang mit wissenschaftlichen Erkenntnissen zu steuern.

2003 veröffentlichte das *National Institute of Health* (NIH) eine „*NIH Data Sharing Policy*“. Nach dieser sind Antragsteller, die eine Zuwendung ab 500.000 US-Dollar beantragen, aufgefordert Aussagen zum „*data sharing*“ zu tätigen (NIH, 2003). Andere Förderer in den Lebenswissenschaften folgen diesem Schritt. So verabschiedete der *Wellcome Trust* (2010) 2007 eine „*Policy on*

8. Siehe dazu Abschnitt 2. Interdisziplinäre *Policies*.

data management and sharing“. 2011 äußeren sich mehrere medizinische Förderorganisationen auf Initiative des *Wellcome Trust* in einem gemeinsamen Statement unter dem Titel „*Sharing research data to improve public health*“ zum Thema. Darin kündigen die Förderer an, ihre zukünftigen Aktivitäten rund um die Zugänglichkeit wissenschaftlicher Daten zu intensivieren (Wellcome Trust, 2011).

Auch die DFG hat sich hier eingereiht. Bereits 2009 veröffentlichte der Unterausschuss Informationsmanagement des Ausschusses für Wissenschaftliche Bibliotheken und Informationssysteme der DFG „Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten“. In diesen wird Mittelempfängern empfohlen, Forschungsdaten „nach Möglichkeit überregional und frei zur Verfügung“ zu stellen (DFG, 2009). 2010 verankerte die DFG das Thema in ihrem „Leitfaden für Antragsteller“. In diesem werden Antragsteller aufgefordert, Maßnahmen zu beschreiben, die „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.“ (DFG, 2010).

Die US *National Science Foundation* (NSF) ist in ihrer Politik noch etwas konkreter und verlangt seit 2011 einen so genannten Daten Management Plan, in der ein Antragsteller beschreiben muss, welche Maßnahmen er zur Umsetzung der „*Data Sharing Policy*“ der NSF trifft (NSF, 2011).

2.1.5 *Journal Policies*

Neben den *Policies* von Forschungsförderern kommt den Richtlinien von Zeitschriften eine besondere Bedeutung zu. Der Zugang zu Daten, die Grundlage einer Publikation sind, ist zum einen im Rahmen der inhaltlichen Qualitätssicherung durch Peer-Review-Verfahren vonnöten, zum anderen fördern Herausgebergremien und Verlage verstärkt die offene Zugänglichkeit von wissenschaftlichen Daten. Je nach disziplinärem Fokus und tradiertem Umgang mit den Daten variieren diese *Policies*.

In der *Editorial-Policy* der Nature-Zeitschriftenfamilie heißt es beispielsweise „a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to readers without preconditions.“ Weiter werden, neben einer Reihe von Spezifizierungen, Hinweise für fachspezifische Besonderheiten gegeben (Nature, 2009b).

Für die Publikationen der *American Geophysical Union* (AGU) gilt eine explizite „*Policy on Referencing Data in and Archiving Data for AGU Publications*“. In dieser werden beispielsweise konkrete Anforderungen an ein „*Data Archive*“ und an die Zitierung von Forschungsdaten beschrieben (AGU, 1996).

Ähnlich konkrete Aussagen zur Veröffentlichung von Daten, die Grundlage einer Textpublikation sind, treffen einige *Open-Access*-Zeitschriften. Im Rah-

men ihres Selbstverständnisses einer offenen Wissenschaftskommunikation fördern diese Zeitschriften häufig eine mögliche Nachnutzung der Daten. So heißt es beim *Open-Access-Flagship* der *Public Library of Science* (PLoS) PLoS ONE unter dem Abschnitt „*Sharing of Materials, Methods, and Data*" in der *Editorial-Policy*: „PLoS is committed to ensuring the availability of data and materials that underpin any articles published in PLoS journals." Weiter werden Hinweise auf geeignete Forschungsdaten-Repositorien gegeben. (PloS ONE, o. J.)

In einigen Fachgebieten der Lebenswissenschaften ist diese Forderung bereits Praxis. So heißt es in der *Policy* der Zeitschrift *Cell* „One of the terms and conditions of publishing in *Cell* is that authors be willing to distribute any materials and protocols used in the published experiments to qualified researchers for their own use." Beispielsweise müssen Nukleotid- und Proteinsequenzen in geeigneten Datenbanken, wie z. B. der *Worldwide Protein Data Bank*, ab dem Zeitpunkt der Veröffentlichung ohne Restriktionen zugänglich sein und durch die Angabe der „*accession number*" der jeweiligen Datenbank identifizierbar sein (Cell, 2011).

Festgehalten werden muss, dass die Aussagen zum Umgang mit Forschungsdaten im Rahmen von *Editorial-Policies* ein komplexes Thema sind. In Abhängigkeit der Disziplinen sind die Herangehensweisen an dieses Thema vielfältig. Mit Blick auf die zentrale Rolle der Zeitschriften kommt der Weiterentwicklung von *Editorial-Policies* zum Umgang mit wissenschaftlichen Daten eine zentrale Rolle zu.

2.1.6 Fazit

Betrachtet man die Vielfalt der empfehlenden und verpflichtenden Aussagen zum Umgang mit wissenschaftlichen Daten, so wird deutlich, dass die Forderung nach Nachprüfbarkeit überwiegt. In Deutschland ist diese durch *Policies* zur guten wissenschaftlichen Praxis verankert (DFG, 1998).

Ausgehend von der wissenschaftspolitischen Forderung der OECD (2007) nach einem freien Zugang zu Forschungsdaten gewinnt darüber hinaus die Forderung nach der Nachnutzung der Daten an Bedeutung. Diese Forderung hat mehrheitlich noch einen empfehlenden Charakter.

Die Verknüpfung der beiden Ansätze, Nachprüfbarkeit zur guten wissenschaftlichen Praxis und Nachnutzung im Kontext von *Open Access*, wird zukünftig wohl zunehmen.

Insbesondere in den Geo-, Lebens- und Sozialwissenschaften mangelt es nicht an *Policies*. Doch die Praxis zeigt, dass es mit der Verabschiedung einer *Policy* nicht getan ist. Untersuchungen zur Bereitschaft von Publizierenden in PLoS-Zeitschriften, die alle über *Policies* zum „*sharing*" verfügen, zeigen, dass auch in den Lebenswissenschaften noch keine Kultur des „*data sharing*" etabliert ist

(Savage & Vickers, 2009). Auch im Rahmen der Geowissenschaften zeigen sich am Beispiel des „Internationalen Polarjahr 2007–2008“, trotz einer verpflichtenden *Policy* und einem „*Data and Information Service*“, die vielfältigen Herausforderungen rund um einen zeitgemäßen Umgang mit den erhobenen Daten (Carlson, 2011).

Auch die adäquate Umsetzung der institutionellen Regeln zur guten Wissenschaft rund um die Empfehlung 7 „Datenhaltung“ der „Vorschläge zur Sicherung guter wissenschaftlicher Praxis“ (DFG, 1998) scheint im wissenschaftlichen Alltag nicht umfassend realisiert zu werden.

Die in den *Policies* beschriebenen Empfehlungen und Verpflichtungen bedürfen einer umfassenden Strategie, die die in den „Grundsätzen zum Umgang mit Forschungsdaten“ der Allianz der deutschen Wissenschaftsorganisationen beschriebenen Themenfelder aufgreift (Allianz, 2010). Dabei kommt der Entwicklung von Anreizmechanismen und unterstützenden Infrastrukturen eine zentrale Rolle zu. Zu oft ist die Aufbereitung von Daten in eine nachprüfbare oder nachnutzbare Form Nebentätigkeit, die, wenn sie denn geschieht, zwischen Abschlussbericht und neuem Projektantrag betrieben wird. Es fehlt in vielen Disziplinen an Kulturen des zeitgemäßen Umgangs mit wissenschaftlichen Daten und damit auch an einer Professionalisierung der entsprechenden Angebote. So müssen z. B. auch Nachwuchswissenschaftler im Rahmen der Ausbildung mit geeigneten Maßnahmen des Forschungsdatenmanagements vertraut gemacht werden.

In wissenschaftlichen Institutionen muss der Umgang mit Forschungsdaten effektiv organisiert und in die wissenschaftlichen Arbeitsabläufe integriert werden. Dies bedeutet auch, dass entsprechende Werkzeuge angeboten werden müssen.

Entsprechend dem von Treloar und Harboe-Ree (2008) beschriebenen „*Data Curation Continuum*“ (vgl. Kap. 1.2) variieren die Bedürfnisse an das Management der Daten je nach Station der Daten in ihrem Lebenszyklus. Während die Daten in vielen Disziplinen zu Beginn ihres Lebenszyklus in einer „privaten Domäne“ gespeichert und organisiert werden, findet die Nachprüfung oder Nachnutzung in einer „dauerhaften Domäne“ statt. Zentral ist die organisatorische und technische Gestaltung der Schnittstellen zwischen den „Domänen“.

Darüber hinaus bedarf es Anreizmechanismen, die die Etablierung von Kulturen des „*data sharings*“ fördern. So sollten mit der Verabschiedung einer *Policy*, ob empfehlend oder verpflichtend immer auch fördernde Maßnahmen verbunden werden.

Jede *Policy* bedarf einer unterstützenden Infrastruktur und führt damit auch zu organisatorischen, technologischen und ökonomischen Konsequenzen. Mit Blick auf die zukünftige Entwicklung eines zeitgemäßen Umgangs mit wissenschaftlichen Daten sollte berücksichtigt werden, dass die Herausforderung weniger in der Verabschiedung einer *Policy* als vielmehr in der Umsetzung einer *Policy* liegt.

Literaturhinweise

Allianz der deutschen Wissenschaftsorganisationen, 2008. *Schwerpunktinitiative „Digitale Information" der Allianz-Partnerorganisationen*. Online: http://www.allianzinitiative.de/fileadmin/user_upload/keyvisuals/atmos/pm_allianz_digitale_information_details_080612.pdf [Zugriff am 09.08.2011].

Allianz der deutschen Wissenschaftsorganisationen, 2010. *Grundsätze zum Umgang mit Forschungsdaten*. Online: http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze/ [Zugriff am 09.08.2011].

AGU (American Geophysical Union), 1996. *Policy on Referencing Data in and Archiving Data for AGU Publications*. Online: http://www.agu.org/pubs/authors/policies/data_policy.shtml [Zugriff am 09.08.2011].

Berliner Erklärung, 2003. *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. Online: http://oa.mpg.de/lang/de/berlin-prozess/berliner-erklarung/ [Zugriff am 09.08.2011].

Carlson, D., 2011. A lesson in sharing. *Nature*, 469(7330), S. 293. doi: 10.1038/469293a.

CDU (Charles Darwin University), 2010. *Research Data Management Guidelines*. Online: http://www.cdu.edu.au/governance/documents/rdmg3228v100jul2010.pdf [Zugriff am 09.08.2011].

Cell, 2011. *Information for Authors*. Online: http://www.cell.com/authors [Zugriff am 09.08.2011].

CESSDA (Council of European Social Science Data Archives), 2011. *Dissemination*. Online: http://www.cessda.org/sharing/dissemination/index.html [Zugriff am 09.08.2011].

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

DFG (Deutsche Forschungsgemeinschaft), 2000. *Task Force legt Abschlußbericht vor*. Online: http://www.dfg.de/service/presse/pressemitteilungen/2000/pressemitteilung_nr_26/index.html [Zugriff am 09.08.2011].

DFG (Deutsche Forschungsgemeinschaft), 2009. *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. Online:

http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf [Zugriff am 09.08.2011].

DFG (Deutsche Forschungsgemeinschaft), 2010. *Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für die Antragstellung.*(DFG-Vordruck 1.02 – 8/10) Online: http://www.dfg.de/download/programme/emmy_noether_programm/antragstellung/1_02/1_02.pdf [Zugriff am 09.08.2011].

DCSD (Danish Committees on Scientific Dishonesty), 2009. *Guidelines for Good Scientific Practice*. Online: http://en.fi.dk/publications/2009/the-danish-committees-on-scientific-guidelines-for-good-scientific-practice/UVVU%20GVP%20ENG%2015052009.pdf [Zugriff am 09.08.2011].

ESF & EUROHORCs, 2008. *The EUROHORCs & ESF Vision on a Globally Competitive ERA and their Road Map for Actions to Help Build it.* (Science Policy Briefing; 33, june 2008) Online: http://www.eurohorcs.org/SiteCollectionDocuments/EUROHORCs_ESF_ERA_RoadMap.pdf [Zugriff am 09.08.2011].

ESF & EUROHORCs, 2009. *EUROHORCs & ESF Vision on a Globally Competitive ERA and their Road Map for Actions*. (Corporate Publication, 22.07.2009) Online: http://www.esf.org/index.php?eID=tx_ccdamdl_file&p[file]=24823&p[dl]=1&p[pid]=4053&p[site]=European%20Science%20Foundation&p[t]=1314215639&hash=08d47b052bbcd4d383a2e14afde64161&l=en [Zugriff am 17.08.2011].

Hey, T. Tansley, S. & Tolle, K., 2009. Jim Gray on eScience. A transformed scientific method. In: S. Tansley & K. Tolle, ed. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond, Wash.: Microsoft Research, S. XVII–XXXI. Online: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf [Zugriff am 09.08.2011].

IPY (International Polar Year), 2008. *International Polar Year 2007–2008 Data Policy, International Polar Year Data and Information Service (IPYDIS).* Online: http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf [Zugriff am 09.08.2011].

Klump, J., 2010. Digitale Forschungsdaten. In: H. Neuroth et al., Hrsg. 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*, S. 17:104–17:115. urn:nbn:de:0008-2010071949.

MPG (Max-Planck-Gesellschaft), 2009. *Regeln zur Sicherung guter wissenschaftlicher Praxis*. Online unter: http://www.mpg.de/229457/Regeln_guter_wiss_Praxis__Volltext-Dokument_.pdf [Zugriff am 09.08.2011].

Nature, 2008. Big Data. *Nature*, 455(7209). Online: http://www.nature.com/news/specials/bigdata [Zugriff am 09.08.2011].

Nature, 2009a. Data Sharing. *Nature*, 461(7261). Online: http://www.nature.com/news/specials/datasharing [Zugriff am 09.08.2011].

Nature, 2009b. *Guide to Publication Policies of the Nature Journals*. Online: http://www.nature.com/authors/gta.pdf [Zugriff am 09.08.2011].

NHMRC (National Health and Medical Research Council), Australian Research Council & Universities Australia, 2007. *Australian Code for the Responsible Conduct of Research*. Canberra: National Health and Medical Research Council. Online: http://www.nhmrc.gov.au/publications/synopses/r39syn.htm [Zugriff am 09.08.2011].

NIH (National Institutes of Health), 2003. *Final NIH Statement on Sharing Research Data was*. Online: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [Zugriff am 09.08.2011].

NSF (National Science Foundation), 2011. *Proposal and Award Policies and Procedures Guide. Grant Proposal Guide*. Chapter II – Proposal Preparation Instructions. Online unter: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp [Zugriff am 09.08.2011].

OECD (Organisation for Economic Co-operation and Development), 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Online: http://www.oecd.org/dataoecd/9/61/38500813.pdf [Zugriff am 09.08.2011].

PLoS ONE, o. J. *PLoS ONE Editorial and Publishing Policies.* Sharing of Materials, Methods, and Data. Online unter: http://www.plosone.org/static/policies.action#sharing [Zugriff am 09.08.2011].

Savage, C.J. & Vickers, A.J., 2009. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE*, 4(9), S. e7078. doi:10.1371/journal.pone.0007078.

Schofield, P.N. et al., 2009. Post-publication sharing of data and tools. *Nature*, 461(7261), S. 171–173. doi:10.1038/461171a.

Science, 2011. *Special Online Collection, Dealing with Data*. Online: http://www.sciencemag.org/site/special/data/ [Zugriff am 09.08.2011].

Smith, D. & Carrano, A., 1996. International Large-Scale Sequencing Meeting. *Human Genome News*, 6(7). Online: http://www.ornl.gov/sci/techresources/Human_Genome/publicat/hgn/v7n6/19intern.shtml [Zugriff am 09.08.2011].

Treloar, A. & Harboe-Ree, C., 2008: Data management and the curation continuum. How the Monash experience is informingrepository relationships. In: *Proceedings of VALA 2008*. Melbourne, Australien 5.–7. Feb. 2008. Online: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf [Zugriff am 09.08.2011].

Toronto International Data Release Workshop Authors, 2009. Prepublication data sharing. *Nature*, 461(7261), S. 168–170. doi:10.1038/461168a.

Universität Siegen, 2001. *Grundsätze und Verfahrensrichtlinien zur Sicherung einer guten wissenschaftlichen Praxis an der Universität Gesamthochschule Siegen*. Online: http://www.uni-siegen.de/start/forschung/kombibox_forschung/grundsaetze_und_verfahrensrichtl_wiss_praxis.pdf [Zugriff am 09.08.2011].

Webster, F., o. J. *Data Access Policies*. Online: http://www.codata.org/data_access/policies.html [Zugriff am 09.08.2011].

Wellcome Trust, 2010. *Policy on data management and sharing*. Online: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm [Zugriff am 09.08.2011].

Wellcome Trust, 2011. *Sharing research data to improve public health: full joint statement by funders of health research*. Online: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm [Zugriff am 09.08.2011].

Wefer, G. Hrsg., 2010. *Dynamische Erde – Zukunftsaufgaben der Geowissenschaften, Strategieschrift*. Bremen: MARUM. Online: http://www.geokommission.de/Dynamische_Erde.html [Zugriff am 09.08.2011].

2.2 Rechtliche Probleme der elektronischen Langzeitarchivierung von Forschungsdaten[1]

Gerald Spindler [1], Tobias Hillegeist [1,2]
[1] Universität Göttingen
[2] Landgericht Lüneburg

2.2.1 Einleitung

Neben der Langzeitarchivierung von Büchern und Zeitschriften gewinnt die Langzeitarchivierung von Forschungsdaten (sog. Rohdaten) in jüngster Zeit eine immer bedeutendere Rolle, da immer mehr Hochschulen und Forschungseinrichtungen dazu übergehen, die von ihnen gewonnenen Daten zu archivieren. Dabei sollen die Daten in den meisten Fällen nicht nur archiviert, sondern auch Dritten, wie beispielsweise anderen Forschungseinrichtungen oder einzelnen Fremdforschern zur Verfügung gestellt werden. Aus rechtlicher Sicht ist dabei vor allem entscheidend, ob die Archivierung dieser Daten eine urheberrechtliche Relevanz aufweist, die Daten also urheberrechtlich geschützt sind und, sofern dies zutrifft, wer Inhaber der erforderlichen Nutzungsrechte ist.

Hinsichtlich der Archivierung personenbezogener Daten können sich darüber hinaus datenschutzrechtliche Probleme stellen, was vor allem für Universitätskliniken relevant ist.

2.2.2 Rechtlicher Schutz nach dem Urheberrechtsgesetz (UrhG)

Sofern an Forschungsdaten ein urheber- oder leistungsrechtlicher Schutz bestünde, dürften diese nur archiviert werden, sofern die archivierende Einrichtung Inhaber der erforderlichen Nutzungsrechte wäre bzw. der jeweilige Rechteinhaber der Einrichtung die Archivierung gestatten würde.

2.2.2.1 Schutz einzelner Daten

Dabei lässt sich zunächst feststellen, dass wissenschaftliche Primärdaten grundsätzlich nicht dem Schutz des Urheberrechtsgesetzes unterliegen. Ein urheberrechtlicher Schutz würde gem. § 2 Abs. 2 UrhG voraussetzen, dass die einzelnen Daten eine persönliche geistige Schöpfung darstellen. Da es bei Forschungsda-

1. Siehe zu diesem Problem auch Spindler / Hillegeist, Langzeitarchivierung wissenschaftlicher Primärdaten, in: NESTOR Handbuch – Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 2.3, Kap. 16:14, abrufbar unter http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf

ten jedoch an der für einen urheberrechtlichen Schutz notwendigen geistigen Schöpfungshöhe fehlt, unterliegen zumindest die einzelnen Daten grundsätzlich nicht dem Schutz des Urheberrechtsgesetzes.

2.2.2.2 Schutz von Datensammlungen

Dies gilt in der Regel auch dann, wenn die Daten in Tabellen oder auf andere Art zusammengefasst werden. Eine solche Zusammenstellung könnte zwar ein Datenbankwerk nach § 4 Abs. 2 UrhG oder eine Datenbank gem. § 87a UrhG darstellen. Zum einen wird es dafür jedoch regelmäßig an der erforderlichen geistigen Schöpfungshöhe fehlen, die bei Datenbankwerken nach § 4 Abs. 2 UrhG in der individuellen Auswahl oder Anordnung der Daten bestehen muss. Eine solche Individualität wird in den vorliegend relevanten Fällen grundsätzlich nicht vorliegen, da die Anordnung nach logischen Gesichtspunkten erfolgen wird.

Zum anderen wird der sui-generis-Schutz der Datensammlung nach § 87a UrhG in den meisten Fällen daran scheitern, dass für die vorliegend in Betracht kommenden Datensammlungen in der Regel keine wesentliche Investition im Sinne der Vorschrift erforderlich ist. Investitionen werden vielmehr bei der Datenerhebung getätigt werden, deren Kosten jedoch im Rahmen des § 87a UrhG gerade nicht zu berücksichtigen sind[2]. Trotzdem sollte im Einzelfall stets genau geprüft werden, ob nicht ausnahmsweise doch eine Datenbank i.S.d. § 87a UrhG vorliegt.

Sofern dies ausnahmsweise zuträfe, wäre gem. § 87a UrhG diejenige Person bzw. die Einrichtung Datenbankhersteller und damit Inhaber der an der Datenbank bestehenden Nutzungsrechte, die diese Investition getätigt und damit das organisatorische und wirtschaftliche Risiko übernommen hat[3]. Dies wird im Regelfall die Hochschule oder das Forschungsinstitut sein, in dessen Einrichtungen die Daten zusammengestellt wurden. Damit würden sich hinsichtlich einer Langzeitarchivierung der Daten keine urheberrechtlichen Probleme ergeben. Zu beachten ist jedoch, dass in den Fällen, in denen Forschungsprojekte durch sogenannte Drittmittel finanziert werden, die finanzierende Einrichtung wohl Trägerin der wesentlichen Investition und damit Datenbankherstellerin im Sinne

2. EuGH GRUR 2005, 254, 256 Tz. 40 ff. – Fixtures-Fußballspielpläne II; EuGH C-46/02 Tz. 44 ff.; EuGH GRUR 2005, 252, 253 – Fixtures-Fußballspielpläne I; siehe auch Erwägungsgrund 9, 10 und 12 der RL 96/9/EG; Leistner, K&R 2007, 457, 460 (nach dem die Kosten für die „Erzeugung" jedoch wie Beschaffungsinvestitionen behandelt werden sollen, sofern die Informationen mit einem vergleichbaren Aufwand hätten beschafft werden können); ders., GRUR Int. 1999, 819, 832; a.A. Czychowski in Fromm/Nordemann, Urheberrecht, 10. Aufl. 2008, § 87a Rn. 19; so auch Spindler, JZ 2004, 150, 152 (Anmerkung zu BGH Urt. v. 17.7.2003 – Paperboy).

3. Czychowski in: Fromm / Nordemann, § 87a Rn. 25; siehe auch Erwägungsgrund 14 der RL 96/9/EG.

des § 87a UrhG wäre, so dass ihr die zur Langzeitarchivierung erforderlichen Nutzungsrechte zustünden. Im Rahmen der elektronischen Archivierung werden dabei vor allem das Recht der Vervielfältigung[4] und das Recht der öffentlichen Zugänglichmachung im Internet betroffen[5]. In diesen Fällen sollten Forschungseinrichtungen in ihren Verträgen mit den Drittmittelgebern vereinbaren, dass eventuell entstehende Nutzungsrechte an Datenbanken, die im Rahmen des finanzierten Forschungsprojektes erstellt werden, der Forschungseinrichtung zumindest als einfache Nutzungsrechte eingeräumt werden. Auf diese Weise wäre sichergestellt, dass die Forschungseinrichtung die anfallenden Daten auch archivieren und Dritten zugänglich machen dürften.

2.2.2.3 Erlangung der Nutzungsrechte aufgrund eines bestehenden Arbeitsverhältnisses

Sofern die Nutzungsrechte ursprünglich bei einem Angestellten der archivierenden Einrichtung entstanden sind, könnte diese sie bereits aufgrund des Arbeitsverhältnisses erlangt haben, da insoweit der Grundsatz gilt, dass Nutzungsrechte, die ein Arbeitnehmer im Rahmen seiner Angestelltentätigkeit erlangt, dem Arbeitgeber zustehen.

Dabei erfolgt die Einräumung der Nutzungsrechte regelmäßig im Voraus bei Abschluss des Arbeits- oder Dienstvertrages[6], spätestens jedoch mit Ablieferung des Werkes[7]. Sofern der Urheber eines Werkes bzw. der Datenbankhersteller oder Lichtbildner in einem Angestellten- oder Dienstverhältnis zur Universität stand, wäre er also gegenüber der Universität grundsätzlich zur Übertragung der Nutzungsrechte verpflichtet. Zu beachten ist allerdings, dass aufgrund der durch Art. 5 Abs. 3 GG verfassungsrechtlich garantierten Wissenschaftsfreiheit diese Grundsätze nicht auf Hochschul-, Honorar- oder Gastprofessoren übertragen werden können, da die Veröffentlichung von Forschungsergebnissen nicht

---

4. Zum Begriff der „Vervielfältigung“ siehe: BT-Drucks. IV/270, 47; BGH NJW 1955, 1276; Dreyer in Dreyer / Kotthoff / Meckel, 2. Aufl. 2009, § 16 Rn. 6; Dustmann in Fromm/Nordemann, Urheberrechtsgesetz, § 16 Rn. 9; Heerma in Wandtke / Bullinger, Praxiskommentar zum Urheberrecht, 3. Aufl. 2009, § 16 Rn. 2;

5. Zum Verhältnis der „öffentlichen Zugänglichmachung“ i.S.v. § 19a UrhG zum Begriff der „öffentlichen Wiedergabe“ i.S.v. § 87b UrhG siehe: Loewenheim in Loewenheim, § 43 Rn. 21; Kotthoff in Dreyer/Kotthoff/Meckel, § 87b Rn. 9; Dreier in Dreier / Schulze, Urheberrechtsgesetz, 3. Aufl. 2008; § 87b Rn. 3; Dannecker, K&R 1999, 529, 536.

6. Dreier in Dreier / Schulze, § 43 Rn. 19.

7. BGH GRUR 1974, 480, 483 – Hummelrechte; A. Nordemann in Fromm / Nordemann, § 43 Rn. 30.

mehr zu deren Aufgabenbereich gehört[8]. Handelt es sich bei dem Urheber des Datenbankwerkes oder dem Datenbankherstellers um einen Professor, so wird die Universität daher nicht aufgrund des bestehenden Arbeitsverhältnisses Inhaberin der entsprechenden Nutzungsrechte[9]. Aus diesem Grund sollte in den von Hochschulen oder Forschungseinrichtungen geschlossenen Arbeitsverträgen grundsätzlich eine Klausel enthalten sein, wonach die Vertragspartner ihrem künftigen Arbeitgeber die Rechte, die sie im Rahmen ihrer Forschungstätigkeit erlangen, zumindest als einfache Nutzungsrechte einräumen. Hinsichtlich des Inhalts einer solchen Klausel ist zu beachten, dass diese aufgrund der sogenannten Zweckübertragungslehre nicht pauschal abgefasst sein darf, sondern vielmehr die genauen Nutzungsrechte und -arten bezeichnen muss.

2.2.2.4 Rechtsgeschäftlicher Erwerb der erforderlichen Nutzungsrechte

Hat die archivierende Einrichtung die erforderlichen Nutzungsrechte nicht bereits aufgrund eines Arbeitsvertrages oder der Vorschrift des § 137 l UrhG erlangt, bleibt ihr nur die Möglichkeit, sich die benötigten Nutzungsrechte von dem Rechteinhaber rechtsgeschäftlich übertragen zu lassen. Relevant ist dies vor allem bei Forschungsprojekten, bei denen auch ehrenamtliche Beiträger mitgewirkt haben und demzufolge in keinem Angestelltenverhältnis zu der archivierenden Einrichtung standen.

2.2.2.5 Zwischenergebnis

Aus dem oben Gesagten folgt, dass Forschungseinrichtungen besonderes Augenmerk auf die Ausgestaltung ihrer Arbeitsverträge legen sollten. Insbesondere sollten sie detaillierte Rechteübertragungsklauseln in ihre Arbeitsverträge einfügen, um spätere Rechtsunsicherheiten hinsichtlich der Inhaberschaft von Nutzungsrechten von vornherein zu vermeiden. Aufgrund des im Urheberrecht geltenden Bestimmtheitsgrundsatzes genügt hierfür die pauschale Vereinbarung, dass „…alle bestehenden Rechte übertragen werden…“ allerdings nicht.

---

8. BGH GRUR 1991, 523, 525 – Grabungsmaterialien, sowie die Vorinstanz OLG Karlsruhe GRUR 1988, 536, 537 – Hochschulprofessor; BGH GRUR 1985, 529, 530 – Happening; Dreier in Dreier/Schulze, § 43 Rn. 12; A. Nordemann in Fromm / Nordemann, § 43 Rn. 43.

9. BGH GRUR 1991, 523, 525 – Grabungsmaterialien, sowie die Vorinstanz OLG Karlsruhe GRUR 1988, 536, 537 – Hochschulprofessor; BGH GRUR 1985, 529, 530 – Happening; Dreier in Dreier / Schulze, § 43 Rn. 12; Rojahn in Schricker, § 43 Rn. 31, 65; A. Nordemann in Fromm/Nordemann, § 43 Rn. 43; Jarass in Jarass / Pieroth, Grundgesetz, 11. Aufl. 2011, Art. 5 Rn. 138.

2.2.3 Datenschutzrechtliche Probleme

Rechtliche Probleme können sich bei der Archivierung von Forschungsdaten ferner aus dem Bundesdatenschutzgesetz[10], den einzelnen Landesdatenschutzgesetzen[11] sowie dem Sozialgesetzbuch X[12] ergeben, sofern die zu archivierenden Daten einen Personenbezug aufweisen. Dies ist vor allem bei medizinischen Forschungsdaten denkbar.

2.2.3.1 Schriftliche Einwilligung des Betroffenen

Aus diesem Grund müssen archivierende Einrichtungen vor der Archivierung personenbezogener Daten eine schriftliche Einwilligung des jeweiligen Betroffenen einholen[13]. Dies bedeutet, dass die Einwilligung gem. § 126 Abs. 1 BGB vom Betroffenen eigenhändig durch Namensunterschrift oder mittels notariell beglaubigten Handzeichens unterzeichnet werden muss. Eine Kopie oder ein Fax genügen demzufolge nicht[14]. Dabei kann die schriftliche Form gem. § 126 Abs. 3 BGB durch die elektronische Form ersetzt werden, was in einigen Datenschutzgesetzen sogar ausdrücklich vorgesehen ist[15]. Sollte eine Einwilligung des Betroffenen nicht eingeholt werden können, könnte eine Archivierung der Daten aufgrund der Erlaubnisnorm des § 10 NDSG oder einer entsprechenden Vorschrift der übrigen Landesdatenschutzgesetze zulässig sein[16]. Dies gilt jedoch nur dann, wenn die betreffenden Daten weder zu einem anderen Zweck als dem der wissenschaftlichen Forschung erhoben worden sind noch im Rahmen eines anderen Forschungsvorhabens gewonnen wurden.

10. Siehe bspw.: § 4 BDSG.
11. Siehe bspw.: § 4 Abs. 1 LDSG B-W; § 4 Abs. 1 BBDSG; Art. 15 Abs. 1 BayDSG; § 6 Abs. 1 BlnDSG; § Abs. 1 BremDSG; § 5 Abs. 1 HmbDSG; § 7 Abs. 1 HDSG; § 4 Abs. 3 LDSG M-V; § 4 Abs. 1 NDSG; § 4 Abs. 1 DSG NRW; § 5 Abs. 1 LDSG R-P; § 4 Abs. 1 SDSG; § 4 Abs. 1 DSG-LSA; § 4 Abs. 1 SächsDSG; § 4 Abs. 2 S. 1 SDSG; § 11 Abs. 1 LDSG S-H; § 4 Abs. 1 ThürDSG.
12. Siehe § 67b Abs. 1 SGB X.
13. Zu den Rechtsfolgen bei Nichteinhaltung der Formvorschriften siehe: Gola / Schomerus, BDSG, 10. Aufl. 2010, § 4a Rn. 13; Roßnagel / Holznagel / Sonntag, Handbuch Datenschutzrecht, 2003, Rn. 40; Simitis in Simitis, Bundesdatenschutzgesetz, 7. Aufl. 2011, § 4a Rn. 26; vgl. auch § 7 Abs. 2 S. 1 HDSG; § 4 Abs. 2 S. 1 NDSG; § 4 Abs. 2 S. 2 ThürDSG.
14. Ellenberger in Palandt, Bürgerliches Gesetzbuch, 70. Aufl. 2011, § 126 Rn. 8; Wendtland in Bamberger / Roth (Hrsg.), Kommentar zum Bürgerlichen Gesetzbuch Bd. 1, 2. Aufl. 2007, § 126 Rn. 6.
15. So etwa in § 4 Abs. 5 SächsDSG; § 4 Abs. 3 S. 2 Alt. 2 ThürDSG.
16. Siehe § 35 LDSG B-W; Art. 23 BayDSG; § 30 BlnDSG; § 28 BBDSG; § 19 BremDSG; § 27 HmbDSG; § 33 HDSG; § 34 LDSG M-V; § 28 LDSG NRW; § 30 DSG R-P; § 30 SDSG; § 27 DSG-LSA; § 36 SächsDSG; § 22 LDSG S-H; § 25 ThürDSG.

2.2.3.2 Entbindung von der ärztlichen Schweigepflicht

Sofern eine wirksame datenschutzrechtliche Einwilligung des Probanden vorliegt, ist darin außerdem gleichzeitig eine (zumindest konkludent erteilte) Entbindung des behandelnden Arztes von seiner ärztlichen Schweigepflicht zu sehen, was vor allem dann relevant ist, wenn medizinische Untersuchungsdaten archiviert werden sollen. Die Entbindung von der Schweigepflicht entspricht dabei konsequenterweise in ihrer Reichweite dem Umfang, in welchem der Proband auch in die datenschutzrechtlich relevante Nutzung seiner Daten eingewilligt hat.

Der Betroffene kann seine Einwilligung allerdings jederzeit widerrufen, womit diese mit Wirkung für die Zukunft entfällt. Des Weiteren ist aus datenschutzrechtlicher Sicht zu beachten, dass die Daten sowohl nach den Vorschriften des BDSG als auch nach den Landesdatenschutzgesetzen zu anonymisieren sind, sobald der Forschungszweck dies zulässt. Sofern dies nicht möglich sein sollte, hat die archivierende Einrichtung die Merkmale, mit denen ein Personenbezug hergestellt werden kann, zumindest gesondert zu speichern.

2.2.3.3 Zwischenergebnis

Bevor personenbezogene Daten archiviert werden, ist also stets zu prüfen, ob die Merkmale, mit denen ein Personenbezug hergestellt werden kann, wirklich noch erforderlich sind. Einer solchen Prüfung sind darüber hinaus regelmäßig die bereits archivierten Daten zu unterziehen. Da die elektronische Archivierung personenbezogener Daten regelmäßig eine automatisierte Datenverarbeitung darstellt, hat die archivierende Einrichtung die jeweiligen datenschutzrechtlichen Vorschriften hinsichtlich der zu ergreifenden technischen und organisatorischen Maßnahmen zu beachten, namentlich des § 9 S. 1 i.V.m. der Anlage zu § 9 S. 1 BDSG sowie des § 7 Abs. 2 NDSG beziehungsweise der entsprechenden Vorschriften der übrigen Landesdatenschutzgesetze.

2.2.3.4 Rechtsfolge bei einem Verstoß gegen datenschutzrechtliche Bestimmungen

Sofern die archivierende Einrichtung gegen die datenschutzrechtlichen Vorgaben verstößt, kann dies zum einen eine Straftat oder Ordnungswidrigkeit darstellen. Darüber hinaus stehen dem Betroffenen unter Umständen auch Schadensersatzansprüche zu, wobei sowohl das BDSG als auch die Landesdatenschutzgesetze in den Fällen einer automatisierten Datenverarbeitung eine verschuldensunabhängige Haftung vorsehen.

In den Fällen einer nicht automatisierten Datenverarbeitung kann sich die archivierende Einrichtung zwar exkulpieren, trägt allerdings die Beweislast für das fehlende Verschulden. Besondere Bedeutung erlangt die verschuldensunabhängige Haftung im Rahmen der Auftragsdatenverarbeitung, da der Anspruchsgeg-

ner in diesen Fällen der Auftraggeber, im Rahmen der Langzeitarchivierung also die archivierende Einrichtung bleibt. Diese sollte aus diesem Grund in die Verträge mit der beauftragten Einrichtung für den Fall eines Verstoßes gegen datenschutzrechtliche Vorschriften eine Freistellungsklausel in den Vertrag aufnehmen.

2.2.4 Fazit

Abschließend lässt sich feststellen, dass der rechtliche Problemschwerpunkt der elektronischen Langzeitarchivierung von Forschungsdaten nicht, wie man zunächst annehmen könnte, urheberrechtlicher sondern datenschutzrechtlicher Natur ist. Diese Probleme lassen sich aus Sicht der archivierenden Einrichtung insoweit relativ einfach vermeiden, indem nur Daten ohne Personenbezug archiviert werden. Nichtsdestotrotz sollte vor jeder Digitalisierung unbedingt genau geprüft werden, ob nicht aus den oben beschriebenen Gründen ausnahmsweise doch ein urheberrechtlicher Schutz an der zu archivierenden Datenbank besteht. Unabhängig davon sollte bei dem Abschluss neuer Arbeitsverträge ein besonderes Augenmerk auf eine klar und detailliert formulierte Rechteübertragungsklausel gelegt werden.

2.3 Datenmanagementpläne

Uwe Jensen
GESIS – Leibniz-Institut für Sozialwissenschaften

2.3.1 Langfristige Ziele und aktuelle Relevanz

Datenmanagementpläne sind Ergebnisse der wissenschaftspolitischen Forderung nach einem zeitgemäßen Management und Austausch von wissenschaftlichen Daten. Der verantwortungsvolle Umgang mit Forschungsdaten wird national und international mit der Erwartung verbunden, dass die Ergebnisse öffentlich geförderter Forschungsprojekte für Analysen und Replikationen dauerhaft und global verfügbar gemacht sind (OECD, 2007; Allianz, 2010).

Für ein Forschungsprojekt ist ein Datenmanagementplan (DMP) von praktischer Relevanz, seit nationale und internationale Förderinstitutionen von Antragstellern erwarten, systematisch zu beschreiben, wie mit Forschungsdaten während der Projektlaufzeit und nach Projektabschluss umgegangen werden soll. Zur Förderung der nachhaltigen Sicherung und Verfügbarkeit von Forschungsdaten hat die DFG 2010 die Erstellung eines Datenmanagementplans in die Antragsbedingungen aufgenommen.

> „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, S. 32)

Bei der Planung entsprechender Maßnahmen stellen Leitfäden und Standards des Datenmanagements der Fachdisziplin wichtige Empfehlungen bereit. Der Umgang mit sozialwissenschaftlichen Daten und Metadaten wird u. a. in den Leitfäden „*Guide to Social Science Data Preparation and Archiving*“ (ICPSR, 2009a) und „*Managing and Sharing Data*“ (Van den Eynden et al., 2011) ausführlich beschrieben. Die dargestellten Prinzipien und Fragestellungen können im Sinne eines Rahmenkonzeptes auch in andere Fachkontexte übertragen und an entsprechende Projektbedingungen angepasst werden.

2.3.2 Kernelemente eines Datenmanagementplans

Ein strukturiertes Konzept zum Datenmanagement ist für jedes Projekt essenziell, um die Arbeitsergebnisse während der Laufzeit zu sichern und zur weiteren Nutzung der Forschergemeinschaft bereitzustellen. In Anlehnung an eine Über-

sicht des *Inter-University Consortium for Political and Social Research* werden im Folgenden Kernelemente eines DMP (ICSPR, 2009b) entlang der Projektphasen des Forschungsdatenzyklus (ICPSR, 2009b) vorgestellt.

2.3.2.1 Projektplanung und Datenmanagement

Einschlägige projektbezogene Anforderungen der Drittmittelgeber an das projektbezogene Datenmanagement und die spätere Zugänglichkeit der Daten (Archivierung, Bereitstellung) nach Projektabschluss sollten eindeutig benannt werden.

Beratungs- und Kooperationsmöglichkeiten im fachlichen Umfeld können genutzt werden, etwa um eingeführte Arbeitsprozesse zu optimieren, indem aktuelle fachliche und technische Empfehlungen zur standardisierten Datendokumenten und Datenspeicherung für das Projekt aufgegriffen werden. In diesem Rahmen kann ein Forscherteam auch frühzeitig Überlegungen zur langfristigen Sicherung und öffentlichen Verfügbarkeit der Daten anstellen. Soweit keine rein projektspezifischen Archivierungs- und Zugangslösungen beabsichtigt oder vorhanden sind, können auch vorhandene Dienstleistungsangebote lokaler oder nationaler Infrastruktureinrichtungen und Datenserviceeinrichtungen in Anspruch genommen werden.

Umgang mit vorhandenen und neuen Forschungsdaten

Welche Forschungsdaten bzw. Metadaten in einem DMP im Detail zu berücksichtigen sind, lässt sich nicht allgemeingültig beantworten. „Forschungsprimärdaten“ kennzeichnet die DFG als „Daten, die im Verlauf von Quellenforschungen, Experimenten, Messungen, Erhebungen oder Umfragen entstanden sind. Sie stellen die Grundlagen für die wissenschaftlichen Publikationen dar“. Wobei diese je nach Fachdisziplin „unterschiedlich zu definieren“ sind und die Wissenschaftler darüber entscheiden, ob „bereits Rohdaten hierzu zählen oder ab welchem Grad der Aggregierung die Daten langfristig aufzubewahren sind“ (DFG, 2009, S. 2). Die aktuellen DFG Förderrichtlinien beziehen sich auf „(Mess-) Daten [...], die für die Nachnutzung geeignet sind“ (DFG, 2010, S. 32).

Im Zusammenhang mit der Forschungsfrage und dem zu entwickelnden Forschungs- und Arbeitsplan stellt sich auch die praktische Frage, ob bzw. welche projektrelevanten Daten bereits vorhanden sind. Weiterhin wäre zu beschreiben, ob und wie sie in das Projekt integriert bzw. benutzt werden, etwa im Zuge von Vergleichen über Zeit oder Raum.

Die Beschreibung der Forschungsdaten, die in einem Projekt benutzt, bearbeitet und / oder neu erzeugt werden, stellen also ein zentrales Element eines Datenmanagementplans dar. Die domainspezifische Herkunft bzw. die Art ihrer Erhebung (Befragung, Beobachtung, Experiment, Simulation) und ihre weitere Bearbeitung spielt für die Planung eine wesentliche Rolle. So erfordern Daten der Umfrageforschung andere Maßnahmen als Simulationen mit umfangreichen

Datenbeständen zur Entwicklung von Klimamodellen. D. h. die Eigenart, die Struktur, der Umfang und der Grad an Komplexität und Veränderlichkeit, sowie die Art der Weiterverarbeitung und Speicherung der Forschungsdaten bestimmen auch Art und Umfang des Datenmanagements in einem Forschungsprojekt. Anschaulich zeigt dies etwa das Projekt IBF zum „Aufbau eines Informationsnetzes für biologische Forschungsdaten von der Erhebung im Feld bis zur nachhaltigen Sicherung in einem Primärdatenrepositorium“ (IBF, 2010).

Metadaten zur Dokumentation der Daten und des Entstehungskontextes

Mit der Datenerzeugung eng verbunden ist die fachspezifische Beschreibung der Daten und des Kontextes ihrer Erzeugung auf Studien-, Datensatz- und Variablenebene während der projektbezogenen Phasen des Forschungsdatenzyklus.

Metadaten zum Projekthintergrund und dem speziellen **Studiendesign** informieren über Ziele, Zwecke einer Untersuchungen und die zugrundeliegenden Fragestellung (Thema, zeitliche und geographische Raum der Untersuchung) und bereits vorhandene Daten und Studien.

Die Dokumentation zum **Methodendesign** der Datenerhebung (Stichprobendesign, Verfahren der Stichprobenziehung), des Messinstrumentes (z. B. technische Sensoren oder Fragebögen) und der geplanten Repräsentation der Daten in strukturierten Datensätzen ist eine Voraussetzung, um die Daten auch langfristig zu nutzen. Dazu zählt entsprechend die präzise Beschreibung der Datensatzstruktur mit ihren Variablen (Name, Label) und deren Ausprägungen (Kategorien und Kodes) sowie die Struktur und Beziehung von Daten in komplexen Datensätze, wie etwa bei der Kumulation von Zeitreihen oder Datensätzen mit unterschiedlichen geographischen Messorten.

Die Beschreibung der Durchführung der **Datenerhebung**, der dabei gewonnenen Daten und die Dokumentation ihrer weiteren Bearbeitung, z. B. durch logische Prüfung der Datenkonsistenz, formale Kontrollen und entsprechende Korrekturen oder Bereinigung der Daten zur Sicherung der Datenqualität, stellen weitere Metadatenfacetten dar, die der längerfristigen Benutzbarkeit von Forschungsdaten dient. Typischerweise unterliegen Datensätze nach ihrer ersten Erzeugung im Laufe der Zeit weiteren Anpassungen, etwa durch die formale Standardisierung, inhaltliche Harmonisierung, zusätzlich konstruierter Variablen oder Daten, die aus weiteren Erhebungswellen integriert werden. Diese Veränderungen sind durch Versionierung der Datensätze zu kennzeichnen und durch entsprechende Metadatensets zu dokumentieren.

Schließlich sind die Daten durch entsprechende Metadaten hinsichtlich ihrer **Vertraulichkeit**, Nutzung und Zugänglichkeit zu kennzeichnen. Dies betrifft sowohl die Projektlaufzeit selbst und als auch das Ende eines Projektes unter dem Aspekt der langfristigen Sicherung und Verfügbarkeit. Im Interesse der erhöhten Sichtbarkeit von Forschungsdaten als eigenständigem Ergebnis wis-

senschaftlicher Arbeit ist auch zu berücksichtigen, wie die Daten durch persistente **Identifikatoren** dauerhaft erreichbar und zitierfähig vorgehalten werden können (DataCite, o. J.).

In welcher Form (was, wie, wann, wo, wer) die Dokumentation von Daten erstellt werden, kann nur im konkreten Projektkontext unter Berücksichtigung von lokalen und überregionalen Dateninfrastrukturen und fachspezifischen Kooperationsmöglichkeiten definiert werden.

Durch die projektspezifische Beratung und den Einsatz von disziplinspezifischen Standards und *Tools* kann bereits frühzeitig das Management von Daten und Metadaten im Projekt strukturiert vorbereitet werden. Durch diese projektspezifischen Prozessvorbereitungen wird das Wissen um die Daten systematisch gesichert und für langfristige Archivierung und Datennutzung vorbereitet. Damit verringern sich auch die Risiken einer nachgelagerten Rekonstruktion von Daten und Metadaten.

Die einmaligen und laufenden **Kosten** des Datenmanagements (Personal- und Sachmittel) für die adäquate Aufbereitung und Dokumentation sowie die dauerhafte Sicherung der Daten und Metadaten während der Projektlaufzeit dürfen nicht unterschätzt werden und sind entsprechend in den Budgetplanungen zu berücksichtigten. Wird die Studie mit den Daten und Dokumentationen an ein Archiv übergeben, sind auch die Kosten für die Vorbereitung und Zusammenstellung aller erforderlichen Materialien und Informationen, z. B. zum Rechtemanagement, mit einzuplanen (vergl. z. B. die Checkliste „*Data Management Costing Tool*“ (UK Data Archive, 2011).

2.3.2.2 Datenmanagement und Datenorganisation während der Projektlaufzeit

Zentrale Anforderung an das Datenmanagement im Projektverlauf ist die Festlegung von formalen Verantwortlichkeiten, organisatorischen Konventionen und technischen Regeln, um die produzierten Daten und Metainformationen zu organisieren, zu kontrollieren, zu sichern und für den Projektbetrieb bereit zu halten.

Vor der Planung der Maßnahmen zur Erhebung, Sicherung, Verarbeitung und Dokumentation von Forschungsdaten sind die grundsätzlichen **Verantwortlichkeiten** und Aufgaben von Primärforschern und Projektmitarbeitern im Datenmanagement des Projektes festzulegen und entsprechend im Datenmanagementplan zu dokumentieren. Weiterhin sind die Prozesse der **Qualitätssicherung** zu beschreiben, die die geschützte Speicherung der digitalen Informationen gewährleisten und die Nutzbarkeit und Qualität der gespeicherten Daten garantieren. Im Zuge der weiteren Datenorganisation sind die Rechte des Datenzugangs für Projektmitarbeiter unter besonderer Berücksichtigung des Zugangs zu und der Bearbeitung von sensiblen Daten und Informationen festzulegen.

Eine wesentliche Aufgabe besteht dann darin, die Organisation und Speicherung von Daten, Metadaten und Dokumenten vorausschauend zu planen und

diesbezügliche Maßnahmen und Verfahren zu dokumentieren. D. h. welche Dateien im Verlauf der täglichen Arbeitsroutinen, mit welchem Status (z. B. Originaldatei, temporäre Arbeitsdatei; Entwurf, Zwischenversion, Endversion), wo (Arbeitsplatz PC, zentraler Dateiserver, Datenbank) und wie lange (temporär, Projektlaufzeit, Langzeitverfügbarkeit) in welchem Format gespeichert werden, sind relevante Informationen für einen Datenmanagementplan.

Die Frage, in welchem technischen **Dateiformat** Daten und Metadaten erfasst, unterhalten, gespeichert und verfügbar gemacht werden sollen, beinhaltet die Prüfung der Zweckmäßigkeit etablierter Standards in den Arbeitsroutinen der verschiedenen Disziplinen. Insbesondere für die Projektzeit ist zu berücksichtigen, ob ein Dateiformat proprietär oder offen dokumentiert bzw. verbreitet verfügbar ist, um etwa im Projekt Daten und Metadaten zu sichern, mit Projektpartnern auszutauschen oder Speichermedien für eine langfristige Sicherung und Weitergabe vorzubereiten. Je nach Disziplin sind entsprechende Empfehlungen zur Anwendung bestimmter Formate für spezielle Arbeitsschritte und Verfahren (z. B. zur Datenerhebung, -dokumentation und -auswertung) und Standardroutinen zur Konversion und den Austausch von Daten und Metadaten zwischen unterschiedlichen Bearbeitungsprogrammen zu berücksichtigen.

Eng verbunden mit der Wahl von technischen Formaten für die Forschungsdaten und deren Metadaten ist die Systematik einer gemeinsamen oder getrennten Speicherung (z. B. SPSS Datei für Daten; XML Datei für Metadaten) von Bedeutung. Angesichts der weiteren Bearbeitungsprozesse im Projektverlauf (z. B. Bereinigung von Fehlern in den Daten) sind geeignete Verfahren zum Erhalt der Konsistenz von Daten und Metadaten zu etablieren. Die dynamische Veränderungen und Anpassungen von Datensätzen, z. B. durch die Integration von Messreihen, erfordern qualifizierte Versionierungsverfahren. Diese Maßnahmen sichern sowohl die Datentransparenz im Projekt als auch die Nachvollziehbarkeit von weiteren Änderungen durch formale Standardisierungen oder Formattransformationen im Zuge der Vorbereitung und Bereitstellung von fertigen Datenprodukten. Als Teil der strukturierten Verwaltung von Dateien, Metadaten und weiteren Informationen (z. B. Methodenbericht, Messinstrument, Messprotokoll) ist die standardisierte Benennung etwa von Dateiverzeichnissen und Dateien durch Namenskonventionen unerlässlich.

Das Thema **Datensicherheit** betrifft alle technischen und organisatorischen Maßnahmen zum Schutz der physischen Daten vor Veränderung, Verlust und Zerstörung, die entsprechend zu erläutern sind. Die Sicherung von Daten und Dokumentationen stellt dabei ein zentrales Element des täglichen Datenmanagements dar. In diesem Zusammenhang sind Speichermethoden, Backupverfahren, notwendige physische Ressourcen sowie automatisierte und administrative Routinen zu planen. Zum Schutz der so gesicherten Informationen vor Veränderung oder Verlust sind weitere Aspekte bezogenen auf die Daten (z. B. Zugangs-

rechte, Virenschutz, Datensignaturen) und die technische Infrastruktur zu beachten (z. B. Ausfallsicherheit des technischen Systems).

2.3.2.3 Langzeitarchivierung und Zugang zu Forschungsdaten

Zum Projektende stellen sich eine Reihe von Fragen zum weiteren Umgang mit den erhobenen Daten, durchgeführten Datenanalysen und -dokumentationen. Frühzeitige Überlegungen zur langfristigen Sicherung und offenen Verfügbarkeit der Daten und zur Kooperationen mit entsprechenden Einrichtungen vereinfachen entsprechende Maßnahmen am Ende des Projektes. Die Planung des konkreten Datenmanagements bezüglich der langfristigen Sicherungs- und Nutzungsperspektive berührt insbesondere folgende Gesichtspunkte.

Hinsichtlich der **Auswahlkriterien** ist darzulegen, welche Daten und Metadaten an welchem Ort wie lange gespeichert und unter welchen Bedingungen zur nachhaltigen Nutzung bereitgestellt werden sollen. Dabei ist systematisch zwischen der zeitlich befristeten Archivierung und der zeitunabhängigen Langzeitarchivierung von Forschungsdaten zu unterscheiden. Ein Datenmanagementplan beschreibt die aktuellen oder geplanten Verfahren der Sicherung innerhalb des Projektkontextes und / oder der Auflösung und Übergabe des Datenbestandes zur langfristigen Sicherung, Pflege und Bereitstellung der Daten, der Metadaten und anderen Forschungsergebnissen. Projektspezifische und fördererrelevante Kriterien, auf deren Grundlage die Auswahl von Daten und Dokumentationen und deren institutionelle Speicherung entschieden wird, sind entsprechend darzulegen (DFG, 2009). Die fachlichen, organisatorischen, rechtlichen, finanziellen und technischen Beurteilungs- und Auswahlkriterien des gewählten Trägers der Langzeitsicherung sind entsprechend zu berücksichtigen (ICPSR, 2009c; Hänger, Huth & Wiesenmüller, 2010). Dies betrifft etwa standardisierte inhaltliche, formale und technische Prüfungen der Daten und Dokumentationen oder Konversionen von Dateien in Formate der Langzeitsicherung. Die datenhaltende Einrichtung muss ihrerseits die Sicherheit, Qualität und Verfügbarkeit der aufgenommenen Studie auf Grundlage disziplinnaher Qualitätsstandards (Dobratz & Schoger, 2010) dauerhaft gewährleisten. Zu grundsätzlichen Fragen der Beurteilung und Auswahl (*Appraisal* and *Selection*) von Forschungsdaten im Rahmen langfristiger Archivierungsstrategien von Forschergruppen und Archiveinrichtungen stellt beispielsweise das *Digital Curation Centre* (DCC) in Großbritannien ausführliche Informationen bereit (DCC, 2010a).

Mit der weiteren Verwertung und Nutzung der Daten und Metadaten nach dem Ende eines Projektes ist bereits während der Projektvorbereitung ein „effektives **Rechtemanagement**“ (Veddern, 2004) einzuplanen. Im Interesse der Rechtssicherheit und -klarheit zwischen Datengeber(n) und Datenserviceeinrichtung sollten die relevanten rechtlichen Aspekte mit einem Archivierungs- und Nutzungsvertrag geregelt werden. Entsprechende Verträge sollten auch eine Nachfolgeplanung für diese Datenbestände vorsehen, falls Datengeber nicht

mehr erreichbar sind oder die datenhaltende Institution aufgelöst wird (ICPSR, 2009d). Zwei wichtige Themen, die in einen DMP zu berücksichtigen sind, werden im Folgenden grob umrissen.

Für Daten, die aufgrund ihres Informationsgehaltes eines besonderen Schutzes vor missbräuchlicher Verwendung bedürfen (z. B. Mikrozensus), müssen entsprechende Gesetze berücksichtigt werden. Prinzipiell muss rechtlich geklärt sein, ob die Daten überhaupt für eine weitere wissenschaftliche Nutzung herangezogen werden dürfen. Zur Einhaltung der Bestimmungen zum **Datenschutz**[1] [Metschke & Wellbrock, 2002] sind fachliche, organisatorische und technische Maßnahmen zum Schutz der Person bei der Erhebung, Verarbeitung, Sicherung und Weitergabe ihrer Daten und deren Durchsetzung zu ergreifen. Anschließend sind Maßnahmen für eine solche Nutzung, wie etwa die Anonymisierung oder Pseudonymisierung, die Einschränkung des Datenzugriffs oder die ausschließliche Nutzung in Räumlichkeiten der datenhaltenden Einrichtung, zu planen. So informiert z. B. die Technologie- und Methodenplattform für die vernetzte medizinische Forschung (TMF e.V.) u. a. über vielfältige rechtlichen und ethischen Anforderungen und Lösungen in diesem Arbeitsfeld (TMF, 2010). Das auch der Datenzugang zu naturwissenschaftlichen Daten nicht voraussetzungslos offen ist, zeigen etwa die „*Special cases*“ der Datenregelungen zum internationalen Polarjahr 2007–2008 (IPY, 2008).

Bezüglich der zu klärenden Nutzungsrechte muss logisch zwischen der Nutzung von Daten innerhalb eines Projektes und Nutzungsrechten durch Dritte im Zusammenhang mit der Archivierung und weiteren Bereitstellung unterschieden werden. Weiterhin sind Rechte von Dritten in Datenmanagementplan zu berücksichtigen, etwa wenn Messinstrumente benutzt werden, die bestehende Schutzrechte geistiger (z B. Urheberecht) oder gewerbliche Art (Patente) berühren. Überlegungen, wie entsprechende Nutzungsrechte im Projekt realisierbar sind, sollten auch berücksichtigen, wie bzw. ob diese Rechte eine geplante Archivierung bzw. die weitere Nutzung der Daten nach Projektende beeinflussen (ICPSR, 2009d; Veddern, 2004). Hilfestellungen zur Regelung der Nutzung und Verwertung von Forschungsdaten im Rahmen von Lizenzierungsmodellen bietet etwa der Leitfadens „*How To License Research Data*“ des *Digital Curation Center* (DCC, 2011). In Nutzungsverträgen mit Langzeitarchiven können Forscher durch Zugangskategorien vereinbaren, dass der offene Datenzugang für begrenzte Zeiträume ausgeschlossen oder zustimmungspflichtig ist, um z. B. die Erstpublikation von Forschungsergebnissen vorzubereiten (GESIS, 2010). Auf Basis der Vereinbarungen zur Datennutzung und Datenweitergabe ist auch zu regeln, ob und unter welchen rechtlichen Bedingungen Daten und Metadaten auch über Datenkataloge und Datenportale öffentlich zugänglich und nutzbar

1. Vgl. den Beitrag von Spindler/ Hillegeist in Kapitel 2.2.

gemacht werden. So nutzt etwa die Helmholtz Gemeinschaft *Creative Commons* Lizenzen, um den offenen Zugang zu Forschungsdaten etwa über die *Scientific Drilling Database* (SDDB) zu fördern (SDDB, 2008).

2.3.3 Wege zur Gestaltung fachspezifischer Datenmanagementpläne

Die beschriebenen Elemente eines Datenmanagementplans unterliegen fach- und projektspezifischen Bedingungen und Gewichtungen. Die konkrete Umsetzung eines Datenmanagementplans ist eng mit der Verfügbarkeit von Dateninfrastrukturen sowie hinreichenden Richtlinien zur Datenpolitik, praktischen Leitfäden und effektiven Werkzeugen verbunden. Europäische Dateninfrastrukturentwicklungen und Projekte, die sich mit domainspezifischen Datenmanagementfragen befassen, werden z. B. in den Berichten *„Infrastructure Planning and Data Curation“* (Ruusalepp & Pryor, 2008) und dem *„Report on Data Management“* (e-IRG, 2009) ausführlich vorgestellt. Mit der Entwicklung und Differenzierung deutscher Dateninfrastrukturen wird die Erstellung von fachspezifischen Konzepten zum Datenmanagement öffentlich gefördert. Bezüglich der zu berücksichtigenden „Unterschiede der wissenschaftlichen Disziplinen“ stellt die Allianz der deutschen Wissenschaftsorganisationen fest:

> „Formen und Bedingungen des Zugangs zu Forschungsdaten müssen gesondert für die jeweiligen Fachdisziplinen unter Berücksichtigung der Art und Weise der Datenerhebung, des Umfangs und der Vernetzbarkeit des Datenmaterials sowie der praktischen Brauchbarkeit der Daten entwickelt werden. Zugleich ist den jeweiligen Lebenszyklen und Nutzungsszenarien der Daten in dem konkreten Forschungsfeld Rechnung zu tragen.“ (Allianz, 2010)

So präsentiert die Konzeptstudie „Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie“ zentrale Themen des Managements von Forschungsdaten aus der Chemie von der Erstellung, über relevante Metadatenstandards bis hin zur Langzeitarchivierung (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010). Das Infrastrukturprojekt „wibaklidama“ beschreibt Umfeld, Anforderungen und Aufgabenfelder beim Management von Klimadaten (Wibaklidama, 2009). Im Projekt BeLab (Beweissicheres elektronisches Laborbuch) wird ein „ein Konzept für die beweissichere elektronische Langzeitarchivierung (LZA) von Forschungsprimärdaten für lange Zeiträume“ entwickelt, mit dem experimentell erzeugte Forschungsdaten entlang des Datenzyklus digital erfasst und dauerhaft gesichert werden (BeLab, 2010).

Wichtige Informationsquellen zum Forschungsdatenmanagement stellen auch die Leitlinien für verschiedenste Fachdisziplinen aus anderen Ländern breit. Die *Association of Research Libraries* (ARL) in den USA bietet zahlreiche fachspe-

zifische Informationen zu Datenmanagementplänen mit Bezug auf die Anforderungen der *National Science Foundation*, z. B. zu Geowissenschaften, Physik und Sozialwissenschaften (ARL, 2010). Das *Digital Curation Center* (DCC) unterstützt nationale Forschergruppen bei der Erstellung von fachbezogenen Datenmanagementplänen in Verbindung mit den spezifischen Anforderungen von britischen Fördereinrichtungen (DCC, 2010b). Die „*Checklist for a Data Management Plan*" fasst anhand von zehn Kernthemen praxisrelevante Aspekte bei der Erstellung eines Datenmanagementplans zusammen (DCC, 2011b). Zum Management von Forschungsdaten im universitären Umfeld entwickelt u. a. die Universität Melbourne Strategien, Ausbildungskonzepte und praktische Handlungshilfen für unterschiedliche Nutzergruppen und unterstützt die Erstellung von Datenmanagementplänen durch strukturierte Anleitungen und Prüflisten (Melbourne University 2007, 2009a, 2009b).

Die Bedeutung des Managements von Forschungsdaten und die Rolle von Datenmanagementplänen hat die Kommission Zukunft der Informationsinfrastruktur im Bericht „*Gesamtkonzept für die Informationsinfrastruktur in Deutschland*" für das Handlungsfeld Forschungsdaten herausgestellt (KII, 2011). Analysen, Handlungsbedarfe (S. B801 ff.) und inhaltliche Empfehlungen (S. 52) beschreiben ein breites Aktionsspektrum, um u. a. Organisation, Finanzen, Technik und Recht in diesem Arbeitsfeld den Anforderungen der Zukunft anzupassen und die Wissensvernetzung durch internationalen Datenaustausch zu fördern.

Literaturhinweise

Allianz Initiative, 2010. *Grundsätze zum Umgang mit Forschungsdaten*. Online: http://www.allianzinitiative.de/fileadmin/user_upload/Home/Video/Grunds%C3%A4tze%20Umgang%20mit%20Forschungsdaten.pdf [Zugriff am 28.07.2011].

ARL (Association of Research Libraries), 2010. *Resources for Data Management Planning*. Online: http://www.arl.org/rtl/eresearch/escien/nsf/nsfresources.shtml#nsfguidancedmp [Zugriff am 28.07.2011]

BeLab (Beweissicheres elektronisches Laborbuch), 2011. *Projektseite.* Online: http://www.belab-forschung.de/belab/ [Zugriff am 28.07.2011].

DataCite, o. J. Online: *Home*. http://datacite.org/ [Zugriff am 28.07.2011]

DCC (Digital Curation Center), 2010a. *Appraise & Select Research Data for Curation*. Online: http://www.dcc.ac.uk/resources/how-guides/appraise-select-research-data [Zugriff am 28.07.2011].

DCC (Digital Curation Center), 2010b. *Data Management Plans*. Online: http://www.dcc.ac.uk/resources/data-management-plans [Zugriff am 28.07.2011].

DCC (Digital Curation Center), 2011. *How to License Research Data*. Online: http://www.dcc.ac.uk/resources/how-guides/license-research-data [Zugriff am 28.07.2011].

DFG (Deutsche Forschungsgemeinschaft), 2009. *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. (Jan. 2009) Online: http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf [Zugriff am 28.07.2011].

DFG (Deutsche Forschungsgemeinschaft), 2010. *DFG-Vordruck 1.02 – 8/10 V.1.02*. Teil „Leitfaden für die Antragstellung" Abschnitt 3.7, S. 32. Online: http://www.dfg.de/download/formulare/1_02/1_02.pdf [Zugriff am 28.07.2011].

Dobratz S. & Schoger, A., 2010: 5.2 Grundkonzepte der Vertrauenswürdigkeit und Sicherheit. In: H. Neuroth et al., Hrsg. 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. (Version 2.3) Online: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949 [Zugriff am 18.08.2011].

e-IRG (e-Infrastructure Reflection Group), 2009. *Data Management Task Force 2009*. Report on Data Management. Online: http://www.e-irg.eu/images/stories/e-irg_dmtf_report_final.pdf [Zugriff am 28.07.2011].

GESIS, 2010. *Beratung und Unterstützung bei der Archivierung eigener Studien*. Online: http://www.gesis.org/unser-angebot/archivieren-und-registrieren/datenarchivierung/ [Zugriff am 28.07.2011].

Hänger, A. Huth, K. & Wiesenmüller, H., 2010. 3.5 Rahmenbedingungen für die LZA digitaler Objekte. In: H. Neuroth et al., Hrsg 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* (Version 2.3) Online: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949 [Zugriff am 18.08.2011].

IBF (Information Network on Biological Research Data gained in the Field up to the Sustainable Storage in a Primary Data Repository), 2010. *Projekt „Aufbau eines Informationsnetzes für biologische Forschungsdaten von der Erhebung im Feld bis zur nachhaltigen Sicherung in einem Primärdatenrepositorium*". Projektseite. Online: http://www.diversitymobile.net/wiki/IBF_Project [Zugriff am 28.07.2011].

ICPSR (Inter-University Consortium for Political and Social Research), 2009a. Guide *to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle.* 4th ed. Ann Arbor, MI. Online: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf [Zugriff am 28.07.2011]

ICPSR (Inter-University Consortium for Political and Social Research), 2009b. *Elements of a Data Management Plan*. Online: http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/dmp/elements.html [Zugriff am 28.07.2011].

ICPSR (Inter-University Consortium for Political and Social Research), 2009c. *Digital Curation.* Selection and Appraisal. Online: http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/selection.jsp [Zugriff am 28.07.2011].

ICPSR (Inter-University Consortium for Political and Social Research), 2009d. *Framework for Creating a Data Management Plan*. Online: http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/dmp/framework.html [Zugriff am 28.07.2011].

IPY (International Polar Year), 2008. *International Polar Year 2007–2008 Data Policy.* Online: http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf [Zugriff am 28.07.2011].

KII (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. (April 2011) Online: http://www.leibniz-gemeinschaft.de/?nid=infrastr [Zugriff am 28.07.2011].

Melbourne University, 2007. *Records Management Advice*. (Stand: 05.05.2011) Online: http://www.unimelb.edu.au/records/ [Zugriff am 28.07.2011].

Melbourne University, 2009a. *Research Integrity – Research Data Management Tool Box*. (Stand: 10.09.2010) Online: http://research.unimelb.edu.au/integrity/conduct/data/review [Zugriff am 28.07.2011].

Melbourne University, 2009b. *Research Data Management for researchers*. (Stand: 21.06.2011) Online: http://www.eresearch.unimelb.edu.au/activities/research_data_management_for_researchers [Zugriff am 28.07.2011].

Metschke, R. & Wellbrock R., 2002. *Datenschutz in Wissenschaft und Forschung. Materialien zum Datenschutz Nr. 28*. 3. überarbeitete Aufl. Berlin. (Dez. 2002) Online: http://www.datenschutz-berlin.de/attachments/47/Materialien28.pdf?1166527077 [Zugriff am 28.07.2011].

OECD (Organisation for Economic Co-Operation and Development), 2007. *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Online: http://www.oecd.org/dataoecd/9/61/38500813.pdf [Zugriff am 28.07.2011].

Ruusalepp, R. & Pryor, G., 2008. *Infrastructure Planning and Data Curation*. Online: http://www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf [Zugriff am 28.07.2011].

SDDB (Scientific Drilling Database), 2008. *The SDDB data policy explained*. Online: http://www.scientificdrilling.org/front_content.php?idcat=239 [Zugriff am 28.07.2011].

TIB (Technische Informationsbibliothek) Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010. *Konzeptstudie „Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie“*. Online: http://www.tib-hannover.de/fileadmin/projekte/primaer-chemie/Konzeptstudie_Forschungsdaten_Chemie.pdf [Zugriff am 28.07.2011].

TMF (Telematikplattform für Medizinische Forschungsnetze), 2010. *Rechtliche und ethische Rahmenbedingungen für die vernetzte medizinische Forschung*. Online: http://www.tmf-ev.de/Themen/Rahmenbedingungen.aspx [Zugriff am 28.07.2011].

UK Data Archive, 2011. *Data Management Costing Tool*. Online: http://www.data-archive.ac.uk/media/257647/ukda_jiscdmcosting.pdf [Zugriff am 28.07.2011].

Van den Eynden, V. et al., 2011. *Managing and Sharing Data: Best Practice For Researchers*. 3. ed. Online: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf. [Zugriff am 28.07.2011].

Veddern, M., 2004. *Multimediarecht für die Hochschulpraxis. Ratgeber zum Urheberrecht, Patentrecht und Onlinerecht mit Verträgen, Verwertungsmodellen und Rechtemanagement*. Centrum für eCompetence in Hochschulen NRW. 2. Aufl. Online: https://eldorado.uni-dortmund.de/bitstream/2003/21358/1/veddern.pdf. [Zugriff am 28.07.2011].

Wibaklidama, 2009. *Projekt „Wissensbasiertes Klima-Datenmanagement“*. Online: http://wibaklidama.fh-potsdam.de/index.php?id=29 [Zugriff am 28.07.2011].

2.4 Metadaten und Standards

Uwe Jensen, Alexia Katsanidou, Wolfgang Zenk-Möltgen
GESIS – Leibniz-Institut für Sozialwissenschaften

Standardisierte Metadaten sind eine notwendige Voraussetzung für die Dokumentation und dauerhafte Sicherung von Forschungsdaten. Als Werkzeug fördern sie nachhaltig die Erschließung und Nutzung datenbasierter Forschungsergebnisse.

Metadaten sind Daten oder Informationen, die in strukturierter Form analoge oder digitale Forschungsdaten (Objekte) dokumentieren. Sie beschreiben, erklären, verorten oder definieren Objekte, Ressourcen und Informationsquellen für die Wissenschaft. Hierdurch helfen sie, Forschungsdaten zu managen, zu erschließen, zu verstehen und zu benutzen (NISO, 2004).

Im Folgenden werden Forschungsdaten vereinfacht als das Objekt verstanden, das von einer übergeordneten Ebene aus mit Metadaten beschrieben wird. Dabei ist die Heterogenität von Daten und Metadaten zwischen den wissenschaftlichen Disziplinen und auch innerhalb von Fachdisziplinen für die Standardisierung eine nicht zu übersehende Herausforderung. Auch deshalb erscheint es in disziplinübergreifenden Kontexten besonders sinnvoll, den fachspezifischen Kontext der Datenentstehung und die Charakteristika dieser Forschungsdaten möglichst eindeutig herauszustellen. Dabei stellt das Grundmuster, Daten und Metadaten entlang des Forschungsdatenzyklus zu organisieren (UK Data Archive, 2011), eine wichtige verbindende Klammer dar. Dieses Prinzip ist für die systematische Auswahl und Anwendung von Metadaten erforderlich, denn die Anzahl der Metadatenstandards ist überwältigend und ihre Beziehungen untereinander verkomplizieren die Sache weiter (Riley, 2009a).

2.4.1 Studiendesign und Erhebung von Forschungsdaten

Die ganzheitliche Planung einer Studie ist wichtig, um alle relevanten Ebenen, wie z. B. Forschungsfrage, Theorierahmen, Forschungsstrategie, Design der Datenerhebung und Analyse der Forschungsdaten zu berücksichtigen. Dabei spielen Metadaten und Standards eine besondere Rolle, weil sie maßgeblich dazu beitragen, die Qualität der Studie zu sichern. So definiert etwa ISO 20252:2006 einheitliche Qualitätskriterien in der Markt-, Meinungs- und Sozialforschung. Die konkrete Projektplanung ist oftmals in spezifische Qualitätssicherungssysteme eingebunden. So werden Daten der amtlichen Statistik durch Qualitätsstandards abgesichert und durch Qualitätsberichte dokumentiert (Körner & Schmidt, 2006). Gleichzeitig entwickelt die wissenschaftliche Praxis Empfehlungen und Standards zur Dokumentation von Studien und Daten (ICPSR, 2009a). Dazu zählen die Dokumentation der Studie als Ganzes durch

eine detaillierte Studienbeschreibung, die Dokumentation der Verfahren und Instrumente der Datenerhebung sowie der Charakteristika der Samples (Studienebene), die Dokumentation der Struktur und Eigenart der Datensätze (Datensatzebene) sowie die Beschreibung der Variablen einer Datendatei (Variablenebene).

Darüber hinaus werden disziplinspezifische Metadaten zur **Instrumentenentwicklung** bzw. zur Durchführung der Messung in entsprechenden Berichten oder Protokollen festgehalten. So ist die Entwicklung eines sozialwissenschaftlichen Fragebogens bereits ein langer Prozess, während dessen mehrere Versionen erstellt werden, bevor das Messinstrument erprobt und im Feld angewendet werden kann. Die Dokumentation solcher Versionierungsprozesse und der Kommentare der Primärforscher durch entsprechende Tools (Hopt et al., 2010) sind als Metainformationen etwa bei methodischen Fragen zur Datenqualität sehr hilfreich. Gleiches gilt etwa für die Entwicklung und Standardisierung von Antwortskalen, die umfangreiche Tests benötigen, bevor sie endgültig geeicht sind und z. B. in Skalenhandbüchern (ZIS, 2010) oder in umfangreichen technischen Berichten (PISA, 2006) dokumentiert werden.

Die **kulturellen Kontexte** und Fragen der Mehrsprachigkeit stellen Primärforscher in international vergleichenden Einstellungsuntersuchungen vor besondere Herausforderungen. Hier sind die verschiedenen Facetten der Studien- und Datendokumentationen je Sprache und die integrierten Datensätze durch umfangreiche Metadaten qualitativ hochwertig, adäquat und nutzerfreundlich (Jinfang Niu, 2009) zu dokumentieren. Dazu zählen auch ausführliche Beschreibungen z. B. der länderspezifischen Stichprobenziehung, der Übersetzungsprozesse während der Fragebogenentwicklung und entsprechender Pretests (Survey Research Center, 2010).

Metadaten aus der Phase der **Datenerhebung** beschreiben einen z. T. nicht reproduzierbaren Kontext, der sehr bedeutsam für die Auswertung und Interpretation der Daten ist. Hierzu gehören auch Paradaten, die den Prozess der Datenerhebung selbst dokumentieren. Dieses können automatisch generierte Daten von Computer-unterstützten Umfragen (z. B. *call record data* oder *keystrokes*) oder auch Informationen sein, die von Interviewern oder anderen technischen Systemen erfasst wurden (z. B. *digital audio recording*) (Couper, 1998). Ein typisches Beispiel für Paradaten sind „*call records*“, die Datum und Uhrzeit eines Kontakts und das Ergebnis (z. B. Interview, Ablehnung) erfassen. Paradaten erlauben eine unabhängige Evaluation des Nichtstichprobenfehlers in Umfragen. Sie können Nonresponse-Korrekturen verbessern und dabei auch die Wahrscheinlichkeit voraussagen, von einer kontaktierten Person ein Interview zu erhalten (Kreuter et al., 2010).

2.4.2 Auswertung von Forschungsdaten

Die erste Datenauswertung wird üblicherweise von den Primärforschern einer Studie durchgeführt, die dazu ihre umfangreichen informellen Kenntnisse über die erhobenen Daten nutzen können. Sekundäre Nutzer hingegen brauchen formale Metadaten für die Beschreibung der Daten, um die Daten verstehen zu können. Daher sollten auch Datenanalysen durch entsprechende Metadaten nachvollziehbar abgesichert und dokumentiert werden. Die Art und Weise, wie Daten analysiert werden und welche Analysemethoden dabei eingesetzt werden, stellen notwendige Informationen über den Analyseprozess dar. Insbesondere sollte die Kodierung der Felddaten in der statistischen Analyse etwa durch Syntaxfiles als eine Form von Metadaten betrachtet werden, die u. a. auch die Qualität der Analysen dokumentiert und im Sinne guter wissenschaftlicher Praxis entsprechende Replikationen erlauben.

In diesem Zusammenhang besitzt die Dokumentation der Prozesse zur **Harmonisierung** von Daten einen besonderen Stellenwert. Datenharmonisierung bedeutet die Transformation von Daten aus unterschiedlichen Quellen in standardisierte Maße, die eine vergleichende Auswertung erlauben. Harmonisierung bedeutet meistens auch die Herstellung von „*conversion keys*“ (Umrechnungsschlüssel), mit denen Werte eines oder mehrerer Quellvariablen in neue Werte einer standardisierten Zielgröße transformiert werden. Beispiele aus den Sozialwissenschaften sind Bildungs- und Einkommensvariablen, die unterschiedliche Ausprägungen in verschiedenen Ländern haben. Ohne Harmonisierung könnten vergleichende Analysen hierzu nicht durchgeführt werden. Dieser Prozess sollte dokumentiert und die Metadaten sollten verfügbar gemacht werden (Quandt et al., 2009). Metadaten zur Erhebung und Analyse von Forschungsdaten, auch in Form von Computercode, werden in Zukunft an Bedeutung gewinnen, etwa wenn Datensätze als eigenständige Publikation veröffentlicht werden (Hanson et al., 2011). Gleichzeitig stellen sie einen Teil der Informationen bereit, um die Daten nachvollziehbar für (Re)-Analysen zu nutzen. Dabei sind internationale transdisziplinäre Forschungsprojekte mit noch weitergehenden Anforderungen und Problemen konfrontiert, z. B. wenn anhand langer Zeitreihen und heterogener Datenbestände aus unterschiedlichen Fachdisziplinen ein hochkomplexer Forschungsgegenstand untersucht werden soll. So beschreibt etwa der *Study Implementation Plan* des 2006 von ESSP (*Earth System Science Partnership*) und WHO (*World Health Organisation*) initiierten Projektes „*Global Environmental Change and Human Health*“ eine Reihe von kritischen Schwachstellen bei der Nutzung von vorhandenen Datenbeständen (Confalonieri & McMichael, 2007).

Es bleibt festzuhalten, dass Metadaten über das Studiendesign und den Prozess der Datenerstellung und ihre weitere Aufbereitungsgeschichte oft lückenhaft oder unstrukturiert gesammelt oder veröffentlicht werden, wie etwa eine Umfrage von Bose und Frew (2005) zeigt. Auch deshalb ist ein strukturiertes

Datenmanagement für Forschungsdaten generierende Projekte wichtig, weil sie damit die verschiedenen Aspekte eines verantwortungsvollen Umgangs mit Daten und Metadaten planen, überprüfen und dokumentieren können. Vor diesem Hintergrund haben auch Einrichtungen zur Forschungsförderung damit begonnen, Förderzusagen mit der Erstellung von Datenmanagementplänen zu verknüpfen (NSF, 2010; DFG, 2010).[1]

2.4.3 Dokumentation von Forschungsdaten

Für die Dokumentation von Forschungsdaten sind – neben der Beschreibung der Genese der Daten durch ein Projekt und dessen Erhebungsinstrumente – die Bedeutung der Daten selbst zentral. Für die Reproduktion einer Datenanalyse oder die Erstellung von Sekundäranalysen muss die Bedeutung von Messreihen und -werten klar definiert werden. Dabei geht es vor allem um die Bedeutung der oft numerischen, aber manchmal auch alphanumerischen Werte eines Datensatzes. Hier sind die disziplinspezifischen Unterschiede sehr groß, da es um die Dokumentation von Messwerten geht, die sachlich unterschiedliche Bezüge und Einheiten haben. So sind in den Naturwissenschaften Messungen etwa der Temperatur in Grad Celsius oder Grad Kelvin, in der Medizin die Messung des Blutzuckergehalts in Millimol pro Liter oder in Milligramm pro Deziliter bekannt, die zu jeweils unterschiedlichen Interpretationen des Datenwerts führen. Für die Volkswirtschaften gibt es z. B. unterschiedliche Indikatoren für die Messung des Wirtschaftswachstums, wie etwa das nominale und reale Bruttoinlandsprodukt als hoch aggregierte Größen, die auf weiteren Werten beruhen. Hier müssen die Definitionen der gewählten Einheiten sehr klar sein oder durch die Metadatendokumentation deutlich gemacht werden. In den Sozialwissenschaften ist es bei der Messung mithilfe von Umfragen nötig, die gestellten Fragen ausführlich zu dokumentieren, da hier kleine Abweichungen große Effekte auf die Antworten haben können. Darüber hinaus ist es üblich, für abstraktere Einstellungen, wie etwa die Meinung zur gegenwärtigen Regierung, Skalen zu bilden. Hierbei ist es für die Interpretation der Werte unerlässlich, eine genaue Dokumentation zur Verfügung zu haben.

Die unterschiedlichen fachspezifischen Anforderungen haben zur Herausbildung verschiedener **Metadaten-Frameworks** geführt, die Inhalte von Metadaten auf verschiedenen Ebenen spezifizieren. Ein Metadaten-Framework kann dabei als ein System sicher ergänzender Standards, Normen und kontrollierter Vokabulare beschrieben werden. Zum Teil handelt es sich um rein inhaltliche Festlegungen für die Metadaten, zum Teil werden bis auf die technische Ebene Formate für die Datendokumentation definiert, wie die Metadaten erfasst und

1. Vgl. Kapitel 2.3 in diesem Band.

ausgetauscht werden können. Damit werden innerhalb der Disziplinen die Ziele der Interoperabilität und Nachvollziehbarkeit erreicht.

Ein Beispiel aus den Sozialwissenschaften für ein konsistentes Metadatensystem zur Dokumentation von Forschungsdaten ist das DDI Format der *Data Documentation Initiative* (DDI). DDI ermöglicht es, alle Metadaten aus den verschiedenen Stadien des Datenlebenszyklus strukturiert zu erfassen (Vardigan et al., 2008). Forschungsdaten werden mit DDI bezüglich ihrer Konzepte, Erhebung, Verarbeitung, Verteilung, Exploration, Analyse, Wiederverwendung und Archivierung beschrieben. Neben Umfragen können auch komplexe Zeitreihen aus nationalen und internationalen Vergleichsstudien, Indikatoren und andere Aggregatdaten sowie geographische Daten beschrieben werden. Die DDI-Spezifikation ist als XML-Schema (Gregory et al., 2009) definiert und ist kompatibel zu anderen Metadatenstandards, etwa mit ISO/IEC 11179 oder Dublin Core. Durch seine stark modular aufgebaute Struktur und die Verwendung von eindeutig identifizierbaren Elementen, versionierbaren Elementen und ganzen Strukturen, die gepflegt werden können („*maintainables*“), unterstützt DDI vor allem die Wiederverwendung und die Vergleichbarkeit von Metadaten.

Ein komplexes Beispiel für ein Metadaten-Framework aus dem Bereich der Klimaforschung ist das Daten- und Metadatenmodell CERA-2 (*Climate and Environmental Retrieval and Archive*). Hier werden Metadaten von Simulationsdaten zur Klimaentwicklung nach den Regeln des WDC-C (*World Data Center for Climate*) festgelegt, um die Daten aus verschiedenen Instituten einheitlich zu beschreiben und miteinander verbinden zu können. Die CERA-2 Metadaten bestehen aus einer Beschreibung des Experiments und einer Beschreibung des Datensatzes. Dabei werden das Projekt und die beteiligten ForscherInnen sowie der Zeitraum, die erhobenen Variablen und ihre Codierung, sowie Struktur und Format der Daten beschrieben. Zur Beschreibung der Variablen werden die CF Konventionen (*Climate and Forecast Metadata Convention*) eingesetzt, so dass über Standardnamen die Vergleichbarkeit der Daten ermöglicht wird. Die Datensatzbeschreibung nimmt auch Bezug auf Skalen oder Codelisten sowie zu Verfahren zur Erzeugung der Daten und ihrer Qualität. Die Daten erhalten *Persistent Identifier*, werden in einem Langzeitarchiv gesichert und im Fall von Veränderungen durch Errata oder neue Datenversionen ergänzt (Toussaint, 1999). Aktuell beteiligt sich eine Arbeitsgruppe des für das WDC-C zuständige deutsche Klimarechenzentrums an dem EU geförderten Projekt METAFOR (*Common Metadata for Climate Modelling Digital Repositories*), das sich die Entwicklung eines gemeinsamen Informationsmodells für die Klimawissenschaften zum Ziel gesetzt hat.

Weitere Beispiele für Metadatenspezifikationen sind ISO/TS 17369:2005 (*Statistical Data and Metadata Exchange SDMX*), das weitverbreitet von statischen Ämtern eingesetzt wird, ISO 19115:2003 für Geographische Information und Dienste sowie ISO/IEC 11179:2003 für Metadaten-Registraturen und

Semantik (Gregory et al., 2009). Aus dem Bereich der Umfrageforschung sind auch die Standards von Triple-S für einen technischen Austausch von Metadaten und die ISO 20252:2006 für die Definition einheitlicher Qualitätskriterien in der Markt-, Meinungs- und Sozialforschung bekannt.

In Ergänzung zum Einsatz internationaler Standards zur Datendokumentation wird zusätzlich die Anwendung von nationalen oder internationalen **Klassifikationssystemen** empfohlen, z. B. ISO Normen für geografische Einheiten (ISO 3166) und Sprachen (ISO 639-1:2002) oder Codes für Berufe (ISCO). Für die Katalogisierung von Forschungsdaten werden entweder fachspezifische oder fachübergreifende Klassifikationen angewendet, etwa die Dewey-Dezimalklassifikation (DDC) zur disziplinübergreifenden Erschließung oder fachspezifische eingesetzt (z. B. *Klassifikation Sozialwissenschaften*, GESIS, 2006).

Darüber hinaus stellen **kontrollierte Vokabulare** einen weiteren wichtigen Baustein zur standardisierten Dokumentation von unterschiedlichsten Objekten und deren Eigenarten bereit. Durch die gezielte Erschließung von Forschungsdaten mithilfe von standardisierten Vokabularen, etwa zur Methodik einer Umfrage (Art, Häufigkeit, Stichprobe der Befragung), wird die Vergleichbarkeit von Datensätzen wesentlich verbessert. Eine besondere Bedeutung haben hier mehrsprachige Vokabulare, wie etwa der mehrsprachige ELSST (*European Language Social Science Thesaurus*), der von europäischen Datenarchiven zur Dokumentation und Erschließung sozialwissenschaftlicher Forschungsdaten (*CESSDA Data Catalogue)* eingesetzt wird. Im Zusammenhang mit der Entwicklung von Datendokumentationsstandards wie dem DDI Standard wird auch die entsprechende Entwicklung und Anwendung kontrollierter Vokabulare vorangetrieben (Jääskeläinen et al., 2009). Die in den Codelisten benutzten Vokabulare werden dazu im Standardformat Genericode erfasst. Genericode ist eine Spezifikation des OASIS *Code List Representation TC*, mit der ein Standardmodell und die Präsentation in XML für kontrollierte Vokabulare bereitstellt wird. Eine weitere Möglichkeit der Repräsentation kontrollierter Vokabulare aus dem Bereich Semantik Web ist das *Simple Knowledge Organization System* (SKOS), das im August 2009 als neuer Standard durch das W3C veröffentlicht wurde.

Für die Erstellung umfangreicher Metadatendokumentationen ist es häufig unerlässlich, über **Managementsysteme** oder Software zur Pflege der Metadaten zu verfügen. Diese helfen bei der Eingabe, Verwaltung und Versionierung der Metadaten. Solche Anwendungen sind meist stark an die Bedürfnisse des Fachgebiets angepasst und benutzen proprietäre Datenformate. Für die **Interoperabilität** von Daten und Metadaten sollten die Anwendungen jedoch Standardformate verwenden, die nicht proprietär sind und die einfache Wiederverwendung von Forschungsdaten ermöglichen. Solche Standards sollten als offene Standards von der wissenschaftlichen Gemeinschaft getragen werden, so wie etwa der DDI Standard in den Sozialwissenschaften oder die *Network Common Data Form* (NetCDF), die in der Erdsystemwissenschaft angewendet wird.

2.4.4 Publikation, Zitation und Recherche von Forschungsdaten

Für die Publikation von Analyse-Ergebnissen auf der Basis von Forschungsdaten werden in der Regel Zeitschriftenaufsätze oder Monographien erstellt. Diese werden mit den bekannten Metadaten für Literatur dokumentiert, für die schon lange Standards wie DublinCore, MARC (*Machine-Readable Cataloguing*), MODS (*Metadata Object Description Schema*), METS (*Metadata Encoding & Transmission Standard*) oder Z39.50 (*Information Retrieval Protocol*) etabliert sind.

Für die Verknüpfung der Publikationen mit den verwendeten Forschungsdaten haben sich bereits einige Initiativen gebildet, die eindeutige und dauerhafte **Identifikatoren** (*Persistent Identifiers*, PI) für digitale Daten vergeben. Als technische Lösungen haben sich verschiedene Systeme etabliert, etwa das Handle System der CNRI (*Corporation for National Research Initiatives*), oder das darauf aufbauende DOI System der *International DOI Foundation*. Weitere PI Systeme sind das URN System, welches durch die IETF (*Internet Engineering Task Force*) definiert ist und für den Namensraum URN:NBN durch Nationalbibliotheken verwendet wird, sowie PURL (*Persistent Uniform Resource Locators*) getragen vom OCLC (*Online Computer Library Center*) und Zepheira.

Die bereits für Literatur benutzten und weit verbreiteten DOI Namen werden von der Initiative DataCite verwendet, um auch Forschungsdaten eindeutig und dauerhaft identifizierbar und zitierbar zu machen. DataCite hat 19 Mitgliedsinstitutionen aus Europa, Nordamerika, Asien und Australien, die es sich zum Ziel gesetzt haben, den Datenpublikationsagenten in ihren Fachbereichen die Vergabe von DOI Namen für Forschungsdaten zu ermöglichen. Neben der wissenschaftlichen Replikationsmöglichkeit der Analysen steht bei DataCite das Ziel im Vordergrund, der Leistung von Wissenschaftlern bei der Herstellung von Forschungsdatensätzen durch die Zitierfähigkeit höhere Sichtbarkeit zu verleihen. Dafür vergeben die datenhaltenden Einrichtungen DOIs und verpflichten sich zur Pflege der Metadaten. Dabei wird die DOI mit einer URL verbunden, so dass die zitierten Datensätze schnell über das Web aufgefunden werden können. Diese Trennung von *Persistent Identifier* und dem eigentlichen Ort des digitalen Objekts ermöglicht die langfristige Stabilität der Zitation. Gleichwohl sichert es auch die Möglichkeit des Zugriffs auf die Daten, solange der verantwortliche Publikationsagent seiner Pflicht nachkommt die Metadaten aktuell zu halten. Einschränkungen des Zugriffs, deren Gründe auf Seiten des Publikationsagenten vorliegen (z. B. Datenschutz, Beschränkung auf wissenschaftliche Verwendung, Gebühren), können weiterhin vom Publikationsagenten angewendet werden.

In Deutschland sind Mitglieder von DataCite die Technische Informationsbibliothek Hannover (TIB), die Deutsche Zentralbibliothek für Medizin

(ZBMED), die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW) und GESIS – Leibniz-Institut für Sozialwissenschaften. GESIS betreibt mit dem Projekt da|ra (Registrierungsagentur für sozialwissenschaftliche Daten) bereits ein Portal, in dem Datenanbieter DOI Namen für ihre Daten registrieren können und die Metadaten für interessierte Wissenschaftler zur Verfügung stellen. Diese Metadaten enthalten neben den für die Zitation wichtigen Elementen auch inhaltliche und formale Beschreibungen des Forschungsdatensatzes, die Datenerhebung und beteiligte Personen und Institutionen. GESIS hat dazu das von DataCite entwickelte Metadatenschema um spezifische Elemente für die Sozialwissenschaften ergänzt und diese auch mit dem für sozialwissenschaftliche Metadaten etablierten Standard DDI abgeglichen. In diesen Metadaten kann recherchiert werden und die Datenverfügbarkeit (direkter Download, Datenbestellung oder nur Vor-Ort Nutzung) wird angezeigt. Ziel ist es, einen möglichst direkten Zugang zu den Forschungsdaten zu bieten. In Zukunft wird durch eine Kooperation von GESIS mit der ZBW da|ra als einheitliches Portal für die Daten der Sozial- und Wirtschaftswissenschaften zugänglich sein.

In der Zukunft werden Metadatenbestände (*Repositories*) oder Verzeichnisse von Metadatenbeständen an Bedeutung gewinnen, da sie technisch die Recherche nach Forschungsdaten erleichtern. Der bereits etablierte CESSDA *Data Catalogue* zeigt, dass standardisierte Metadaten für die Recherche in Forschungsdatenbeständen eine große Erleichterung darstellen. Die im DataCite Verbund registrierten Daten des Projekts PANGAEA (*Publishing Network for Geoscientific & Environmental Data*) zeigen auch die Möglichkeit der Verbindung von wissenschaftlicher Literatur und den verwendeten Forschungsdaten durch die Verwendung von persistenten Identifikatoren. Weitere in DataCite geplante Anwendungen sind die Erstellung von Zitationsstatistiken für Forschungsdaten, die einen Anreiz zur Herstellung von Datensätzen mit hoher Daten- und Metadatenqualität bieten, und die Integration der Datenregistrierung in den Arbeitsablauf bei der Veröffentlichung von wissenschaftlichen Artikeln.

2.4.5 Langzeitarchivierung von Forschungsdaten

Metadaten und Standards zur Archivierung von Forschungsdaten haben in der Diskussion über deren dauerhafte Verfügbarkeit einen besonderen Stellenwert.

Die Vereinheitlichung von technischen und organisatorischen Abläufen ist eines von vielen Zielen zur Entwicklung von Standards zur langfristigen Sicherung von Forschungsdaten. Bei der Entwicklung von Langzeitarchiven besitzt das **Referenzmodell** für ein *Open Archival Information System (OAIS)* (CCSDS, 2002) seit langem eine zentrale Leitfunktion. Als ISO-Standard (ISO 14721:2003) anerkannt, stellt das Modell einen offenen Standard bereit, der auf Grundlage eines umfassendes Begriffs-, Struktur- und Organisationskonzeptes beschreibt, welche Aspekte zu berücksichtigen sind, damit ein Archiv OAIS

konform ist. Die Art und Weise, wie ein Langzeitarchiv organisatorisch oder technisch zu etablieren ist und welche Daten oder Metadaten dabei berücksichtigt werden, ist dann für die jeweiligen Anwendungskontexte speziell zu entwikkeln. Offen heißt auch, dass der Standard durch internationale Reviewverfahren veränderbar ist oder Anstoß für weitere Entwicklungen geben kann (CCSDS, 2009; Brübach, 2010).

In Anlehnung oder in expliziter Übereinstimmung mit OAIS wurde auf internationaler Ebene auch damit begonnen, Kernanforderungen an vertrauenswürdige digitale Langzeitarchive zu entwickeln (Ten Principles, 2008). Praxisorientierte Normen, Methoden und Werkzeuge zur Beurteilung der **Vertrauenswürdigkeit** von Langzeitarchiven stellen z. B. TRAC (*Trustworthy Repositories Audit & Certification: Criteria and Checklist*), DRAMBORA (*Digital Repository Audit Method Based on Risk Assessment*) und DAS (*Data Seal of Approval*) bereit. Das deutsche Kompetenznetzwerk zur digitalen Langzeitarchivierung NESTOR entwickelte auf Grundlage nationaler und internationaler Arbeitsergebnisse einen *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* (NESTOR, 2008).

Gleichzeitig greifen Forschungsdatenarchive diese Kriterien auf, indem sie den Grad der Übereinstimmung mit dem OAIS Rahmenkonzept oder weiteren qualitätsorientierten Kriterienkatalogen ausweisen. So zeigt etwa das Langzeitdatenarchiv ICPSR mit ihrem *Digital Preservation Policy Framework* anhand von sieben Kriterien (Übereinstimmung mit OAIS, Administrative Verantwortung, Funktionsfähigkeit, Finanzen, Angemessenheit von Prozessen und Technologien, Systemsicherheit, Prüfverfahren), wie es die Bedingungen für ein vertrauenswürdigen Datenarchiv erfüllt (ICPSR, 2007). Auf diesem Hintergrund entstand auch ein umfassender Leitfaden zur Datensicherung, der für das IHSN (*International Household Survey Network*) entwickelt wurde und sich insbesondere an die Verantwortlichen in Statistikämtern richtet (ICPSR, 2009b).

Aus den Anforderungen der Langzeitarchivierung stellt sich die Frage nach den notwendigen Metadaten sowie deren **Standardisierung** und **Interoperabilität** neu.

PREMIS (*Preservation Metadata: Implementation Strategies*) ist ein international anerkannter Standard von Metadaten für die Langzeitarchivierung. Die aktuelle Version 2 des Datenlexikons beschreibt in Form semantischer Einheiten Informationen, die über das Archivsystem und dessen Management bekannt sein sollten, um den Prozess der Langzeitarchivierung zu unterstützen. D.h. auch, dass PREMIS mit seinen Kernelementen nur eine Teilmenge aller Metadaten zur Langzeitarchivierung beschreibt. Gleichzeitig wird nicht vorgegeben, wie diese semantischen Informationen in technischen Systemen repräsentiert werden sollen. Sie können aber mit den Metadatenelementen des seit 2008 vorliegenden PREMIS XML Schemas verknüpft werden (Library of Congress, 2009). Die PREMIS Managing Agency betreut die Entwicklung des Standards

und stellt Diskussionslisten, umfangreiche Informationen und Empfehlungen zur Anwendung des Standards bereit. Das Projekt LMER (*Long-term preservation Metadata for Electronic Resources*) der Deutschen Nationalbibliothek stellte 2005 ein Metadatenschema zur Langzeitarchivierung auf Grundlage des Metadatenmodells der Neuseeländischen Nationalbibliothek vor.

METS (*Metadata Encoding and Transmission Standard*) ist ein XML Schema basierter Standard zur Dokumentation von Metadaten digitaler Objektsammlungen dient. Die Objekte werden in 7 Hauptabschnitten dokumentiert, z. b. die innere Struktur des Objektes (structural map), die Gruppierung von zusammengehörigen Dateien (file section) und administrative Metadaten (technische, Metadaten, Rechteinformation sowie Metadaten über das Ausgangsmaterial). Gleichzeitig können andere Metadatentypen, z. B. zur Erschließung oder Langzeitarchivierung (PREMIS, Dublin CORE, MARC, MIX, usw.) mit dem METS Dokument verknüpft oder in METS integriert werden. Im Sinne von OAIS kann ein METS Dokument als Informationspaket für die Einlieferung, Archivierung oder Bereitstellung verwendet werden (Library of Congress, 2005). Die Homepage des METS Standard stellt umfangreiche technische Dokumentationen und Projekt- und Anwenderinformationen bereit.

Die Auswahl von tauglichen technischen Formaten für die nachhaltige Archivierung und Bereitstellung von digitalen Forschungsdaten stellt eine dauerhafte Herausforderung für das Management eines Langzeitarchivs dar. Bei der Auswahl geeigneter Formate werden TIFF, XML oder PDF/A als langfristig nutzbare Dateiformate herausgestellt (Library of Congress, 2007). Gleichzeitig entstehen immer neue Anforderungen bei der Migration, Emulation, Konversion oder Kapselung von medienspezifischen Forschungsdaten, die im Zuge der Software- und Hardwareentwicklung oder in neuen Anwendungszusammenhängen gelöst werden müssen. So erfordern etwa Datenanalysen oder Harmonisierungen von Skalen in virtuellen Arbeitsumgebungen auch Lösungen zur Langfristsicherung von neu produzierten Daten oder Metadaten (Dickmann et al., 2010).

Konkrete Strategien zum Einsatz von Metadaten in der Langzeitarchivierung müssen die Eigenarten der fachspezifischen Forschungsdaten berücksichtigen. So werden weitere Standards des Bibliothekswesens, z. B. LMER, MIX, textMD, im Zusammenhang mit METS und PREMIS angewendet (Dappert & Enders, 2010).

Domainspezifische Anforderungen an Metadaten zur Langzeitarchivierung, die nicht von METS und PREMIS abgedeckt werden, stellen z. B. geographische Forschungsdaten dar. Hier etablieren sich disziplinspezifische Initiativen, um nach adäquaten Lösungen für diese Herausforderungen zu suchen (McGarva u.a., 2009). Die Entwicklung von Dateninfrastrukturen mit domainspezifischen Daten wie z. B. im Projekt GENESI-DR (*Ground European Network for Earth Science Interoperations – Digital Repositories*) zeigt die komplexe Vielfalt von

Standardisierungsanforderungen. GENESI-DR berücksichtigt etwa die Datenerschließung, die Interoperabilität von Metadaten, den gesicherten, lizenzbasierten Zugang oder spezifische Archivierungsformate, z. B. SAFE (*Standard Archive Format for Europe*) für Geodaten der ESA. Speziell auf die Community der Meeresforscher ausgerichtet, stellt das Projekt MMI (*Marine Metadata Interoperability*) einen umfangreichen Katalog mit fachspezifischen Vokabularen und Standards bereit, um die komplexe Welt der Metadaten für diese Disziplin überschaubar zu machen.

Angesichts der Vielzahl von Standards (Riley, 2009b) und ihrer unterschiedlichen Zielsetzungen bedarf es praktikabler Anwendungskonzepte, die Metadaten zum Verstehen, Präsentieren und Erschließen von Daten mit den Metadaten zur langfristigen Sicherung und Verfügbarkeit verknüpfen. Erst damit werden die konzeptuellen Ansprüche, den gesamten Lebenszyklus von Forschungsdaten abzudecken, einlösbar.

Erwähnte Standards und Metadaten Schemata – Modelle, Kriterien, Projekte

CCSDS, (2002). 650.0-B-1: *Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1*, January 2002. Der Vorschlag wurde als ISO 14721:2003 anerkannt. Online: http://public.ccsds.org/publications/archive/650x0b1.PDF [Zugriff am 27.07.2011] und http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683 [Zugriff am 27.07.2011].

CCSDS, (2009). 650.0-P-1.1 *Reference Model for an Open Archival Information System* (*OAIS*), Draft Recommended Standard, Issue 1.1 August 2009. Online: http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf [Zugriff am 27.07.2011].

CERA-2. *Climate and Environmental Retrieval and Archive*. Online: http://www.mad.zmaw.de/wdc-for-climate/cera-data-model/ [Zugriff am 27.07.2011].

CESSDA *Data Catalogue*. Council of European Social Science Data Archives. Online: http://www.cessda.org/accessing/catalogue/ [Zugriff am 27.07.2011].

CF. *Climate and Forecast Metadata Convention*. Online: http://cf-pcmdi.llnl.gov/ [Zugriff am 27.07.2011].

da|ra**.** *Registrierungsagentur für sozialwissenschaftliche Daten***.** Online: http://www.gesis.org/dara/ [Zugriff am 27.07.2011].

DAS. *Data Seal of Approval*. Online: http://www.datasealofapproval.org/ [Zugriff am 27.07.2011].

DataCite**.** Online: http://www.datacite.org [Zugriff am 27.07.2011].

DDC. *Dewey-Dezimalklassifikation. Deutsche Informationsseite*. Online: http://www.ddc-deutsch.de/index.htm [Zugriff am 27.07.2011].

DDI. *The Data Documentation Initiative*. DDI ALLIANCE., Secretariat ICPSR (Interuniversity Consortium for Political and Social Research), Michigan, Ann Arbor. Online: http://www.ddialliance.org/ [Zugriff am 27.07.2011].

DOI System. International DOI Foundation. Online: http://www.doi.org/ [Zugriff am 27.07.2011].

DRAMBORA. *Digital Repository Audit Method Based on Risk Assessment*. Online: http://www.repositoryaudit.eu/ [Zugriff am 27.07.2011].

Dublin Core. *The Dublin Core® Metadata Initiative*. Online: http://dublincore.org/ [Zugriff am 27.07.2011].

ELSST. *European Language Social Science Thesaurus*. Online: http://elsst.esds.ac.uk/ [Zugriff am 27.07.2011].

Genericode. *A standard format for defining code lists*. Online: http://www.genericode.org/ [Zugriff am 27.07.2011].

GENESI-DR. *Ground European Network for Earth Science Interoperations – Digital Repositories*. Siehe Nachfolgeprojekt GENESI-DEC (Digital Earth Community) – Technical Info and Tutorials. Online: http://www.genesi-dec.eu/news/ [Zugriff am 27.07.2011].

GESIS, 2006. Klassifikation Sozialwissenschaften. Online: http://www.gesis.org/download.php?url=/fileadmin/upload/dienstleistung/tools_standards/klass.pdf [Zugriff am 27.07.2011].Library of Congress, 2005. *METS: Überblick und Anleitung* (Übersetzung: A. Menne-Haritz, Juli 2005). Online: http://www.loc.gov/standards/mets/METSOverview.v2_de.html [Zugriff am 27.07.2011].

Handle System. CNRI Corporation for National Research Initiatives. Online: http://handle.net/.[Zugriff am 27.07.2011] und Online: http://www.cnri.reston.va.us/ [Zugriff am 27.07.2011],

ICPSR, (2007). *Digital Preservation Policy Framework*. Online: http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/preservation/policies/dpp-framework.jsp [Zugriff am 27.07.2011].

ICPSR, (2009a). Inter-university Consortium for Political and Social Research, 2009. *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (4th ed.). Ann Arbor, MI.. Online: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf [Zugriff am 27.07.2011].

ICPSR, (2009b). Inter-university Consortium for Political and Social Research (ICPSR). 2009. *Principles and Good Practice for Preserving Data, International Household Survey Network*, IHSN Working Paper No 003, December 2009. Online: http://www.ihsn.org/home/download.php?file=IHSN-WP003.pdf [Zugriff am 27.07.2011].

ISCO. *International Standard Classification of Occupations*. International Labour Organisation. Online: http://www.ilo.org/public/english/bureau/stat/isco/index.htm [Zugriff am 27.07.2011].

ISO 639-1:2002. *Codes for the representation of names of languages*, Part 1: Alpha-2 code. Online: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22109 [Zugriff am 27.07.2011].

ISO 3166. *Standard for country codes*. Maintenance agency. Online: http://www.iso.org/iso/country_codes.htm [Zugriff am 27.07.2011].

ISO 14721:2003. *Space data and information transfer systems* -- Open archival information system -- Reference model. Online: http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683 [Zugriff am 27.07.2011].

ISO 19115:2003. *Geographic information – Metadata*. (Online) http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020 [Zugriff am 27.07.2011].

ISO 20252:2006. *Market, opinion and social research -- Vocabulary and service requirements*. Online: http://www.iso.org/iso/catalogue_detail.htm?csnumber=39339 [Zugriff am 27.07.2011].

ISO/IEC 11179:2003. *Information Technology -- Metadata registries*. Online: http://metadata-stds.org/11179/ [Zugriff am 27.07.2011].

ISO/TS 17369:2005. *Statistical data and metadata exchange* (SDMX). Online: http://www.iso.org/iso/catalogue_detail.htm?csnumber=40555 [Zugriff am 27.07.2011].

Library of Congress, 2007. *Sustainability of Digital Formats*. Online: http://www.digitalpreservation.gov/formats/ [Zugriff am 27.07.2011].

Library of Congress, Caplan, P. Network Development & MARC Standards Office, 2008. *PREMIS verstehen*. Übersetzung: Tobias Beinert, Bayerische Staatsbibliothek. Online: http://www.loc.gov/standards/premis/understanding_premis_german.pdf [Zugriff am 27.07.2011].

LMER. *Long-term preservation Metadata for Electronic Resources*. Online: http://www.d-nb.de/standards/lmer/lmer.htm [Zugriff am 27.07.2011].

MARC. *Machine-Readable Cataloguing*. Online: http://www.loc.gov/marc/ [Zugriff am 27.07.2011].

METAFOR. *Common Metadata for Climate Modelling Digital Repositories*. Online: http://metaforclimate.eu/ [Zugriff am 27.07.2011].

MMI. *Marine Metadata Interoperability.* Online: http://marinemetadata.org/conventions/ [Zugriff am 27.07.2011].

METS. *Metadata Encoding & Transmission Standard*. Online: http://www.loc.gov/standards/mets/ [Zugriff am 27.07.2011].

MODS. *Metadata Object Description Schema*. Online: http://www.loc.gov/standards/mods/ [Zugriff am 27.07.2011].

NESTOR – Kompetenznetzwerk Langzeitarchivierung, 2008. *NESTOR-Kriterien – Kriterienkatalog vertrauenswürdige digitale Langzeitarchive* (Version II). Frankfurt am Main: NESTOR. Online: http://nbn-resolving.de/urn:nbn:de:0008-2008021802 [Zugriff am 27.07.2011].

NetCDF. *Network Common Data Form*. Online: http://www.unidata.ucar.edu/software/netcdf/ [Zugriff am 27.07.2011].

OASIS *Code List Representation TC*. Online: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=codelist [Zugriff am 27.07.2011].

PANGAEA *Publishing Network for Geoscientific & Environmental Data*. Online: http://www.pangaea.de/ [Zugriff am 27.07.2011].

PREMIS. *Preservation Metadata: Implementation Strategies*. The PREMIS maintenance activity. Homepage http://www.loc.gov/standards/premis/ [Zugriff am 27.07.2011].

PURL. *Persistent Uniform Resource Locators*. Träger OCLC (Online Computer Library Center) und Zepheira http://purl.org/ – http://www.oclc.org/ – http://zepheira.com/ [Zugriff am 27.07.2011].

Riley J. & Becker D., 2009a. *Seeing Standards: A Visualization of the Metadata Universe*. Online: http://www.dlib.indiana.edu/~jenlrile/metadatamap/ [Zugriff am 27.07.2011].

Riley J. & Becker D., 2009b. *Glossary of Metadata Standards*. Online: http://www.dlib.indiana.edu/~jenlrile/metadatamap/seeingstandards_glossary_pamphlet.pdf [Zugriff am 27.07.2011].

SAFE. *Standard Archive Format for Europe within ESA Earth Observation*. Online: http://earth.esa.int/SAFE/ [Zugriff am 27.07.2011].

SDMX – *Statistical Data and Metadata Exchange*. SDMX Initiative. Online: http://sdmx.org/ [Zugriff am 27.07.2011].

SKOS. *Simple Knowledge Organization System*. Online: http://www.w3.org/TR/skos-reference/ [Zugriff am 27.07.2011].

Survey Research Center, 2010. *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Online: http://ccsg.isr.umich.edu/pdf/00FullGuidelines3.pdf [Zugriff am 27.07.2011].

Ten Principles, 2007. *Ten basic characteristics of digital preservation repositories*. The Digital Curation Center (U.K), DigitalPreservationEurope, NESTOR (Germany), Center for Research Libraries (North America). Online: http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re [Zugriff am 27.07.2011].

TRAC. *Trustworthy Repositories Audit & Certification: Criteria and Checklist.* Online: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf [Zugriff am 27.07.2011].

Triple S. *Survey Interchange Standard*. The Triple-S Group. Online: http://www.triple-s.org/ [Zugriff am 27.07.2011].

URN:NBN Namensraum durch Nationalbibliotheken z. B. für die Deutsche Nationalbibliothek. Online: http://www.persistent-identifier.de/?link=3352 [Zugriff am 27.07.2011].

URN System. IETF. Internet Engineering Task Force. Online: http://www.ietf.org/ [Zugriff am 27.07.2011].

WDC-C. *World Data Center for Climate*. Online: http://www.mad.zmaw.de/wdc-for-climate/ [Zugriff am 27.07.2011].

Z39.50. *Information Retrieval Protocol* (Z39.50/ISO 23950 – ANSI/NISO Z39.50) Online: http://www.loc.gov/z3950/agency/ [Zugriff am 27.07.2011].

ZIS 2010. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*, Version 14.00, 2010. Online: http://www.gesis.org/unser-angebot/studien-planen/zis-ehes/ [Zugriff am 27.07.2011].

Literaturhinweise

Bose, J. & Frew J., 2005. Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys*, 37(1), March 2005, S. 1–28. Online: http://dx.doi.org/10.1145/1057977.1057978. [Zugriff am 27.07.2011]

Brübach, N., 2010. Die Überarbeitung und Ergänzung des OAIS. In: H. Neuroth et al., Hrsg. 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. (Version 2.3.) Göttingen. Kap.4:13, Kap.4:15- Kap.4:16. Online: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949 [Zugriff am 27.07.2011].

Couper, M., 1998. Measuring survey quality in a CASIC environment. In: *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. Online: http://www.amstat.org/sections/srms/proceedings/papers/1998_006.pdf [Zugriff am 27.07.2011].

Dappert, A. & Enders, M., 2010. Digital Preservation Metadata Standards. *Information Standards Quarterly* (ISQ), 22(2), Special Issue: Digital Preservation. Online: http://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf [Zugriff am 27.07.2011].

DFG (Deutsche Forschungsgemeinschaft), 2010. *Leitfaden für Antragstellung*. (DFG-Vordruck 1.02 – 8/10. S. 32, Abschnitt 3.7 Umgang mit den im Projekt erzielten Forschungsdaten) Online: http://www.dfg.de/download/programme/emmy_noether_programm/antragstellung/1_02/1_02.pdf [Zugriff am 27.07.2011].

Dickmann, F. Henke, H. & Harms, P., 2010. *Technische Evaluation der Grid-Technologie für das Modellprojekt. Kollaborative Datenauswertung und virtuelle Arbeitsumgebung – VirtAug.* (SOEB Arbeitspapier 2010-1) Online: http://www.wissgrid.de/publikationen/Expertise_VirtAug.pdf [Zugriff am 27.07.2011].

Confalonieri, U. & McMichael, A. eds., 2007. *Global Environmental Change and Human Health: Science Plan and Implementation Strategy.* (Earth System Science Partnership (DIVERSITAS, IGBP, IHDP, and WCRP) Report No.4; Global Environmental Change and Human Health Report, No. 1), S. 68–72. Online: http://www.gechh.unu.edu/FINAL_GECHH_SP_UPDATED.pdf [Zugriff am 27.07.2011].

Gregory, A. Heus, P. & Ryssevik, J., 2009. *Metadata*. Working Paper Series of the Council for Social and Economic Data. (RatSWD Working Paper No. 57) (March 2009) Online: http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_57.pdf [Zugriff am 27.07.2011].

Hanson, B. Sugden, A. & Alberts, B., 2011. Making Data Maximally Available. *Science*, 331(6018), S. 649. Online: http://dx.doi.org/10.1126/science.1203354 [Zugriff am 27.07.2011]

Hopt, O. et al., 2010. *Questionnaire management and DDI: The QDDS case*. DDI Working Paper Series. Online: http://www.ddialliance.org/sites/default/files/QuestionnaireManagementAndDDI-TheQDDSCase.pdf [Zugriff am 27.07.2011].

Jääskeläinen, T. Moschner, M. & Wackerow, A., 2009. Controlled Vocabularies for DDI 3: Enhancing Machine-Actionability. *IASSIST Quarterly*, 33 (1/2), S 34 -39. Online: http://www.iassistdata.org/downloads/iqvol3312wackerow_0.pdf [Zugriff am 27.07.2011].

Jinfang Niu, 2009. *Perceived Documentation Quality of Social Science Data*. Ph. D The University of Michigan. Online: http://deepblue.lib.umich.edu/handle/2027.42/63871. [Zugriff am 27.07.2011].
Siehe auch
Jinfang Niu, 2009. *Overcoming Inadequate Documentation*. Online: http://deepblue.lib.umich.edu/handle/2027.42/78325 [Zugriff am 27.07.2011].

Körner, A. & Schmidt, J., 2006. *Qualitätsberichte – ein neues Informationsangebot über Methoden, Definitionen und Datenqualität der Bundesstatistiken*. Statistisches Bundesamt, Wirtschaft und Statistik 2/2006, Wiesbaden. Online: http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Querschnittsveroeffentlichungen/WirtschaftStatistik/AllgemeinesMethoden/Qualitaetsberichte,property=file.pdf [Zugriff am 27.07.2011]

Kreuter, F. & Casas-Cordero, C., 2010. *Paradata*. Working Paper Series of the Council for Social and Economic Data (RatSWD Working Paper No. 136) (April 2010) Online: http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_136.pdf [Zugriff am 27.07.2011].

Kreuter, F., et al., 2010. Using proxy measures and other correlates of survey outcomes to adjust for nonresponse: Examples from multiple surveys. *Journal of the Royal Statistical Society, Series A*. Abstract Online: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2009.00621.x/abstract. [Zugriff am 27.07.2011]

McGarva, G. Morris, S. & Janée, G., 2009. *Technology Watch Report 09-01: Preserving Geospatial Data*. Digital Preservation Coalition. Online: http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee [Zugriff am 27.07.2011].

NSF (National Science Foundation), 2010. *Dissemination and Sharing of Research Results – Data Management Plan Requirements*. Online: http://www.nsf.gov/bfa/dias/policy/dmp.jsp [Zugriff am 27.07.2011].

PISA, 2006. *Technical Report*. Online: www.oecd.org/dataoecd/0/47/42025182.pdf [Zugriff am 27.07.2011].

Quandt, M. Agache, A. & Friederichs, M., 2009. How to Make the Unpublishable Public. The Approach of the CESSDA Survey Data Harmonisation Platform. *The 5th International Conference on e-Social Science*. Köln, Deutschland 24.-26. June 2009. Online: www.ncess.ac.uk/resources/content/papers/Quandt.pdf [Zugriff am 27.07.2011].

Toussaint, F., 1999. *Wissenschaftliches Datenmanagement: Das „CERA-2 Daten- und Metadatenmodell"*. Online: http://www.mad.zmaw.de/uploads/media/9911_ptb_01.pdf [Zugriff am 27.07.2011].

UK Data Archive, 2011. *Research Data Lifecycle*. Online: http://www.data-archive.ac.uk/create-manage/life-cycle [Zugriff am 27.07.2011].

Vardigan, M. Heus, P. & Thomas, W., 2008. Data Documentation Initiative: Toward a Standard for the Social Sciences. *The International Journal of Digital Curation*, 3(1). Online: http://www.ijdc.net/index.php/ijdc/article/view/66/45 [Zugriff am 27.07.2011].

2.5 Forschungsdaten-Repositorien

Andreas Aschenbrenner [1], Heike Neuroth [2]
[1] Österreichische Akademie der Wissenschaften
[2] Niedersächsische Staats- und Universitätsbibliothek Göttingen

2.5.1 Einleitung

Vorangegangene Kapitel haben die zentrale Bedeutung und Rolle von Forschungsdaten in der Wissenschaft beschrieben. Die vertrauenswürdige Archivierung und Verfügbarkeit dieser Daten ist eine der Grundvoraussetzungen des wissenschaftlichen Diskurses. Repositorien spielen eine wichtige Rolle in diesem Kontext, so sind sie für die Langzeitarchivierung von Forschungsdaten verantwortlich, dienen der gemeinsamen Datenhaltung sowie ihrem Austausch und kollaborativen Nutzung innerhalb einer wissenschaftlichen *Community.*

Wissenschaftliche Daten unterlaufen in ihrem Lebenszyklus je nach wissenschaftlicher Methodik und Fach-*Community* unterschiedliche Stationen mit jeweils spezifischen Anforderungen an das Datenmanagement. Ebenso stellen die *Community* oder die Öffentlichkeit Anforderungen wie die Verifikation, Reproduzierbarkeit und Nachnutzbarkeit wissenschaftlicher Ergebnisse. Dieses Kapitel analysiert Repositorien aus technischer, organisatorischer und Nutzersicht. Angelehnt an die NESTOR Definition eines Langzeitarchivs (Dobratz & Schoger, 2010) verstehen die Autoren dieses Kapitels ein *Repository* als eine Organisation (bestehend aus Personen und technischen Systemen), die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit digitaler Objekte sowie für ihre Interpretierbarkeit zum Zwecke der Nutzung durch eine bestimmte Zielgruppe (vgl. „*designated community*“ des *Open Archival Information Systems* (OAIS) (NSSDC, o. J.) übernommen hat. Allerdings zeigt der heutige Stand, dass es sowohl weltweit, als auch national noch nicht für alle Fachdisziplinen entsprechende Repositorien gibt. Ein zumindest in den Naturwissenschaften erfolgreicher Ansatz stellt das *World Data System* (ICU WDS, 2010) dar, das aus dem *World Data Center System* (WDC) hervorgegangen ist (NGDC, o.J.). Auch hier soll eine Zertifizierung der existierenden *World Data Centers* (NGDC, 2009) für definierte organisatorische, politische, technische und inhaltliche Kriterien sorgen, damit Forschungsdaten vertrauenswürdig und nachhaltig vorgehalten werden. Auch in Deutschland gibt es eine Reihe von Forschungsdaten-Repositorien (vgl. Kap. 3.1), jedoch ist die Langzeitarchivierung von Forschungsdaten über alle wissenschaftlichen Disziplinen zurzeit noch nicht gesichert. Erste entscheidende Impulse für einen konzentrierten nationalen Ansatz kommen sicherlich von der GWK Initiative „Kommission Zukunft der Informationsinfrastruktur“ (WGL, 2011), deren im April 2011 vorgelegter Abschlussbericht als Basis für die in Vorbereitung befindlichen grundlegenden

Empfehlungen des Wissenschaftsrates zur Forschungsinfrastruktur in Deutschland dienen wird.

Es steht außer Frage, dass ohne fachspezifische Repositorien, die zum Beispiel auch komplexe Objektmodellierungen (z. B. in den Geisteswissenschaften bei kritischen Editionen oder bei Daten aus der Archäologie) oder verschiedene Versionen von Daten berücksichtigen, die Wissenschaft in den heutigen IT-gestützten Forschungsprozessen nicht optimal versorgt ist. Gerade der immer größer werdende Einsatz von Virtuellen Forschungsumgebungen für bestimmte Forschungsfragen und vernetzt arbeitende Forschergruppen zeigt, dass die Wissenschaft im Forschungsdatenmanagement unterstützt werden muss, hier spielen fachliche Repositorien eine entscheidende Rolle.

2.5.2 Definition, Funktionen und Aufgaben von Repositorien

Repositorien finden sich in den unterschiedlichsten Kontexten und mit den unterschiedlichsten Funktionsanforderungen (Aschenbrenner & Kaiser, 2005). Sie haben sich meist unabhängig voneinander entwickelt und noch heute ist der Bereich keineswegs überschaubar. Es gibt daher keine universelle Definition oder zeitlose Standards, auf die zurückgegriffen werden kann.

Heery und Anderson (2005) beschreiben Kernfunktionen von Repositorien als die technisch robuste sowie organisatorisch nachhaltige und vertrauenswürdige Verwaltung von (datei-basierten) Daten und zugehörigen Metadaten sowie die organisatorische und technische Einbettung der Schnittstellen für Ablage und Zugriff. In dieser Definition der Kernfunktionen wird das Zusammenspiel aus Technik und organisatorischen Maßnahmen deutlich.

Es ist auch eine klare Trennung zu verwandten Systemen wie *Code-Repositories* (vgl. Apache Subversion, Git), *Registries* (vgl. oft Datenbank-basierte Kataloge wie *Service Registries*, Metadaten-*Registries*) und Anderen. Ausschlaggebend für die Unterscheidung dieser Systeme ist zumeist die Art der Daten, die sie beherbergen, und wie sie mit ihnen umgehen. Im Kontext von Repositorien für Forschungsdaten arbeitet man oft mit dem Begriff der „digitalen Objekte". Digitale Objekte sind digitale Daten, die als intellektuelle Einheiten aus (einer oder mehreren) Dateien, zugehörigen Metadaten sowie einem Netzwerk aus anderen Objekten bzw. referenzierbaren Informationen bestehen können. Ein Beispiel wäre ein digitalisierter Brief mit der zugehörigen Transkription in Volltext, die jeweils beschrieben und mit anderen Briefen zu einer Korrespondenz verknüpft sind. Objekte können alle Arten von Daten umfassen – strukturiert, semi-strukturiert (z. B. XML-basiert) oder unstrukturierte Daten wie z. B. Bilder oder Videos.

Repositorien-Systeme decken je nach Fokus und Zielgruppe unterschiedliche Funktionen[1] ab, die sich oft auch in spezifischen Bezeichnungen spiegeln (z. B. „*institutional repositories*" für Publikationsserver, „*trusted repositories*" für

Langzeitarchivierungsumgebungen, oder „*open access repositories*“ für frei zugängliche Daten):

- Verwaltung von Informationsobjekten (Speicherkonzepte, Datenarten z. B. Publikationen in PDF, Bilder über 100 MB, *stream*-bare Videos)
- Metadatenverwaltung zur Identifikation, Administration und langfristigen Erhaltung von Informationsobjekten sowie deren Einbettung in einen inhaltlichen, intellektuellen Kontext
- Vernetzung bzw. (standardisierte) Verknüpfung der Objekte untereinander mit Kontextdaten
- *Workflow*-Unterstützung zur Registrierung von Informationsobjekten (manueller *Ingest-Workflow* und automatischer Datentransfer)
- Zugang zu und Nachnutzung von Forschungsdaten durch persistente Identifikation, Suchmechanismen, Schnittstellen (z. B. *Open Archives Initiative* (OAI)[2])
- Präsentation, Einbettung in Nutzungsumgebungen, Unterstützung von kollaborativen und kooperativen Arbeitsformen
- Analyse der Nutzung (Nutzungsstatistiken) und Archivinhalte (z. B. *Text Mining*, Visualisierung)
- Berücksichtigung von rechtlichen Rahmenbedingungen (Datenschutz, Urheberrecht etc.)
- Mechanismen zur Langzeitarchivierung

Systeme können sich zum Teil erheblich darin unterscheiden, wie sie diese Kernfunktionen umsetzen und welche Zusatzfunktionalitäten sie anbieten. Gerade im Aufbau einer *Repository*-basierten Forschungsumgebung, die mitunter spezifisch auf den jeweiligen Anwendungsfall und Forschungskontext zugeschnitten sein muss, ist daher oft viel Anpassungsarbeit oder Eigenentwicklung nötig.

2.5.3 Auswahl Software

Während früher ein *Repository* eher verwendungsspezifisch und häufig ad-hoc entwickelt wurde, stellt sich die Situation heutzutage deutlich verändert dar. Eine breite *Community* teilt ähnliche Anforderungen an solche Systeme, tauscht

1. Diese kurze Auflistung kann nicht vollständig sein und listet nur einige Kern-Funktionalitäten unterschiedlicher Fokusgruppen und Ziele. Für weitere technische Funktionen siehe z. B. den ISO Standard zu einem „*Open Archival Information System*“ (OAIS) (CCSDS, 2002), das DELOS *Reference Model* (DELOS, o.J.) und andere.
2. http://www.openarchives.org/ [Zugriff am 14.08.2011].

ihre Erfahrungen hierzu aus und entwickelt gemeinschaftlich und nach dem *Open Source* Prinzip entsprechende Softwaresysteme.

Vor allem im Bereich von Publikationsservern zeichnet sich eine gewisse Konvergenz der Technologien ab. Bereits in den 90er Jahren sind erste Gesamtpakete für Repositorien aufgekommen, darunter der CERN *Document Server*[3] oder der Hochschulschriftenserver der Universität Stuttgart OPUS[4]. Andere Institutionen haben eigene Systeme entwickelt oder bestehende Systeme aufgegriffen und für ihre Bedürfnisse angepasst, wo dies sinnvoll und möglich war.

Heute gibt es eine Vielzahl von *Repository* Systemen, wie z. B. die Auflistung von OSI (2004) oder die Überblicksarbeit von Borghoff, et al. (2005) zeigen. Die ebenso weit verbreiteten *Web-Content-Management*-Systeme (z. B. Plone[5], Drupal[6], Joomla![7]) eignen sich üblicherweise nicht als Datenrepositorien, da sie oft Workflows für Metadaten-Beschreibungen nicht unterstützen bzw. aus Langzeitarchivierungssicht nicht robust genug sind. Besonders gefragt sind zurzeit vor allem folgende drei *Repository* Systeme, die auch auf der internationalen *OpenRepositories*[8] Konferenz stark vertreten sind:

- ***EPrints***[9]. *Out-of-the-Box* Komplettsystem für Publikationen mit weitgehend vorgegebenen Strukturen und einfacher Verwaltung.
- ***DSpace***[10]. Komplettsysteme für Publikationen mit einem vorstrukturierten *Workflow*-System zur Eingabe von Metadaten, etc. beim *Ingest*.
- ***Fedora***[11]. *Middleware* zur Modellierung und Verwaltung von Daten, wobei unterschiedliche Projekte auch spezifischere Nutzerumgebungen (z. B. eSciDoc[12], Fez[13], Muradora[14]) auf Fedora aufsetzen.

Anfang 2011 weist das Verzeichnis OpenDOAR[15] z. B. über 1.800 laufende *Repository*-Installationen nach, davon nutzen ein Drittel DSpace gefolgt von

3. http://cds.cern.ch/ [Zugriff am 14.08.2011], http://www.cern.ch, [Zugriff am 14.08.2011].
4. http://elib.uni-stuttgart.de/opus/ [Zugriff am 14.08.2011].
5. http://plone.org/ [Zugriff am 14.08.2011].
6. http://www.drupal.de/ [Zugriff am 14.08.2011].
7. http://www.joomla.de/ [Zugriff am 14.08.2011].
8. http://www.openrepositories.org/ [Zugriff am 14.08.2011].
9. http://www.eprints.org/ [Zugriff am 14.08.2011].
10. http://www.dspace.org/ [Zugriff am 14.08.2011].
11. http://www.fedora-commons.org/ [Zugriff am 14.08.2011].
12. http://www.escidoc.org/ [Zugriff am 14.08.2011].
13. http://sourceforge.net/projects/fez/ [Zugriff am 14.08.2011].
14. http://www.muradora.org/ [Zugriff am 14.08.2011].
15. http://www.opendoar.org/ [Zugriff am 14.08.2011].

EPrints. *DSpace* wurde ursprünglich für das *Massachusetts Institute of Technology* (MIT)[16] entwickelt, wird inzwischen durch eine große *Community* („DSpace Federation“) weiterentwickelt und durch die Firma HP auch kommerziell vertrieben. Neben diesen drei *Open Source* Systemen hat jüngst auch z. B. Microsoft mit einem eigenen Produkt, dem Publikationsserver Zentity[17], aufhorchen lassen.

Diese Softwarepakete sind zwar als Publikationsserver weit verbreitet, aber für Forschungsdaten sind nicht alle einsetzbar. *Workflows* und Datenmodelle in *EPrints* und *DSpace* sind primär auf dokument-artige Publikationen (z. B. Dissertationen, Journale, Berichte) ausgelegt und für andere Arten von Forschungsdaten (z. B. veränderliche Objekte, bestehend aus mehreren Dateien mit komplexen Metadaten) ungeeignet.

Von den genannten Systemen ist nur Fedora so flexibel, dass es ideal für die Verwaltung und Archivierung von Forschungsdaten dienen kann. Zwei Eigenschaften seien hier speziell herausgehoben:

- (1) Die Fedora Service-Architektur[18] ist die Basis einer offenen, evolutionären Umgebung für wissenschaftliche *Workflows*, und
- (2) Fedora-Mechanismen zur Metadatenmodellierung (vgl. *Content Model Architecture* (Fedora Commons, 2007)) ermöglichen die Beschreibung unterschiedlichster Datenarten, wie es beispielsweise das Fedora-basierten eSciDoc[19] für die unterschiedlichen Disziplinen in der Max-Planck-Gesellschaft umsetzt.

Neben Fedora seien noch zwei weitere *Repository*-Pakete genannt: iRODS und Tupelo. Diese Systeme eignen sich besonders für Forschungsdaten, da sie (a) für große Datenmengen skalieren, (b) Modellierbarkeit von Daten und Metadaten unterstützen und (c) die Systeme aus Langzeitarchivierungssicht robust genug sind.

- ***iRODS***[20] – stammt von Datenzentren und ist besonders zur effizienten Verwaltung von sehr großen Datenmengen geeignet. iRODS ist ein weitgehend monolithisches System und mit zumeist proprietären Schnittstellen, wächst aber durch eine weltweite *Open Source Community.*

16. http://web.mit.edu/ [Zugriff am 14.08.2011].
17. http://research.microsoft.com/en-us/projects/zentity/ [Zugriff am 14.08.2011].
18. Vgl. z. B. das Konzept der „Disseminatoren“ im ursprünglichen Architekturkonzept (Payette & Lagoze, 1998)
19. http://www.escidoc.org/ [Zugriff am 14.08.2011].
20. http://irods.sdsc.edu/index.php/Main_Page [Zugriff am 14.08.2011].

- ***Tupelo***[21] – ist eine kleine Initiative mit einer leichtgewichtigen Software, die sich primär auf die Daten- und Metadatenmodellierung mithilfe semantischer Technologien konzentriert.

2.5.4 Architektur, Technologien, Standards

Trotz der unterschiedlichen Systeme und der Dynamik in der *Repository-Community* mit immer neuen Entwicklungen gibt es bei allen Software-Paketen einen deutlichen technischen Trend zu Offenheit und Interoperabilität. Dieser Trend entsteht nicht nur durch eine gemeinsame Ideologie der Software-Macher, sondern begründet sich auf die Anforderungen der Organisationen, die *Repository*-Systeme betreiben, sowie der Endnutzer, die (mitunter mehrere) *Repositories* und Zusatzdienste für ihre wissenschaftliche Arbeit benötigen. Somit betreffen die im Folgenden vorgestellten Architekturkonzepte und Standards durchaus alle *Repository*-Systeme – auch kommerzielle, wie die von Microsoft oder andere Eigenentwicklungen.

Abgeleitet von den in Abschnitt „Definition, Funktionen, Aufgaben" vorgestellten Anforderungen, kann man generell drei konzeptuelle Schichten in *Repository*-Systemen unterscheiden: *Storage*, Datenmanagement und Nutzung.

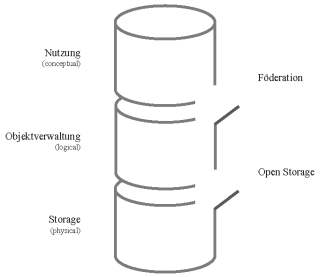

Abb. 1: Schichten-Architektur mit den drei konzeptuellen Schichten – *Storage*, Objektverwaltung, und Nutzung – angelehnt an die 3 Ebenen von Thibodeau (2002). Rechts: Bezeichnung der Interoperabilitätsebenen „Föderation" und „*Open Storage*".

2.5.4.1 Architekturschicht: Storage

Die *Storage*-Ebene beherbergt digitale Objekte – also Daten gemeinsam mit zugehörigen Metadaten. Aus Gründen der Stabilität entscheiden sich *Repository*-Systeme auf dieser Ebene zumeist für eine datei-basierte Ablage (also nicht in Datenbanken), und ermöglichen die Rekonstruktion aller Informationen aus den Dateien.

Während kleinere Repositorien mit einem lokalen Server ihre kompletten *Storage*-Anforderungen abdecken können, entscheiden sich manche Repositorien zur Auslagerung der Daten in ein Datenzentrum bzw. Rechenzentrum. Gerade für Forschungsdaten liegt ein wesentlicher Vorteil bei der Auslagerung

21. http://tupeloproject.ncsa.uiuc.edu/ [Zugriff am 14.08.2011].

des *Storage* darin, dass ggf. größere Datenmengen verwaltet werden können, mehrere Repositories auf eine gemeinsame *Storage*-Ebene zugreifen können und dass Aufgaben zur *Bit-Preservation* (z. B. Datenreplikation, *Tape-Backup*, Integritätstests) gekapselt werden können.[22]

2.5.4.2 Architekturschicht: Objektverwaltung

Das Datenmanagement in Repositorien verknüpft Daten und Metadaten zu Objekten, beschreibt Relationen zwischen Objekten, versioniert Objekte, verknüpft sie mit unterschiedlichen Darstellungs- und Zugriffsmechanismen und bettet sie in (existierende) Softwareumgebungen ein. Verbreitete Standards schließen Daten- und Metadatenbeschreibungsformate (z. B. Dublin Core[23], METS[24]) wie auch Standards für APIs (vgl. z. B. *Common Repository Interfaces Group* (CRIG)[25]) mit ein. Gerade Forschungsdaten verlangen oft eine große Flexibilität und Ausdrucksfähigkeit in der Daten- und Metadaten-Modellierung. Anforderungen an z. B. Zugriffsrechte und Veränderbarkeit der Daten können sich zwischen Forschungskontexten und Forschungsprojekten stark unterscheiden.

2.5.4.3 Architekturschicht: Nutzung

Während Publikationsrepositorien primär auf die Einfuhr und die Suche von Publikationen ausgerichtet sind, ist die Bandbreite der Nutzungsszenarien bei Forschungsdaten-Repositorien wesentlich breiter. Je nach Forschungskontext sollten Daten z. B. direkt von Messinstrumenten in das *Repository* überführt, in wissenschaftliche Workflows eingebettet oder in bestehende Forschungsapplikationen integriert werden.

Aufgrund dieser Bandbreite an Nutzungsszenarien und Forschungskontexten ist es kaum sinnvoll, generelle technische Standards auf einer Nutzungsebene zu erarbeiten. Beratungsangebote und Leitfäden wie die von WissGrid (2011) können allerdings wertvolle Erfahrungen zum Aufbau spezialisierter Forschungsumgebungen und Ratschläge zur Nachnutzung und Vernetzung von existierenden Werkzeugen geben.

---

22. Für *Cross-Repository* Interoperabilität reicht eine *Storage*-Ebene zur Dateiablage nicht aus. Die *Repository-Storage*-Ebene bezieht auch standardisierte Mechanismen zur Ablage von Metadaten, Datenversionierung, *Locking*, etc mit ein. Vgl. z. B. *Fedora High Level Storage* (Fedora Repository Development, 2007).
23. http://dublincore.org/ [Zugriff am 14.08.2011].
24. http://www.loc.gov/standards/mets/ [Zugriff am 14.08.2011].
25. http://www.ukoln.ac.uk/repositories/digirep/index/CRIG [Zugriff am 14.08.2011].

2.5.4.4 Offene Repository-Umgebungen

Technisch gesehen eröffnet der Trend zu Offenheit und Interoperabilität ganz neue Möglichkeiten, die vor allem im Umfeld von Forschungsdaten noch weiter erforscht werden müssen. Dieser Trend wird allein schon dadurch gefördert, dass manche Institutionen mehrere Installationen von unterschiedlichen Systemen bei sich führen, um unterschiedlichen Anforderungen in ihrer Organisation gerecht zu werden. Aber auch die Sichtbarkeit der *Open Access* Bewegung (Berliner Erklärung, 2003) und aufkommende *e-Science* Mechanismen zur Vernetzung unterschiedlichster Daten und Dienste untereinander[26] fördern die Offenheit und Interoperabilität von *Repository* Systemen.

Für die Interoperabilitäts-Ebene „*Open Storage*“ (vgl. Abb. 1: Schichten-Architektur mit den drei konzeptuellen Schichten – *Storage*, Objektverwaltung, und Nutzung – angelehnt an die 3 Ebenen von Thibodeau (2002). Rechts: Bezeichnung der Interoperabilitätsebenen „Föderation“ und „*Open Storage*“.) gibt es derzeit noch keine eindeutigen Standards. Derzeit arbeitet z. B. das *Duraspace*-Projekt (Minton Morris, 2008) an einer generellen *Cloud*-basierten *Storage*-Ebene für Fedora und DSpace, die für den Produktivbetrieb geeignet ist und auch Anforderungen der Langzeitarchivierung (bzw. zumindest *Bit-Preservation*) abdecken wird.

Förderationsstandards wie OAI-PMH (Open Archives, o.J.), OAI-ORE (Pepe et al., 2009) und Zing[27] verschränken das Datenmanagement unabhängiger Repositorien zu einem übergreifenden, virtuellen Repositorium. Nutzer von Föderationen wie DRIVER (*Digital Repository Infrastructure Vision for European Research*)[28] oder Europeana[29] haben dadurch unmittelbaren Zugriff zu einer Vielzahl von institutionellen und thematischen Repositorien. Auch im Bereich von Forschungsdaten werden diese Standards bereits vereinzelt eingesetzt (WissGrid, 2010). Allerdings werden erst die Entwicklungen der nächsten Jahre zeigen, wie diese Standards für neue Anwendungen im Kontext von Forschungsdaten eingesetzt werden können – z. B. Analyse und Visualisierung von Forschungsdaten sowie Rechtemanagement und Aufgabensteuerung für Forschergruppen – und wie Repositorien-basierte Infrastrukturen den Aufbau und die Vernetzung von virtuellen Forschungsumgebungen verändern (Aschenbrenner et al., 2010).

26. Zum Beispiel die Verknüpfung von Publikationen mit den zugrunde liegenden wissenschaftlichen Rohdaten und Diensten zur Analyse der Daten. Vgl. DRIVER (2009).
27. Im Rahmen der ZING-Initiative (Z39.50 International Next Generation) entstand der technische Standard SRU Search/ Retrieval via URL (Library of Congress, 2011).
28. http://www.driver-repository.eu/ [Zugriff am 14.08.2011].
29. http://www.europeana.eu/ [Zugriff am 14.08.2011].

2.5.5 Weitere Aspekte

Neben technologischen Aspekten gibt es eine Reihe weiterer Überlegungen, die frühzeitig berücksichtigt werden müssen und Einfluß nehmen auf den Aufbau und die (Weiter-) Entwicklung von Forschungsdaten-Repositorien.

Dies beinhaltet zum Beispiel Vorüberlegungen[30] zu Strategie und Management und umfaßt Definition (*mission statement*), Zielgruppe(n), notwendige Kooperationen (z. B. Rechenzentrum, Bibliothek) und Regelungen für den potentiellen Nachfolgebetrieb im Notfall. Sogenannte *Service-Level-Agreements* (SLA) müssen ausgearbeitet werden und die verschiedenen Stufen des Angebotes (von *bitstream preservation* bis hin zu „echter *data curation*“) verständlich und transparent dokumentiert sein. Ein Betriebsplan, der auch Qualitätskontrolle und Überwachung im Sinne von Monitoring umfaßt, ist ebenfalls integraler Bestandteil eines Repositoriums. Ein stabiler Finanzierungsplan und mittel- bis langfristige Überlegungen zu Personalplanungen inklusive Aufbau notwendiger Qualifikationen und Kompetenzen gehören ebenfalls dazu.

Angaben über die zu archivierenden Sammlungen und Objekte müssen dokumentiert sein inklusive notwendiger Standards (z. B. Metadatenstandards) und rechtlicher Rahmenbedingungen. Die Anforderungen zum Beispiel in Bezug auf Authentizität, Integrität, Nachnutzbarkeit, Sicherheit und Verfügbarkeit sind klar zu definieren. Ein stetiger Abgleich der Anforderungen mit dem bestehendem Dienstleistungsangebot ist zu leisten. Vereinbarungen und Verträgen über Rechte, Verpflichtungen, Haftungen und Umsetzungen zwischen den unterschiedlichen Akteuren sind zu treffen und zu dokumentieren. Die einzelnen Arbeitsabläufe sind mit klarer Rollenverteilung und Festlegung von Verantwortlichkeiten zu regeln. Die Erfordernisse bei der Umsetzung durch eine IT-Infrastruktur und Technologie inklusiver langfristiger Technologiestrategie sind festzulegen.

Die hier beschriebenen Aspekte geben nur einen kleinen Einblick in die nötigen (Vor-)Überlegungen wieder und zeigen auf, dass ein wesentlicher Bereich im Vorfeld, abhängig von den unterschiedlichen Beteiligten und den organisatorischen sowie strukturellen Rahmenbedingungen, zu klären ist. Die demnächst veröffentlichten DIN[31] Richtlinien und ISO Standard[32] im Bereich der vertrauenswürdigen Zertifizierung von Repositorien geben einen umfassenden Einblick. Beispiele für Forschungsdaten-Archive in Deutschland wie das Deutsche

---

30. Nach Ludwig, J. & Strathmann, S.: „Zehn-Punkte-Plan zum Aufbau eines Angebots zur Langzeitarchivierung und zum Forschungsdatenmanagement“, Veröffentlichung in Vorbereitung.
31. DIN 31644, vgl. auch NESTOR (2010).
32. ISO 16363 für vertrauenswürdige Langzeitarchive.

Fernerkundungszentrum (DFD[33]), Pangaea[34] für die Geo- und Umweltwissenschaften oder die *World Data Center* (WDC MARE[35], WDC Climate[36], WDC RSAT[37]) zeigen, dass die intensive Zusammenarbeit mit den jeweiligen Fachdisziplinen unerläßlich für die Akzeptanz solcher Repositorien ist. Einerseits müssen die Fachwissenschaftler eng bei der Formulierung der Anforderungen eingebunden werden, andererseits müssen sie klar den Nutzen und den Mehrwert solcher Langfrist-Archive erkennen, um ihre Daten dort abzulegen. Die Aufgabe der Langzeitarchivierung von Forschungsdaten muss als *Community*-Aufgabe verstanden werden. Nicht umsonst finden sich in bereits gut organisierten, zum Teil international vernetzten Fachdisziplinen mit einem in der Regel überdurchschnittlich hohen Aufkommen von Forschungsdaten bereits erste stabile Ansätze von Forschungsdaten-Repositorien.

2.5.6 Aktuelle Entwicklungen, Diskussionen und Ausblick

In den letzten Jahren hat es eine Reihe von Aktivitäten, Entwicklungen und Diskussionen im Bereich von Forschungsdaten gegeben. So hat zum Beispiel die Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen im Juni 2010 im Rahmen der Arbeitsgruppe Forschungsdaten (Allianz, o.J.) Grundsätze (Allianz, 2010) zum Umgang mit Forschungsdaten veröffentlicht, die unter anderem von den Organisationen Deutsche Forschungsgemeinschaft (DFG), Fraunhofer-Gesellschaft, Helmholtz-Gemeinschaft, Hochschulrektorenkonferenz (HRK), Leibniz-Gemeinschaft, Max-Planck-Gesellschaft und Wissenschaftsrat unterschrieben wurden. Diese Grundsätze beginnen mit einer Präambel, in der festgehalten wird, dass „Qualitätsgesicherte Forschungsdaten ... einen Grundpfeiler wissenschaftlicher Erkenntnis [bilden] und ... unabhängig von ihrem ursprünglichen Erhebungszweck vielfach Grundlage weiterer Forschung sein [können]“. Weiter heißt es „Die nachhaltige Sicherung und Bereitstellung ... bildet eine strategische Aufgabe, zu der Wissenschaft, Politik und andere Teile der Gesellschaft gemeinsam beitragen müssen“. Die Eckpunkte der Grundsätze beinhalten Sicherung und Zugänglichkeit, Unterschiede der wissenschaftlichen Disziplinen, Wissenschaftliche Anerkennung, Lehre und Qualifizierung, Verwendung von Standards sowie Entwicklung von Infrastrukturen.

---

33. http://www.dlr.de/ [Zugriff am 14.08.2011].
34. http://www.pangaea.de/ [Zugriff am 14.08.2011].
35. http://www.wdc-mare.org/ [Zugriff am 14.08.2011].
36. http://www.mad.zmaw.de/wdc-for-climate/ [Zugriff am 14.08.2011].
37. http://wdc.dlr.de/ [Zugriff am 14.08.2011].

Im Jahr 2010 wurde die *„Kommission Zukunft der Informationsinfrastruktur"* (WLG, 2011) gebildet mit dem Auftrag, ein nationales Gesamtkonzept für die Informationsinfrastruktur in Deutschland zu erarbeiten und 2011 vorzulegen. Zu den insgesamt acht eingesetzten thematischen Arbeitsgruppen findet sich auch eine AG Forschungsdaten, die im Oktober 2010 dem Steuerungsgremium der KII einen Bericht vorgelegt hat, der Aspekte wie Status Quo in Deutschland, internationaler Kontext, Nutzererwartungen, Handlungsbedarf, Visionen, Querschnittsthemen, Ressourcenabschätzung und Aufgaben und Rahmenbedingungen abdeckt. Letztendlich sollen daraus auch für den Themenbereich Forschungsdaten Handlungsempfehlungen für den Gesamtbericht[38] der KII abgeleitet werden, die darüber Auskunft geben, wie in Deutschland das Thema Forschungsdaten und Forschungsdaten-Repositorien gesamtheitlich angegangen und umgesetzt werden kann. Bei diesen Diskussionen hat sich klar herauskristallisiert, dass jede datenintensive Disziplin einen Datenmanagementplan entwickeln sollte und dass eine Initial- und Grundfinanzierung für den Aufbau und den Betrieb von Dateninfrastrukturen nötig ist. Die daraus abgeleiteten Handlungsempfehlungen umfassen technische (z. B. Diensten für die Zitierbarkeit von Forschungsdaten), organisatorische (z. B. Festlegung von klaren Verantwortlichkeiten und organisatorischen Strukturen), finanzielle (z. B. Grundfinanzierung) rechtliche (z. B. transparente rechtliche Regelungen) und sonstige Aspekte (z. B. Etablierung von Anreizsystemen für die Wissenschaftler). Dabei ist die Anerkennung der Forschungsdaten als nationales Kulturgut eine wesentliche Grundbedingung.

Insgesamt kann festgehalten werden, dass sich bei dem Thema Forschungsdaten-Repositorien in Deutschland in den letzten Jahren viel bewegt hat, auf fachwissenschaftlicher, technologischer und politischer Ebene. Dabei hat sich auch gezeigt, dass die Technologie nur eine Seite der Herausforderungen darstellt. Die andere Seite besteht darin, sowohl die politischen als auch strukturellen Rahmenbedingungen für den Aufbau und den dauerhaften Betrieb von fachwissenschaftlichen Forschungsdaten-Repositorien zu schaffen, als auch die Fachwissenschaftler sowie die weiteren Akteure (Infrastruktureinrichtungen wie Rechenzentren und Bibliotheken) in einem organisatorischen Gesamtkonzept sinnvoll einzubeziehen. Es bleibt abzuwarten, wie die Öffentlichkeit und die Politik auf den Gesamtbericht der KII reagieren und welche konkreten Maßnahmen in Deutschland ergriffen und umgesetzt werden.

38. Der Bericht der Arbeitsgruppe „Forschungsdaten" ist im „Gesamtkonzept" publiziert, vgl. WGL, 2011.

Literaturhinweise

Allianz der deutschen Wissenschaftsorganisationen, 2010. *Grundsätze zum Umgang mit Forschungsdaten*. Online. http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze/ [Zugriff am 14.08.2011].

Allianz der deutschen Wissenschaftsorganisationen, o.J. *Forschungsprimärdaten*. Online: http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/ [Zugriff am 14.08.2011].

Aschenbrenner, A. & Kaiser, M., 2005. *White Paper on Digital Repositorys. reUSE! Deliverable*. Online: http://www2.uibk.ac.at/reuse/docs/reuse-d11_whitepaper_10.pdf [Zugriff am 14.08.2011].

Aschenbrenner, A. Blanke, T. Küster, M. W. & Pempe, W., 2010. Towards an Open Repository Environment. *Journal of Digital Information* (JoDI), 11(1).

Berliner Erklärung, 2003. *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.* (Stand: 22.10.2003) Online: http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/ [Zugriff am 09.08.2011].

Borghoff, U. M. et al., 2005. *Vergleich bestehender Archivierungssysteme*. (NESTOR-Materialien, 3) Online: http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20050117016 [Zugriff am 14.08.2011].

CCSDS (Consultative Committee for Space Data Systems), 2002. *Reference Model for an Open Archival Information System (OAIS)*. (CSSDS 650.0-B-1) (Jan. 2002) Online: http://public.ccsds.org/publications/archive/650x0b1.pdf [Zugriff am 14.08.2011].

DELOS, o.J. *A Reference Model for Digital Library Management Systems*. Online: http://www.delos.info/index.php?option=com_content&task=view&id=345&Itemid= [Zugriff am 14.08.2011].

Dobratz, S. & Schoger, A., 2010. Kapitel 8.3 Evaluierung der Vertrauenswürdigkeit digitaler Archive. In: Heike Neuroht et al., Hrsg. 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* (Version 2.3.) Online: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_78.pdf [Zugriff am 14.08.2011].

DRIVER (Digital Repository Infrastructure Vision for European Research), 2009. *Enhanced Publications*. Online: http://www.driver-repository.eu/Enhanced-Publications.html [Zugriff am 14.08.2011].

Fedora Commons, 2007. *The Fedora Content Model Architecture (CMA)*. (Version 3.0 Beta 1) Online: http://www.fedora-commons.org/

documentation/3.0b1/userdocs/digitalobjects/cmda.html [Zugriff am 14.08.2011].

Fedora Repository Development, 2010. *High Level Storage.* (Stand: 07.12.2010) Online: https://wiki.duraspace.org/display/FCREPO/High+Level+Storage [Zugriff am 14.08.2011].

Heery, R. & Anderson, S., 2005. *Digital Repositorys Review.* Online: http://www.jisc.ac.uk/uploaded_documents/digital-Repositorys-review-2005.pdf [Zugriff am 16.08.2011].

ICU WDS (International Council for Science World Data System), 2010. *ICSU World Data System (Home)*.Online: http://icsu-wds.org/ [Zugriff am 14.08.2011].

Library of Congress, 2011. *SRU Search / Retrieval via URL*. (Stand: 04.08.2011) Online: http://www.loc.gov/standards/sru/ [Zugriff am 14.08.2011).

Minton Morris, C., 2008. *DSpace Foundation and Fedora Commons Receive Grant from the Mellon Foundation for DuraSpace*. (Stand: 11.11.2008, 9:21 am) Online: http://expertvoices.nsdl.org/hatcheck/2008/11/11/dspace-foundation-and-fedora-commons-receive-grant-from-the-mellon-foundation-for-duraspace/ [Zugriff am 14.08.2011].

NESTOR, 2010. *AG Vertrauenswürdige Archive – Zertifizierung (aufgegangen in DIN NABD 15)*. (Stand: 14.12.2010) Online: http://www.langzeitarchivierung.de/arbeitsgruppen/agkritkat.htm [Zugriff am 14.08.2011].

NGDC (National Geophysical Data Center), o.J. *World Data System*. Online: http://www.ngdc.noaa.gov/wdc/wdcmain.html [Zugriff am 14.08.2011].

NGDC (National Geophysical Data Center), 2009. *List of current WDCs*. (Last Revised: 30.06.2006) Online: http://www.ngdc.noaa.gov/wdc/list.shtml [Zugriff am 14.08.2011].

NSSDC (National Space Science Data Center), o. J. *ISO Archiving Standards*. Online: http://nssdc.gsfc.nasa.gov/nost/isoas/ [Zugriff am 16.8.2011].

Open Archives, o.J. *Open Archives Initiative – Protocol for Metadata Harvesting*. Online: http://www.openarchives.org/pmh/ [Zugriff am 14.08.2011].

OSI (Open Society Institute), 2004. *Guide to Institutional Repository Software*. 3. ed. Online: http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf [Zugriff am 14.08.2011].

Payette, S. & Lagoze, C., 1998. Flexible and Extensible Digital Object and Repository Architecture (FEDORA). In: Nikolaou, C., ed. 1998. Research and advanced technology for digital libraries, *Second European Conference on Research and Advanced Technology for Digital Libraries*. (LNCS 1513) Heraklion, Kreta, Griechenland 21.-23. Sept. 1998. Berlin: Springer, S. 41–59. Online: http://www.cs.cornell.edu/payette/papers/ECDL98/FEDORA.html [Zugriff am 14.08.2011].

Pepe, A., Mayernik, M., Borgman, C. L. & Van de Sompel, H., 2009. From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *JASIST Journal of the American Society for Information Science and Technology,* 61(3). Online: http://arxiv.org/ftp/arxiv/papers/0906/0906.2549.pdf [Zugriff am 14.08.2011].

Thibodeau, K., 2002. *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years.* Online: http://www.clir.org/pubs/reports/pub107/thibodeau.html [Zugriff am 14.08.2011].

WissGrid, 2010. *WissGrid-Spezifikation: Grid-Repository.* Online: http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.5.2-grid-repository-spezifikation.pdf [Zugriff am 14.08.2011].

WissGrid, 2011. *Grid für die Wissenschaft*. (Stand: 18.04.2011) Online: http://www.wissgrid.de [Zugriff am 14.08.2011].

WGL (Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz e.V./ Leibniz Gemeinschaft), 2011. *Informationsstruktur*. Online: http://www.wgl.de/?nid=infrastr&nidap=&print=0 [Zugriff am 14.08.2011].

2.6 Langzeiterhaltung digitaler Forschungsdaten

Jens Klump
HelmholtzZentrum Potsdam Deutsches GeoForschungsZentrum GFZ – Zentrum für GeoInformationstechnologie (ZeGIT)

2.6.1 Herausforderung Langzeiterhaltung

Digitale Forschungsdaten sind sowohl die unverzichtbare Basis als auch das Resultat wissenschaftlicher Arbeit, wie sie sich seit dem 17. Jahrhundert entwickelt hat. Jede Disziplin hat ihre eigenen technischen und inhaltlichen Ansprüche, ihre Datenbestände und Konventionen, so dass Forschungsdaten alle Formate und Medientypen umfassen: von einfachen Tabellen im Textformat für z. B. physikalische Messdaten über semantisch ausgezeichnete Textsammlungen der Literaturwissenschaften bis hin zu interaktiven 3D-Modellen von Ingenieuren (Altenhöner et al., 2009).

Allen ist aber gemeinsam, dass sie eine wertvolle, jedoch schwierig zu handhabende Ressource darstellen. Forschungsdaten werden aufwendig produziert, im Falle von Beobachtungsdaten wie z. B. Wetterbeobachtungen können sie einmalig und nicht wiederherstellbar sein, sie können riesige Datenmengen umfassen. Dies gilt insbesondere für unwiederholbare Messungen, wie z. B. Umweltmessdaten (Pfeiffenberger, 2007), denn diese gehören meist zu Sammlungen, die aktiv genutzt werden und in Teilen als Referenzdaten für andere Arbeiten dienen (National Science Board, 2005). Andere Datensätze dienen lediglich der Sicherung im Sinne einer guten wissenschaftlichen Praxis und werden mit der Zeit immer weniger nachgefragt (Severiens & Hilf, 2006).

Wie alle digitalen Daten werden sie ohne langfristig wirksame Gegenmaßnahmen technisch veralten, nicht mehr nutzbar sein und damit verloren gehen. Hinzu kommt, dass keine der Infrastrukturen für eine digitale Langzeitarchivierung sich dauerhaft betreiben lässt wenn es keine Nutzer gibt, denn erst wenn eine Nachfrage der Wissenschaft nach einer digitalen Langzeitarchivierung besteht, können dauerhafte Strukturen entstehen. Für ein Forschungsdatenarchiv ist es daher eine der wichtigsten Aufgaben, seine Zielgruppe zu definieren, denn an ihr müssen sich die angebotenen Gruppen, Dienstleistungen, organisatorischen und technischen Prozesse orientieren (Parsons & Duerr, 2005).

Die nachhaltige Speicherung und auch Nachnutzung von Forschungsdaten[1] ist eine strategische, organisatorische und technische Aufgabe, die nur gemeinsam von wissenschaftlichen Anwendern und Datenzentren gelöst werden kann. (Altenhöner et al., 2009)

2.6.2 Der Daten-Lebenszyklus – Vom Labor bis ins Archiv

Im Laufe des digitalen Lebenszyklus von Forschungsdaten werden in den verschiedenen Phasen sehr unterschiedliche Anforderungen an die Persistenz der Daten und der Werkzeuge zum Umgang mit Forschungsdaten gestellt. Zwischen dem Entstehen der Daten in wissenschaftlichen Arbeitsprozessen und der sicheren, nachnutzbaren Archivierung der Daten besteht ein breites Spektrum von teilweise gegensätzlichen Anforderungen, auch *Digital Curation Continuum* genannt. Organisatorisch ist ein Kontinuum allerdings nicht handhabbar, weswegen es notwendig ist, innerhalb einer Organisation zu bestimmen, wer in welcher Phase des Lebenszyklus von Forschungsdaten für deren Pflege verantwortlich ist. Auf Grund des vorhandenen Kontextwissens reicht in den Phasen vor der Speicherung in der dauerhaften Domäne ein eingeschränktes Metadatenprofil aus, das bei der Überführung in die nächste Domäne angereichert werden muss, da in der nachfolgenden Domäne dieses Kontextwissen meist fehlt. Der Prozess der Anreicherung der Metadaten kann teilweise bis vollautomatisch erfolgen (Treloar et al., 2007; Treloar & Harboe-Ree, 2008).

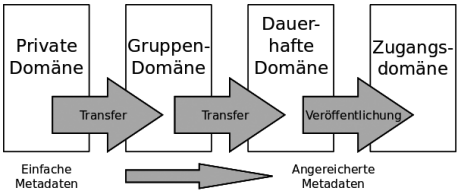

Abb. 1: Gliederung des *Data Curation Continuum* in vier Verantwortungsdomänen. Abhängig von den vorliegenden institutionellen Strukturen sind auch andere Gliederungen denkbar. Im Prozess des Datentransfers werden die vorliegenden Metadaten um wei-

1. In der Vergangenheit wurde oft von „Primärdaten“, bzw. „Forschungsprimärdaten“ (DFG 2009, 1998) gesprochen. Der Begriff „Primärdaten“ bedarf jedoch einer Klärung, denn für den einen sind „Primärdaten“ der Datenstrom aus einem Gerät, z. B. einem Satelliten, für andere sind „Primärdaten“ zur Nachnutzung aufbereitete Daten, wieder andere differenzieren nicht nach Grad der Verarbeitung sondern betrachten alle Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren, als Primärdaten. Der begrifflichen Klarheit wegen sollte daher das Präfix „Primär-“ nicht mehr verwendet und statt dessen nur noch die Begriffe „wissenschaftlichen Daten“ oder „Forschungsdaten“ verwendet werden.

tere Elemente angereichert. Dies geschieht mit der Unterstützung von Informationsfachleuten und mit informationstechnischen Werkzeugen. (Abbildung Klump, 2009)

Organisatorisch und technisch müssen sich diese Prozesse möglichst nahtlos in die wissenschaftlichen Arbeitsabläufe eingliedern. Allerdings sind wissenschaftliche Daten geprägt durch ihre Herkunft aus experimentellem Vorgehen, d.h. anders als Daten aus Arbeitsabläufen der Industrie oder Verwaltung stammen Forschungsdaten überwiegend aus informellen Arbeitsabläufen, die immer wieder ad hoc an die untersuchte Fragestellung angepasst werden (Barga & Gannon, 2007).

Der kritischste Moment im Lebenszyklus von Forschungsdaten ist, wenn das Projekt endet, denn hier endet meistens auch die Finanzierung weiterer Maßnahmen zur Datenerhaltung und das Interesse der Forscher ist bereits auf das nächste Projekt gerichtet. Um diese Klippe zu umschiffen ist es daher notwendig die zu archivierenden Daten möglichst früh in das Archiv zu überführen. Dieser Schritt ist im gesamten Betrieb eines Forschungsdatenarchivs der aufwendigste und damit auch teuerste, der ca. 45% der Gesamtkosten der Langzeitarchivierung eines digitalen Objekts verursacht (Beagrie et al., 2010). Zudem sind die Risiken unkontrollierbarer Kostensteigerungen hier am höchsten (Digital Preservation Testbed, 2005).

Da in den meisten Fällen keine Formatvorgaben bestehen werden Forschungsdaten in einer Vielfalt von Dateiformaten hergestellt. Häufig sind sie semantisch uneinheitlich strukturiert und nur lückenhaft mit Metadaten beschrieben.[2] Diese Faktoren stellen für die digitale Langzeitarchivierung von Forschungsdaten eine größere Herausforderung dar (Abrams, 2007; Klump, 2008; Lormant et al., 2005). Ein Forschungsdatenarchiv sollte daher Vorgaben machen können, welche Datenformate es für die Langzeitarchivierung annehmen kann. Auch die vom Archiv angenommenen Datenformate müssen langfristig überwacht werden um rechtzeitig Maßnahmen ergreifen zu können, sollte ein Datenformat wegen seiner Software- oder Hardwarevoraussetzungen zu veralten drohen (Curtis et al., 2007).

Ungeachtet des in der „Berliner Erklärung“ durch die Universitäten, Wissenschafts- und Forschungsförderungsorganisationen geleisteten Bekenntnisses zum offenen Zugang zu wissenschaftlichem Wissen gibt es Gründe, aus denen manche Daten nicht offen zugänglich gemacht werden können. Anders als im kommerziellen Bereich dienen Zugriffsbeschränkungen hier nicht in erster Linie der Sicherung von Verwertungsrechten oder dem Schutz vor Betriebsspionage, sondern sie sind entweder gesetzlich vorgeschrieben (z. B. Datenschutz) oder dienen dem Schutz von Personen oder Objekten, die durch eine Veröffentlichung der Daten gefährdet würden. Auch andere Gründe, die eine Zugriffsbe-

2. Siehe Kapitel 2.4 Metadaten und Standards

schränkung rechtfertigen, sind denkbar. In der digitalen Langzeitarchivierung von Forschungsdaten sind deshalb Zugriffsbeschränkungen ein wichtiges Thema. Für geschützte Datenobjekte müssen Verfahren und Richtlinien entwickelt werden, die auch über lange Zeiträume hinweg zuverlässig die Zugriffsrechte regeln und durchsetzen können. Die Bedingungen sollten jedoch so gestaltet sein, dass dadurch weder der wissenschaftliche Erkenntnisgewinn behindert wird, noch notwendige Maßnahmen der digitalen Langzeiterhaltung der Daten blockiert werden. Die bisher üblichen „Identitätssilos" bergen auf lange Sicht die Gefahr, dass deren Inhalte veralten oder dass der Aufwand, die Inhalte zu pflegen, enorm hoch wird. Abhilfe könnten vernetzte Systeme schaffen, indem sie einen zertifikatbasierten Ansatz oder Techniken sozialer Netzwerke nutzen (Choi et al., 2006). Ebenso muss der Umgang mit „verwaisten" Datenbeständen geregelt werden. Zusätzlich sollten die Lizenzen auch maschinenlesbar hinterlegt werden, um den Umgang mit komplexen Zusammenstellungen von Daten zu erleichtern.

2.6.3 Kriterienkataloge für digitale Archive und Zertifizierung

Große Teile der wissenschaftlichen Überlieferung in Form von Daten und Texten aus wissenschaftlicher Forschung liegen heute digital vor, und in vielen Fällen ausschließlich in digitaler Form. Sie zu erhalten, sollte nicht allein Aufgabe der Wissenschaft sein, diese muss auch durch Fachleute aus den Gedächtnisorganisationen unterstützt werden, zu denen inzwischen zunehmend auch Datenzentren und andere digitale Archive gehören. Wie lässt sich jedoch die Vertrauenswürdigkeit eines digitalen Archivs feststellen?

Die Notwendigkeit systematischer Datenarchivierung wurde bereits früh erkannt. Bereits 1995 beauftragte die *International Organization for Standardization* (ISO) das *Consultative Committee for Space Data Systems* (CCSDS) damit eine Norm für die Langzeitarchivierung von Daten aus Weltraummissionen zu erarbeiten (ISO 14721:2003). Im Laufe der Arbeiten am Normentwurf wurde klar, dass ein Referenzmodell für die Entwicklung weiterer, davon abgeleiteter Normen notwendig wäre. Das Ergebnis dieser Arbeiten war das *Open Archival Information System Reference Model* (OAIS-RM), das heute den meisten als „OAIS-Modell" bekannt ist, und seit dem mehrfach überarbeitet wurde (CCSDS, 2009; Rank et al., 2010). Auch in Deutschland wurden umfangreiche Arbeiten zur Langzeiterhaltung von Forschungsdaten geleistet, insbesondere im Kontext des Kompetenznetzwerks Langzeitarchivierung (NESTOR) (z. B. Altenhöner et al., 2009; Dobratz et al., 2008; Klump, 2009; Severiens & Hilf, 2006).

Das OAIS-RM war auch die Grundlage für eine Reihe von Kriterienkatalogen, mit denen die Vertrauenswürdigkeit eines digitalen Langzeitarchivs ermit-

telt werden soll. Trotz scheinbar paralleler Arbeiten sind die wichtigsten Kataloge (TRAC[3], NESTOR[4] und *Data Seal of Approval*[5]) auf einander abgestimmt. Zwischen den CCSCS[6], DIN NABD 15[7], AG Vertrauenswürdige Archive und DANS / *Data Seal of Approval* wurde im Sommer 2010 eine weitere Zusammenarbeit und Abstimmung der Aktivitäten vereinbart (Klump, 2011).

2.6.4 Zusammenfassung

Umfangreiche Vorarbeiten der vergangenen Jahre haben dazu geführt, dass uns eine ganze Reihe von Werkzeugen für die Langzeiterhaltung von Forschungsdaten zur Verfügung steht. Diese Werkzeuge sind jedoch noch unzureichend in die wissenschaftlichen Arbeitsabläufe integriert, die daraus entstehenden Medienbrüche verursachen immer noch einen Mehraufwand, der selten geleistet wird. In einzelnen Fachgebieten gibt es jedoch bereits vorbildliche Lösungen, die sich auch auf andere Fachgebiete übertragen lassen.

Um das Ziel einer nachhaltigen digitalen Langzeitarchivierung von Forschungsdaten zu erreichen, muss eine Strategie verfolgt werden, die Langzeitarchivierung von Daten zu einem anerkannten Beitrag zur wissenschaftlichen Kultur macht und in den institutionellen Arbeitsabläufen verankert. Diese organisatorische Strategie muss gleichzeitig von einer technischen Strategie unterstützt werden, die den Akteuren für die digitale Langzeitarchivierung von wissenschaftlichen Forschungsdaten geeignete Werkzeuge in die Hand gibt. Wissenschaft und Wissenschaftsorganisationen haben dies erkannt, so dass in den nächsten Jahren mit erheblichen Fortschritten gerechnet werden darf.[8]

---

3. Trustworthy Repositories Audit & Certification: Criteria and Checklist (Ambacher et al., 2007).
4. (Dobratz et al., 2008).
5. (Sesink et al., 2008).
6. Consultative Committee for Space Data Systems, s.o..
7. DIN Normenausschuss Bibliotheks- und Dokumentationswesen 15, Schriftgutverwaltung und Langzeitverfügbarkeit digitaler Informationsobjekte
8. Verweis auf Kapitel 3.4 Archivierung von Forschungsdaten

Literaturhinweise

Abrams, S., 2007. File Formats. In: S. Ross & M. Day, S., Hrsg. 2007. *DCC Digital Curation Manual*., Glasgow: Digital Curation Centre, 53 S. Online: http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/file-formats/file-formats.pdf [Zugriff am 18.08.2011].

Altenhöner, R. et al., 2009. *Digitale Forschungsdaten bewahren und nutzen – für die Wissenschaft und für die Zukunft*. Göttingen: Niedersächsische Staats und Universitätsbibliothek. urn:nbn:de:0008-2009071031

Ambacher, B. et al., 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC).* Chicago, IL.: CRL Center for Research Libraries. Online: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf [Zugriff am 18.08.2011].

Barga, R. & Gannon, D. B., 2007. Scientific versus business workflows. In: I. J. Taylor, E. Deelman, D. B. Gannon, & M. Shields, Hrsg. 2007. *Workflows for e-Science*. London: Springer, S. 9–16. doi:10.1007/978-1-84628-757-2_2

Beagrie, N., Lavoie, B. F. & Woollard, M., 2010. *Keeping research data safe 2*. Bristol: Joint Information Systems Committee (JISC). Online: http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx [Zugriff am 18.08.2011].

CCSDS (Consultative Committee for Space Data Systems), 2009. *Audit and certification of trustworthy digital repositories.* Draft Recommended Practice, Red Book. Grenbelt, MD: CCSDS. Online: http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206520R1/Attachments/652x0r1.pdf [Zugriff am 18.08.2011].

Choi, H. et al., 2006. Trust Models for Community Aware Identity Management. *WWW2006.* Edinburgh. Online: http://hdl.handle.net/10379/488 [Zugriff am 19.09.2011].

Curtis, J. et al., 2007. AONS – An obsolescence detection and notification service for Web archives and digital repositories. *New Review of Hypermedia and Multimedia*, 13(1), S. 39–53. doi:10.1080/13614560701423711

DFG (Deutsche Forschungsgemeinschaft), 1998. *Sicherung guter wissenschaftlicher Praxis*. Online: http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf [Zugriff am 18.08.2011].

DFG (Deutsche Forschungsgemeinschaft), 2009. *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. Online: http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/veroeffentlichungen/dokumentationen/download/ua_inf_empfehlungen_200901.pdf [Zugriff am 18.08.2011].

Digital Preservation Testbed, 2005. *Costs of Digital Preservation*. From digital volatility to digital permanence. Den Haag: Nationaal Archief. Online: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf [Zugriff am 18.08.2011].

Dobratz, S. et al., 2008. *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive*. NESTOR-Materialien. 2. Aufl. Frankfurt am Main: Deutsche Nationalbibliothek. urn:nbn:de:0008-2008021802.

Klump, J., 2008. *Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten*. NESTOR-Materialien. Expertise. Frankfurt am Main: Kompetenznetzwerk Langzeitarchivierung (NESTOR). urn:nbn:de:0008-2008040103.

Klump, J., 2009. Digitale Forschungsdaten, In: H. Neuroth et al., Hrsg. 2009. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. S. 104–115. urn:nbn:de:0008-20100305375.

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

Lormant, N. Huc, C. Boucon, D. & C. Miquel, 2005. How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information? *Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005)*. Edinburgh, 11 S. Online: http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/003.pdf [Zugriff am 18.08.2011].

National Science Board, 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Washington, DC.: National Science Foundation. Online: http://www.nsf.gov/pubs/2005/nsb0540/ [Zugriff am 18.08.2011].

Parsons, M. A., & R. Duerr, 2005. Designating user communities for scientific data: challenges and solutions. *Data Science Journal*, (4), S. 31–38. doi:10.2481/dsj.4.31.

Pfeiffenberger, H., 2007. Offener Zugang zu wissenschaftlichen Primärdaten. *Zeitschrift für Bibliothekswesen und Bibliographie*, 54(4-5), S. 207–210. hdl:10013/epic.28454.d001

Rank, R. H., Cremidis, C. & K. R. McDonald, 2010. Archive Standards: How Their Adoption Benefit Archive Systems, In: L. Di & H. K. Ramapriyan, Hrsg. 2010. *Standard-Based Data and Information Systems for Earth Observation.* Berlin: Springer, S. 127–142. doi:10.1007/978-3-540-88264-0_8.

Sesink, L. Horik, R. van & Harmsen, H., 2008. *Data Seal of Approval*. Den Haag: Data Archiving and Networked Services (DANS). Online: http://www.datasealofapproval.org/

Severiens, T. & Hilf, E. R., 2006. *Langzeitarchivierung von Rohdaten.* (NESTOR-Materialien 6). Frankfurt am Man: NESTOR – Kompetenznetzwerk Langzeitarchivierung. urn:nbn:de:0008-20051114018

Treloar, A. Groenewegen, D. & C. Harboe-Ree, 2007. The Data Curation Continuum – Managing Data Objects in Institutional Repositories. *D-Lib Magazine*, 13(9/10), S. 13. doi:10.1045/september2007-treloar.

Treloar, A. & Harboe-Ree, C., 2008. Data management and the curation continuum: how the Monash experience is informing repository relationships. *VALA2008 14th Biennial Conference*. Melbourne, Australien 5.-7. Feb. 2008. Online: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf [Zugriff am 18.08.2011].

2.7 Systeme und Systemarchitekturen für das Datenmanagement

Matthias Razum
FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur

2.7.1 Einführung

Datenmanagementsysteme (DMS) stellen die technische Basis für die Erfassung, Anreicherung und Bereitstellung von Forschungsdaten dar. Sie umfassen typischerweise neben der eigentlichen Speicherung der Datenobjekte weitere Dienste, etwa zur Registrierung, Suche oder Verwaltung von Zugriffsrechten. Die Publikation von Daten und damit verbundene Dienste gehören im Allgemeinen nicht dazu. Treloar, Groenewegen und Harboe-Ree unterscheiden in „*The Data Curation Continuum*“ (2007) zwei grundlegende Domänen der Verwaltung von Forschungsdaten:

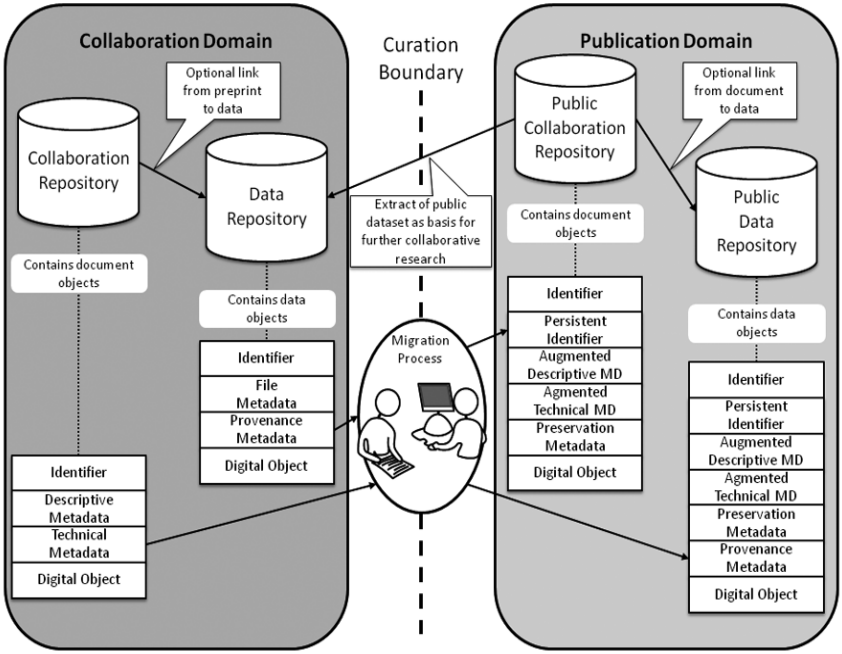

Abb. 1: Collaboration, Publication, and the Curation Boundary (Quelle: nach Treloar, 2007)

DMS decken meist die linke Domäne ab, also das Management von Daten für die eigentliche Forschung und die Zusammenarbeit in (verteilten) Forschungs-

gruppen. Die Publikationsdomäne auf der rechten Seite wird meist durch spezialisierte Systeme abgedeckt.

Publizierte Daten müssen langfristig (perspektivisch: *ad infinitum*) verfügbar bleiben, wie herkömmliche Publikationen zitiert und damit die Grundlage weiterer Publikationen bilden können. Gleichzeitig sind diese Daten statisch, d.h. sie unterliegen keiner Änderung mehr. Derartige Eigenschaften werden zunehmend als Grundvoraussetzung publizierter Daten eingefordert. Entsprechend entwickeln sich zurzeit Zertifizierungsverfahren für Datenzentren, um diese Leistungen zuzusichern, etwa durch *DataCite*[1], aber auch aus der Praxis heraus wie zum Beispiel bei der *Open Access* Datenpublikation *Earth System Science Data*[2], die Standards für die Repositorien einfordert, in denen die publizierten Daten hinterlegt sind.

Systeme zum Datenmanagement verwalten dagegen potenziell alle Daten, auch solche, die nie publiziert werden. Sie berücksichtigen Daten mit einer begrenzten Lebensdauer und eingeschränkter Sichtbarkeit. Sie stellen diese für die weitere Verwendung innerhalb eines Instituts oder einer Forschergruppe bereit, etwa zur Analyse, Aggregation oder für Vergleiche mit anderen Daten. Diese Daten können sich noch ändern bzw. in neueren Versionen gespeichert werden. Hier sind Funktionen wie *Lifecycle Management*, z. B. zur automatischen Überwachung von Haltefristen, etwa gemäß den Empfehlungen der Deutschen Forschungsgemeinschaft (DFG, 1998), Versionierung und Autorisierung gefragt. Oft wählen Autoren nur einen Bruchteil der hier verwalteten Daten aus und machen sie in der „*Publication Domain*“ allgemein verfügbar. Der Übergang von Daten aus einem DMS über die *Curation Boundary* zu einem Datenzentrum sieht z. B. das neue *World Data System* der ICSU (2008) mit der Unterscheidung von „*Data Collection and Processing Facilities*“ und den „*Data Archiving and Publication Facilities*“ vor.

Datenmanagement bildet damit eine grundlegende Voraussetzung für die Datenpublikation, ist aber für sich alleine genommen schon eine der meistgeforderten Dienstleistungen im wissenschaftlichen Alltag, wie aktuelle Befragungen zeigen (Kroll & Forsman, 2010; TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010).

2.7.2 Funktionale Anforderungen an Systeme zum Datenmanagement

Die Anforderungen an ein Datenmanagementsystem (DMS) hängen von den verwalteten Datentypen und den Einsatzszenarien des Systems ab. Es lassen

1. http://www.datacite.org/ [Zugriff am 13.08.2011].
2. http://earth-system-science-data.net/ [Zugriff am 13.08.2011].

sich allerdings einige grundlegende Funktionen benennen. Die folgende Aufzählung erhebt weder Anspruch auf Vollständigkeit, noch müssen immer alle funktionalen Anforderungen für ein spezifisches DMS implementiert sein. Allerdings findet man diese Funktionen in vielen Systemen wieder; sie bilden damit generisch Anforderungen an Systeme zum Datenmanagement ab.

2.7.2.1 Verknüpfung von Daten und Metadaten

Daten ohne Beschreibung sind meist wertlos. Messdaten müssen mit Einheiten verknüpft sein, es muss klar sein, was und wie gemessen wurde. Zu den Messdaten gehören Konfigurations- und Kalibrierungsdaten der Instrumente. Die Entstehungs- und Bearbeitungsgeschichte („*Provenance*“) ist wichtig: wer hat wann was mit den Daten gemacht? Sind die Daten Rohdaten, bereits um Fehler bereinigt oder gar aggregiert? Welche Algorithmen oder Webservices wurden zur Lemmatisierung von Texten herangezogen? Wer hat einen altsprachlichen Text transkribiert oder übersetzt? Alle diese Informationen sind für das Verständnis und die Nachnutzung der Forschungsdaten entscheidend, und entsprechend müssen diese Metadaten gemeinsam mit den eigentlichen Daten gespeichert und verwaltet werden. Dabei können die Metadaten unterschiedliche Ausprägungen haben:

- Technische Metadaten (z. B. Dateiformat, Dateigröße, Mime-type)
- *Provenance* Metadaten (z. B. gemäß *Open Provenance Model* (Moreau et al., 2007) oder PREMIS (Caplan & Guenther, 2005)
- Deskriptive Metadaten (fachspezifisch)
- Lizenz-Metadaten

Aus der Vielzahl der möglichen Profile, insbesondere bei den fachspezifischen deskriptiven Metadaten ergibt sich die Anforderung an ein DMS, mehrere Metadatensätze gemäß beliebiger Profile mit den Forschungsdaten verwalten zu können. Während ein konkreter Metadatensatz durchaus über ein Profil definiert sein sollte, empfiehlt sich ein schemafreier Ansatz für das darunterliegende DMS.

Die Erstellung von Metadaten ist ein aufwändiger und im wissenschaftlichen Alltag nur schwer zu bewältigender Prozess. Gerade bei Datensätzen, die wahrscheinlich nie publiziert werden, unterbleibt vielfach die Beschreibung durch Metadaten. Damit verlieren diese Daten allerdings dramatisch an Wert, da sie quasi nicht mehr auffindbar und nur durch den Urheber noch zu interpretieren sind. Insofern sind für DMS Verfahren zur automatischen Metadatengenerierung interessant. Im Bereich der technischen Metadaten gibt es vielversprechende Ansätze, etwa das *File Information Tool Set* FITS (Stern & McEwen, 2009). Werden die Forschungsdaten bereits in einem sehr frühen Stadium – am besten zum Zeitpunkt der Entstehung – in einem DMS gespeichert, kann das DMS die *Provenance*-Metadaten weitgehend automatisch erfassen.

2.7.2.2 Versionierung

Daten verändern sich gegebenenfalls in den ersten Stadien ihres Lebenszyklus. Sie können um Fehler bereinigt, aggregiert und umformatiert werden. Aber nicht nur die Daten selbst, auch die sie beschreibenden Metadaten können in dieser Phase vor der eigentlichen Publikation ergänzt und korrigiert werden. Wichtig ist, Daten und Metadaten auch hier als Einheit zu betrachten und als Einheit zu versionieren.

Gerade bei verteilt arbeitenden Forschungsgruppen erlaubt eine versionierte Speicherung die Nachvollziehbarkeit von Änderungen. Sie können Personen und Zeitpunkten zugeordnet werden und bilden damit die Grundlage für *Provenance*-Metadaten – und damit Informationen, die die Einschätzung der Daten durch Dritte hinsichtlich Vertrauenswürdigkeit und Korrektheit erleichtern.

2.7.2.3 Datenformate

Forschungsdaten kommen in einer Vielzahl von Dateiformaten vor. Der Versuch einer Vereinheitlichung oder Beschränkung kollidiert mit der Dynamik des Forschungsprozesses. Ein DMS wird aber nur akzeptiert werden, wenn es die Forschenden unterstützt, statt sie einzuschränken. Darüber hinaus ist es oftmals sinnvoll, ein Datenobjekt in mehreren Repräsentationen abzuspeichern, etwa einen Scan eines Manuskripts in voller Auflösung im Dateiformat TIFF, eine reduzierte Auflösung für die Darstellung im Web im JPEG-Format und schließlich eine Vorschau als GIF-Datei. Ein anderer Anwendungsfall mehrerer Repräsentationen kann das Originalformat und eine in ein Standardformat migrierte Version sein. Also sollten DMS möglichst keine Annahmen hinsichtlich von Datenformaten treffen und die Verwaltung mehrerer Repräsentationen eines digitalen Objekts unterstützen.

Bei einer Vielzahl von zu unterstützenden Datenformaten kann es sinnvoll sein, einen Dienst zur Charakterisierung der Formate einzusetzen, um automatisch das Format zu bestimmen und geeignete technische Metadaten zu extrahieren. Ein Beispiel für einen solchen Dienst ist das bereits erwähnte FITS.

2.7.2.4 Semantische Relationen zwischen Datenobjekten

Datenobjekte stehen selten alleine und losgelöst da. Sie stehen in Beziehung zu anderen Datenobjekten. Messwerte sind mit Kalibrierungs- bzw. Konfigurationsdaten von Instrumenten verknüpft, Transkripte und Übersetzungen mit dem Originaltext, Fotos von archäologischen Artefakten mit Informationen zur Ausgrabungsstätte.

Meist drückt man diese Beziehungen durch Techniken des *Semantic Webs*[3] aus; insbesondere das *Resource Description Framework* RDF[4] und die *Web*

3. http://www.w3.org/2001/sw/ [Zugriff am 13.08.2011].
4. http://www.w3.org/RDF/ [Zugriff am 13.08.2011].

Ontology Language OWL[5] spielen hierbei eine wichtige Rolle. In den meisten Projekten oder zumindest Disziplinen kommen mehrere Ontologien zum Einsatz. Über „*(Open) Linked Data*“ (Campbell & MacNeill, 2010) entstehen zurzeit in Teilbereichen Standardisierungsbestrebungen, die sinnvollerweise modular aufgebaut und für einen konkreten Anwendungsfall evaluiert und entsprechend kombiniert werden müssen. Für DMS bedingt das die Unterstützung einer Vielzahl von Ontologien[6]. Auch müssen sie sowohl Relationen innerhalb des DMS als auch darüber hinaus berücksichtigen. Sinnvoll ist es, wenn ein DMS diese Relationen nicht nur verwalten, sondern über eine entsprechende Komponente (z. B. einen *Triple Store*) und eine standardisierte Abfragesprache (etwa SPARQL[7]) durchsuchbar machen.

2.7.2.5 Lebenszyklus von Datenobjekten

Datenobjekte durchlaufen verschiedene Phasen von ihrer Entstehung bis zur Publikation oder gegebenenfalls Löschung. Beispiele solcher Phasen umfassen die Entstehung des Datenobjekts, dessen Anreicherung mit Metadaten, das Durchlaufen einer (gegebenenfalls mehrstufigen) Qualitätssicherung, die Archivierung, die Auswahl und Publikation oder die Löschung des Datenobjekts nach einer festgelegten Haltefrist. Diese Phasen zusammengenommen bilden den Lebenszyklus eines Datenobjekts. Sie unterscheiden sich in ihren konkreten Ausprägungen zwischen Datentypen und Disziplinen, aber es lassen sich grundlegende Anforderungen ableiten:

- Die DFG empfiehlt in ihrer Denkschrift zur guten wissenschaftlichen Praxis eine Haltefrist von 10 Jahren für Daten, die die Basis für eine (herkömmliche) Publikation bilden (DFG, 1998). Grundsätzlich werden viele Daten nicht notwendigerweise auf Dauer gespeichert. Vielfach gibt man eine Haltefrist vor, nach deren Erreichen man über das weitere Vorgehen mit den Daten entscheidet.
- Bereitstellung einer Benachrichtigungsfunktion, die bei definierten *Trigger Events* (z. B. Ende einer Haltefrist, Ende einer Embargofrist, Freigabe eines Datensatzes) die Besitzer der Objekte bzw. die Administratoren informieren.
- Da Haltefristen oftmals weit über die Laufzeit von Projekten hinausgehen, sind die ursprünglichen Erzeuger der Objekte gegebenenfalls nicht mehr greifbar. In diesen Fällen ist ein *Ownership Management* sinnvoll, bei dem der Besitz an Objekten (und damit die Verfügungsgewalt, etwa hinsichtlich weiterer Haltefristen oder Löschungen) an andere Personen übergehen

5. http://www.w3.org/TR/owl-guide/ [Zugriff am 13.08.2011].
6. http://linkeddata.org/ [Zugriff am 13.08.2011].
7. http://www.w3.org/TR/rdf-sparql-query/ [Zugriff am 13.08.2011].

kann. Hierbei kommen neben technischen allerdings auch rechtliche Fragen ins Spiel, etwa im Zusammenhang von Urheber- und Nutzungsrecht: wem gehören die Daten, nachdem ein Mitarbeiter eine Institution verlässt?

2.7.2.6 Registrierung

Die Vergabe von *Persistent Identifiern* ist für die Publikation von Daten eine wichtige Voraussetzung. Üblicherweise vergibt man diese erst im Verlauf des Publikationsprozesses, also in der „*Publication Domain*“, und damit nur für wenige ausgewählte Datensätze und zu einem späten Zeitpunkt im Lebenszyklus der Datenobjekte. Es gibt allerdings Szenarien, in denen eine frühere Vergabe sinnvoll erscheint, gegebenenfalls sogar für eine Vielzahl von Datenobjekten oder -fragmenten. In der Computerlinguistik will man z. B. innerhalb eines Textkorpus häufig bis auf Wort- (bzw. *Token*) oder sogar Phonemebene hinab Fragmente über *Persistent Identifier* dauerhaft zitierfähig machen. In anderen Disziplinen gibt es enorm große Datensätze (z. B. in der Klimaforschung oder bei der Sequenzierung von Genomen), die oftmals nur in Ausschnitten zitiert werden sollen. In diesen Fällen kann es sinnvoll sein, schon im DMS der „*Collaboration Domain*“ für diese Fragmente *Persistent Identifier* zu vergeben. DMS sollten also Schnittstellen zu entsprechenden Systemen vorsehen.

2.7.2.7 Content Models

In den vorangegangenen Abschnitten wurde die Flexibilität immer wieder hervorgehoben, so dass ein DMS letztlich als BLOB-*Store* (*Binary Large Object*) erscheinen könnte. Tatsächlich erschwert aber eine fehlende Typisierung sowohl die Entwicklung fachspezifischer Anwendungen als auch die Validierung der gespeicherten Objekte. Insofern sollten DMS die Typisierung von Objekten erlauben. Ähnlich wie ein relationales Datenbanksystem erst durch ein Datenmodell eine sinnvolle Nutzung erlaubt, sollten DMS einen ähnlichen Mechanismus vorsehen: *Content Models*. Dies ist insbesondere dann relevant, wenn innerhalb eines DMS mehrere *Content Models* nebeneinander zum Einsatz kommen sollen. Datenobjekten weist man einem *Content Model* zu, dem sie „gehorchen“ und damit auch gewisse Eigenschaften zusichern.

Was genau ein *Content Model* vorschreibt, hängt einerseits vom konkreten Einsatzzweck ab, zum anderen aber auch von der zugrundeliegenden Architektur. Typischerweise will man aber zu verwendende Metadatenprofile, Dateiformate oder auch Formate für einzelne Eigenschaften festlegen können. Weiterhin können *Content Models* auch die Darstellung von Datenobjekten steuern.

2.7.2.8 Authentifizierung und Autorisierung

Wissenschaft findet heute meist in verteilten Arbeitsgruppen statt, die sich auch über Institutionsgrenzen hinweg erstrecken können. Oft findet man in diesem Zusammenhang den Begriff „virtuelle Organisation“, die z. B. Projektstrukturen

abbildet und gegebenenfalls nur eine kurze Lebensdauer hat. Virtuelle Organisationen stehen typischerweise orthogonal zur internen Aufbauorganisation der beteiligten Institutionen. Verteilte Authentifizierungssysteme wie etwa *Shibboleth* (Scavo & Cantor, 2005), aber auch *OpenID* (Recordon & Reed, 2006) und ähnliche kommerzielle Ansätze spielen hier eine zunehmend wichtige Rolle.

Forschungsdaten können explizit oder implizit sensitive Information enthalten. Sie geben Einblick in Arbeitsweisen, eingesetzte Verfahren und noch nicht publizierte Erkenntnisse. Daten können ohne hinreichende kontextuelle Informationen zu Fehlinterpretationen einladen, andere enthalten personenbezogene Informationen, die besonders zu schützen sind. Entsprechend sind nicht alle Forschungsdaten von Anfang an frei zugreifbar. Vielfach werden überhaupt nur ausgewählte, aggregierte oder anonymisierte Datensätze nach Veröffentlichung eines Artikels publiziert. In anderen Fällen werden aber auch schon sehr frühe Versionen öffentlich zugänglich gemacht, um Dritten zu erlauben, auf Basis dieser Daten eigene Publikationen zu erarbeiten (z. B. *Sloan Digital Sky Survey*[8]). Selbst innerhalb einer Institution oder Forschergruppe sind Zugriffsbeschränkungen durchaus üblich. Die Entscheidung über die entsprechenden Regeln sollte bei den Wissenschaftlern liegen.

Die Regeln selbst können sich sowohl am Zugreifenden und seinen Rechten als auch am Objekt und seinen Eigenschaften festmachen. Beispiele möglicher Kriterien für Zugriffsregeln umfassen:

- Unterscheidung von Daten und Metadaten
- Unterscheidung der verschiedenen Repräsentationen eines Objekts
- Embargofristen eines Objekts
- Status eines Objekts im Lebenszyklus
- Status eines Benutzers als Mitglied einer Arbeitsgruppe oder virtuellen Organisation

Die Komplexität der möglichen Regeln und zu berücksichtigenden Eigenschaften führt zur nächsten Herausforderung: Wissenschaftler müssen verstehen, welche Zugriffsrechte tatsächlich aus ihren Einstellungen resultieren. Dies erfordert eine Benutzungsoberfläche, die die Konsequenz einer Regeländerung verständlich darstellt, also den Spagat zwischen der Umsetzung komplexer Funktionalitäten und einfacher Bedienung schafft.

2.7.2.9 Vertrauen

Inwieweit kann man Daten Dritter vertrauen? In dieser Frage liegt eine der großen Herausforderungen für die Nachnutzung von Daten, die auch die *High Level Expert Group on Scientific Data* in ihrem Bericht (Giaretta et al., 2010) hervor-

8. http://www.sdss.org [Zugriff 13.08.2011].

hebt. Vertrauen ist weitgehend ein sozio-kultureller Prozess und kann nur teilweise durch technische Systeme unterstützt werden.

Etablierte Prozesse wie *Peer Reviewing* finden zunehmend auch bei der Publikation von Daten statt (z. B. bei der Zeitschrift *Earth System Science Data*[9]). Diese Zeitschrift definiert auch Standards, die von den DMS zugesichert werden müssen, um dort hinterlegte Daten für eine Publikation zu akzeptieren. Darunter finden sich etwa Anforderungen an persistente *Identifier*, *Open Access*, langfristige Archivierung und eine wissenschaftsfreundliche Lizenz. Alle diese Anforderungen beziehen sich auf die „*Publication Domain*". Doch auch in der „*Collaboration Domain*" kann durch technische Maßnahmen das Vertrauen in die Daten erhöht werden. Beispiele hierfür sind

- interne Freigabeprozesse mit Qualitätssicherung, siehe auch *Object Lifecycle*
- *Checksums*, um die Integrität der Daten abzusichern
- *Audit Trails*, um Änderungen an Daten nachverfolgen zu können

Besonders wichtig für die Einschätzung der Vertrauenswürdigkeit von Daten ist deren Entstehungsgeschichte. Es gibt einige Ansätze (Razum et al., 2010; Rajbhandari, Hedges & Fabiane, 2010), die Daten zum Zeitpunkt ihrer Entstehung in einem DMS zu erfassen und dabei kontextuelle Informationen automatisch einzubeziehen, etwa Zeitstempel, die Daten des angemeldeten Benutzers, Verknüpfungen zu Konfigurations- und Kalibrierungsdaten eingesetzter Instrumente oder verwendete Programme. Die Datenmanagementgruppe am IFM-GEOMAR in Kiel[10] arbeitet beispielsweise daran, die Untersuchung von Proben über *Workflows* abzubilden. Jeder Schritt eines *Workflows* (etwa das Reinigen einer Probe, die Durchführung einer Messung, usw.) erfasst neben den Daten auch automatisch Metadaten, die nachvollziehbar machen, welche Prozessschritte wann, von wem und wie oft durchlaufen wurden und wie es schließlich zum Endergebnis kam. DMS müssen in der Lage sein, solche Informationen zu verwalten und möglichst deren (semi-)automatische Erfassung zu unterstützen.

2.7.3 Grundlegende Architekturen

Die Funktionalität von DMS lässt sich in Schichten aufteilen:

- eine Persistenzschicht für die eigentliche Speicherung der Daten,
- eine Kernschicht für die zentralen Funktionen eines DMS sowie
- eine Diensteschicht mit erweiterten Funktionen.

9. http://earth-system-science-data.net/ [Zugriff am 13.08.2011].
10. https://portal.ifm-geomar.de/web/guest/about-us [Zugriff am 13.08.2011].

Diese Aufteilung ist logisch zu verstehen; je nach Art des Systems können diese Schichten explizit oder eher implizit vorhanden sein:

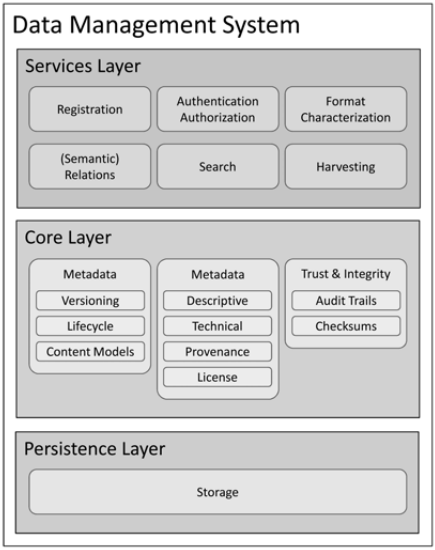

Abb. 2: Schichten und Komponenten einer DMS-Architektur

Bei der Implementierung der Persistenzschicht eines DMS muss man zwischen datensatzorientierten und dateiorientierten Systemen unterscheiden. Erstere verwenden typischerweise eine relationale Datenbank, letztere basieren meist auf sogenannten *Digital Object Repositories* (DOR). Vielfach lassen sich Daten sowohl datensatz- als auch dateiorientiert darstellen. Eine Entscheidung zwischen beiden Optionen fällt meist im Spannungsfeld zwischen langfristiger Archivierung von Daten (dateiorientiert) und interaktivem Arbeiten mit den Daten (datensatzorientiert).

Weiterhin kann man Einzelsysteme von verteilten Systemen unterscheiden. Zu letzteren gehören *Grid*-basierte Systeme und zunehmend auch *Cloud*-basierte Ansätze. Die folgenden Abschnitte führen jeweils einige Vor- und Nachteile der unterschiedlichen Ansätze an und geben Beispiele für derartige Systeme.

Für die Implementierung einer Benutzungsoberfläche für DMS in der „*Collaboration Domain*“ kommen häufig pragmatische Ansätze zum Tragen, etwa die Verwendung von Wikis oder *Content Management* Systeme wie z. B. Plone oder Drupal (ein Beispiel hierfür ist Islandora[11]).

11. http://islandora.ca/ [Zugriff am 13.08.2011].

2.7.3.1 Datenbanken

Datenbanken, und hier besonders relationale Datenbanken, bieten sich für strukturierte Daten an. Der größte Vorteil von relationalen Datenbanksystemen sind die seit Jahren bewiesene Leistungsfähigkeit und Zuverlässigkeit der Software sowie die vorhandene umfangreiche Erfahrung mit ihnen. Funktionen wie Transaktionalität, Clusterfähigkeit und hohe Geschwindigkeit sowie etablierte Werkzeuge zur Datensicherung und -wiederherstellung sprechen für Datenbanken. Wichtig ist darüber hinaus die Möglichkeit, auch aus sehr großen Datenmengen über *Queries* beliebige Teilmengen zu extrahieren. Beispiele für derartige Systeme sind *Data Warehouses* in der Bioinformatik[12] oder das *World Data Center* WDC-MARE[13], das Daten aus dem Bereich der Meereskunde vorhält.

Die Skalierbarkeit von Datenbanken in den Terabyte-Bereich erfordert allerdings ein großes technisches Verständnis und ist mit hohem administrativem Aufwand verbunden. Eine weitere Herausforderung von Datenbank-basierten Systemen ist die Nachverfolgung von Änderungen, die Versionierung und die Erfassung von *Provenance*-Daten. Dies ist nicht technisch bedingt, sondern resultiert aus der meist pragmatisch erfolgten Datenmodellierung, die diese Aspekte in der Regel nicht beachten. Gerade bei aufwändig aufgebauten Datensammlungen kann diese Unterlassung den Wert der Daten, insbesondere hinsichtlich einer Nachnutzung, einschränken. Buneman spricht hier von der Notwendigkeit sogenannter „*Curated Databases*“ (Buneman, Cheney, Tan & Vansummeren, 2008), die vom Design und der Implementierung her aber deutlich aufwändiger sind und Erfahrung mit den grundlegenden funktionalen Anforderungen (s.o.) erfordern.

Neben relationalen Datenbanksystemen können auch spezialisierte Systeme zum Einsatz kommen, z. B. XML-Datenbanken oder sogenannte No-SQL-Datenbanken[14]. Letztere skalieren sehr gut und lassen sich einfach verteilen, bieten aber nur eingeschränkte Konsistenz („*eventual consistency*“). Das *Compact Muon Solenoid* (CMS) Experiment, Teil des Large *Hadron Colliders* am CERN, nutzt zum Beispiel für sein Datenmanagement *Couch*-DB[15]. In beiden Fällen handelt es sich aber um relativ neue Technologie, zu der noch wenig Erfahrung existiert.

12. Beispiele für derartige Systeme sind die Lipase *Engineering Database* (http://www.led.uni-stuttgart.de/ [Zugriff am 13.08.2011]) bzw. die CYP450 *Engineering Database* (http://www.cyped.uni-stuttgart.de/ [Zugriff am 13.08.2011]) am Institut für Technische Biochemie an der Universität Stuttgart.

13. http://www.wdc-mare.org/ [Zugriff am 13.08.2011].

14. Beispiele für No-SQL-Datenbanken sind Cassandra (http://cassandra.apache.org/ [Zugriff am 13.08.2011]) oder *Couch*-DB (http://couchdb.apache.org/ [Zugriff am 13.08.2011]).

15. http://www.couch.io/case-study-cern [Zugriff am 13.08.2011].

2.7.3.2 Digital Object Repositories

In vielen Fällen liegen Daten aber nicht strukturiert, sondern semi- oder unstrukturiert als Dokumente oder Dateien vor. Hier sind Schema-basierte Ansätze umständlich und unflexibel. *Digital Objekt Repositories* unterstützen die dateiorientierte Speicherung und haben sich, ausgehend von der Bibliothekswelt, zunehmend auch im Bereich der Forschungsdaten etabliert. Bekannte Vertreter sind etwa EPrints[16], aDORe[17] und Fedora[18] (*Flexible Extensible Digital Object Repository Architecture*) zusammen mit darauf aufsetzenden Systemen wie etwa eSciDoc[19] oder EASY[20]. Die meisten dieser Systeme orientieren sich am OAIS-Referenzmodel *Consultative Committee for Space Data Systems* (CCSDS) (2002). Grundsätzlich können auch herkömmliche Dokumentenmanagement-Systeme wie z. B. Alfresco[21] oder *Content Repositories* wie JackRabbit[22] zum Einsatz kommen, die allerdings nur Teile der oben genannten Anforderungen erfüllen und damit nur eingeschränkt für das wissenschaftliche Datenmanagement eignen.

Vorteile des dateiorientierten Ansatzes sind Schemafreiheit, Unterstützung beliebiger Dateiformate für den *Content* und ausgezeichnete horizontale Skalierbarkeit. Ein *Repository* verhält sich – stark vereinfacht – wie ein Webserver: es ist sehr einfach, weitere Webserver (bzw. *Repositories*) hinzuzufügen und Daten auf die verschiedenen Instanzen aufzuteilen. Die Sicherung oder Replizierung der Daten kann auf Dateisystemebene erfolgen, was bis für mittlere *Repository*-Größen einfach zu bewerkstelligen ist. Wächst die Anzahl der Dateien zu stark an, kann sich das aber insbesondere hinsichtlich der Sicherung und vor allem der Wiederherstellung ins Gegenteil verkehren. Durch technische Maßnahmen (*Snapshot*-fähige Dateisysteme, Verteilung der Daten auf diverse Dateisysteme, Einsatz von *Cloud*- oder *Grid*-Technologie) lassen sich diese Probleme umgehen.

2.7.3.3 Grid und Cloud

Eine weitere Architekturoption für DMS stellen *Grid*-Systeme dar, also Systeme zum verteilten Rechnen und zur Speicherung von Daten über mehrere (räumlich verteilte) Knoten. Dabei stellt eine *Middleware* (z. B. *Globus Toolkit*, gLite oder UNICORE) Dienste zur Verfügung, die aus Sicht des Anwenders die

16. http://www.eprints.org/ [Zugriff am 13.08.2011].
17. http://african.lanl.gov/aDORe/projects/adoreArchive/ [Zugriff am 13.08.2011].
18. http://www.fedora-commons.org/ [Zugriff am 13.08.2011].
19. http://www.escidoc.org/ [Zugriff am 13.08.2011].
20. https://easy.dans.knaw.nl/dms [Zugriff am 13.08.2011].
21. http://www.alfresco.com/ [Zugriff am 13.08.2011].
22. http://jackrabbit.apache.org/ [Zugriff am 13.08.2011].

Komplexität der *Grid*-Infrastruktur hinter einer Dienste-Schicht verbirgt. *Grid*-Systeme können sowohl Rechen- wie auch Speicherressourcen zur Verfügung stellen. Je nach Schwerpunkt spricht man auch von *Computational Grids* und *Storage* oder *Data Grids*, wobei letzterer Begriff eher aus der Industrie kommt und momentan durch (*Private*) *Storage Clouds* verdrängt wird. In *Grid*-Systemen bieten Nutzer von Diensten bzw. Ressourcen meist auch eigene Ressourcen an. In Deutschland koordiniert *D-Grid* (Gentzsch, 2006) den Aufbau und den Betrieb einer solchen Infrastruktur. Im Bereich der Geisteswissenschaften zeigt das Projekt *TextGrid* (Gietz, et al., 2006), dass *Grid*-Technologie durchaus auch für die langfristige Speicherung umfangreicher Textkorpora (als Primärdaten z. B. für Linguisten) geeignet ist. Einen etwas generischeren Ansatz verfolgt *Diligent* (A *Digital Library Infrastructure on Grid Enabled Technology*)[23].

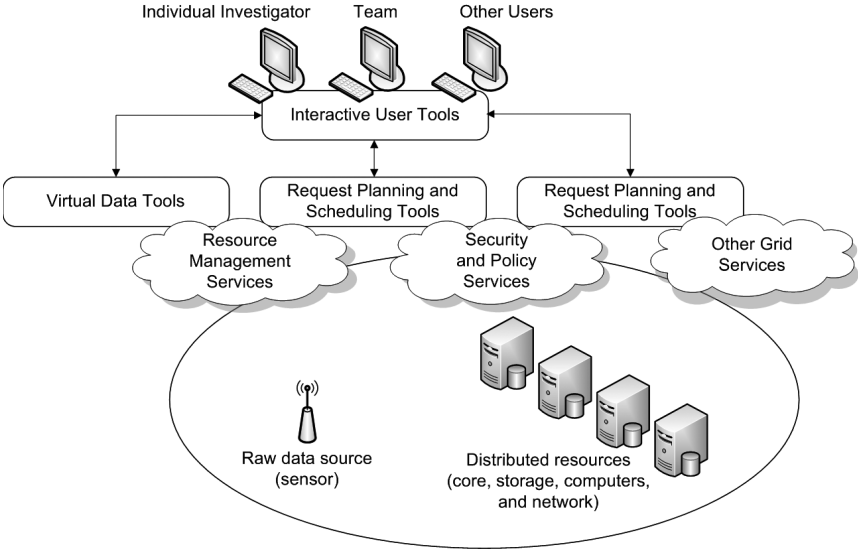

Abb. 3: Beispiel für den grundlegenden Aufbau eines *Data Grids* am Beispiel des *Grid Physics Network* (Quelle: nach Foster, 2003)

Cloud Computing basiert ebenfalls auf verteilten Diensten; hier bietet aber im Gegensatz zum *Grid* ein zentraler Anbieter Kunden seine Infrastruktur an. Neben kommerziellen Angeboten bauen Firmen und Universitäten inzwischen vielfach sogenannte „*Private Clouds*“ auf, bei denen die Ressourcen auf eigenen Systemen vor Ort vorgehalten werden. Hier lässt sich eines der größten Hemmnisse für den breiten Einsatz von *Grid*- und *Cloud*-Technologien umgehen,

23. http://diligent.ercim.eu/ [Zugriff am 13.08.2011].

nämlich das notwendige Vertrauen in die Sicherheit, Vertraulichkeit, Zuverlässigkeit und Langfristigkeit der angebotenen Leistungen.

Duraspace, die Organisation hinter Fedora und *DSpace*, arbeitet an einem *Cloud*-basierten *Storage Layer* für *Digital Object Repositories* namens *Duracloud*[24].

2.7.3.4 Mischformen

In der Praxis kommen meist Mischformen der vorgestellten Systeme zum Einsatz, also z. B. die Kombination von DOR und relationalen Datenbanken. Sei es, dass die Metadaten in einem RDBMS, die Daten aber im Dateisystem liegen, oder dass eine Mischung von strukturierten und unstrukturierten Daten vorliegt. Auch die Kombination von DOR und *Grid*-Technologie via SRB oder iRODS kommt vor, insbesondere im Umfeld von Fedora, für das entsprechende Plugins bereitstehen. Das oben erwähnte *TextGrid*-Projekt arbeitet an einer solchen Kombination. Hier empfehlen sich pragmatische Entscheidungen für eine geeignete Architektur für ein DMS, die sich insbesondere an den Anforderungen der Wissenschaftler, der Strukturiertheit der Daten und den Anforderungen an *Provenance* und Langzeitarchivierung orientieren sollte.

2.7.3.5 Ausblick

Die Heterogenität der Daten- und Dateiformate, der Größe und Menge der Daten, die Vielzahl der unterschiedlichen intendierten Anwendungsfällen eines DMS, die Verschiedenheit der disziplinspezifischen Anforderungen machen es unmöglich, eine übergreifende Architektur zu beschrieben. Neue Technologien wie z. B. NoSQL-Datenbanken erweitern die möglichen Komponenten einer DMS-Architektur. Insofern werden Mischformen zunehmend an Bedeutung gewinnen, in denen unterschiedliche technische Lösungen für die verschiedenen Datentypen und Anforderungen kombiniert und zu einem System zusammengefasst werden – hoffentlich transparent für Wissenschaftler, der seine Daten möglichst einfach verwalten, archivieren, publizieren und mit Mitarbeitern und Kollegen austauschen möchte.

---

24. http://www.duraspace.org/duracloud.php [Zugriff am 13.08.2011].

Literaturhinweise

Berners-Lee, T. Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American,* 284(5), S. 34–43.

Buneman, P. Cheney, J. Tan, W.-C. & Vansummeren, S, 2008. Curated databases. *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. Vancouver, Kanada 9.-12. Juni 2008. New York, NY: ACM, S.1–12.

Campbell, L. M. & MacNeill, S., 2010. *The Semantic Web, Linked and Open Data*. Bolton: JISC CETIS. Online: http://wiki.cetis.ac.uk/images/1/1a/The_Semantic_Web.pdf [Zugriff am 27.01.2011].

Caplan, P. & Guenther, R. S., 2005. Practical Preservation: The PREMIS Experience. *Library Trends*, 54 (1), S. 111–124.

CCSDS (Consultative Committee for Space Data Systems), 2002. *Reference Model for an Open Archival Information System (OAIS)*. Washington, DC: National Aeronautics and Space Administration.

DFG (Deutsche Forschungsgemeinschaft), 1998. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft" (Denkschrift).* Weinheim: Wiley-VCH.

Foster, I., 2003. The Grid: A New Infrastructure for 21st Century Science. In: F. Berman, G. Fox & T. Hey, ed. 2003. *Grid Computing: Making the Global Infrastructure a Reality.* Chichester: John Wiley & Sons. doi: 10.1002/0470867167.ch2.

Gentzsch, W., 2006. D-Grid, an E-Science Framework for German Scientists. *ISPDC '06: Proceedings of the Proceedings of The Fifth International Symposium on Parallel and Distributed Computing*. Timisoara, Rumänien 6.-9. Juli 2006. Los Alamitos, Calif.: IEEE Computer Society, S. 12–13.

Giaretta, D. et al., 2010. *Riding the wave – How Europe cain gain from the rising tide of scientific data. Final report of the high level expert group on scientific data*. Online: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf [Zugriff am 18.12.2010].

Gietz, P. et al., 2006. TextGrid and eHumanities. *Second IEEE International Conference on e-Science and Grid Computing.* Amsterdam, Niederlande 4.-6. Dez. 2006. Los Alamitos, Calif.: IEEE Computer Society, S. 133–141.

ICSU (International Council for Science), 2008. *Ad hoc Strategic Committee on Information and Data – Final Report to the ICSU Committee on Scientific*

Planning and Review. Online: http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf [Zugriff am 12.12.2010].

Kroll, S. & Forsman, R., 2010. A Slice of Research Life: Information Support for Research in the United States. *Report commissioned by OCLC Relesearch in support with the RLG partnership.* Online: http://www.oclc.org/research/publications/library/2010/2010–15.pdf [Zugriff am 10.02.2010].

Moreau, L. et al., 2007. *The Open Provenance Model*. Online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.7394 [Zugriff am 23.11.2010].

Rajbhandari, S. Hedges, M. & Fabiane, S., 2010. BRIL – Capturing Experiments in the Wild. *5th International Conference on Open Repositories.* Madrid, Spanien 6.-9. Juli 2010. Online: http://or2010.fecyt.es/Resources/documentos/GSabstracts/BRIL.pdf [Zugriff am 15. 01.2011].

Razum, M. et al., 2010. Research Data Management in the Lab in the Lab. *5th International Conference on Open Repositories.* Madrid, Spanien 6.-9. Juli 2010. Online: http://or2010.fecyt.es/Resources/documentos/GSabstracts/ResearchDataManagamentIinTheLab.pdf [Zugriff am 15.01.2011].

Razum, M. Schwichtenberg, F. Wagner, S. & Hoppe, M., 2009. eSciDoc Infrastructure: A Fedora-Based e-Research Framework. In M. Agosti et al., ed., Research and Advanced Technology for Digital Libraries, *13th European Conference, ECDL 2009.* (LNCS 5714). Korfu, Griechenland, 27. Sept.-2. Okt. 2009. Berlin: Springer, S. 227–238.

Recordon, D. & Reed, D., 2006. OpenID 2.0: a platform for user-centric identity management. *Proceedings of the second ACM workshop on Digital identity management*. Alexandria, VA, USA 30. Okt.-3. Nov. 2006. New York, NY: ACM, S. 11–16.

Scavo, T. & Cantor, S., 2005. Shibboleth Architecture – Technical Overview. *Working Draft: draft-mace-shibboleth-tech-overview-02*.

Stern, R. & McEwen, S., 2009. FITS – The File Information Tool Set. Poster. *4th International Conference on Open Repositories.* Atlanta, USA 18.-21. Mai 2009. Online*:* http://hdl.handle.net/1853/28508 [Zugriff am 27.01.2011].

TIB (Technische Informationsbibliothek) Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010. *Konzeptstudie „Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie“*. Online:

http://www.fiz-chemie.de/fileadmin/user_upload/PDF_DE/abstract_Konzeptstudie_Forschungsdaten_Chemie.pdf [Zugriff am 13.08.2011].

Treloar, A. Groenewegen, D. & Harboe-Ree, C., 2007. The Data Curation Continuum. *D-Lib Magazine*, 13(9/10). http://dx.doi.org/ doi:10.1045/september2007-treloar.

2.8 Datenanalyse und -visualisierung

Bettina Berendt [1], Joaquin Vanschoren [2], Bo Gao [1]
[1] K.U. Leuven, Belgien
[2] Universiteit Leiden, Niederlande

Das Management von Forschungsdaten wird dann interessant, wenn die Datenhaltung nicht nur eine Dokumentationsfunktion erfüllt, sondern es erlaubt, Daten in neuen Weisen wieder- und weiterzunutzen. Repositorien für das Forschungsdatenmanagement (FDM) stellen daher idealerweise nicht nur Mechanismen für das Speichern und Finden von Daten zur Verfügung, sondern auch die Grundlagen oder auch die Tools für Analysen auf diesen Daten, die über die schon im „Ursprungsprojekt" der Daten geleisteten hinausgehen.

Visualisierungen erscheinen im FDM in zweierlei Form: Zum einen können Visualisierungen *Objekte*, *Materialien* oder *Ergebnisse* der dokumentierten Forschung sein – beispielsweise kann eine medien-, kunst- oder sozialwissenschaftliche Datenbank Fotografien oder Filme als Studienobjekt enthalten, als in einer empirischen Untersuchung eingesetzter Stimulus, oder als Dokumentation einer Interviewreihe. Zum anderen können Visualisierungen *Methode* der dokumentierten Forschung sein – oder auch Methode der Forschung, die erst durch FDM möglich wird.

Eine typische Methode der dokumentierten Forschung sind etwa Datengraphiken, die vorrangig der *Präsentation* von Messdaten oder Informationen dienen. Visualisierungen können aber darüber hinaus auch der *Exploration* dienen. Erst durch FDM möglich wird die Exploration der *Gesamtheit* der verwalteten Daten. Visuelle wie nicht-visuelle explorative Datenanalysen spielen eine wesentliche Rolle in der Generierung neuer Forschungsideen und Hypothesen (z. B. Tukey, 1977; Hand, Smyth & Mannila, 2001) und sind damit eine der Kernmotivationen, FDM zu betreiben.[1] Wir fokussieren daher auf Datenvisualisierung als eine Form der Datenanalyse und sprechen allgemein von Analyse/Visualisierung.

Die Frage, was Analyse/Visualisierung für das Forschungsdatenmanagement ist, lässt sich nun in zweierlei Weise konkretisieren: einerseits, was *alles* Analyse/Visualisierung für das FDM sein kann; andererseits, was Analyse/Visualisierung *speziell* für das FDM ist. Daher muss weiter gefragt werden, (a) was Forschungsdaten sind, (b) welche Daten *speziell* in Forschungsdaten-Datenbanken auftauchen, (c) welche Anforderungen das FDM-Ziel „Analyse" an Daten und Visualisierungsformen stellt.

1. Natürlich können auch Datenanalysen Objekt, Material oder Methode der dokumentierten Forschung sein.

2.8.1 Welche Anforderungen stellt die FDM-Aufgabe der (u.a. visuellen) Datenanalyse an Daten und an Visualisierungsformen?

Zur Beantwortung der Frage nach den Anforderungen führen wir zunächst den Begriff des FDM auf den allgemeineren des Datenmanagement zurück: „das Datenmanagement ist die Menge aller methodischen, konzeptionellen, organisatorischen und technischen Maßnahmen und Verfahren zur Behandlung von ‚Daten' [mit dem Ziel] der Gewährleistung einer maximalen Unterstützung auszuführender Geschäftsprozesse …" (Schenk, 2010, S. 316). Diese allgemeine Definition muss instanziiert werden: Wir betrachten als die zu unterstützenden „Geschäftsprozesse" verschiedenste Prozesse der *eScience* und betrachten in diesem Artikel speziell Fragen zu methodischen, organisatorischen und technischen Aspekten einer hierfür geeigneten Infrastruktur und dabei besonders Methoden zur Auswertung der Daten mit dem Ziel der Wissensgewinnung sowie Aspekte der technischen (Implementation) und organisatorischen (Web-Basierung) Realisierung dieser Methoden. Schließlich sind die ‚Daten' des *Forschungs*datenmanagements natürlich *Forschungs*daten. Forschungsdaten umfassen u.a. (i) Dokumente, z. B. Textdokumente oder Tabellenkalkulationsdokumente, (ii) Laborbücher, Feldaufzeichnungen, weitere Logbücher, (iii) Fragebögen, Transkripte, Codebücher, (iv) Audio- und Videoaufzeichnungen, (v) Fotografien und Filme, (vi) Testantworten, (vii) Dias, Artefakte, Muster, Proben, (viii) Sammlungen digitaler Objekte, die während des Forschungsprozesses angelegt wurden, (ix) Daten-Dateien, (x) Datenbankinhalte verschiedener Formate, (xi) Bestandteile einer Anwendung (Input, Output, Logfiles für die Analyse-Software, Simulationssoftware, Schemata), (xii) Methodologien und Workflows, und (xiii) Standardvorgehensweisen und Protokolle. Des Weiteren kann es sinnvoll sein, Daten wie z. B. Korrespondenz oder Projektfinanzierungsanträge über den konkreten Forschungszusammenhang, in dem sie entstanden sind, aufzubewahren (University of Edinburgh Information Services, 2009).

FDM kann dann einen Mehrwert erzeugen, wenn solche Daten nicht nur existieren, sondern auch zugreifbar, auffindbar/recherchierbar und (meta-)analysierbar sind. Hieraus ergeben sich Anforderungen für Infrastruktur und Methoden inkl. möglicher Visualisierungen, die jedoch für die meisten der oben genannten Datenformen nicht spezifisch für das FDM sind. So stellen beispielsweise Dokumente Anforderungen an eine FDM-Datenbank, die kaum von denen normaler Websuchmaschinen abweichen: Datenzugriff kann beispielsweise durch die Zuweisung persistenter URLs an die Dokumente und Verwendung von HTTP gewährleistet werden, *Information-Retrieval*-Methoden ermöglichen Auffindbarkeit und Recherchierbarkeit, und Standardverfahren z. B. des *Text Mining* und *Visual Text Mining* (vgl. Feldman & Sanger, 2007) können wahlweise auf einem Server oder lokal von einem Client-Tool angeboten werden, um die Dokumente zu analysieren. (Handelt es sich nicht oder nicht vor-

wiegend um Textdokumente, so können weitere Verfahren z. B. zum *Visual Information Retrieval* und *Multimedia-Mining* (Enser, 2008) zur Anwendung kommen.) Somit kann diese Beobachtung mindestens auf FDM-Daten der Formen (i), (ii), (iii), (iv), (v), (vi), (viii) und (xiii) verallgemeinert werden.

Auch weitere Formen von Forschungsdaten können von Standardtechnologie profitieren: Sollen Inhalte von (je nach am FDM beteiligten Partner wahrscheinlich unterschiedlicher) Datenbanken zugänglich gemacht werden, so eignen sich Techniken für Föderierte Datenbanken (z. B. Conrad, 1997) bzw. für das (*Semantic*) *Web*. Auch Dokumente können Inhalte in diesem Sinn sein. Daten als solche (ob in Datei- oder Datenbankform) können von verschiedenen Formen der Informationsvisualisierung profitieren (z. B. Schumann & Müller, 1999). Für einige Formen von FDM-Daten wie z. B. Workflows gibt es Standards für interoperable Notation und auch für die Visualisierung einzelner Instanzen sowie Meta-Analysen.

Was also ist *anders* im FDM? Wir wollen hier die Grundidee der *eSciences* verfolgen: der Wissenschaften mit hohen Anforderungen an Rechenleistung, die das Internet als einen globalen, kollaborativen Arbeitsplatz nutzen, s. z. B. (Nielsen, 2008). Hierbei nehmen die Verwaltung und (Nach-)Nutzung von Daten (z. B. Messdaten) sowie Bestandteilen einer Anwendung (Input, Output, Prozessdaten) eine wichtige Rolle ein. Somit können z. B. Beobachtungs- oder Experimentaldaten mit hoher Detailtiefe notiert, ausgetauscht und kollaborativ vervollständigt werden, da viel einfacher und/oder systematischer als beispielsweise bei einer Literaturstudie sichtbar wird, welche Parameterkombinationen eines bestimmten Experimentallaufs untersucht worden sind und mit welchem Ergebnis, und welche noch nicht untersucht worden sind.[2]

2.8.2 Experimentaldatenbanken als ein typisches Element von *eScience* und FDM

Es ist daher nicht erstaunlich, dass in allen e*Sciences Online*-Infrastrukturen gebaut werden, die dem Austausch von *Experimenten* dienen. Solche Infrastruk-

2. Selbstverständlich gilt für jegliche Offenlegung wissenschaftlicher Daten – genau wie für Daten anderer Datenproduzenten oder -halter auch – dass diese u.U. Privacy relevant sein können. Es muss daher in jedem Einzelfall sorgfältig geprüft werden, ob datenschutzrechtliche Fragen hinreichend berücksichtigt sind. Hierbei ist auch zu beachten, dass „anonymisierte" Daten derzeit zwar nicht oder nur vermindert unter das Datenschutzrecht fallen, dass solche Anonymität aber in vielen Fällen durch – häufig einfache – *Analyse*operationen hinfällig wird und somit de facto schwerwiegende Verletzungen der Privatsphäre möglich werden (z. B. Sweeney, 2002; Barbaro & Zeller, 2006). Das Forschungsgebiet des *Privacy-preserving Data Publishing* analysiert derartige Probleme und Lösungsansätze; für einen Überblick, s. Fung et al., 2010). Zu juristischen Fragen des FDM s.a. Spindler und Hillegeist in diesem Band.

turen beruhen auf drei wesentlichen Komponenten: (a) einer formale Beschreibungssprache, (b) Ontologien zur Gewährleistung gemeinsamer Semantik und (c) einem durchsuchbaren Repositorium. Visualisierungen der Ontologien können das Verständnis der Ontologien befördern (s. Neher & Ritschel, in diesem Band, zu Ontologien und Katifori et al., 2007, zu ihrer Visualisierung), und in Anfragen an das Repositorium können visuelle Anfragesprachen genutzt werden. Auch hier unterscheidet sich das FDM nicht von anderen Anwendungsgebieten der Ontologie-Visualisierung oder der visuellen Anfragesprachen.

Entscheidenden Mehrwert können Repositorien und Visualisierungen im FDM dann schaffen, wenn sie eine Zusammenschau und Verdichtung von Daten aus unterschiedlichen Quellen bieten, aus der neue Beobachtungen dieser Daten und Ideen oder Hypothesen für Anschlussuntersuchungen gewonnen werden können. Die Zusammenschau wird durch interoperable Syntax und Semantik, realisiert in den Ontologien/Schemata des Repositoriums, möglich gemacht. Es können dann beispielsweise statistische (Meta-)Analysen auf der Kombination von Datensätzen aus unterschiedlichen Messreihen unterschiedlicher Forschungsprojekte durchgeführt werden; oder das Data-Mining-Tool eines FDM-Partners kann die Daten eines anderen FDM-Partners direkt laden und auswerten. Die Generierung einer Datenvisualisierung unterscheidet sich hierin vom Gesichtspunkt der Software her nicht von anderen Arten der Analyse wie beispielsweise einer statistischen Auswertung. Gerade für explorative Analysen, die immer nur teilautomatisiert sein können und daher auf menschliche Wahrnehmung und Verarbeitung angewiesen sind, können Visualisierungen jedoch oft besser verdichten als numerische oder textuelle Darstellungsformen.

Ein wesentliches Charakteristikum typischer Forschungsdaten sind ihre großen Mengen und eine oft hohe Dimensionalität der einzelnen Datensätze. Damit kommt Techniken des *Data Mining* (durch seinen Fokus auf Skalierbarkeit von Analysen) und Techniken der *Dimensionalitätsreduktion*[3] (durch ihren Fokus auf Darstellbarkeit in 2D/3D/4D-Medien[4]) eine entscheidende Rolle zu. Anders ausgedrückt: Methoden zur Visualisierung hochdimensionaler Daten haben verschiedene Anwendungsgebiete, sind aber im FDM zum Zwecke der Informationsstrukturierung, -filterung und -kondensierung noch unverzichtbarer als

3. Beispiele sind *PCA*, *Classical Scaling*, *Isomaps*, *LLE*, *Laplacian Eigenmaps*, *Diffusion Maps*, *Kernel PCA* und *Sammon Maps* (s. Gao, 2010).

4. 3D ist heute noch überwiegend 2.5D auf 2D-Ausgabegeräten wie z. B. Standardbildschirmen; eine weitere Dimension und damit bis zu 4D lassen sich durch Animation visualisieren. Aktuelle Entwicklungen wie z. B. stereoskopische Verfahren („3D-Kino“) können die Dimensionalität der Medien erhöhen. Unabhängig von der Dimensionalität des *Mediums* können Visualisierungs*formate* auch höherdimensional sein. So können z. B. Parallelkoordinaten-Darstellungen beliebig viele Dimensionen auf einem 2D-Medium darstellen.

anderswo, wobei regelmäßig Daten und Metadaten über Inputs, Outputs und prozedurale Elemente kombiniert werden müssen.

Wir wollen diese allgemeinen Beobachtungen an einem konkreten Beispiel zeigen und hierbei auch illustrieren, welche Design-Entscheidungen die Wahl einer geeigneten Visualisierung bestimmen können.

Die Datenbank enthält Experimente zum maschinellen Lernen, einem Teilgebiet der Informatik, in der die Entwicklung von Experimentaldatenbanken noch nicht so weit fortgeschritten ist wie in anderen naturwissenschaftlichen Disziplinen etwa der Bioinformatik, der Astronomie oder der Physik (Vanschoren et al., im Druck). Das Beispiel ist jedoch hinsichtlich der Frage des aktuellen Artikels besonders interessant, da hier Datenanalysetechniken nicht nur Gegenstand des FDM, sondern auch Gegenstand der dokumentierten Forschung sind. Aus Platzgründen werden wir uns auf zwei der in diesem Abschnitt als zentral identifizierten Elemente konzentrieren: Beschreibung der Daten in Form einer Ontologie und Visualisierung auf Basis von Dimensionalitätsreduktion.

2.8.3 Fallbeispiel: Eine Experimentaldatenbank für maschinelles Lernen / Data Mining

Das maschinelle Lernen beschäftigt sich mit der Induktion von beschreibenden und vorhersagenden Modellen von Daten. Diese Induktion ist ein wichtiger Bestandteil des Prozesses des Data Mining, der Wissensentdeckung in Daten. Eine typische Aufgabe im maschinellen Lernen ist das Klassifikationslernen. Hier ist eine große Zahl von Dateninstanzen gegeben, die mit Hilfe von *Features* (Merkmalen) beschrieben werden und von denen man für einige weiß, zu welcher *Klasse* sie gehören – dieses aber auch für die anderen vorhersagen will. Beispielsweise kann ein Mailprovider oder eine Suchmaschine das Ziel haben, Mails oder Webseiten als Spam (oder Nicht-Spam) zu klassifizieren. Dieses Klassenlabel kann für bestimmte *Datenmengen* gegeben sein (beispielsweise weil Nutzer frühere Mails als Spam gekennzeichnet haben); interessant ist natürlich die Vorhersage für neue, noch unbekannte und unklassifizierte Dokumente. Als Features eignen sich hier u.a. Textmerkmale (z. B. die verwendeten Wörter), Metadaten (z. B. die URL) und weitere automatisch extrahierbare Eigenschaften (z. B. die Verlinkung zu anderen Inhalten). Ein Klassifikationslern-*Verfahren* (z. B. Entscheidungsbaumlernen) kann verschiedene *Parameter* haben (z. B. wie „korrekt“ ein Blatt des Entscheidungsbaums klassifizieren muss). Ein parametrisiertes Verfahren hat auf einer gegebenen Datenmenge eine bestimmte *Performanz* (z. B. die Genauigkeit: wie viele Datensätze korrekt klassifiziert wurden). Auch die Datenmengen selbst haben Eigenschaften, beispielsweise wie die Klassen in ihnen verteilt sind. (Dieser Parameter ist wichtig, weil z. B. auf einer Datenmenge, die zu 99% aus Nicht-Spam besteht, sogar ein primitives Verfahren, das alle Mails als Nicht-Spam klassifiziert, bis zu 99% korrekt sein kann.) Die voll-

ständige Beschreibung eines Experiments kann weitere Features beinhalten (z. B. zur Beschreibung des Testverfahrens, mit dessen Hilfe die Genauigkeit bestimmt wurde). Die derzeit größte Datenbank zur Beschreibung solcher Experimente im maschinellen Lernen inklusive der Ontologie, die ihrem Schema zugrunde liegt, ist unter http://expdb.cs.kuleuven.be zu finden; sie enthält Datensätze über mehr als 650,000 Experimente (Vanschoren et al., im Druck).

Die Ontologie wird in Form der *Experiment Markup Language* ExpML notiert und mit Instanzen gefüllt und in der Datenbank ExpDB gespeichert. Sie enthält die Konzepte und Strukturen des Data-Mining-Experimentalprozesses sowie die dabei verwendeten Algorithmen und Datenmengen und ihre Details. Wie andere Ontologien dient sie einerseits der Lesbarkeit sowohl durch Menschen als auch durch Software-Agenten. Im Geiste des *Ontology-Engineerings* wurden soweit wie möglich Konzepte und Relationen aus verwandten und erprobten Ontologien wiederverwendet, insbesondere die Top-Level-Klassen und -Relationen einschlägiger Ontologien aus der Bioinformatik und dem Data Mining (u.a. BFO, OBI, IAO sowie OntoDM und DMOP; für Details s. Vanschoren et al., im Druck). Abbildung 1 zeigt eine (manuell erstellte) Visualisierung der Ontologie.

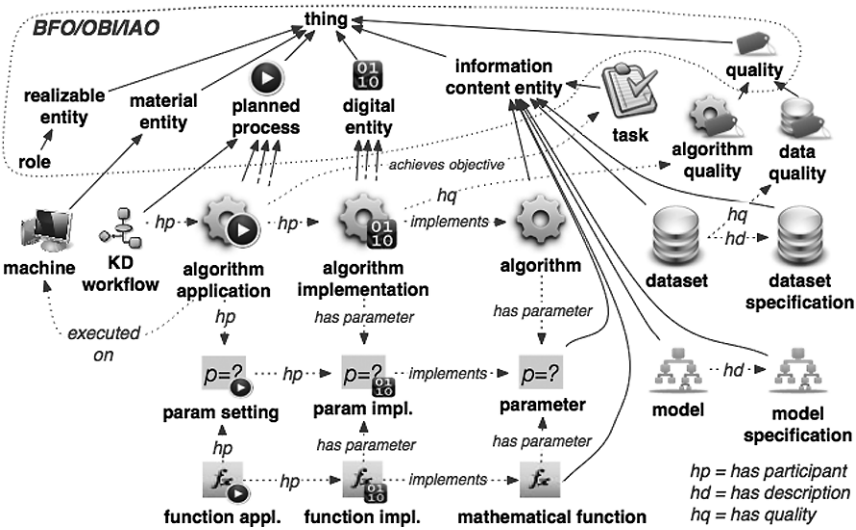

Abb. 1: Top-Level-Konzepte der Data-Mining-Ontologie.

Eine typische Datenbankabfrage zu Experimenten im Klassifikationslernen produziert große und daher leicht unübersichtliche Tabellen. Ein Plot etwa der Korrektheit nach Basis-Lernverfahren ist leicht (etwa mit Excel) erstellt, gibt aber nur unzureichenden Einblick in die reichhaltige Information dieser Datenbank.

Daher wurde ein Plugin entworfen und implementiert, das bestehende und neue Formen zur Visualisierung hochdimensionaler Daten einsetzt (Gao, 2010).

Ein typisches Anwendungsbeispiel ist die Datenmenge „Bagging“, die 1447 Experimentalläufe verschiedener Verfahren des Bagging auf Standard-Datensets enthält. (Bagging ist eine Kombination der Anwendung eines Klassifikationslerners auf unterschiedliche Stichproben der Basisdaten, ein Ansatz, der zu wesentlich besserer Vorhersagequalität führt.) Jeder Lauf ist durch 15 Features spezifiziert, die teils die Basisdaten beschreiben, teils die Algorithmen und teils die Ergebnisse. Zwei der Features sind nominalskaliert, der Rest intervallskaliert. Mithin müssen 1447 Punkte in einem 15-dimensionalen Raum so dargestellt werden, dass auch die Effekte der verschiedenen Dimensionen inspizierbar werden.

Die erste Teilaufgabe, die sich hier stellt, ist die Suche nach einer Verteilung der 1447 Punkte über die Ebene (in der vorliegenden Arbeit kamen nur 2D-Visualisierungen zur Anwendung; die Frage würde sich aber auch bei einer 3D-Visualisierung fast unverändert stellen), die die Distanzen im ursprünglichen 15-dimensionalen Raums möglichst wenig verzerrt, also einander nahe Punkte (ähnliche Experimentalläufe) möglichst beieinander lässt und einander ferne voneinander trennt. Zu diesem Zweck wurden die von acht Verfahren der Dimensionalitätsreduktion mit jeweils verschiedenen Parameterwerten produzierten 2D-Projektionen durch die Maße *trustworthiness* und *continuity* (Venna, 2007) bewertet und das Visualisierungsverfahren gewählt, das auf beiden Maßen die besten Ergebnisse lieferte. Es handelt sich um *Diffusion Maps* (Coifman et al., 2005) mit den Parametern $t=\sigma=1$. Die zweite Teilaufgabe ist, wie man anhand der entstandenen Basistopologie interessante Abhängigkeiten zwischen Features visualisieren kann. Zu diesem Zweck wurde jedes Feature nach [0;1] normalisiert, und für jedes Feature wurde eine eingefärbte Variante des Punkteplots erstellt, deren Farbskala von dunkelblau (0) bis rot (1) läuft. Auf kleinem Raum zeigen also 15 3D-Bilder (2 räumliche Dimensionen plus Farbe) Information, aus der Metaaussagen über alle Experimentalläufe und ihre Variablen abgeleitet werden können.

Abbildung 2 zeigt das Ergebnis. Die vom Visualisierungsverfahren gebildeten vertikalen „Stäbchen“ von Punkten entsprechen im Wesentlichen den Basis-Datenmengen, wie man an den homogenen Farben der „Stäbchen“ z. B. für Feature 3 (dataset), 12 (number of missing values – diese sind natürlich jeweils gleich für verschiedene Experimentalläufe auf denselben Daten), 14 (default accuracy – Anteil der Instanzen, die zur Mehrheitsklasse gehören) oder auch 4, 5, 6, 7 sieht. Innerhalb dieser visuellen Cluster hat sich als Y-Achse eine Dimension ergeben, die v.a. der Zahl der Iterationen entspricht (1), und als X-Achse eine Dimension, die gut mit der Performanz der Landmarker-Lernverfahren (8–10) übereinstimmt, was sich auch in der Performanz des Gesamtverfahrens (15) widerspiegelt, ohne aus den dem Bagging zugrunde liegenden Lernverfahren (2) vorhersagbar zu sein.

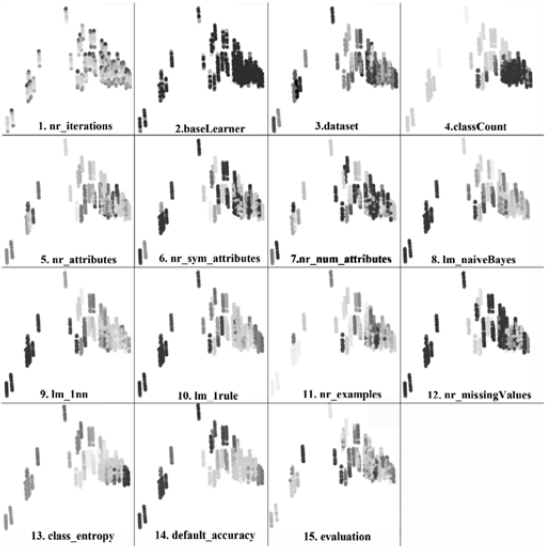

Abb. 2: Analyse der 1447 Experimentalläufe zum Bagging hinsichtlich 15 Parametern (Eine farbige Variante dieser Abbildung ist auf http://www.berendt.de/FDM.png zu finden.)

2.8.4 Zusammenfassung

Zukunftsorientiertes Forschungsdatenmanagement geht über die Dokumentation von Forschungsergebnissen und -prozessen hinaus und ermöglicht neue Formen der Wieder- und Weiternutzung der gespeicherten Daten. Dabei spielen Re- und Meta-Analysen dieser Daten eine besondere Rolle, und Visualisierungen als Form der explorativen Datenanalyse können wertvolle Erkenntnisse liefern und zu neuen Forschungsfragen führen. In diesem Artikel haben wir die Interoperabilität der Daten (durch gemeinsame Semantik und Syntax, realisiert durch Ontologien und Markup-Sprachen) als wichtige Voraussetzung für Analysen und den Einsatz geeigneter Verfahren zur Dimensionalitätsreduktion als Kernbestandteil geeigneter Visualisierungen identifiziert und mit einem Beispiel, einer Experimentaldatenbank zum Data Mining / maschinellen Lernen, illustriert. Die Entwicklung und Bereitstellung solcher Tools werden zu den wichtigsten Schritten der nächsten Zukunft gehören, um das Forschungsdatenmanagement offen und damit wissenschaftlich generativ zu gestalten.

Literaturhinweise

Barbaro, M. & Zeller, T., 2006. A face is exposed for AOL searcher no. 4417749. *New York Times*, 9. Aug.

Coifman, R.R. et al., 2005. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21), S. 7426–7431.

Conrad, S., 1997. *Föderierte Datenbanksysteme: Konzepte der Datenintegration.* Berlin: Springer.

Enser, P., 2008. The evolution of visual information retrieval. *Journal of Information Science*, 34, S. 531–546.

Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data.* Cambridge, UK: Cambridge University Press.

Fung, B.C.M. Wang, K. Chen, R. & Yu, P.S., 2010. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*, 42(4). DOI=10.1145/1749603.1749605, http://doi.acm.org/10.1145/1749603.1749605.

Gao, B., 2010. *Advanced Visualizations of Machine Learning Behaviour.* Masters Thesis. K.U. Leuven, Department of Computer Science. UDC: 681.3*I20.

Hand, D.J. Smyth, P. & Mannila, H., 2001. *Principles of Data Mining.* Cambridge, MA: MIT Press.

Katifori, A. et al., 2007. Ontology visualization methods – a survey. *ACM Computing Surveys*, 39(4). DOI=10.1145/1287620.1287621, http://doi.acm.org/10.1145/1287620.1287621.

Nielsen, M., 2008. The future of science: Building a better collective memory. *APS Physics*, 17(10), 8. Online: http://www.aps.org/publications/apsnews/200811/upload/November-2008-Volume-17-No-10.pdf [Zugriff am 13.08.2011].

Schenk, M., Hrsg., 2010. *Instandhaltung technischer Systeme. Methoden und Werkzeuge zur Gewährleistung eines sicheren und wirtschaftlichen Anlagenbetriebs.* Berlin: Springer.

Schumann, H. & Müller, M., 1999. *Visualisierung: Grundlagen und allgemeine Methoden.* Berlin: Springer.

Sweeney, L., 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), S. 557–570.

Tukey, J.W., 1977. *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

University of Edinburgh Information Services, 2009. *Defining Research Data.* Online:http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/data-mgmt/research-data-definition [Zugriff am 02.02.2011].

Vanschoren, J. Blockeel, H. Pfahringer, B. & Holmes, G. (im Druck). Experiment databases. A new way to share, organize and learn from experiments. *Machine Learning.* Online: https://lirias.kuleuven.be/handle/123456789/297378 [Zugriff 13.08.2011].

Venna, J., 2007. *Dimensionality Reduction for Visual Exploration of Similarity Structures.* PhD Thesis. Helsinki University of Technology, Department of Computer Science and Engineering. Online: http://lib.tkk.fi/Diss/2007/isbn9789512287529/ [Zugriff am 13.08.2011].

3.1 Institutionalisierte „*Data Curation Services*“

Michael Lautenschlager
ICSU World Data Center Climate
Deutsches Klimarechenzentrum GmbH

Der Lebenszyklus wissenschaftlicher Daten (*Data Life Cycle*) durchläuft die Stadien Erzeugung, Bearbeitung, Archivierung und Wiederverwendung. Die Wiederverwendung von Daten führt zur Erzeugung neuer Daten und leitet den nächsten Zyklus ein.

Die Archivierung von Daten ist hier nicht zu verstehen als Zwischenspeicherung von Ergebnissen im Rahmen eines wissenschaftlichen Bearbeitungsprozesses, sondern als Langzeitarchivierung mit dem Ziel, die archivierten Daten auch nach vielen Jahren für wissenschaftliche, interdisziplinäre Nachnutzung bereitzustellen (Wiederverwendung). Im Rahmen der Regeln zur Sicherung guter wissenschaftlicher Praxis der Forschungsgesellschaften werden hier Zeiträume von 10 Jahren und mehr gefordert. Zentrale Forderung in der Langzeitarchivierung ist die Sicherstellung der Datenintegrität. Im Zeitalter der elektronischen Speicherung von Daten ist das ein aktiver Prozess, der kontinuierliche Pflege der archivierten Daten erfordert.

Wesentliche Bestandteile der Integritätssicherung elektronischer Daten sind:

- Sicherstellung der Unversehrtheit (*Bit-stream Preservation*)
- Sicherstellung der Lesbarkeit
- Sicherstellung der Interpretierbarkeit

Für eine Wiederverwendung wissenschaftlicher Daten müssen diese identisch zum Original sein, aus dem *Bit-stream* müssen Informationen (z. B. Ziffern und Buchstaben) auslesbar sein und diese Ziffern und Buchstaben müssen interpretiert werden können, um die ursprünglichen Informationen wieder zu gewinnen. „*Data Curation Services*“ unterstützen die Integritätssicherung elektronischer Daten. Sicherstellung der Les- und Interpretierbarkeit elektronischer Daten ist eine Voraussetzung für eine zukünftige Überprüfung wissenschaftlicher Ergebnisse. Damit unterstützen „*Data Curation Services*“ direkt die Regeln zur guten wissenschaftlichen Praxis, wie sie von den Wissenschaftsgesellschaften formuliert wurden.

3.1.1 Sicherstellung der Unversehrtheit

Sicherstellung der Unversehrtheit oder auch „*Bit-stream Preservation*“ garantiert die Bit-genaue Erhaltung der archivierten Daten ohne Ansehen ihres Inhalts. Die Umsetzung dieser Anforderung wird durch technische und organi-

satorische Maßnahmen im Arbeitsablauf der Archivierung und der Archivbetreuung erreicht.

Einem möglichen Datenverlust wird vorgebeugt durch Sicherheitskopien der archivierten Datenentitäten, die mit unterschiedlichen Technologien erstellt und an unterschiedlichen Orten gespeichert werden. Die Anzahl der Sicherheitskopien hängt ab, vom Grad der Sicherheit, der erreicht werden soll, dem Aufwand, den man treiben kann, und zum Teil auch von gesetzlichen Vorschriften die erfüllt werden müssen. Wird an dieser Stelle noch Dokumentensicherheit gefordert, muss zusätzlich noch sichergestellt werden, dass die Datenentitäten nachträglich nicht mehr verändert werden können.

Einem möglichen Datenverlust bei Speicherung auf Festplatten wird häufig durch Speicherung in der *Raid* (*Redundant Array of Independent Disks*) Architektur vorgebeugt. Dabei wird in der Praxis unterschieden zwischen *Raid* 0, keine Sicherung, *Raid* 1, komplette Spiegelung der Daten, und *Raid* 5 bzw. *Raid* 6, bei denen redundante Informationen zur Rekonstruktion der Daten bei Verlust auf einer oder zwei zusätzlichen Festplatten abgespeichert werden.

Die Unversehrtheit von Datenentitäten kann geprüft werden durch Prüfsummen bzw. *Check Sums*. Dabei wird jeder Datenentität während ihrer Erzeugung eine Bit-genaue Prüfsumme mitgegeben, deren Unveränderlichkeit die Unversehrtheit einer Datenentität nach Kopiervorgängen oder Transport über das Netzwerk garantiert. Ist die Prüfsumme vor und nach der Operation identisch, ist auch die zugehörige Datenentität identisch. Ein einfaches Beispiel für eine Prüfsumme ist die Quersumme der Ziffern einer Zahl. Allerdings werden mit diesem Verfahren beispielsweise „Zahlendreher“, also ein häufig vorkommender Fehler in der Eingabe von numerischen Informationen durch Menschen, nicht erkannt.

Eine weitere Maßnahme zur Sicherung der Unversehrtheit von Datenentitäten ist die regelmäßige Erneuerung von elektronischen Speichermedien. Alle Speichermedien unterliegen einem Alterungsprozess, der gemessen wird in Standzeit oder Nutzungsintensität. Wird das „Haltbarkeitsdatum“ überschritten, werden in Langzeitarchiven die betroffenen Datenentitäten automatisch und weitgehend transparent für den Nutzer auf ein frisches Medium kopiert. Bei diesem Prozess findet auch häufig ein Generationswechsel im Speichermedium mit z. B. höherer Kapazität statt. Die Unversehrtheit der kopierten Datenentitäten kann wieder mit Hilfe ihrer Prüfsummen sichergestellt werden. Im Fehlerfall wird der automatische Prozess unterbrochen und manuelles Eingreifen durch einen Betreuer des Datenarchivs (Daten-Kurator) wird erforderlich.

3.1.2 Sicherstellung der Lesbarkeit

Abhängig vom archivierten Datenvolumen und der wissenschaftlichen *Community* werden zwei Strategien zur Sicherstellung der Lesbarkeit von Datenentitäten verfolgt: Formatkonvertierung oder Migration der Lese-Werkzeuge.

Kleinere Datenvolumina werden häufig mit Werkzeugen erzeugt, deren Quelle für die Wissenschaft nicht zugreifbar ist und damit auch deren Ausgabeformate nicht im Detail bekannt oder beeinflussbar sind. Neue Generationen von Werkzeugen bieten häufig neue Ausgabeformate mit Lesbarkeit der oder des alten Formats. Diese Lesbarkeit alter Formate ist aber nicht auf Dauer garantiert. Beispiele solcher Werkzeuge sind die die typischen *Office*-Anwendungen oder Geographische Informationssysteme (GIS), bei denen nicht das Programm selbst gekauft wird, sondern Nutzungsrechte. Zur Sicherstellung der Lesbarkeit dieser Formate über lange Zeiträume müssen *Data Curation Services* existieren, die Datenentitäten mit abgekündigten Formaten aufspüren und in neue Formate wandeln. An dieser Stelle ist die *Bit-stream* Sicherung verletzt (Prüfsummen sind nicht identisch) und zur Sicherung der Datenintegrität sind Qualitätssicherungen erforderlich, die Unversehrtheit des Inhalts sicherstellen. Die Entwicklung solcher Dienste steht noch ganz am Anfang und speziell die wissenschaftlichen Bibliotheken bemühen sich im Rahmen der elektronischen Informationsversorgung um Systematisierung bestehender Ansätze. Formate dieser Kategorie und deren Eignung für die Langzeitarchivierung werden im Detail im NESTOR Handbuch diskutiert. Ein Beispiel für einen aktuellen Standard für elektronische Dokumente ist das PDF (*Portable Document Format*).

Die zweite Kategorie sind große bis sehr große Datenmengen aus den Bereichen numerische Modellierung, Satellitendaten, *Monitoring*-Systeme, Daten aus Biodiversitätsuntersuchungen oder Hochenergiephysik. Allen diesen Daten ist ihre maschinelle Erzeugung gemeinsam, bei der die Datenproduktion nur durch die Leistungsfähigkeit der Maschinen beschränkt wird. Die diskutierten Archivvolumina wachsen alle 3 bis 5 Jahre um eine Größenordnung von Terabyte über aktuell Petabyte zu Exabyte Datenarchiven. Die Archive verwalten zwar riesige Datenmengen, aber sie sind weitgehend homogen und in einer überschaubaren Anzahl von Formaten gespeichert. Diese Formate sind in der jeweiligen Wissenschafts-*Community* definiert, sind Speicherplatz sparend und enthalten häufig Angaben zur weiteren, maschinellen Verarbeitung der Daten (z. B. Variable, Einheit, Raum-Zeit-Bezug). Es sind selbstbeschreibende Binärformate, die nicht wie ASCII Daten direkt lesbar sind, sondern für deren Erzeugung und weiteren Verarbeitung spezielle Computerprogramme (*Libraries* oder Bibliotheken) benötigt werden. Diese Format-Bibliotheken sind in der Verantwortung der jeweiligen Wissenschaftsdisziplin. Aufgrund der Datenmenge verbietet sich hier der oben skizzierte Weg der Formatkonvertierung zur Sicherstellung der Lesbarkeit. Für diesen Typ Daten wird der Weg der Migration der Format-Bibliotheken

von einer Rechnergeneration auf die nächste gewählt, wobei darauf geachtet wird, dass bei Weiterentwicklung des Formatstandards auch alte Daten noch gelesen werden können (Abwärtskompatibilität). Beispiele für diese Formate sind die den Naturwissenschaften verwendeten Formate NetCDF (*Network Common Data Form*) und HDF (*Hierarchical Data Format*). Neben der Unterstützung der *Data Curation Services* erleichtern derartige Standardformate den Datenaustausch und die maschinelle Verarbeitung.

3.1.3 Sicherstellung der Interpretierbarkeit

Dieser Teil der Integritätssicherung ist der am schwierigsten zu standardisierende Bereich, da es hier um (wissenschaftliche) Inhalte von Datenentitäten geht. Während *„Data Curation Services“* die Unversehrtheit und Lesbarkeit elektronischer Daten auch über Disziplingrenzen hinweg sicherstellen, ist eine Disziplin übergreifende Interpretierbarkeit konzeptionell schwierig und erst in Ansätzen realisiert (Beispiel: Verwendung elektronischer Daten in der Klimafolgenforschung). Zu unterschiedlich erscheinen die Begriffswelten und Beschreibungstraditionen, die über die Wissenschaftsdisziplinen hinweg aufeinander abgebildet werden müssten, um eine breitere Umsetzung zu etablieren.

Interpretierbarkeit und Wiederverwendung von Daten ist eng verknüpft mit Kontextinformation bzw. Metadaten („Daten über Daten“). Hier sind Informationen enthalten zum Inhalt der Datenentität, zur technischen Verarbeitung, zur Qualität und zur Datenhistorie (*data provenance*). Die Suche nach Datenentitäten zu bestimmten Begriffen kann ebenfalls durch Metadaten unterstützt werden. Dafür werden die Metadaten in Katalogen zusammengefasst und in Repositorien abgelegt. Standardisierung von Metadaten in Schemata erleichtert die Suche nach Schlüsselbegriffen und steigert die Effizienz. Häufig werden für die Suche in standardisierten Datenkatalogen relationale Datenbanksysteme verwendet. Sowohl Metadaten Schema (auch Datenmodell genannt) als auch der Inhalt der Metadaten sind abhängig von der Wissenschaftsdisziplin (also dem Dateninhalt selbst) und dem Anwendungsgebiet bzw. der Nutzergruppe dieser Metadaten.

Ein Ziel der Langzeitarchivierung muss sein, die Nachnutzbarkeit von Daten sicher zu stellen. Dafür müssen die Metadaten so vollständig sein, dass Datenentitäten unabhängig vom Erzeuger auf ihre Verwendbarkeit hin beurteilt und letztlich auch verwendet werden können. Die so definierte Vollständigkeit ist dabei nicht unabhängig von der Nutzergruppe des Datenarchivs. Grob unterschieden wird

- Fachnutzung durch Wissenschaftler aus der entsprechenden Fachdisziplin,
- interdisziplinäre Nutzung durch Wissenschaftler aus unterschiedlichen Fachdisziplinen und
- Datennutzung im Rahmen Öffentlichkeit und Politik.

Neben dem Dateninhalt selbst stellen diese Nutzergruppen unterschiedliche Anforderungen an die Komplexität von Metadaten.

Metadaten werden häufig in strukturierten Datenmodellen abgelegt zur Unterstützung der Vergleichbarkeit der beschriebenen Datenentitäten. Feste Wertelisten für Schlüsselbegriffe (wie z. B. für Variable oder die Raum-Zeit Zuordnung) unterstützen Datensuche, Datenaustausch und Vergleichbarkeit. Strukturierte Datenmodelle formalisieren und standardisieren die Datendokumentation und erleichtern die Sicherstellung der Interpretierbarkeit im Rahmen der Datenpflege (*Data Curatio*n) im Langzeitarchiv. Im Zuge von Datenpflege und Qualitätssicherung müssen auch Verarbeitungsschritte und Historie der Daten (*data provenanc*e) in den Metadaten aktualisiert werden.

Die Definition strukturierter Datenmodelle und die Festlegung von Wertelisten für Schlüsselbegriffe erscheinen durchführbar innerhalb einer Wissenschafts-*Community*, die mit gleichen Datentypen in einer definierten Begriffswelt arbeitet. Eine interdisziplinäre Verwendung stellt Anforderungen an die Abbildung von Begrifflichkeiten verschiedener Wissenschaftsdisziplinen aufeinander (Ontologien). Namen für bestimmte Größen und Sachverhalte sind durchaus unterschiedlich in verschiedenen Wissenschaftsdisziplinen. Werden Datenarchive für den interdisziplinären Zugriff geöffnet, sollte der Fachspezifik der jeweiligen Ontologien und Begriffswelten Rechnung getragen werden, damit bei einer Datensuche auch die erwarteten Größen geliefert werden.

Eine weitere Komplexitätsstufe in der Sicherstellung der Interpretierbarkeit ergibt sich im Übergang von Datenmanagement zum Informationsmanagement. Im Informationsmanagement werden strukturierte Information (Metadaten und Daten) verknüpft mit unstrukturierter Information wie Zeitschriftenveröffentlichungen, Texten und Grafiken. Während die strukturierten Informationen als Tabellen und Hierarchien in relationalen Datenbanken effizient organisiert werden können, bieten sich für unstrukturierte Informationen XML-Datenbanken und das *Ressource Description Framework* (RDF) an, die erlauben, eine Netzwerktopologie der Information frei zu definieren. Elektronische Datenentitäten können so flexibel verknüpft werden mit Zusatzinformationen wie z. B. graphischen Darstellungen, Literatur oder verwandten Datenquellen. Das so entstehende Informationsnetzwerk unterscheidet sich deutlich von der Matrix-Struktur für Informationen in relationalen Datenbanksystemen.

3.1.4 Data Curation Services

Das Referenzmodell für ein *Open Archival Information System* (OAIS) definiert Informationen, Services und Prozesse, die in einem Langzeitdatenarchiv implementiert werden sollten. Die hier gewählte und diskutierte Schichteneinteilung beginnend mit der Maschinen bezogenen Ebene (Sicherstellung der Unversehrtheit), fortschreitend mit der syntaktischen Ebene (Sicherstellung der Lesbarkeit)

und endend mit der semantischen Ebene (Sicherstellung der Interpretierbarkeit) ist verbunden mit einer Abnahme von Gemeinsamkeiten über die Wissenschafts-Communities hinweg. Während die Sicherstellung der Unversehrtheit als *Curation Service* für alle Daten auf der Ebene der Rechenzentren organisiert wird, wird für die ‚Sicherstellung der Lesbarkeit' bereits in den *Curation Services* zwischen heterogenen Daten geringen Volumens und homogenen Daten großen Volumens unterschieden. In der Ebene ‚Sicherstellung der Interpretierbarkeit' sind bisher erst in einzelnen Wissenschaftsdisziplinen *Curation Services* für Metadaten eingerichtet. Definition und Entwicklung von Ontologien sind in Forschungsgebieten mit inter-disziplinärer Ausrichtung zu beobachten. Die Vereinheitlichung von Strukturen und Metdatenmodellen zur Beschreibung von Daten ist parallel zur Entwicklung von Ontologien in einzelnen Forschungsgebieten zu erkennen. Ein Beispiel ist die EU-Direktive INSPIRE zur Beschreibung und Organisation von Raum bezogenen Daten nicht nur im Forschungsbereich, sondern gerade auch für Behörden und Ämter in den europäischen Mitgliedsstaaten.

Das ICSU *World Data Center Climate* (WDCC) hat als Arbeitsschwerpunkt die Langzeitarchivierung von Klimamodelldaten und verwandter Beobachtungsdaten. Das WDCC unterstützt mit seinen „*Data Curation Services*" alle drei Ebenen der Integritätssicherung elektronischer, wissenschaftlicher Daten. „*Bitstream Preservation*" ist im Rahmen des Massenspeichersystems am Deutschen Klimarechenzentrum (DKRZ) realisiert. Es werden zwei Magnetbandkopien gespeichert und die Anzahl der Bandzugriffe wird protokolliert. Erreicht der Zugriffszähler seinen Maximalwert, wird der Bandinhalt automatisch auf ein frisches Medium kopiert und das alte Magnetband wird im Silo ausgetauscht. Lesbarkeit der Daten im WDCC wird im Rahmen der Qualitätssicherung kontrolliert. Es wird geprüft, ob die gespeicherten Daten lesbar sind und die ausgelesenen Werte den Erwartungswerten der gespeicherten Variablen entsprechen. Die Interpretierbarkeit wird mit dem verwendeten, umfangreichen Metadatenmodell umgesetzt. Ziel ist, dass im WDCC gespeicherte wissenschaftliche Daten auch nach 10 Jahren und mehr direkt und ohne Nachfrage beim Datenautor in wissenschaftlichen Arbeiten verwendet werden können. Besondere Anforderungen an die Datenbeschreibung stellt die Klimafolgenforschung mit ihrem interdisziplinären Ansatz.

Literaturhinweise

CCSDS (Consultive Committee for Space Data Systems), 2002. *OAIS Reference Model for an Open Archival Information System (OAIS).* Blue Book. (Jan. 2002) Online: http://public.ccsds.org/publications/archive/650x0b1.pdf [Zugriff am 09.08.2010].

DFG (Deutsche Forschungsgemeinschaft), 1998. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis.* Denkschrift. Weinheim: Wiley-VCH.

European Commission, o.J. *INSPIRE Directive*. Online: http://inspire.jrc.ec.europa.eu/index.cfm [Zugriff am 14.08.2011].

Neuroth, H. et al. Hrsg., 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. (Version 2.3.) Online: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949.

Wikipedia, 2011a. *RAID*. (Version vom 12.08.2011, 18:34 h) Online: http://de.wikipedia.org/w/index.php?title=RAID&oldid=92366907 [Zugriff am 09.08.2011].
(Originaldokument: Patterson, D. A. Gibson, G. & Katz, R. H., 1988. *A Case for Redundant Arrays of Inexpensive Disks*. Online: http://www-2.cs.cmu.edu/~garth/RAIDpaper/Patterson88.pdf [Zugriff am 09.08.2011].)

Wikipedia, 2011b. *Prüfsumme*. (Version vom 19.04.2011, 21:45 h) Online: http://de.wikipedia.org/w/index.php?title=Pr%C3%BCfsumme&oldid=87905903 [Zugriff am 14.08.2011].
(Rivest, R., 1992. *The MD5 Message-Digest Algorithm*. (Network Working Group – Request for Comments: 1321). Online: http://people.csail.mit.edu/rivest/Rivest-MD5.txt [Zugriff am 09.08.2011].)

Wikipedia, 2011c. *PDF*. (Version vom 08.08.2011, 15:31 h) Online: http://de.wikipedia.org/w/index.php?title=Portable_Document_Format&oldid=92200868 [Zugriff am 09.08.2011]
(Adobe Systems, 2011. *PDF Reference and Adobe Extensions to the PDF Specification*. Online: http://www.adobe.com/devnet/pdf/pdf_reference.html [Zugriff am 09.08.2011].)

Wikipedia, 2011d. *NetCDF*. (Version vom 07.04.2011, 10:09 h) Online: http://de.wikipedia.org/w/index.php?title=NetCDF&oldid=87398911 [Zugriff am 14.08.2011]
(Unidata Program Center, o.J. *NetCDF (Network Common Data Form)*. Online: http://www.unidata.ucar.edu/software/netcdf/ [Zugriff am 14.08.2011].)

Wikipedia, 2011e. *HDF.* (Version vom 18.06.2011, 9:08 h) Online: http://de.wikipedia.org/w/index.php?title=Hierarchical_Data_Format&oldid=90174104
(HDF Group, o.J. *Welcome*. Online: http://www.hdfgroup.org/ [Zugriff am 14.08.2011].)

Wikipedia, 2011f. *Resource Description Framework.* (Version vom 16.06.2011, 10:14 h) Online: http://de.wikipedia.org/w/index.php?title=Resource_Description_Framework&oldid=90099766 [Zugriff am 09.08.2011]. (W3C, 2010. Resource Description Framework (RDF). (Stand: 07.03.2010, 7:34 h) Online: http://www.w3.org/RDF/ [Zugriff am 14.08.2011].)

Wissenschaftlicher Rat der Max-Planck-Gesellschaft, 2001. *Verantwortliches Handeln in der Wissenschaft – Analysen und Empfehlungen.* (Max-Planck-Forum Bd. 3). München: Max-Planck-Gesellschaft.

3.2 Strategien bei der Veröffentlichung von Forschungsdaten

Sünje Dallmeier-Tiessen
CERN / Humboldt Universität zu Berlin

Zusammenfassung

Forschungsdaten liegen in Abhängigkeit der Disziplinen in vielfältigen Formen und Formaten vor. Sie sind in allen Disziplinen Teil des wissenschaftlichen Erkenntnisprozesses. Als digitales Informationsobjekt sind sie komplex und bislang wenig studiert. Mit den Möglichkeiten neuer Informationstechnologien wurden in den letzten Jahren neue Wege in der Publikation von Forschungsdaten beschritten. Mit Fokus auf die Naturwissenschaften werden im Folgenden drei Publikationsmodelle beschrieben: Die Veröffentlichung von Forschungsdaten als eigenständiges Objekt in einem Forschungsdatenrepositorium, die Veröffentlichung von Forschungsdaten mit textueller Dokumentation und die Veröffentlichung von Forschungsdaten als Anreicherung einer interpretativen Text-Publikation.

3.2.1 Einführung

Das Verständnis des Begriffes Forschungsdaten variiert je nach Disziplin. Möchte man sich dem Begriff disziplinübergreifend annähern, so lässt sich dies z. B. über die Positionierung der Forschungsdaten im wissenschaftlichen Arbeitsprozess versuchen. Forschungsdaten sind der zentrale Gegenstand des wissenschaftlichen Erkenntnisprozesses. Ausgehend von bereits veröffentlichten Ergebnissen der Community werden sie ausgewertet und eingeordnet. Im Rahmen des Publikationsprozesses werden sie anschließend beschrieben. Diese Beschreibung in textueller Form stellt traditionell eine Interpretation und Diskussion der Daten dar. Dabei sind je nach Disziplin unterschiedliche Kulturen etabliert, die auf den heterogenen Formen und Formaten der Daten basieren. Während in den Naturwissenschaften beispielsweise oftmals ein Untersuchungsgegenstand durch unterschiedliche Methoden erhoben und ausgewertet wird, wird in den Geisteswissenschaften häufig auf bereits publizierte Untersuchungsgegenstände wie historische Text- und Bildmaterialien Bezug genommen, die traditionell unter dem Begriff „Quellenmaterial“ gefasst werden.

Am Beispiel der Biologie wird die Heterogenität auch innerhalb der Disziplinen deutlich: In dieser Disziplin können z. B. numerische Daten von Laborexperimenten, audiovisuelle Objekte wie Fotos oder Videosequenzen von Tierbeobachtungen, komplexe Modelle von Simulationen (z. B. von Umweltveränderungen) als Forschungsdaten bezeichnet werden.

Da eine detaillierte interdisziplinäre Definition des Begriffs Forschungsdaten schwerfällt, werden für die folgende konzeptionelle Betrachtung der Publikation von wissenschaftlichen Daten, Forschungsdaten als digitale Informationsobjekte betrachtet, die während des Forschungsprozesses entstehen und als Grundlage einer interpretativen wissenschaftlichen Textpublikation dienen.

Mit der rasanten Entwicklung der Informationstechnologien haben sich in den letzten Jahren neue Wege im Umgang mit wissenschaftlichen Daten eröffnet, die nach dem Verständnis von Jim Gray zu einem neuen Paradigma des wissenschaftlichen Arbeitens geführt haben:

„The world of science has changed, and there is no question about this. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, *fourth paradigm* for scientific exploration." (Hey et al., 2009, S. xix) Mit diesem „*fourth paradigm*" entstehen neue Wege im Umgang mit wissenschaftlichen Daten.

3.2.2 Diskussion um den zeitgemäßen Umgang mit Forschungsdaten

In Deutschland wird der Umgang mit Forschungsdaten vornehmlich in Bezug auf die Nachprüfbarkeit der Daten geregelt. Ein zentraler Punkt sind dabei die von der Deutschen Forschungsgemeinschaft (DFG) 1998 veröffentlichten „Vorschläge zur Sicherung guter wissenschaftlicher Praxis". Diese sehen vor, dass „Primärdaten als Grundlagen für Veröffentlichungen […] auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden [sollen]." (DFG, 1998). Der Aspekt der Nachnutzung gewinnt aktuell an Relevanz. 2010 hat die DFG diesen Aspekt im „Leitfaden für Antragsteller" verankert. In diesem heißt es: „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." (DFG, 2010) Andere Förderer sind in ihren Anforderungen noch konkreter, so erwartet die *National Science Foundation* (NSF) einen Datenmanagementplan von Antragsstellern (NSF, 2011).

Auf internationaler Ebene wurden in den letzten Jahren zahlreiche Empfehlungen und Vorgaben zum Umgang mit Forschungsdaten verabschiedet. Zentral sind hier die 2007 von der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD, 2002) publizierten „*Principles and Guidelines for Access*

to Research Data from Public Funding“. In den Disziplinen variiert die Bereitschaft des „*data sharings*“ deutlich. In der Genomforschung wurden 1996 die „*Bermuda Principles*“ im Rahmen des Humangenomprojektes formuliert (z. B. Marshall, 2001), nach denen Gensequenzen frei zugänglich gemacht werden sollen. Auch Fachgesellschaften wie z. B. die *American Geophysical Union* (AGU, 1993) haben Richtlinien zum Umgang mit Forschungsdaten formuliert. Die AGU-Richtlinie bezieht sich auf Daten, die Grundlage einer Publikation sind, die in einem Organ der Fachgesellschaft veröffentlicht werden. Darüber hinaus haben auch Journale, wie z. B. PLoS ONE, in ihren Veröffentlichungsrichtlinien Aussagen zu dem Thema formuliert. PLoS ONE fordert unter dem Punkt „*Sharing of Materials, Methods, and Data*“ die Zugänglichkeit der Daten, die Basis einer Publikation sind (PLoS ONE, o. J.).

3.2.3 Publikationsmodelle

Auf Basis der beschriebenen Diskussion und der Annäherung an den Begriff „Forschungsdaten“ können diese Daten als zentraler Baustein des wissenschaftlichen Diskurses betrachtet werden.

Ausgehend von der Zugänglichkeit der Daten können unterschiedliche Aspekte einer wissenschaftlichen Nachnutzung herausgestellt werden, die auch auf die Publikationsstrategie der Daten wirken. Dabei sind die folgenden Ansätze zu unterscheiden:

- Die Publikation von Daten um eine Nachprüfbarkeit der Daten zu ermöglichen.
- Die Publikation von Daten um eine Nachnutzung der Daten zu ermöglichen.

Der häufig anzutreffende Begriff des „data sharing“ spezifiziert, dass eine Publikation von Forschungsdaten zum Zwecke der Nachnutzung durch Dritte erfolgt. Nach Borgman (2010) sprechen folgende Gründe für das „*sharing*“ wissenschaftlicher Daten: „to make the results of publicly funded data available to the public, to enable others to ask questions of extant data, to advance the state of science and to reproduce research“. Die beiden erstgenannten Argumente sind aus der Perspektive der Nachnutzung zu betrachten, deren Anspruch durch öffentliche Interessen gesteuert ist. Die beiden letztgenannten Argumente sind aus Sicht der Datenproduzenten interessant und greifen das Interesse der jeweiligen wissenschaftlichen Community auf (nach Borgman, 2010).

Zentraler Akteur bei der Publikation von Forschungsdaten ist der Wissenschaftler oder das Team, das die Daten erhebt und interpretiert.

Wird eine Veröffentlichung der Daten angestrebt, so bedarf es häufig einer erweiterten textuellen Aufbereitung der Daten von Seiten der Wissenschaftler, die im Folgenden unter dem Begriff „Dokumentation“ gefasst wird. In der

Dokumentation werden die Kontextinformationen bereitgestellt, um Personen, die an der ursprünglichen Datenproduktion nicht beteiligt waren, eine Nachnutzung der Daten zu ermöglichen.

Da es in vielen Disziplinen bisher keine etablierte Kultur der Veröffentlichung wissenschaftlicher Daten gibt, ist der Aufwand einer nachnutzbaren oder nachprüfbaren Veröffentlichung der Daten eine aufwendige und zusätzliche Tätigkeit, die über den „normalen" Wissenschaftsalltag hinausgeht. Bei der Publikation der Daten müssen z. B., je nach Disziplin, rechtliche und ethnische Rahmenbedingungen berücksichtigt werden. So wirken beispielsweise in den Sozialwissenschaften datenschutzrechtliche Aspekte auf die Publikationsmöglichkeiten und -strategien. Mit Blick auf die kompetitive Forschungslandschaft mag es dem einzelnen Wissenschaftler außerdem notwendig erscheinen, die Veröffentlichung von Daten nur eingeschränkt oder zeitverzögert anzustreben.

Ziel der folgenden übergreifenden Darstellung ist es, einen Überblick auf die gegenwärtige Landschaft der Forschungsdatenpublikation zu schaffen und damit eine Grundlage für zukünftige Diskussionen zu bieten. Hierbei soll vor allem herausgestellt werden, wie die wissenschaftlichen Daten in dem traditionellen Publikationsprozess eingebunden sind.

Mit Blick auf die Naturwissenschaften können drei Publikationsmodelle unterschieden werden[1].

- Die Veröffentlichung von Forschungsdaten als eigenständiges Objekt in einem Datenrepositorium
- Die Veröffentlichung von Forschungsdaten mit einer textuellen Dokumentation
- Die Veröffentlichung von Forschungsdaten als Anreicherung einer interpretativen Text-Publikation.

Alle Modelle setzen bestimmte Rahmenbedingungen in Bezug auf die genutzten Infrastrukturen voraus, insbesondere in Bezug auf eine vertrauenswürdige Publikationsumgebung, die beispielsweise Standards der Langzeitarchivierung berücksichtigen.

Die unterschiedlichen Modelle ziehen unter Umständen unterschiedliche Zeitpunkte der Forschungsdatenpublikation nach sich, insbesondere in Bezug zum interpretativen Artikel. Dieser Aspekt soll hier nicht im Detail ausgeführt werden, er kann jedoch bei der Wahl des Publikationsmodells entscheidend sein.

3.2.3.1 Eigenständige Veröffentlichung im Forschungsdaten-Repositorium

Ein Publikationsmodell ist die Veröffentlichung von wissenschaftlichen Daten in einem Forschungsdaten-Repositorium. Das Repositorium ermöglicht es, For-

1. Darüber hinaus gibt es sicher weitere individuelle Herangehensweisen, auf die in diesem Papier nicht eingegangen werden kann.

schungsdaten zeitlich und örtlich unabhängig von einer interpretativen Veröffentlichung zu veröffentlichen. Die Daten werden als eigenständiges Objekt in ein Datenrepositorium abgelegt. Je nach Disziplin kann dies z. B. nach einer automatisierten Prozessierung aus einem „System“ heraus erfolgen. Dieses Publikationsmodell ist in vielen naturwissenschaftlichen Disziplinen verbreitet. Je nach Disziplin und Organisationstruktur in den Disziplinen variieren die Anforderungen an die Standardisierung und Bereitstellung. Dieses Modell kann auch als Baustein verstanden werden, welches – Standardisierung und Interoperabilitiät vorrausgesetzt – mit anderen Arbeitsabläufen und Produkten der Wissenschaft kombiniert werden kann, siehe dazu auch die Modelle in 3.2.3.2 und 3.2.3.3. Beispielhaft seien hier vier Repositorien für dieses Publikationsmodell genannt, die die Diversität und die Möglichkeiten des Modells widerspiegeln:

- PANGAEA (http://www.pangaea.de): In diesem Datenrepositorium werden geowissenschaftliche Forschungsdaten veröffentlicht. Dank der Nutzung des *Digital Object Identifier* (DOI) wird eine persistente Adressierung der Daten ermöglicht. Komplexere Forschungsdatensätze können als „*Collection*“ publiziert werden.
- GenBank (http://www.ncbi.nlm.nih.gov/genbank): In dieser Datenbank werden Forschungsdaten nach den disziplinären Regeln der Genforschung publiziert. Eine „*accession number*“ dient zur persistenten Adressierung. Nach Vorgaben der *Editorial-Policies* vieler lebenswissenschaftlicher Zeitschriften muss dieser bei einer Interpretation der Daten in einem Aufsatz angegeben werden.
- Integrated Ocean Observing System (http://www.ioos.gov): Hier werden automatisch erfasste Messdaten der marinen Umweltforschung automatisiert aus dem System heraus veröffentlicht.
- Dryad (http://datadryad.org): Dieses Datenrepositorium der Biologie vergibt DOIs für die veröffentlichten Datensätze. Dryad konzentriert sich auf Datensätze, die Grundlage einer interpretativen Publikation sind. Es gibt eine Zitierempfehlung für den Datensatz, welche zusätzlich zu einer Zitierung des interpretativen Artikels erfolgt.

3.2.3.2 Veröffentlichung mit einer textuellen Dokumentation

Im Hinblick auf die Nachnutzung und Nachvollziehbarkeit von Forschungsdaten benötigen diese oftmals eine erweiterte Dokumentation, die zwar keine Interpretation, aber einen Kontext zur Erhebung der Daten liefert und Interessierten eine Nachnutzung ermöglicht. Vor diesem Hintergrund werden neuere Konzepte entwickelt. So werden z. B. Forschungsdaten mit einer begutachteten Dokumentation publiziert.

Im Folgenden sollen zwei Beispiele dieser Publikationsstrategie vorgestellt werden, bei denen Forschungsdaten mit einer nicht-interpretativen Dokumentation angereichert werden (siehe Abbildung 1).

Abb. 1: Die Publikation von Forschungsdaten mit einer erweiterten (begutachteten) Dokumentation und ggf. weitere Materialien, welche die Nachnutzung ermöglichen.Weitere Service können an die Publikation gekoppelt werden. Zusätzliche Analysen und (mehrere) interpretative Artikel können auf den Forschungsdaten aufbauen.

Beispiel: Earth System Science Data (ESSD)[2]

Die *Open-Access*-Zeitschrift *Earth System Science Data* (ESSD) widmet sich der erweiterten und qualitätsgesicherten Dokumentation von geowissenschaftlichen Forschungsdaten. Die Zeitschrift wird von *Copernicus Publications* verlegt und nutzt ein zweistufiges *Peer-Review*-Verfahren unter der Nutzung des „*Public Peer-Reviews*“ (Pöschl, 2010) um einen Datensatz und dessen Dokumentation zu veröffentlichen und zu begutachten.

Der Pubklikationsprozess in ESSD ist wie folgt organisiert: Ein Publizierender verfasst eine Dokumentation eines Datensatzes. In dieser beschreibt er beispielsweise die verwendeten Instrumente und angewandte Verfahren bei der Bearbeitung der Daten. Diese Beschreibung, die wie ein klassischer Aufsatz gestaltet wird, reicht er dann unter Nennung eines Persistent Identifier des Datensatzes bei der Zeitschrift ein. Darüber hinaus veröffentlicht er die beschriebenen Daten auf einem frei zugänglichen Daten-Repositorium. Nach einem erfolgreichen Begutachtungsprozess wird dann die Datendokumentation publiziert. Nach diesem Ansatz sind Datenrepositorium und Datenjournal deutlich getrennt. Das Ambiente der Zeitschrift ESSD erinnert stark an die Journale, welche „interpretativen“ Artikel publizieren.

2. http://earth-system-science-data.net/ [Zugriff am 18.08.2011].

Das Datenjournal ist unabhängig von Repositorien, hat jedoch eigene Kriterien für die Akzeptanz der Repositorienwahl festgelegt (Pfeiffenberger, 2011), welche sich nach den gegenwärtigen Standards für vertrauenswürdige Repositorien (z. B. von NESTOR, 2009) richten.

Für die Wissenschaftler ist der Prozess der Publikation ein bekannter, obwohl das zentrale Publikationsobjekt Forschungsdaten und deren Dokumentation in diesem Rahmen neu ist. Der zusätzliche Aufwand einer Einreichung der Publikationsobjekte zu zwei Publikationsplattformen (Zeitschrift und Datenrepositorium) mag für den Wissenschaftler durch den Erhalt eines zitierfähigen Artikels in einem bekannten Journalformat kompensiert werden.

Beispiel: Overlay Journal Infrastructure Meteorological Sciences (OJIMS)[3]

Unter dem Begriff „*Overlay Journal*“ fasst man Publikationsmodelle zusammen, welche auf Textpublikationen zurückgreifen, die auf verschiedenen Plattformen *Open Access* publiziert sind. Das Konzept beinhaltet eine thematische Aggregation, meist mit einem weiteren Service kombiniert, z. B. einem Begutachtungsprozess- oder Kommunikationsprozess. Das Konzept kann auch auf die Publikation von Forschungsdaten übertragen werden. Das *Overlay Journal Infrastructure Meteorological Science* (OJIMS) wendet dieses für meteorologische Daten an. Die meteorologischen Daten werden auf einem Forschungsdaten-Repositorium gespeichert, die Dokumentation in einem Dokumentenrepositorium. Beides wird anschließend von Fachwissenschaftlern begutachtet. Auf Basis dieser qualitätsgesicherten Publikation können dann Interpretationen der Daten vorgenommen werden.

3.2.3.3 (Interpretative) Publikation plus Forschungsdaten

Traditionell werden wissenschaftliche Ergebnisse in einer Textpublikation beschrieben und diskutiert. Über Disziplingrenzen hinweg lässt sich dabei seit Jahrhunderten eine ähnliche Artikelstruktur verfolgen, die auf die Darstellung eines Ergebnisses fixiert ist. Neben dem Text dienen Tabellen und Abbildungen der Beschreibung des dargestellten Sachverhaltes. Zusätzliche Materialien werden oft in den Anhang dargestellt (siehe Abbildung 2). Mit der Digitalisierung der Wissenschaftskommunikation und der Umstellung auf elektronische Zeitschriften können diese Materialien, z. B. Messreihen im CSV-Format oder Proteinsequenzen im FASTA-Format nachnutzbar und nachprüfbar publiziert werden. Die Publikation solcher Daten, die insbesondere zur Unterstützung einer interpretativen Publikation dienen, wird heute insbesondere auf den folgenden Wegen umgesetzt.

3. http://proj.badc.rl.ac.uk/ojims [Zugriff am 18.08.2011].

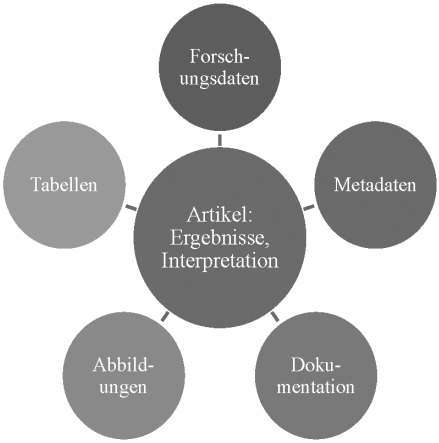

Abb. 2: Die Publikation von Forschungsdaten als Anhang einer interpretativen Publikation. Dieser wird beispielsweise in einem Journal publiziert und die zugrunde liegenden Forschungsdaten werden zusammen mit dem Artikel bereitgestellt. Im Anhang zu dem interpretativen Artikel finden sich traditionell auch oftmals Abbildungen, Tabellen und detailliertere Angaben zur Methodik.

Die Daten werden auf der gleichen Plattform wie auch die Textpublikation veröffentlicht, aber nicht individuell persistent und standardisiert adressiert. Dies kann auf einer Verlags- oder auch Repositorienplattform sein, welche vorwiegend für Textpublikation ausgelegt ist. Die Daten sind somit Bestandteil des Artikels.

In einigen Disziplinen werden Forschungsdaten als eigenständiges Objekt, das referenzierbar ist, auf der Publikationsplattform des textuellen Artikels gespeichert. Dank der Referenzierbarkeit, die über eine persistente Adressierung gegeben ist, kann der Datensatz zitiert werden. In dem interpretativen Artikel wird dann diese Identifizierung angegeben.

Der Datensatz kann in diesem Modell auch in einem externen Datenrepositorium (siehe 3.2.3.1) zeitlich parallel zu dem interpretativen Artikel publiziert werden. Dies bedeutet für den Wissenschaftler, dass dieser den Forschungssdatensatz und den interpretativen Artikel in zwei unterschiedlichen Orten einreicht. Eine koordinierte Zusammenarbeit der involvierten Infrastruktureinrichtungen, oft Verlage und Datenrepositorien, wie sie in 3.2.3.1. beschrieben worden sind, ist dabei wichtig, um zusätzlichen Aufwand auf Seiten der Wissenschaftler zu vermeiden und um den Anforderungen einer vertrauenswürdigen Langzeitarchivierung zu entsprechen. In der Nachnutzung können Textpublikation und Forschungsdaten individuell genutzt und zitiert werden (siehe z. B. die Zitierempfehlung des Datenrepositoriums Dryad). Kooperationen zwischen den

Publikationsagenten Verlag / Journal und Datenrepositorium lassen sich bereits in einigen Disziplinen beobachten. Es sollen hier zwei Beispiele vorgestellt werden.

- PANGAEA und Elsevier: Der Verlag Elsevier und das Datenrepositorium PANGAEA sind eine Kooperation eingegangen (Elsevier, 2010), die erlaubt, dass Leser von Artikeln auf Science Direct direkten Zugang zu Forschungsdaten haben, welche einem geowissenschaftlichen Artikel zugeordnet sind und auf PANGAEA abgelegt wurden. Der Link zu dem Datensatz ist prominent auf der Artikelseite platziert und führt den Leser direkt zum Datensatz.
- Dryad und verschiedene Verlage / Journale: Das Datenrepositorium Dryad kooperiert mit verschiedenen Verlagen und Journalen, insbesondere aus den biologischen Disziplinen. Das Dryad Konsortium hat eine „*Joint Data Archiving Policy*“ entwickelt, welche die Ablage von Forschungsdaten für Einreichungen in den Partnerjournalen regelt. Es wird explizit von Forschungsdaten gesprochen, welche Grundlage der Interpretation im Artikel sind: „*data supporting the results in the paper*“ (Datadryad, 2010).

Einen interessanten Überblick über unterschiedliche, disziplinspezifische Ausprägungen dieses Modells sind in den geförderten SURF-Projekten zu finden, welche an der disziplinspezifischen Umsetzung der sogenannten „*enhanced publications*“ arbeiten (SURF Foundation, 2010). SURF definiert „*enhanced publications*“ als „a publication – usually a text – that has been enhanced with additional material [...] The supplementary material may consist, for example, of research data, illustrative images, metadata sets, or post-publication data such as comments or rankings.“ Sechs disziplinspezifische Projekte werden von SURF seit Januar 2011 gefördert, u.a. aus der Ökonomie und der Linguistik.

3.2.4 Fazit

Die Entwicklung von Publikationsmodellen für Forschungsdaten findet sich aktuell in einer dynamischen Entwicklung, was eine Typisierung der unterschiedlichen Ansätze schwierig macht.

In dieser Übersicht wurden drei Modelle der Publikation von Forschungsdaten beschrieben. Im erstgenannten Modell werden Forschungsdaten zitierfähig auf einem Repostorium veröffentlicht. In einem zweiten Modell werden Forschungsdaten mit einer „erweiterten“ Dokumentation einem Begutachtungsprozess unterzogen. Zusätzliche und qualitätsgesicherte Kontextinformationen erleichtern dann die Nachnutzung. Ein bereits in zahlreichen Disziplinen umgesetztes Modell ist die Forschungsdatenpublikation als Anhang zu einem interpretativen Artikel. Hierbei kann der Datensatz z. B. in einem externen Datenrepositorium abgelegt werden und dann in dem Artikel referenziert werden.

Die weitere Entwicklung der Publikationsmodelle wird stark von den Publikationskulturen der einzelnen Disziplinen abhängig sein. Aufgrund der hohen disziplinspezifischen Charakteristika der Forschungsdaten sind angepasste Lösungen, die den Anforderungen der Disziplinen entsprechen, nötig. Der Kooperation zwischen Wissenschaft, Infrastruktureinrichtungen und Verlagen kommt dabei eine entscheidende Rolle zu.

Literaturhinweise

AGU Publications Committee, 1993. *Policy on Referencing Data in and Archiving Data for AGU Publications*. (Revised: March 1994, Dec. 1995, Oct. 1996) Online: http://www.agu.org/pubs/authors/policies/data_policy.shtml [Zugriff am 22.02.2011].

Bermuda Principles, o. J. *Policies on Release of Human Genomic Sequence Data*. (Last modified: 29.10.2003) Online: http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml [Zugriff am 22.02.2011].

Borgman, C. L., 2010. Research Data: Who will share what, with whom, when, and why? *5th China-North American Library Conference*. Peking, China 8.-12. Sept. 2010. Online: http://works.bepress.com/borgman/238/ [Zugriff am 22.02.2011].

Dallmeier-Tiessen, S. & Pfeiffenberger, H., 2008. Peer Reviewed Data Publication in Earth System Sciences. In: C. Puschmann & D. Stein, D., Hrsg. 2008. *Towards Open Access Scholarship*. Düsseldorf: Düsseldorf University Press, S.77–84.

DataDryad, 2010. *Joint Data Archiving Policy (JDAP)*. Online: http://datadryad.org/jdap [Zugriff am 22.02.2011].

DFG (Deutsche Forschungsgemeinschaft), 1998. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*. Denkschrift. Weinheim: Wiley-VCH.

DFG (Deutsche Forschungsgemeinschaft), 2010. *Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für die Antragstellung.*(DFG-Vordruck 1.02 – 8/10) Online: http://www.dfg.de/download/programme/sachbeihilfe/antragstellung/1_02/1_02.pdf [Zugriff am 18.08.2011].

Elsevier, 2010. *„Elsevier and PANGAEA Take Next Step in Connecting Research Articles to Data“*. Online: http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01616 [Zugriff am 22.02.2011].

Hey, T. Tansley, S. & Tolle, K., 2009. Jim Gray on eScience: a transformed scientific method. In: S. Tansley & K. Tolle, Hrsg. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond, Wash.: Microsoft Research, S. XVII–XXXI. Online: http://www.fourthparadigm.org [Zugriff am 18.08.2011].

Marshall, E., 2001. Bermuda Rules: Community Spirit, With Teeth. *Science*,291(5507), S. 1192. DOI*:* 10.1126/science.291.5507.1192

NSF (National Science Foundation), 2011. *National Grant Proposal (GPG)*. Chapter II – Proposal Preparation Instructions, C. 2.j. Data Management. Online: http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp, 2011 [Zugriff am 22.02.2011].

NESTOR – Kompetenznetzwerk Langzeitarchivierung, 2009. *NESTOR-Kriterien – Kriterienkatalog vertrauenswürdige digitale Langzeitarchive.* (Version II). Frankfurt am Main: NESTOR.

OECD (Organisation for Economic Co-operation and Development), 2007. *Principles and Guidelines for Access to Research Data from Public Funding*. Online: http://www.oecd.org/dataoecd/9/61/38500813.pdf [Zugriff am 20.02.2011].

Pfeiffenberger, H. & Carlson, D., 2011. Earth System Science Data (ESSD) – A Peer Reviewed Journal for Publication of Data. *D-Lib Magazine*, (17)1/2.

PLoS ONE, o. J. *PLoS ONE Editorial and Publishing Policies.* Sharing of Materials, Methods, and Data. Online: http://www.plosone.org/static/policies.action#sharing [Zugriff am 22.02.2011].

Pöschl, U.: Interactive Open Access Publishing and Peer Review: The Effectiveness and Perspectives of Transparency and Self-Regulation in Scientific Communication and Evaluation. In: Liber Quarterly 19, 2010, Nr. 3/4.

SURF Foundation, 2010. *Enhanced Publication.* Online: http://www.surffoundation.nl/nl/publicaties/Documents/SHAREflyer_verrijkte_publicatie_pdfversie_def_ENG.pdf [Zugriff am 22.02.2011].

Zacharias, M. C., 2010. *Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans.* Online: http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928 [Zugriff am 18.08.2011].

3.3 Semantische Vernetzung von Forschungsdaten

Günther Neher [1], Bernd Ritschel [2]
[1] Fachhochschule Potsdam. Fachbereich Informationswissenschaften
[2] Deutsches GeoForschungsZentrum GFZ, Potsdam

3.3.1 Einführung

In diesem Kapitel soll im Gesamtkontext des Forschungsdatenmanagements die Frage diskutiert werden, welche Bedeutung den aktuellen Entwicklungen im Bereich des sog. „*Semantic Web*“ als Komponente innerhalb einer zukünftigen Forschungsdateninfrastruktur zukommt, welche Anwendungsbereiche sinnvoll erscheinen und welche übergreifenden infrastrukturellen Maßnahmen hierzu als notwendig erachtet werden. Neben einer kurzen Einführung in die grundlegenden Konzepte und technologischen Bausteine des *Semantic Web*, sowie einer Darstellung des wichtigen Aspekts der Ontologieentwicklung, wird auf das sich aktuell sehr stark entwickelnde Gebiet der *Linked Open Data* näher eingegangen, Das Potential für den Bereich der Forschungsdaten wird erläutert und spezifische für den Wissenschaftsbereich notwendige Anforderungen definiert. Abschließend wird ein grobes Vorgehensmodell für die Transformation von Datenbeständen und Datenmodellen in die *Semantic Web*-Infrastruktur skizziert und abgerundet durch eine kurze Übersicht zu Entwicklungswerkzeugen und Middleware. Als roter Faden werden die Darstellungen zur besseren Illustration durchgängig durch ein Projektbeispiel aus dem Bereich der Geoforschungsdaten begleitet (diese Textpassagen erscheinen jeweils *kursiv* gedruckt).

Die Erforschung komplexer geowissenschaftlicher Phänomene, wie zum Beispiel globale und regionale Klimaveränderungen im System Erde, ist nur durch die integrative Nutzung von Forschungsdaten verschiedener Wissenschaftsdisziplinen möglich. Die zunehmende Verwendung von Metadaten, die sowohl den Inhalt als auch den Kontext von Forschungsdaten beschreiben, ermöglicht die Nutzung von semantischen Methoden zu ihrer Vernetzung. Sind die Daten und Metadaten als strukturierte Objekte im Web weltweit eindeutig referenzierbar, so können hierfür Methoden und Techniken des Semantic Web genutzt werden. Das Semantic Web stellt dabei die für Menschen und Anwendungsprogramme notwendige Infrastruktur für die interoperable Vernetzung von Daten, Informationen und Hintergrundwissen zur Verfügung. Eine Voraussetzung für das effektive Teilen und Zusammenführen von Wissen ist dessen Repräsentation in einer standardisierten Form. Die Modellierung des Domänenwissens erfolgt in Ontologien unter Verwendung von Semantic Web Sprachen, wie RDF, RDF-S und OWL. Derartige Ontologien ermöglichen sowohl die Repräsentation und persistente Speicherung des eigenen Wissens und damit dessen Wiederverwendung als auch den Austausch und die Vernetzung mit anderen Wissensressourcen im

Web, die in gleicher Art formalisiert vorliegen. (Zur besseren Pflege und möglichst universellen Nachnutzung von Ontologien ist dabei eine Separierung von Inhalts- und Applikationsanteilen möglich. Die konsequente Anwendung des formalen Rahmens von Semantic Web Standards garantiert, dass diese Wissensressourcen im Internet dauerhaft, eindeutig referenzierbar und nachnutzbar sind.

3.3.2 Praxisbeispiel: GFZ ISDC[1] – Semantic Web Projekt

Zielsetzung des nachfolgend zur Illustration dienenden „fiktiven" GFZ ISDC – *Semantic Web* Projekts ist die semantische Integration unterschiedlichster geowissenschaftlicher Daten zur Generierung neuen Wissens unter Verwendung von *Semantic Web* Technologien. Dabei werden Metadaten zu geowissenschaftlichen Datenprodukten, kontextrelevante Zusatzinformationen, sowie übergreifendes strukturiertes objektbasiertes Wissen miteinander verknüpft. Die obengenannten Quellen und ihre Verbindungen zueinander sind in der ISDC-Ontologie modelliert[2]. Die ISDC-Ontologie (vgl. Abb. 5: ISDC-Ontologie (Ausschnitt)) bzw. ihre Repräsentation in der *Web Ontology Language* (OWL) sollen später als Grundlage dienen für die Entwicklung von *Semantic Web* Applikationen, die geeignet sind für

- eine globale Verknüpfung von Daten- und Informationsobjekten im *Semantic Web*
- die integrative wissensbasierte Recherche innerhalb des ISDC-*Repository* und außerhalb in Ressourcen des *Semantic Web*
- die Repräsentation / Visualisierung von geowissenschaftlichem Domänenwissen und Rechercheergebnissen.

Ziel der Entwicklung ist die semantische Integration möglichst vieler geowissenschaftlicher Daten- und Metadatenquellen, aller Arten von Publikationen sowie Applikationen, die sich sowohl innerhalb des ISDC-*Repository* als auch extern und global im WWW und dessen Erweiterung im *Semantic Web* befinden. Zusätzlich zu den bereits existierenden ISDC-Entitäten, wie Projekt, Plattform, Instrument, Produkt Typ und Institution wurden im *top-down* Ansatz die Entitäten Geophenomenon, *Personell* und *Country* und *Keyword* ins Modell aufgenommen. Die Semantik der zu den Entitäten (Klassen) gehörigen, im *bottom-up* Ansatz modellierten Individuen und deren Eigenschaften spiegeln sich in einem ISDC-eigenen *Ontology*-Vokabular wider. Die Verwendung von standardisierten *Keywords* aus dem inhaltlichen GCMD / DIF-Vokabular[3] zur Cha-

1. ISDC = Information System and Data Center.
2. Die aus dem Modell erzeugte OWL-Datei ist unter dem URL:http://isdc.gfz-potsdam.de/ontology/isdc_1.0.owl [Zugriff am 15.08.2011] abrufbar.
3. GCMD = Global Change Master Directory. DIF = Directory Interchange Format.

rakterisierung aller Individuen innerhalb der ISDC-Ontologie ermöglicht eine zu starren schlüsselbasierten Relationen komplementäre Verknüpfung. Diese Verknüpfung wird u.a. genutzt, um neues Wissen über Geophänomene, wie zum Beispiel dem oben erwähnten Klimawandel, aus dem Inhalt des ISDC-*Repository* mittels ontologiebasierter *Reasoning*-Methoden zu generieren.

3.3.3 Grundlagen: Semantic Web[4]

„*The Semantic Web is not a separate Web, but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*“ (Berners-Lee, Hendler & Lassila, 2001).

Dieser häufig zitierte Satz, welcher die Zielsetzung und Vision des „*Semantic Web*“ greifbar machen soll, beschreibt das *Semantic Web* als eine Weiterentwicklung vom aktuellen Zustand eines „*Web* der (verknüpften) Dokumente“ in ein „*Web* der (verknüpften) Daten“. Speziell für den Bereich des Forschungsdatenmanagements birgt diese Entwicklungsrichtung des Web ein enormes Nutzungspotential.

Eine Charakteristik des heutigen Web besteht darin, dass Informationen in Form von Dokumentinhalten vorwiegend in einer rein auf den menschlichen Leser zugeschnittenen Form vorliegen, z. B. auf Basis von (X)HTML. Die Bedeutung des jeweiligen Dokumentinhalts wird vom Leser intellektuell auf der Basis von Kontextwissen und Sprachverständnis „verstanden“. Für ein Anwendungsprogramm („die Maschine“) ist der Dokumentinhalt, abgesehen von evtl. Formatierungs-Tags, in der Regel nicht mehr als eine Abfolge von Zeichenketten, ohne jegliche Möglichkeit der Bedeutungsinterpretation. Dieses „Unverständnis“ des Dokumentinhalts ist eine der Hauptursachen für die i.d.R. immer noch unbefriedigende *Precision* der Suchergebnisse bei der Informationsrecherche im Web. Unter einem pragmatisch-operationalen Gesichtspunkt kann das *Semantic Web* daher zunächst als Toolset „semantischer Technologien“, d.h. als Infrastruktur bestehend aus einer Anzahl aufeinander aufbauender W3C-Standards[5] betrachtet werden, mit dem Ziel Webinhalte in ihrer semantischen Bedeutung maschineninterpretierbar und interoperabel zu machen. Abb. 1 zeigt die Komponenten der Semantic Web-Infrastruktur als schematisches Schichtenmodell (sog. „*Semantic Web Layer Cake*“).

4. Im Rahmen dieses Handbuchs können die zugrundeliegenden Konzepte nur sehr verkürzt dargestellt werden. Für eine genauere Darstellung sei auf die einschlägige einführende Literatur verwiesen (z. B. Pellegrini, 2006; Hitzler, 2008; Allemang, 2008; Antoniou, 2008).

5. World Wide Web Consortium (W3C), http://www.w3.org/2001/sw [Zugriff am 15.08.2011].

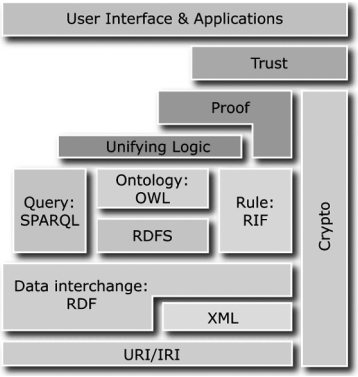

Abb. 1: „*Semantic Web Layer Cake*“ – Schichtenmodell der *Semantic Web*-Standards. (Bildquelle: W3C)

Die Komponenten URI / IRI[6], XML[7] und RDF[8] ermöglichen als „Paket“, Aussagen (Statements) über Ressourcen in standardisierter maschineninterpretierbarer Form zu repräsentieren. Als Ressource gilt dabei alles, was sich über einen URI referenzieren lässt.

Eine Aussage im Sinne des Datenmodells von RDF (*Resource Description Framework*) besteht aus einem Tripel der Form Subjekt-Prädikat-Objekt, wobei jedes Objekt einer Aussage wiederum Subjekt einer neuen Aussage sein kann und somit durch „Verkettung“ die Repräsentation komplexerer Aussagen ermöglicht. Diese einfache Form der Verknüpfbarkeit von RDF-Statements bedeutet eine extreme Flexibilität und ist Grundlage des später beschriebenen *Linked Data* Konzepts mit bisher nicht gekannten Möglichkeiten in Bezug auf Interoperabilität. Entscheidend für die eindeutige maschinelle Interpretierbarkeit von RDF-*Statements* ist die Tatsache, dass das Prädikat, welches die semantische Beziehung zwischen Subjekt und Objekt ausdrückt, ebenfalls in Form eines URI repräsentiert wird, i.d.R. qualifiziert durch ein XML-Namensraumpräfix.

Abb. 2 zeigt als Beispiel die mögliche Darstellung der Aussage „Die unter dem DOI xx.xxxx / nnnn verfügbaren Daten zum Thema Klimawandel wurden vom Geoforschungszentrum Potsdam publiziert. Die Daten umfassen den Zeitraum von 2007.“ als RDF-Statement in N3-Serialisierung und als gerichteter Graph[9].

6. Uniform Resource Identifier (RFC 3986) / Internationalized Resource Identifier (RFC 3987).
7. Extensible Markup Language, http://www.w3.org/XML/ [Zugriff am 15.08.2011].
8. Resource Description Framework, http://www.w3.org/RDF/ [Zugriff am 15.08.2011].
9. Die gezielte semantische Suche innerhalb dieses Netzwerks von RDF-*Statements* erfolgt dann über die Abfragesprache SPARQL, auf die hier nicht näher eingegangen wird, vgl. z. B. http://www.w3.org/TR/rdf-sparql-query/ [Zugriff am 15.08.2011].

```
@prefix gn:<http://www.geonames.org/ontology#> .
@prefix dct:<http://purl.org/dc/terms/> .
@prefix foaf:<http://xmlns.com/foaf/0.1/> .
@prefix gnd:<http://d-nb.info/gnd/> .

<http://dx.doi.org/xx.xxxx/nnnn> dct:publisher <http://d-nb.info/gnd/10347719-6> .
<http://dx.doi.org/xx.xxxx/nnnn> dct:temporal "2007^^xsd:date" .
<http://dx.doi.org/xx.xxxx/nnnn> dct:subject "Klimawandel" .
<http://d-nb.info/gnd/10347719-6> gnd:preferredName "Deutsches GeoForschungsZentrum
Potsdam".
<http://d-nb.info/gnd/10347719-6> foaf:homepage <http://www.gfz-potsdam.de/portal/> .
```

Abb. 2a: RDF-Statements in N3/Turtle-Notation

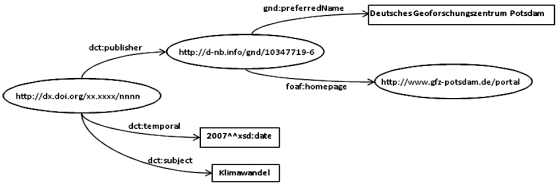

Abb. 2b: RDF-Statements als gerichteter Graph

Während RDF gewissermaßen als Trägerformat für Aussagen über konkrete Instanzen einer Domäne betrachtet werden kann (Datenebene), erlauben die Sprachstandards RDF-Schema (RDF-S) und *Web Ontology Language* (OWL)[10] als übergeordnete Schichten die Definition eigener domänenspezifischer Prädikate und darauf aufbauende Ontologien (vgl. Abb. 3)

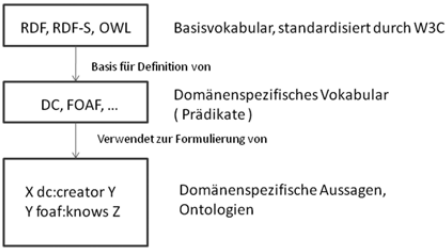

Abb. 3: Definition domänenspezifischer Vokabulare und Aussagen (Schematische Darstellung)

Abb. 4 zeigt in tabellarischer Form eine Auswahl der in den Sprachstandards RDF, RDF-S und OWL definierten Sprachkonstrukte auf deren Basis eigene

10. Web Ontology Language, http://www.w3.org/2004/OWL/ [Zugriff am 15.08.2011].

Vokabulare entwickelt werden können[11]. Speziell die in den Sprachstandards RDF-S und OWL definierten Sprachkonstrukte, wie z. B. rdfs:domain, rdfs:range, owl:sameAs, owl:TransitiveProperty, owl:SymmetricProperty, etc. erlauben die Repräsentation komplexer Domänenmodelle (Ontologien) bis hin zur Möglichkeit maschineller Inferenz auf Basis von Beschreibungslogik (*Description Logic*, DL)

RDF (25)	RDFS (16)	OWL (Auswahl)
rdf:RDF	rdfs:Class	owl:AllDifferent
rdf:about	rdfs:Resource	owl:Class
rdf:Description	rdfs:Datatype	owl:FunctionalProperty
rdf:Property	rdfs:domain	owl:InverseFunctionalProperty
rdf:type	rdfs:range	owl:Restriction
rdf:resource	rdfs:Container	owl:SymmetricProperty
rdf:Statement	rdfs:isDefinedBy	owl:Thing
rdf:subject	rdfs:equivalentClass	owl:TransitiveProperty
rdf:predicate	rdfs:subClassOf	owl:allValuesFrom
rdf:object	rdfs:subPropertyOf	owl:cardinality
rdf:value	rdfs:seeAlso	owl:complementOf
rdf:Bag	rdfs:ContainerMembershipProperty	owl:disjointWith
rdf:Seq	rdfs:member	owl:equivalentClass
rdf:List	rdfs:comment	owl:equivalentProperty
rdf:Alt	rdfs:label	owl:hasValue
rdf:XMLLiteral	rdfs:Literal	owl:intersectionOf
rdf:ID , rdf:nodeID		owl:inverseOf
rdf:parseType		owl:sameAs
rdf:datatype		owl:unionOf

Abb. 4: Sprachkonstrukte der Sprachstandards RDF, RDF-S und OWL

3.3.4 Grundlagen: Ontologien

Das Wort Ontologie stammt aus dem Griechischen und kann mit der „Lehre oder Beschreibung vom Sein“ übersetzt werden. Der ursprünglich nur in der Philosophie genutzte Begriff wird in den Informationswissenschaften und in der Informatik zur Beschreibung von allgemeinem und spezifischem Wissen in digitaler und formalisierter Form verwendet. Ontologien geben Ideen, Konzepten und Wissen eine Struktur, bestehend aus Entitäten, Eigenschaften und Instanzen. Dabei können Entitäten, also die Daten und Informationen über Gegenstände sowie Charakteristiken und Prozesse als Klassen modelliert werden. Eigenschaften von

11. Die vollständigen Spezifikationen der Sprachstandards finden sich beim W3C: zu RDF: http://www.w3.org/RDF/ [Zugriff am 15.08.2011] zu RDF-S: http://www.w3.org/TR/2004/REC-rdf-schema-20040210/ [Zugriff am 15.08.2011], zu OWL: http://www.w3.org/2004/OWL/ [Zugriff am 15.08.2011].

Klassen, wie Beziehungen zu anderen Klassen werden mittels Mengenrelationen und Vererbungsmechanismen abgebildet. Weitere Merkmale, die sich auf Klassen und / oder Instanzen beziehen, finden ihren Niederschlag in spezifischen Eigenschaftsstrukturen der Ontologie. Wissen, dass nicht abgeleitet werden kann muss über Axiome abgebildet werden. Die Ontologie und instanziierte Individuen können auch als Wissensbasis einer Domäne betrachtet werden, wobei eine keine klare Zuordnung der Individuen zur Ontologie oder zur Wissensbasis nicht immer möglich ist. Der Funktion nach lassen sich Ontologien in zwei große Gruppen einteilen. Zu den Ontologien zum Verstehen und Nachnutzen von Wissen gehören dabei allgemeine oder *upper-level* Modelle, wie die *General Formal Ontology* (GFO) und domänenspezifische oder *domain-level* Ontologien, wie *General Formal Ontology – Bio* (GFO-Bio) aus dem medizinischen Bereich. Zur zweiten Gruppe zählen Applikationsontologien, die das Zusammenführen anwendungsspezifischer *Use Cases* mit Domainwissen erlauben, wie beispielsweise im *Agricultural Ontology Service* (AOS) realisiert. Nach der Entscheidung über die Funktion der Ontologie, sollten vor dem eigentlichen Prozess der Modellierung *Use Cases* und Fragen, die durch das Modell beantwortet werden sollen sowie der *Scope* eruiert werden. In Abhängigkeit von der Erfahrung des Modellierers schließt sich nun eine Phase der Recherche nach wiederverwendbaren aus der gleichen Domäne stammenden Ontologien oder Teilen davon an.

Die ISDC-Ontologie repräsentiert einen Ausschnitt geowissenschaftlichen Wissens, welches sowohl geodätische und geophysikalische Daten im ISDC beschreibt als auch den Prozess zur Gewinnung von Forschungsdaten am GFZ Potsdam abbildet. Neben dem Repräsentationsteil besitzt die Ontologie auch Strukturen und Daten zur Realisierung von *Use Cases* zur Recherche und zum Zusammenführen von Wissen. *Use Cases* und Fragen die mit der ISDC-Ontologie beantwortet werden sollen, sind zum Beispiel:

a. Ich bin an Daten zum Klimawandel interessiert.
   Zeige alle Produkttypen, die auch durch die Schlüsselwörter vom Geophänomen „*Climate Change*“ beschrieben sind.
b. Ich arbeite im GGP-Projekt.
   Zeige alle Personen, die an Institutionen in den USA arbeiten und das GGP-Projekt betreuen.
c. Ich beschäftige mich mit globalen Änderungen des Schwerefelds der Erde.
   Zeige alle Produkttypen, Instrumente, Plattformen, Projekte und Institutionen, die mit den Schlüsselwörtern *Satellite*, *Earth Science* und *Gravity* beschrieben sind.

Der Designprozess beginnt mit dem Suchen und Finden geeigneter Begriffe, die später als Klassen modelliert werden können. Die Eigenschaften der Klassen werden dann über Terme oder Begrifflichkeiten abgebildet. Die Klassenbegriffe und die Eigenschaftsterme bilden das Grundgerüst für das formale Sprachvoka-

bular der Ontologie. Vererbungsrelationen zwischen Klassen werden in der Ontologie über eine Hierarchie abgebildet. Dabei entstehen Super- oder Subklassenkonstrukte, die neben der „is-a“ Relation auch transitive Eigenschaften besitzen können. Komplementär zum Sprachvokabular steht ein möglichst standardisiertes thematisches Vokabular für das Reflektieren von spezifischen Eigenschaften von Klassen und / oder Individuen der Ontologie. Dieses Vokabular kann zum Beispiel aus einem Glossar, einer Taxonomie oder einem Thesaurus stammen.

Die Hauptentitäten oder Klassen der ISDC-Domainontologie (siehe Abbildung 5*) sind: ISDC, Project, Platform, Instrument, Product Type und Institution. Zur Abbildung der Use Cases der ISDC-Ontologie, wie oben im Beispiel gezeigt, wurden zusätzlich die Klassen Geophenomenon, Personell, Country und Keyword modelliert. Die Klasse ISDC steht dabei als „abstrakte“ Superklasse, die den Prozess der Generierung von Forschungsdaten und damit in großen Teilen den Data Life Cycle begleitet. „Is-a“ Relationen können dabei folgendermaßen ausgedrückt werden: Platform, Instrument und Product Type gehören zum Prozess der Forschungsdatenerhebung. Klassische „is-a“ Beziehung in der ISDC-Ontologie sind zum Beispiel, ein Superconducting Gravimeter (SG) ist ein Instrument oder eine SG Platform ist eine Platform. In diesem Zusammenhang wirkt die transitive Relationseigenschaft „isPartOf“ und invers dazu „supplies“ als Klammer, die z. B. Instrument als Teil von Project beschreibt, ohne dass diese Eigenschaft dem Instrument explizit zugewiesen werden muss. Das thematische Vokabular zur inhaltlichen Beschreibung der Elemente der ISDC-Ontologie stammt aus dem kontrollierten Wortgut des Katalogs des GCMD der NASA.*

Um eine Ontologie und deren Persistenzform, zum Beispiel die zugehörige OWL-Datei universell im *Semantic Web* nutzen zu können, müssen alle Elemente der Ontologie in ihrem zugehörigen Namensraum per URI referenziert werden. Alle Terme des Sprachvokabulars der Ontologie sollten ebenfalls möglichst ausführlich kommentiert und über das *Web* zugreifbar sein. Die Instanzen der Wissensbasis lassen sich entweder direkt innerhalb der OWL-Datei abbilden oder bei Massendaten in einer geeigneten Datenbank, einem sog. *Triplestore* speichern.

Die ISDC-Ontologie und zugehörige Instanzen sind gemeinsam in einer OWL-Datei gespeichert. Diese OWL-Datei definiert einen eigenen Namensraum isdc (http://isdc.gfz-potsdam.de/ontology/isdc_1.0.owl#) und lässt sich, wie im Abschnitt *„Praxisbeispiel: GFZ ISDC – Semantic Web Projekt“ bereits erwähnt, über den Link: http://isdc.gfz-potsdam.de/ontology/isdc_1.0.owl aufrufen und damit global und universell weiterverwenden.*

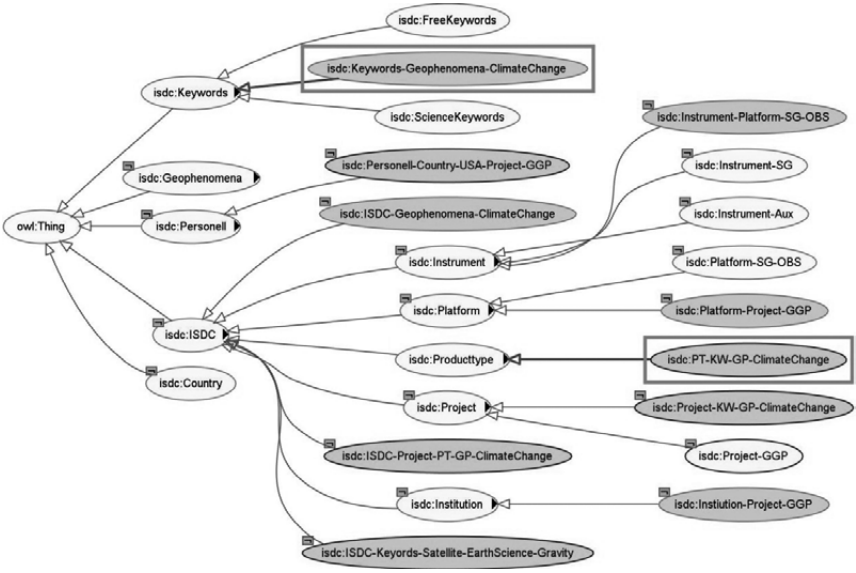

Abb. 5: ISDC-Ontologie (Ausschnitt)

Abb. 5 zeigt einen Ausschnitt aus der ISDC-Ontologie. Neben den Hauptklassen (Füllfarbe: hellgrau) und ihren Beziehungen werden insbesondere die für die Umsetzung der *Use Cases* der ISDC-*Ontology* definierten Klassen (Füllfarbe: dunkelgrau) dargestellt. Die für die Beantwortung der Fragestellung a) notwendigerweise erzeugten Klassen sind umrahmt. Individuen der Ontologie sind in der Grafik nicht dargestellt. Die Abbildung wurde mit dem Protégé-Plugin OWLViz erzeugt.

3.3.5 Linked Open Data[12]

Das Datenmodell von RDF in Kombination mit dem generischen URI-Konzept, welches online verfügbare Datenressourcen und abstrakte Beziehungskonzepte einer Domänenontologie gleichermaßen weltweit eindeutig referenzierbar macht, bildet die Grundlage des Konzepts „*Linked Open Data* (LOD)“, einem Paradebeispiel für Interoperabilität: Isolierte Datenbestände einer Fachdomäne („Datensilos“), die z. B. in relationalen Datenbanken vorliegen, werden durch die Transformation nach RDF „geöffnet“ und können danach mit ebenfalls in RDF vorliegenden Datenbeständen anderer Domänen in generischer Weise über RDF-Statements wechselseitig verknüpft und dadurch angereichert werden. Unterstellt

12. Zentrale Anlaufstelle für Einstieg und weiterführende Informationen zu *Linked Data* ist die Website http://linkeddata.org/ [Zugriff am 15.08.2011].

man für bestimmte Anwendungsbereiche die näherungsweise Gültigkeit einer „Gleichung“ der Form Wissen = Daten + Ontologie, so wird das Potential des *Linked Open Data*-Konzepts bei Anwendung auf Massendaten deutlich, als erster Schritt in Richtung eines Web der Daten anstelle eines Web der Dokumente. Abb. 6 zeigt die Visualisierung derart verknüpfter RDF-Datenbestände als sog. *Linked Data Cloud*[13]. Dabei repräsentiert jeder Kreis den RDF-Datenbestand einer bestimmten Domäne, beispielsweise könnte die Domäne der ISDC Datenprodukte inklusive der zugehörigen ISDC-Ontologie nach Abschluss des Projekts in die *Linked Data Cloud* integriert werden und dann dort als Kreis erscheinen.

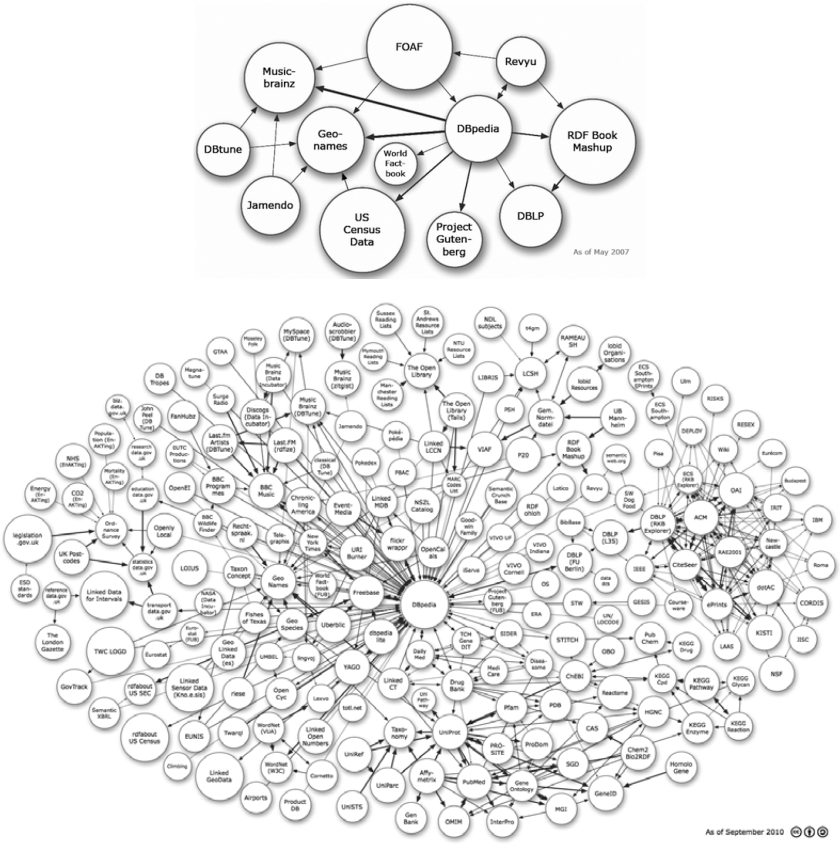

Abb. 6: Entwicklung der *Linked Data Cloud* von Mai 2007 (oben) bis September 2010 (unten). Bildquelle: http://richard.cyganiak.de/2007/10/lod/ [Zugriff am 15.08.2011].

13. Für Detailinformation zur *Linked Data Clouds*. http://richard.cyganiak.de/2007/10/lod/ [Zugriff am 15.08.2011].

Die Größe der Kreise symbolisiert den Umfang der in den jeweiligen Domänen enthaltenen RDF-Tripel und reicht von einigen Zehntausend bis zu Milliarden von Tripeln. Pfeile (uni- oder bi-direktional) stehen für eine semantische Verknüpfung der jeweiligen Domänen in Form von RDF-Statements, d.h. RDF-Tripeln deren Subjekt im Namensraum von Domäne A und deren Objekt im Namensraum von Domäne B liegt. Die Dicke der Pfeile symbolisiert dabei die Anzahl derartiger domänenverbindender Statements und reicht von mindestens fünfzig bis zu über hunderttausend. Das starke Wachstum der *Linked Data Cloud* im Vergleich von Mai 2007 (links) und September 2010 (rechts) ist ein Indikator für die zunehmende Bedeutung von *Linked Open Data* als wesentliche Komponente einer zukünftigen Webinfrastruktur im Bereich daten- und wissensintensiver Anwendungen. Als schon aktuell sehr stark in der *Linked Data Cloud* vertretene Domänen seien hier beispielhaft genannt der Bereich bibliographischer Informationen[14], der Bereich Lebenswissenschaften, inklusive Medizin und Bioinformatik[15], sowie – insbesondere in Großbritannien und den USA – der Bereich der öffentlichen Hand[16].

Auf Basis der bereits vorhandenen *Use Cases* kann kein Zweifel daran bestehen, dass das Konzept *Linked Open Data* für den Bereich von Forschungsdaten enormes Potential besitzt. Dies betrifft sowohl die Öffnung von „Datensilos" und damit den generischen Zugang zu Forschungsdaten, als auch im Sinne der Interdisziplinarität die Möglichkeit einer semantisch präzisen kontextbezogenen Anreicherung von Forschungdaten durch Informationen und Fakten aus anderen Wissenschaftsdomänen. Die im Wissenschaftsbereich bestehenden hohen Qualitätsanforderungen in Bezug u.a. auf Dauerhaftigkeit, Vertrauenswürdigkeit und semantische Präzision, stellen zum gegenwärtigen Zeitpunkt allerdings noch eine große Herausforderung dar, sprich es gibt noch eine Reihe offener Fragen, die im Rahmen des Aufbaus einer *Linked Data*-Infrastruktur für Forschungsdaten beantwortet werden müssen. Abb. 7 zeigt in grob schematischer Darstellung ein denkbares Szenario für eine solche Infrastruktur, wobei vorausgesetzt wird, dass jeder der vier dargestellten Bereiche in den Standards des *Semantic Web-Layer Cake* (vgl. Abb. 1) umgesetzt ist und damit die verbindenden Pfeile im wesentlichen als semantische Verknüpfungen auf Basis von RDF-Statements interpretiert werden können.

---

14. Einen guten Überblick zu Aktivitäten in diesem Bereich bietet die *Website* der *W3C Library Linked Data Incubator Group*: http://www.w3.org/2005/Incubator/lld/ [Zugriff am 15.08.2011].
15. s. z. B. http://www.w3.org/wiki/HCLSIG [Zugriff am 15.08.2011].
16. s. z. B. http://www.data.gov/ [Zugriff am 15.08.2011] (USA) und http://data.gov.uk/ [Zugriff am 15.08.2011] (Grossbritannien).

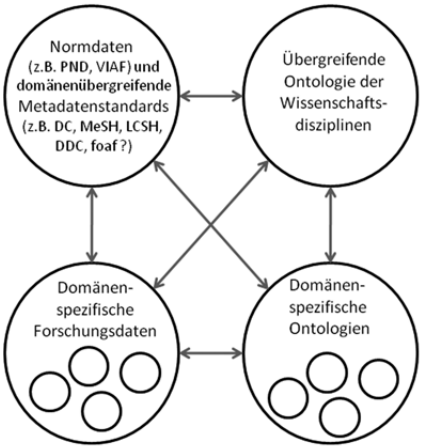

Abb. 7: Komponenten einer denkbaren *Linked Data*-Infrastruktur für Forschungsdaten (grob schematische Darstellung)

Die Komponente „Ontologie der Wissenschaftsdisziplinen“ (Abb. 7, rechts oben) in welcher die verschiedenen Wissenschaftsdisziplinen und Forschungsmethoden in ihrer wechselseitigen Abhängigkeit abgebildet sind, nach dem Beispiel des leider aktuell nicht weitergeführten DFG-Projekts OWD[17], könnte in der Rolle einer *Upper-Level Ontology* das Auffinden vielversprechender Verknüpfungspunkte von Forschungsdaten aus unterschiedlichen Domänen erleichtern und damit die interdisziplinäre Nutzung von Forschungsdaten befördern. Die konkreten semantischen Verknüpfungen zwischen den Datenbeständen würden dann selbstverständlich über die jeweiligen domänenspezifischen Ontologien wie z. B. die ISDC-Ontologie, erfolgen (Abb. 7, rechts unten). Für die Komponente Forschungs(roh)daten (Abb. 7, links unten) existiert in Form der Verbundinitiative *DataCite*[18] und der von der TIB Hannover betriebenen zugehörigen DOI-Registrierungsagentur bereits[19] seit längerem eine Infrastruktur für die weltweit eindeutige und dauerhafte Referenzierbarkeit (und Zitierbarkeit) von wissenschaftlichen Primärdaten. Potentiell problematisch im Sinne der *Linked Data* Philosophie ist dabei die Verwendung des proprietären *Identifiersystems* DOI (*Digital Object Identifier*), welches nicht auf HTTP-URIs basiert und daher den „Umweg“ des Resolverdienstes http://dx.doi.org erfordert, um als

17. http://owd.hu-berlin.de/ [Zugriff am 15.08.2011].
18. http://datacite.org/ [Zugriff am 15.08.2011].
19. http://www.tib-hannover.de/de/die-tib/doi-registrierungsagentur/ [Zugriff am 15.08.2011].

Ressource im Sinne von Subjekt oder Objekt eines RDF-Tripels referenzierbar zu sein. Dies ist prinzipiell unproblematisch, solange in einem *Linked Data* Kontext auf den durch den DOI referenzierten Gesamtdatensatz Bezug genommen wird. Problematisch wird es, wenn in bestimmten Kontexten eine feinere Granularität der Referenzierung (im hypothetischen Extremfall jeder einzelne Messpunkt) erforderlich wäre. Als vierte und letzte Komponente in der in Abb. 7 skizzierten LOD-Infrastruktur für Forschungsdaten stellen Normdaten und etablierte Metadatenstandards sowie die dahinter stehenden Institutionen einen wichtigen Eckpfeiler dar. In diesem Bereich haben sich in den letzten Jahren gravierende Entwicklungen im positiven Sinne[20] vollzogen, wie beispielsweise bei der Zusammenführung und Verfügbarmachung internationaler Normdaten im Rahmen der Initiative VIAF (*Virtual International Authority File*)[21], wodurch die weltweit eindeutige Referenzierbarkeit von Personen, Körperschaften und Tagungen ermöglicht wird, oder die *Semantic Web*-konforme Umsetzung von zahlreichen kontrollierten Standardvokabularen und Klassifikationssystemen wie z. B. MeSH (*Medical Subject Headings*), LCSH (*Library of Congress Subject Headings*), DDC (*Dewey Decimal Classification*) und anderen auf Basis des RDF-Vokabulars SKOS[22].

3.3.6 Vorgehensmodelle und *Tools*

Angesichts der Tatsache, dass aktuell der überwältigende Teil aller digitalen Datenbestände, inklusive Forschungsdaten, z. B. in relationalen Datenbanken, d.h. in *Semantic Web*-inkompatibler Form vorliegt, stellt sich die Frage nach welcher Systematik gegebenenfalls ein erster Schritt in Richtung der Transformation vorhandener Datenbestände in eine *Semantic Web* / *Linked Data*-kompatible Form bewerkstelligt werden kann. Es gibt hier sicher keine „*One-Size-Fits-All*"-Lösung, aber ein grobes pragmatisches Vorgehensmodell nach Art einer nullten Näherung soll hier in Anlehnung an die schematische Darstellung von Abb. 7 dennoch skizziert werden. Existiert noch keine in der jeweiligen Fachcommunity allgemein akzeptierte domänenspezifische Ontologie, so können wie am Beispiel der ISDC-Ontologie beschrieben, aus den typischen Nutzungsszenarien der Datenbestände (*Use Cases*) wichtige Eckpunkte einer Ontologie abgeleitet und modelliert werden. Liegen die Daten bereits in relationaler Form strukturiert vor, so kann prinzipiell das relationale Datenmodell nach RDF-S und OWL „übersetzt" werden, wobei in erster Näherung die Attribute des rela-

20. Einen guten Überblick zu Aktivitäten in diesem Bereich bietet die Website der *W3C Library Linked Data Incubator Group*: http://www.w3.org/2005/Incubator/lld/ [Zugriff am 15.08.2011].

21. http://www.oclc.org/research/activities/viaf/ [Zugriff am 15.08.2011].

22. http://www.w3.org/2004/02/skos/ [Zugriff am 15.08.2011].

tionalen Datenmodells als RDF-Vokabular (Prädikate) fungieren, verbunden mit der Festlegung eines eigenen Namensraums. Um die spätere Verknüpfbarkeit mit schon vorhandenen Datenbeständen in der *Linked Data Cloud* zu erleichtern, sollte in einem zweiten Schritt versucht werden, zumindest Teilaspekte der Ontologie (z. B. Personen, geographische Orte oder Publikation und deren Beziehungen untereinander) durch bereits etablierte Ontologien und deren Namensraumvokabular abzudecken, z. B. FOAF[23], *Geonames* oder *Dublin Core*). Falls für die Instanzen bestimmter Klassen des eigenen Datenbestandes (z. B. Personen, Körperschaften oder geographische Orte) Normdaten in Form von URIs existieren, sollten diese in jedem Fall anstelle von Literalen verwendet werden (z. B. <http://d-nb.info/gnd/10347719-6> als weltweit eindeutiger Identifier der Deutschen Nationalbibliothek für die Körperschaft „Deutsches GeoForschungsZentrum Potsdam“. Abb. 8 zeigt in Anlehnung an das fiktive Beispiel zu Abb. 2 exemplarisch die Verwendung etablierter Namensraumvokabulare für bestimmte Teilaspekte der eigenen Domäne.

```
@prefix gn:<http://www.geonames.org/ontology#> .
@prefix dct:<http://purl.org/dc/terms/> .
@prefix foaf:<http://xmlns.com/foaf/0.1/> .
@prefix gnd:<http://d-nb.info/gnd/> .

<http://dx.doi.org/xx.xxxx/nnnn> dct:publisher <http://d-nb.info/gnd/10347719-6> .
<http://dx.doi.org/xx.xxxx/nnnn> dct:temporal "2007^^xsd:date" .
<http://dx.doi.org/xx.xxxx/nnnn> dct:subject "Klimawandel" .
<http://d-nb.info/gnd/10347719-6> gnd:preferredName "Deutsches GeoForschungsZentrum
Potsdam".
<http://d-nb.info/gnd/10347719-6> foaf:homepage <http://www.gfz-potsdam.de/portal/> .
```

Abb. 8: Verwendung etablierter Namensraumvokabulare für Teilaspekte der eigenen Domäne

Existiert innerhalb der eigenen Domäne kontrolliertes Wortgut in Form eines Thesaurus oder Klassifikationssystems, so muss dieses in SKOS transformiert werden[24] um für die Beschreibung des Datenbestandes im RDF-Kontext nutzbar zu sein. Sind all diese Schritte vollzogen, so ist der Datenbestand prinzipiell „*Linked Data ready*“, kann „publiziert“ und semantische Bezüge zu anderen Datenbeständen hergestellt werden und damit potentiell als „vernetzter Kreis“ innerhalb der *Linked Data Cloud* in Abb. 6 erscheinen.

Bedingt durch die steigende Wahrnehmung des *Semantic Web*-Potentials und die daraus resultierende wachsende Zahl der Akteure im Feld, ist die Palette an

23. FOAF („*friend-of-a-friend*“) ist ein viel genutztes RDF-Vokabular zur Beschreibung von Personen.

24. Eine umfangreiche Sammlung entsprechender Tutorials findet sich beim W3C unter http://www.w3.org/2004/02/skos/references [Zugriff am 15.08.2011].

Werkzeugen zur Unterstützung der oben genannten Schritte mittlerweile sehr umfangreich, teilweise sehr professionell und erfreulicherweise zum großen Teil frei verfügbar und *OpenSource*[25]. Das Spektrum reicht von *Tools* zur automatischen RDF-Konvertierung von relationalen Datenbankbeständen[26], über sog. *Triple Stores* zur persistenten Speicherung von Milliarden von RDF-Tripeln und den performanten Zugriff darauf über sog. SPARQL-*Endpoints*[27], bis hin zu Modellierungs- und Visualisierungswerkzeugen unterschiedlicher Komplexität für den Bereich der Ontologieentwicklung. Daneben existieren spezialisierte Suchmaschinen[28] für das Auffinden von im Web bereits vorhandenen und möglicherweise direkt nachnutzbaren Ontologien, ebenso wie *Browser-Plugins*[29] für das Echtzeit-Browsen von RDF-Graphen. Werkzeuge beispielsweise zur softwareunterstützten Evaluation und Qualitätssicherung von Ontologien, *Tools* zur semiautomatischen Zusammenführung von Ontologien (*Ontology Merging*), semiautomatische Erstellung von Ontologien auf Basis automatischer Textanalyse domänenspezifischer Textkorpora oder der automatischen Erstellung semantischer Verknüpfungen in der *Linked Data Cloud* sind Gegenstand aktueller Forschung.

Nachfolgend soll der Bereich der Ontologieentwicklungswerkzeuge aufgrund seiner Relevanz für die Entwicklung der ISDC-Ontologie näher dargestellt werden. Hier reicht das Spektrum von einfachen Tools zur Modellierung von Ideen und Konzepten, über renommierte Ontologieentwicklungswerkzeuge, wie Protégé (http://protege.stanford.edu) oder Ontoverse (http://www.ontoverse.org), bis hin zu vollständigen, modular aufgebauten Software-Entwicklungsumgebungen, bestehend aus Modellier-, Programmier- und spezifischen *Semantic Web Framework* Bestandteilen. Einige dieser Werkzeuge werden im Folgenden näher vorgestellt. Die Software *Cmap Tools* von IHMC (http://cmap.ihmc.us) eignet sich gut zur grafischen Darstellung von Konzepten, die sich in Entitäten und zugehörige Relationen zerlegen lassen. Der zugehörige *Cmap Tools Ontology Editor* (CEO) (http://www.ihmc.us/groups/coe) erlaubt die komfortable Modellierung von Domainwissen in einem Strukturdiagramm mit benannten Graphen. Dabei werden die Entitäten als Knoten abgebildet. Diese können Klas-

---

25. Eine gute Übersicht befindet sich beim W3C unter: http://www.w3.org/2001/sw/wiki/Tools [Zugriff am 15.08.2011].
26. http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/ [Zugriff am 15.08.2011].
27. Einer der bekanntesten Vertreter ist der *Virtuoso Universal Server* der Firma *OpenLink*: http://virtuoso.openlinksw.com/ [Zugriff am 15.08.2011]. Die *OpenSource*-Version der Virtuoso-Serversoftware kann unter http://sourceforge.net/projects/virtuoso/ [Zugriff am 15.08.2011] heruntergeladen werden
28. z. B. die Suchmaschine Watson: http://watson.kmi.open.ac.uk/ [Zugriff am 15.08.2011].
29. z. B. OpenLink Data Explorer: http://ode.openlinksw.com/ [Zugriff am 15.08.2011].

sen, Individuen oder Literale repräsentieren und werden als verschiedenartige Rechtecke dargestellt. Die Relationen zwischen den Knoten werden als Kanten, also gerichtete Verbindungslinien gezeichnet. In Teilen können die Kanten mit OWL-Syntax, wie zum Beispiel der Abbildung hierarchischer Beziehungen, oder der Definition von *Domain*- und *Range*-Bereichen, versehen werden. Ein sehr mächtiges Werkzeug zur Modellierung von Ontologie und zum Design von *Reasoning*-Applikationen ist die *Open Source* Software Protégé (http://protege.stanford.edu). Das im *Standford Center for Biomedical Informatics Research* innerhalb verschiedener Projekte über zwei Jahrzehnte in JAVA ständig weiterentwickelte Werkzeug ist modular aufgebaut und unterstützt das Modellieren von Ontologien mit Frames, die kompatibel zum *Open Knowledge Base Connectivity protocol* (OKBC) sind und / oder in einer OWL-Umgebung. Die aktuellen Arbeiten haben zu einer Aufspaltung von Protégé in Versionen der Kategorien 3.x und 4.x geführt, die beide weiterentwickelt werden. Die weltweite Nutzung von Protégé und die offene plug-in Architektur sowie eine JAVA-basierte API haben zur Entwurf duzender zusätzlicher Module (http://protegewiki.stanford.edu/wiki/Protege_Plugin_Library) aus allen Bereichen des Ontologieentwurfs und der Applikationsentwicklung im *Semantic Web* geführt. Das Protégé OWL-*Plug-In* stellt alle zur Modellierung von Ontologie notwendigen Tabs, wie *Metadata*, *OWL Classes*, *Properties* und *Individuals* sowie zahlreiche die Modellierung unterstützende *Widgets* und *Wizards* in einer einheitlichen GUI bereit. Weitere integrierbare *third-party Tabs*, wie zum Beispiel OWLViz oder Jambalaya ermöglichen unterschiedliche Formen der Visualisierung des Ontologiemodells, die zur Analyse und Wissensrepräsentation, wie in Abbildung x dargestellt, genutzt werden können. Durch die Integration verschiedener *Reasoner*, wie zum Beispiel Pellet oder FaCT kann bereits während des Modellierungsprozesses die *Ontology* auf Konsistenz überprüft werden. Weiterhin erlauben die *Reasoner* „auf Knopfdruck“ eine (Neu)Klassifikation der Ontologie, das heißt das Ableiten oder Rückschließen eines neuen Subsumtionsbaumes, was zur Generierung nicht explizit ausgedrückter Zusammenhänge und damit letztlich zu neuem Wissen führen kann. Die Möglichkeit in Protégé SWRL- und SPARQL-Konstrukte einbinden zu können, führt zu einer Erweiterung der durch OWL gegebenen logischen Verknüpfungen und kann somit für komplexe Aufgaben des automatischen Schlussfolgerns genutzt werden. Für die Ausgabe und Speicherung der modellierten Ontologie stellt Protégé je nach Version verschiedene Ausgabeformate, wie zum Beispiel N-Triple, N3, Turtle, RDF, OWL, Manchester OWL-Syntax (Protégé Version 4) oder auch UML (UML *Backend Plug-In*), zur Verfügung. In einer Datenbank gespeicherte Individuen einer Wissensbasis können über die Protégé-API direkt mit der zugehörigen Ontologie verbunden werden.

Im Rahmen des ISDC Semantic Web Projekts wurde die Software CEO für das Erstellen und die Visualisierung erster Konzeptstudien der ISDC-Ontologie

verwendet. Desweiteren diente CEO der Weiterentwicklung des strukturellen Aufbaus der ISDC-Produkttypen und anderer ISDC-Entitäten. Die vollständige Modellierung der ISDC-Ontology sowie die beispielhafte Integration von Individuen erfolgte mit dem Programm Protégé in Version 3.4.4 unter Verwendung des OWL-Plugins. Zuerst wurden die ISDC-Entitäten *Project*, *Platform*, *Product Type* und *Institute*, als disjunkte Klassen modelliert und erste Individuen angelegt. Die Klassen Platform und Instrument verfügen über die Unterklasse Platform-SG-OBS respektive die Unterklassen Instrument-Aux und Instrument-SG. Diese weitere Klassifizierung und die Einführung weiterer disjunkter Klassen, wie Geophenomenon und *Keyword* mit seinen Unterklassen *FreeKeyword* und *ScienceKeyword* sowie *Personell* und *Country* dient der Erfüllung der zuvor definierten *Use Cases* und Fragestellungen. Die Eigenschaften der ISDC-Entitäten und Individuen wurden mit entsprechenden *Property*-Konstrukten für *Objects* und *Data Types* mit den zugehörigen Bereichen für *Domain* und *Range* realisiert. Die meisten *Object Properties* haben inverse, einige transitive Charakteristika. Die Bevölkerung der ISDC-Wissensbasis mit weiteren Individuen erlaubt nun die Ausführung der *Use Cases* mit Reasoningtechniken innerhalb von Protégé, auf die im Abschnitt Applikationen näher eingegangen wird.

Eine vollständige *Semantic Web* Entwicklungsumgebung benötigt neben den bereits vorgestellten Modellierungswerkzeugen und *Ontology-Reasonern* ein *Semantic Web Programming Framework* (SWPF) und Programmentwicklungskomponenten. Das SWPF beinhaltet die Elemente *Query Engine*, *Storage*, *Ontology Management* und eventuell *Rule Engine* und einen integrierten *Reasoner*, wie im Jena *Semantic Web Framework* realisiert. Modelliert man die Ontologie mit Protégé, so bietet sich für die Applikationsentwicklung in JAVA *Eclipse* als integrierte Entwicklungsumgebung an.

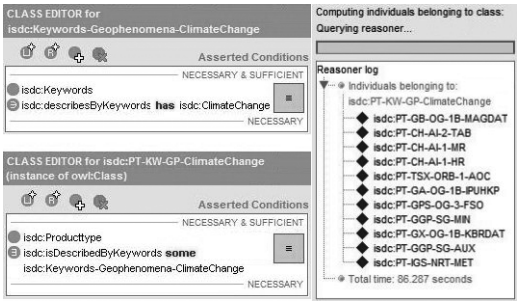

Abb. 9: Beispiel für automatisches Reasoning in der ISDC-Ontologie

Abb. 9 zeigt auf der linken Seite die zur Erfüllung des *Use Cases* a) kreierten Unter-Klassen isdc: *Keywords-Geophenomena-ClimateChange* (1) und isdc: PT-KW-GP-ClimateChange (2) der ISDC-Ontology. Rechts ist die Ergebnisse-

menge an *Product Type* Individuen abgebildet. Die Sub-Klasse (1) hat die Funktion, die für das Geophänomen *Climate Change* gültigen *Keywords* zu ermitteln. Sub-Klasse (2) verwendet die Ergebnissemenge aus (1), um damit alle Individuen der Klasse *Product Type* zu errechnen, die auch durch die entsprechenden *Keywords* beschrieben sind. Die Grafik wurde aus *Screenshots* der Protégé 3.4.4 GUI generiert.

3.3.7 Applikationen

Semantic Web Applikationen sind Anwendungen im WWW, die Sprachen, wie RDF, RDF-S oder OWL, Methoden, wie Klassifizierung und *Reasoning*, sowie Techniken, wie im *Semantic Web* Framework definiert, nutzen. Das betrifft sowohl die Abbildung des zugrunde liegenden Daten- und Informationsmodells als auch die Implementierung der Funktionalität. *Semantic Web* Applikationen greifen auf persistente, eindeutig referenzierte und formal strukturierte (verteilte) Wissensbasen zu. Neben den in vorangegangenen Abschnitten bereits erwähnten Anwendungen zur Repräsentation, Vernetzung und Recherche von Wissen, überdecken *Semantic Web* Applikationen auch die Bereiche Speicherung und Publikation von Daten. Für die grafische, zum Teil interaktive Repräsentation von Domainwissen und den zugrundeliegenden Organisationsstrukturen eigenen sich insbesondere Visualisierungen auf Basis von *Treemaps*, wie in Abbildung 11 unten-links dargestellt. Dabei werden verschieden große, unterschiedlich eingefärbte und auch verschachtelte Rechtecke, Kreise oder Polygone zur Darstellung der Entitäten und zugehöriger Eigenschaften benutzt. Der Vorteil gegenüber den häufig verwendeten sternförmig vernetzten Anordnungen von Entitäten wie Abbildung 11 oben-links zeigt, in 2D, oder auch 3D liegt in der besseren Raumausnutzung. Der Vorteil bei Anwendung von Entitäten-abbildenden Kreissegmenten (siehe Abbildung 11 unten rechts), wie im Visualisierungstool *eyePlorer* (http://eyeplorer.com) realisiert, liegt in einem besonders leicht verständlichen Zugang zu Entitäten und Strukturen.

Die ISDC-Ontologie bedient aktuell die Anwendungsfelder Repräsentation und Vernetzung von Daten und Informationen innerhalb der Domäne Geowissenschaften. Die Struktur und das Vokabular der Ontologie erlaubt darüber hinaus die Kopplung mit Wissensbasen anderer Fachdisziplinen, wie zum Beispiel Meteorologie oder Biologie. Das Modell ermöglicht die Beantwortung wissenschaftlicher Fragestellungen, wie beispielhaft in Abschnitt Ontologie definiert. Das geschieht durch Nutzung eines Reasoners mittels automatischer Schlussfolgerungs- und Inferenzmethoden. Die Recherche nach Individuen erfolgt mit standardisiertem und freiem *Keyword*-Vokabular innerhalb der ISDC-Wissensbasis. Die Referenz für die standardisierten *Keywords* ist das Vokabular der *Science Keywords* des GCMD der NASA. Das Ergebnis aus *Use Case* a) ist in Abb. 9 dargestellt. Mit Hilfe von zwei zusätzlich modellierten Klassen werden

nach dem Ontologie-Reasoning die gesuchten Produkttypen über die Funktion *Compute Individums* ermittelt. Eine mögliche Realisierungsform eines Tools für die grafische Repräsentation von ISDC-Wissensstrukturen in Kombination mit Recherchefunktionalität zeigt Abb. 10. In Anlehnung an das Konzept des eyePlorer wurden weitere Entitäten aus den Bereichen Wissenschaft und Gesellschaft aber auch der Raum- und Zeitbezug in das Konzept aufgenommen. Damit werden domainübergreifende Abfragen möglich, und die Applikation öffnet sich neuen Nutzerkreisen.

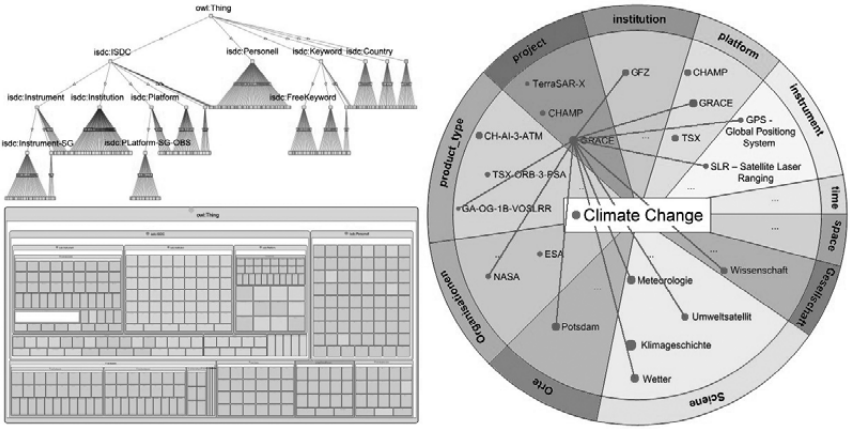

Abb. 10: Visualisierungsbeispiele der ISDC-Ontologie

Abb. 10 stellt links-oben in Bild die ISDC-Ontologie in einer Baumstruktur mit Klassen und Individuen dar. Links-unten ist die gleiche Ontologie als *Nested Tree-Map* gezeichnet. Beide Bilder wurden mit dem Jambalaya-*Plug-In* in Protégé generiert. Rechts im Bild sind ISDC- und weitere Entitäten als Kreissegmente in Anlehnung an die *eyePlorer*-GUI dargestellt. Der Begriff (Konzept) *Climate Change* wird hier im Ergebnis einer Recherche im *Semantic Web* mit zugehörigen Begriffen (Konzepten) aus ausgewählten ISDC-eigenen und fremden Entitäten, wie z. B. Gesellschaft, Wissenschaft oder auch Ort und Zeit dargestellt. Die Auswahl des Begriffs GRACE in der Entität *Project* führt zur Anzeige von mit GRACE in Beziehung stehenden Begriffen in weiteren Entitäten.

3.3.8 Fazit

Die generischen Konzepte der *Semantic Web*-Infrastruktur bieten ein enormes Potential im Bereich wissensintensiver Anwendungsdomänen, welches insbesondere im Bereich der interdisziplinären Nachnutzung wissenschaftlicher For-

schungsergebnisse bei weitem nicht ausgeschöpft ist. Aufgrund der spezifischen Qualitätsanforderungen in Bezug auf dauerhafte Verfügbarkeit von Ressourcen (insbesondere in Bezug auf und Stabilität und Eindeutigkeit von *Identifiern*), der Sicherstellung der Vertrauenswürdigkeit von Daten und der notwendigen Präzision von semantischen Verknüpfungen (RDF-Statements) sind zweifellos noch intensive Forschungsarbeiten und infrastrukturelle Anstrengungen notwendig. Als mehr oder weniger gelöst können Aspekte betrachtet werden, wie die Konversion von Massendaten in RDF-*Tripel* und deren persistente Speicherung in *TripleStores*. Prinzipiell lösbar als kooperative Anstrengung erscheint die Aufgabe der Etablierung vertrauenswürdiger globaler Clearingstellen für bestimmte URI-referenzierbare Normdaten nach dem Beispiel von VIAF oder *DataCite* als Fixpunkte innerhalb der *Linked Data Cloud*. Als große intellektuelle Herausforderung verbleibt aber die Modellierung der individuellen Fachdomänen und die Herstellung von semantischen Bezügen zu anderen Fachdisziplinen. Dies erfordert zum einen die Domänenkompetenz der Fachwissenschaftler und Kollaboration auf globaler Ebene[30] und zum anderen die domänenunabhängige Fachkompetenz im Bereich der *Semantic Web*-Technologien, sprich die adäquate Abbildung von Domänenwissen in RDF-S und OWL, die Kenntnis existierender nachnutzbarer Ontologien und RDF-Vokabulare deren Integration, bzw. Erstellung geeigneter Konkordanzen, sowie die semantisch präzise Verknüpfung mit anderen Wissensdomänen durch geeignete OWL-Sprachkonstrukte. Möglicherweise könnten diese Kompetenzen zum Qualifikationsprofil eines zukünftigen „*Data Librarian*“ oder „*Semantic Web Specialist*“ gehören, um in Zusammenarbeit mit Fachwissenschaftlern das volle Potential des *Semantic Web* im Wissenschaftsbereich auszuschöpfen.

30. Als vielversprechendes Beispiel für eine solche Anstrengung im global betrachtet recht heterogenen Bereich bibbliographischer Daten sei auf die W3C *Library Linked Data Incubator Group* verwiesen. http://www.w3.org/2005/Incubator/lld/ [Zugriff am 15.08.2011].

Literaturhinweise

Allemang, D. & Hendler, J. A., 2008. *Semantic web for the working ontologist*. Modeling in RDF, RDFS and OWL. Amsterdam: Morgan Kaufmann / Elsevier.

Antoniou, G. & Van Harmelen, F., 2008. *A semantic Web primer*. 2nd ed. Cambridge, Mass.: MIT Press (Cooperative information systems).

Berners-Lee, T. Hendler, J. & Lassila. O., 2001. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5), S. 34–43.

Daconta, M. C. Obrst, L. J. & Smith, K. T., 2003. *The Semantic Web: A guide to the future of XML, Web Services and Knowledge Management*. Indianapolis, Ind.: Willey. Online: http://www.wiley.com/legacy/compbooks/daconta/sw/ [Zugriff am 15.08.2011].

Hebeler, J. Fisher, M. Blace, R. & Perez-Lopez, A., 2009. *Semantic Web Programming*. Indianapolis, Ind.: Wiley. Online: http://semwebprogramming.org/ [Zugriff am 15.08.2011].

Herre, H. et al., o. J. General Formal Ontology (GFO): *A Foundational Ontology Integrating Objects and Processes*. Part I: Basic Principles. Research Group Ontologies in Medicine (Onto-Med), University of Leipzig. Online: http://www.onto-med.de/ontologies/gfo.owl [Zugriff am 16.08.2011], http://www.onto-med.de/ontologies/gfo-basic.owl [Zugriff am 16.08.2011].

Hitzler, P., 2008. *Semantic Web*. Grundlagen. 1. Aufl. Berlin: Springer (eXamen.press).

Hoehndorf, R. et al., 2008. GFO-Bio: A biological core ontology. *Applied Ontology*, 3(4), S. 219–227. Online: http://onto.eva.mpg.de/gfo-bio/gfo-bio.owl [Zugriff am 15.08.2011], http://onto.eva.mpg.de/gfo-bio/gfo-bio-meta.owl [Zugriff am 15.08.2011].

Mende, V. et al., 2008: Directory Interchange Format (DIF) Metadata and Handling at the German Research Center for Geosciences' Information System and Data Center. In: S. R. Brady, A. K. Sinha & L. C. Gundersen., eds. 2008. Proceedings. *Geoinformatics 2008—Data to Knowledge*. Potsdam, Deutschland, 11.–13. Juni 2008. (U.S. Geological Survey Scientific Investigations Report 2008–5172), S. 43–46. Online: http://pubs.usgs.gov/sir/2008/5172/sir2008-5172.pdf [Zugriff am 17.08.2011].

Noy, N. F. & McGuinness, D. L., 2001. *Ontology Development 101: A Guide to Creating Your First Ontology.* Online: http://protege.stanford.edu/

publications/ontology_development/ontology101.pdf [Zugriff am 15.08.2011].

Pfeifer, S., 2008. *Verknüpfung von geowissenschaftlichen ISDC-Produkten mit Semantic-Web-Techniken zur Erschließung inhaltlicher Zusammenhänge*. Bachelorarbeit. Hochschule Neubrandenburg, Studiengang Geoinformatik. Online: urn:nbn:de:gbv:519-thesis2008-0262-5, http://digibib.hs-nb.de/file/dbhsnb_derivate_0000000063/Bachelorarbeit-Pfeiffer-2008.pdf [Zugriff am 16.08.2011].

Pfeifer, S., 2010. *Entwicklung einer Ontologie für die wissensbasierte Erschließung des ISDC-Repository und die Visualisierung kontextrelevanter semantischer Zusammenhänge*. Masterarbeit. Hochschule Neubrandenburg, Studiengang Geoinformatik. Online: urn:nbn:de:gbv:519-thesis2010-0139-4, http://digibib.hs-nb.de/file/dbhsnb_derivate_0000000780/Masterarbeit-Pfeiffer-2010.pdf [Zugriff am 16.08.2011].

Pfeifer, S., 2010. ISDC?eyePlorer Konzept, M 107 – Anwenderprojekt, *Studiengang Geoinformatik*. Hochschule Neubrandenburg. Online: http://isdc.gfz-potsdam.de/index.php?name=UpDownload&req=getit&lid=558 [Zugriff am 17.08.2011].

Ritschel, B. et al., 2008. The German Research Center for Geosicences Information System and Data Center – Portal to Geoscientific Data, Information and Knowledge. In: S. R. Brady, A. K. Sinha, & L. C. Gundersen., eds. 2008. Proceedings. *Geoinformatics 2008—Data to Knowledge*. Potsdam, Deutschland, 11.–13. Juni 2008. (U.S. Geological Survey Scientific Investigations Report 2008–5172), S. 33–35. Online: http://pubs.usgs.gov/sir/2008/5172/sir2008-5172.pdf [Zugriff am 17.08.2011].

Ritschel, B. Gericke, L. Kopischke, R.& Mende, V., 2010. *The CHAMP/GRACE User Portal ISDC, System Earth via Geodetic-Geophysical Space Techniques Advanced Technologies in Earth Sciences*, Part 1, S. 15–28. Online: DOI: 10.1007/978-3-642-10228-8 2, http://www.springerlink.com/content/p1w5l77vwth387l1/fulltext.pdf [Zugriff am 15.08.2011].

Ritschel, B. Pfeifer, S. Mende, V. & Freiberg, S., 2008. Semantic Web Technologies for Value Added Services at the German Research Center for Geosciences' Information System and Data Center. In: S. R. Brady, A. K. Sinha, & L. C. Gundersen., eds. 2008. Proceedings. *Geoinformatics 2008—Data to Knowledge*. Potsdam, Deutschland, 11.–13. Juni 2008. (U.S. Geological Survey Scientific Investigations Report 2008–5172), S. 66–69. Online: http://pubs.usgs.gov/sir/2008/5172/sir2008-5172.pdf [Zugriff am 17.08.2011].

3.4 Archivierung von Forschungsdaten

Erich Weichselgartner, Armin Günther, Ina Dehnhard
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)

Forschungsprojekte enden in der Regel nicht, weil das Forschungsziel bereits komplett erreicht, die erhobenen Daten umfassend ausgewertet und die Befunde ausführlich kommuniziert worden sind. Vielmehr wird die Arbeit eingestellt, weil die Finanzierungen ausgelaufen, das (befristet beschäftigte) Personal abgewandert oder andere Forschungsthemen in den Vordergrund getreten sind. Selbst bei planmäßigem Ende empirischer Forschungsprojekte können die erhobenen Daten in der Regel nicht erschöpfend ausgewertet werden. Gründe hierfür sind begrenzte Ressourcen und der zielorientierte Charakter von Forschungsvorhaben. Datenanalysen außerhalb des engeren Zielfokus und abweichende Fragestellungen, evtl. sogar aus anderen Disziplinen, können im Sinne dieser Forschungsökonomie kaum Beachtung finden. Zudem bleibt solchen Re- oder Metaanalysen ein Riegel vorgeschoben, wenn die Daten von Forschern für andere unzugänglich aufbewahrt werden, wie es in den meisten Disziplinen nach wie vor gängige Praxis ist (vgl. PARSE.Insight, 2009, S. 33–34, mit Ergebnissen einer europaweiten Umfrage). Bei teuren oder einmaligen Datenerhebungen ist dies ein Ärgernis zum Schaden der gesamten Wissenschaftsgemeinschaft sowie der Gesellschaft als Ganzes. Dabei gibt es althergebrachte Gegenentwürfe. Zum Beispiel hat die Reanalyse astronomischer und geodätischer Forschungsdaten aus dem Jahre 1796 zur Entdeckung psychologisch-kognitionswissenschaftlicher Gesetzmäßigkeiten geführt, die wiederum für die Ergonomie große Bedeutung haben: Die persönliche Gleichung (Bessel, 1822) und das *Law of Prior Entry* (Titchener, 1908).

Während im 19. Jh. Forschungsdaten noch im Anhang von Publikationen abgedruckt wurden (z. B. Donders, 1868), gab man diese Praxis im 20. Jh. auf, nicht zuletzt aus Gründen der Praktikabilität bei großen Datenmengen. Alternativ haben sich Forscher verpflichtet, die Daten für eine bestimmte Frist aufzubewahren (die Deutsche Forschungsgemeinschaft empfiehlt 10 Jahre; DFG, 1998) und bei Bedarf interessierten Fachkolleginnen zugänglich zu machen (in der Psychologie z. B. gemäß den ethischen Richtlinien der Deutschen Gesellschaft für Psychologie (DGPs); DGPs, 2004). In der Praxis ist es mit diesen Grundsätzen aber nicht weit her, wie Wicherts, Borsboom, Kats und Molenaar (2006) feststellen mussten, die nur in jedem fünften Fall der Datennachfrage erfolgreich waren. Dieser Missstand hat zu einem kritischen Editorial im Wissenschaftsmagazin *Nature* (Nature, 2006) geführt und mit dazu beigetragen, über die Fachgrenzen hinweg Aufmerksamkeit auf die Archivierung und Zugänglichkeit von Forschungsdaten zu lenken. Dazu trug auch die digitale Revolution bei, durch die sich die Rahmenbedingungen änderten und die nachhaltige Sicherung und

Bereitstellung von Forschungsdaten weltweit zu einem wichtigen Thema für die wissenschaftliche Infrastruktur wurde.

Im Folgenden soll aufgezeigt werden, warum die Archivierung und Bereitstellung von Forschungsdaten wichtig ist, wo Probleme liegen, welche Selektionskriterien angemessen sind, welche Herausforderungen zu bewältigen sind und welche Trends sich abzeichnen.

3.4.1 Nutzen von Datenarchivierung und -bereitstellung

Intrinsische Gründe. Datenerhebungen sind mitunter sehr teuer und der geleistete Einsatz soll einen möglichst hohen wissenschaftlichen Gewinn bringen, d.h. die Daten sollen möglichst *umfassend ausgewertet* werden. Ein weiterer Grund, der gerade bei Untersuchungen an Mensch und Tier relevant wird, ist die *Vermeidung redundanter Datenerhebungen*, da diese immer auch eine Belastung sein können (vgl. Ethical Principles of Psychologists and Code of Conduct; APA, 1992). Darüber hinaus sind manche Daten einmalig und können *nicht repliziert* werden (z. B. sozial- und kulturwissenschaftliche Daten, die unmittelbar nach Beendigung der deutschen Teilung erhoben wurden). Weitere Gründe sind die Möglichkeiten der Analyse von Datensätzen unter *neuen Fragestellungen* und Perspektiven, der Reanalyse von Datensätzen mit *anderen Analysemethoden* (u.U. solche, die zum Zeitpunkt der Datenerhebung noch nicht bekannt waren), der *vergleichenden Analyse* verschiedener Datensätze zur Ermittlung der Robustheit von Ergebnissen oder zur Ermittlung der Generalisierbarkeit empirischer Gesetze, der *Ermittlung historischer Veränderungen* und der präziseren *Replikation von Untersuchungen*, die basierend auf den publizierten Forschungsberichten nur in Annäherung möglich wären.

Extrinsische Gründe. Für eine Datenarchivierung sprechen zudem die *Erfüllung der Grundsätze zur Sicherung guter wissenschaftlicher Praxis* und die Möglichkeit der *Überprüfung* publizierter Datenanalysen. Für Letzteres besteht gerade im Zusammenhang mit der Aufdeckung von Fällen wissenschaftlichen Fehlverhaltens durchaus ein Bedarf. Des Weiteren soll die Bereitstellung von Forschungsdaten der *Verstärkung des wissenschaftlichen Austauschs* dienen und die Rezeption der Forschung durch die Fachöffentlichkeit unterstützen. Nach einer Untersuchung von Piwowar und Chapman (2008) haben Zeitschriften mit einer expliziten Richtlinie zur Datenbereitstellung zudem einen höheren *Impact Factor.* Nicht zuletzt soll das wissenschaftliche Gratifikationssystem die Arbeit der Datenerhebung stärker (gerechter) würdigen und Forschungsdaten sollen deshalb *zitierbar* werden (siehe das DataCite-Projekt; DataCite, 2011).

3.4.2 Unterschiede zwischen Disziplinen

Grundsätzlich besteht über die Wissenschaftsdisziplinen hinweg weitgehend Einigkeit, dass Forschungsdaten erhalten und anderen zugänglich gemacht werden sollen. In der Praxis des Forschungsdatenmanagements aber, so haben vergleichende Untersuchungen etwa des *Research Information Network* (RIN, 2008) oder des *Digital Curation Centre* (Key Perspectives, 2010) gezeigt, bestehen große Unterschiede zwischen den Disziplinen oder innerhalb einer Disziplin zwischen unterschiedlichen Forschungsfeldern.

In der Astronomie existiert nicht nur eine hohe Bereitschaft der Forschenden, ihre Daten zu publizieren, sondern es besteht auch eine gut entwickelte Infrastruktur (*Open Access* Datenbanken bzw. Datenzentren). Ähnliches gilt für die Genforschung. Demgegenüber stellt sich die Situation in den Wirtschafts-, Sozial- und Verhaltenswissenschaften deutlich anders dar. Hier ist die Archivierung und Veröffentlichung von Forschungsdaten eher die Ausnahme, denn die Regel. Zwar gibt es auch in diesen Disziplinen große, häufig nationale Umfragestudien, bei denen die Daten gut dokumentiert, veröffentlicht und intensiv von anderen Forschenden genutzt werden (wie z. B. das Sozio-oekonomische Panel (SOEP, o. J.). Dem steht aber eine sehr viel größere Zahl von kleineren Projekten und Studien gegenüber, bei denen die Forschungsdaten lediglich als Basis für Publikationen der projektbeteiligten Forscher dienen, darüber hinaus aber nicht systematisch archiviert und öffentlich zugänglich gemacht werden.

3.4.3 Anreizstruktur

Wie lassen sich diese Unterschiede in der Praxis der Datenarchivierung und -veröffentlichung erklären und – wenn man die Gründe hierfür kennt – verringern? Von grundlegender Bedeutung dürfte die Anreizstruktur der jeweiligen Disziplin oder des jeweiligen Forschungsfeldes sein, d. h. die zu erwartenden „Kosten“ und „Nutzen“, die mit der Archivierung und Veröffentlichung von Forschungsdaten (in der Erwartung der Forschenden) verbunden sind. Zu den *Kosten* gehören:

- Zeit- und Geld-Aufwand für die Archivierung und Veröffentlichung von Forschungsdaten
- Nachteile im wissenschaftlichen Wettbewerb, indem andere die Forschungsdaten wissenschaftlich auswerten (vgl. Haeussler, 2010)
- reputationsschädigende Aufdeckung methodischer Unzulänglichkeiten der Datenauswertung
- Nachteile bei der ökonomischen Nutzung von Forschungsdaten
- rechtliche Probleme durch die Veröffentlichung von Daten (Verletzung des Datenschutzes oder von Eigentumsrechten z. B. bei geschützten Messverfahren)

- die Beschäftigung mit Dingen (wie der „Verwaltung“ von Daten), die dem Selbstverständnis der Forschenden nicht entsprechen

Den Kosten stehen *positive Anreize*, die eigenen Forschungsdaten zu archivieren und zu veröffentlichen, gegenüber. Diese müssen nicht einmal altruistischer Natur sein (Dienst am wissenschaftlichen Fortschritt), sondern können durchaus auch im Eigeninteresse der Forschenden liegen:

- Zugang zu Ressourcen (z. B. Fördermitteln)
- direkter wissenschaftlicher Reputationsgewinn durch die Veröffentlichung von qualitativ hochwertigen Forschungsdaten
- erhöhte Sichtbarkeit der eigenen (weiteren) Forschungsarbeit
- Anerkennung in der *scientific community* (Erfüllung sozialer Normen)
- Eröffnen neuer Kooperations- und Publikationsmöglichkeiten (z. B. Co-Autorenschaft mit Datennutzern)
- wissenschaftliche Anregungen durch die gemeinschaftliche Arbeit mit denselben Daten

Wie die Bilanz von Kosten und Nutzen ausfällt, hängt wesentlich von den institutionellen Strukturen einer Disziplin oder eines Forschungsfeldes ab, wobei auch regionale Unterschiede zum Tragen kommen. In dem Maße, in dem beispielsweise der Zugang zu Fördermitteln von der verpflichtenden Archivierung der in einem Förderprojekt erhobenen Forschungsdaten abhängig gemacht wird, verschiebt sich die Relation von den Kosten (Zeit- und Geldaufwand der Archivierung) hin zum Nutzen (Zugang zu Ressourcen/Fördermitteln). Beispielsweise verlangt das amerikanische nationale Gesundheitsinstitut (NIH) ab Fördersummen von 0,5 Mio. Dollar einen Plan zur Datenarchivierung und -weitergabe (NIH, 2003). Ebenso verschiebt sich die Kosten-Nutzen Relation in dem Maße, in dem Infrastruktureinrichtungen und Dienstleistungen zur Verfügung stehen, die Forschende bei der Archivierung von Daten unterstützen und damit Arbeitskosten verringern. Hiermit einher geht die Entwicklung von Normen und Standards, von Prozessen und *best practices*, von Werkzeugen und Maschinen zur Datenarchivierung, so dass Forschende, die „willig“ sind, auch auf Strukturen zurückgreifen können, die ihnen überhaupt erst eine sinnvolle Datenarchivierung ermöglichen.

Allerdings ist die Existenz einer solchen Infrastruktur alleine noch keine Garantie dafür, dass sie auch genutzt wird, wie RIN (2008, S. 14) für den Fall der *Social and Public Health Sciences* feststellt. Auch „weichere“ Faktoren spielen eine nicht unwesentliche Rolle. In vielen Disziplinen gibt es für die Veröffentlichung von Daten zwar keine in formalisierten Evaluierungsverfahren institutionalisierte Anerkennung als wissenschaftliche Leistung, die der Veröffentlichung eines Aufsatzes vergleichbar wäre. Aber in einigen Disziplinen, wie der Astronomie oder der Genforschung, gibt es offenbar eine klare Erwartung,

dass Forschungsdaten dennoch archiviert und veröffentlicht werden. Wer diese Erwartung nicht erfüllt, gefährdet sein Ansehen in der *community*. In anderen Disziplinen wie z. B. der Psychologie, gibt es einen solchen sozialen Druck nicht. Hier ist es üblich und normal, seine Daten nicht zu veröffentlichen.

Auch die Art der Daten, die in einer Disziplin oder einem Forschungsfeld vorwiegend produziert werden, spielt eine wichtige Rolle. Dort, wo die Deskription natürlicher oder sozialer Systeme und Prozesse im Vordergrund steht, sind die entsprechenden Forschungsdaten oftmals für die *community* eine unverzichtbare Arbeitsgrundlage und werden entsprechend auch nachgefragt. Dort, wo eher das Testen von Hypothesen oder Modellen im Vordergrund steht, wird die Archivierung und Veröffentlichung von Forschungsdaten vielfach als weniger wichtig angesehen. In der Klimaforschung sind beide Fälle anzutreffen: Während bei Beobachtungsdaten aus der Fernerkundung zumindest die Erwartung besteht, dass die Daten der *community* dauerhaft zur Verfügung gestellt werden, wird für Daten aus Modellrechnungen eine Archivierung nur begrenzt (auch in zeitlicher Hinsicht!) für sinnvoll angesehen (vgl. RIN, 2008, S. 64).

3.4.4 Selektion

Kaum eine Archiv- oder Forschungseinrichtung sieht sich in der Lage, alle in der jeweiligen Disziplin produzierten Daten aufzubewahren, so dass sich die Frage nach der Selektion stellt. Besonders in Wissenschaften wie der Astronomie, den Geowissenschaften oder der Hochenergiephysik, bei denen in machen Forschungsstationen oder Datenzentren täglich mehrere Terabyte Daten eingehen, macht der große Datenumfang Selektionskriterien erforderlich. Abgesehen von der Datenmenge liegen Beschränkungen für die Archivierung von Daten im Aufbereitungsaufwand begründet. In den meisten Wissenschaften ist eine ausführliche Beschreibung der zu archivierenden Daten unverzichtbar, damit die Daten für eine spätere Nachnutzung interpretierbar bleiben. Diese Datendokumentationen erfordern hohe personelle und z. T. auch technische Ressourcen, wenn die Apparaturen und Messgeräte der Datenerhebung ebenfalls archiviert werden müssen.

Leitendes Auswahlkriterium sollte die Bedeutsamkeit der Daten für potentielle Nachnutzer sein. Dementsprechend enthalten auch die Selektionsstrategien verschiedener Datenzentren und -archive immer den Aspekt des Nutzens für nachfolgende Analysen. Das *UK Data Archive* (2009, S. 2f.) stellt als grundlegendes Auswahlprinzip fest: „The UKDA collects data, information and other electronic resources of long-term interest and use across the range of social science and historical disciplines.“ Als Hinweis auf ein besonderes Analysepotential wird ein großer Stichprobenumfang genannt. Das *Earth Resources Observation and Science Center* (EROS, 2007) stellt als zusätzliches Prüfkriterium die Frage, welche Konsequenzen es hätte, wenn gerade diese Daten nicht archiviert

werden würden. Studien unter historisch einmaligen Rahmenbedingungen, wie z. B. Untersuchungen im Zusammenhang zur deutschen Einheit oder des Ausbruchs des Vulkans *Eyjafjallajökull*, sind aufgrund ihrer Einmaligkeit prädestiniert für die Archivierung. PsychData (o.J.), ein Forschungsdatenzentrum für die Psychologie, definiert ebenfalls Kriterien in Bezug auf die Bedeutsamkeit von Daten im Forschungsfeld der Psychologie: So wird psychologischen Längsschnittstudien und umfangreichen Querschnittstudien aufgrund ihres Datenumfangs und ihrer Repräsentativität ein besonders hohes Analysepotential zugesprochen (vgl. Weichselgartner, 2008). Auch wenn die Bedeutsamkeit von Daten ein unstrittiges Auswahlkriterium zu sein scheint, so bereitet seine Einschätzung doch Schwierigkeit, insbesondere wenn man den zukünftigen wissenschaftlichen Wert von Forschungsdaten antizipieren möchte. Wie kann heute beurteilt werden, was morgen (in einigen Jahren, Jahrzehnten oder sogar darüber hinaus) prominente Forschungsthemen sein werden? Deswegen verwenden Datenarchive zusätzliche, konkreter definierte Mechanismen für die Datenauswahl. Da die meisten Archivangebote disziplinspezifisch und teilweise innerhalb der Disziplinen noch themenspezifisch ausgerichtet sind, ergibt sich eine gewisse Vorauswahl der Daten bereits hier. Die Erweiterung und Passung eines neuen Datensatzes zu bereits vorhandenen Datenkollektionen ist ebenso von Belang wie die Abdeckung eines Themengebiets, zu dem bisher noch keine archivierten Daten vorhanden sind (siehe UK Data Archive, 2009, S. 3 f. und EROS, 2007). Weitere Selektionskriterien für die Aufnahme von Datensätzen können geografische (z. B. Deutschland, Europa) oder zeitlich-historische (z. B. Periode Weimarer Republik) Beschränkungen bei der Auswahl von archivierungswürdigen Datensätzen sein.

Wesentliches Auswahlkriterium stellt aber disziplinübergreifend die Qualität und Beschaffenheit der zu archivierenden Daten dar. Zum einen wird anhand der Art und der Formate der Daten entschieden, ob entsprechende Techniken und *Know-How* für ihre Aufbereitung und Archivierung vorhanden sind. Zum anderen stellt die Dokumentation der Daten eine wesentliche Voraussetzung für ihre Aufbereitung und Bereitstellung dar. Eine unzureichende und lückenhafte Datendokumentation bedeutet immer, dass Daten für die Archivierung abgelehnt werden, da die Daten nicht mehr interpretierbar und analysierbar sind. Weitere Qualitätskriterien stellen die Authentizität, Integrität und Konsistenz der Daten dar, welche vor Übernahme in ein Archiv überprüft werden.

Ebenfalls zu berücksichtigen bei der Auswahl von Datensätzen sind rechtliche Aspekte. Fragen des Datenschutzes spielen besonders in den Sozial- und Wirtschaftswissenschaften eine wesentliche Rolle bei der Datenselektion (vgl. APA, 1992). Durch Mechanismen wie Anonymisierungsverfahren oder beschränktem Datenzugriff kann in vielen Fällen eine Archivierung und Bereitstellung zwar realisiert werden, jedoch teilweise nicht ohne Informationsverluste. Mögliche

rechtliche Ansprüche auf die Daten müssen eindeutig geklärt sein, bevor die Übergabe von Daten an Datenzentren möglich ist.

3.4.5 Herausforderungen

Datendokumentation. Bei der Langzeiterhaltung von Forschungsdaten stellt eine auf langfristige Interpretierbarkeit ausgerichtete Dokumentation der Daten eine wesentliche Notwendigkeit dar. Forschungsdaten selbst sind ohne eine zugehörige Dokumentation, welche die Erhebungsbedingungen, den Kontext und die Forschungsinstrumente beschreibt, nicht interpretierbar. Je nach Forschungskontext und untersuchtem Objekt sind die Anforderungen an den Umfang und die Detailliertheit der Dokumentation unterschiedlich. Besonders in solchen Disziplinen, in denen es sich um einmalige, nicht wiederholbare Ereignisse handelt, die mit speziellen Messverfahren erfasst werden, ist eine genaue Beschreibung des Forschungskontextes notwendig. So muss in der Ökologie für eine adäquate Interpretation der Daten vielfach der Erhebungskontext detailliert beschrieben werden (vgl. Zimmerman, 2008), während in der Chemie eine stärker standardisierte Beschreibung möglich ist.

In vielen Disziplinen ist die Datendokumentation nicht in den Forschungs- und Publikationsablauf integriert. Datendokumentationen bedeuten einen hohen zusätzlichen Zeit- und Arbeitsaufwand, der meist nicht entsprechend „entlohnt" wird. Sowohl fehlende wissenschaftliche Honorierungen (*credits*, *incentives*) als auch die unzureichende Einplanung der Datendokumentation in der Budgetplanung eines Forschungsvorhabens führen zu einer Vernachlässigung des Datenmanagements im Forschungsalltag. Der Beginn neuer Forschungsprojekte erscheint attraktiver als eine Aufbereitung alter Forschungsprojekte für die Archivierung und Nachnutzung (Esanu et al., 2004). Hinzu kommen in einigen Disziplinen das Fehlen einheitlicher allgemein anerkannter Dokumentationsstandards sowie die fehlende Vermittlung von Standards und Praktiken des Forschungsdatenmanagements in der wissenschaftlichen Ausbildung. Die Implementierung des Dokumentationsprozesses in den eigentlichen Forschungsprozess stellt also eine besondere Herausforderung dar. Als optimal hat sich eine forschungsbegleitende Dokumentation von Beginn des Forschungsvorhabens an herausgestellt, da alle notwendigen Informationen noch direkt verfügbar sind. *Tools*, die die Datendokumentation bereits während des Forschungsprozesses unterstützen, kommt hier eine wichtige Funktion zu (vgl. z. B. Dehnhard und Weiland (2011) für das Dokumentationstool des Datenzentrums PsychData).

Entwicklung von Standards. Doch auch eine für sich genommen „gute" Datendokumentation, die die nach gegenwärtigem Wissenstand wesentlichen Metadaten zum Verständnis der Forschungsdaten enthält, ist möglicherweise nur von begrenztem Nutzen, wenn diese Dokumentation nicht in einer standardisierten Form erfolgt. Bei dem rapide wachsenden Volumen an archivierten

und veröffentlichten Forschungsdaten wird eine Standardisierung der Datendokumentation zunehmend wichtiger, damit diese Daten zunächst archiviert und dann gefunden, mit Daten aus anderen Forschungsprojekten verknüpft und wiederverwendet werden können. Es sind daher in verschiedenen Disziplinen Initiativen entstanden mit dem Ziel, für bestimmte Forschungsgebiete Datenstandards zu entwickeln, die festlegen, welche Informationen mindestens bei einer Datendokumentation in diesem Forschungsbereich enthalten sein sollen („*minimum reporting standards*“ oder „*minimum information specification*“) und wie diese Informationen dokumentiert werden sollen (vgl. z. B. für den Bereich der Biologie OMICS, 2006).

Entwicklung von Werkzeugen. Standardisierung ist ein Aspekt einer Herausforderung, die ebenfalls mit dem wachsenden Volumen und der wachsenden Komplexität von Daten verbunden ist: die Entwicklung von *Tools* und Verfahren, genau die Daten zu finden, die in einem Forschungskontext gebraucht werden können. Bereits bei einer einzelnen psychologischen oder sozialwissenschaftlichen Studie beispielsweise können Daten in erheblichem Umfang und Komplexität (Anzahl und Struktur der erhobenen Variablen, abgeleitete Daten usw.) anfallen, die es einem projektfremden Nutzer schwer machen, diese Daten zu entdecken, abzurufen und zu verwenden. Riesige Mengen an Daten anzuhäufen macht nur dann Sinn, wenn es Hilfsmittel gibt, die in diesen Datenmengen enthaltenen Schätze auch zu heben – ansonsten entstehen lediglich gigantische Datenfriedhöfe.

3.4.6 Ausblick

Ein wichtiger Trend, zumindest in der Theorie des Datenmanagements, ist die Entwicklung weg von einer am Ende des Forschungsprozesses stehenden oder gar nachträglichen Archivierung von Forschungsdaten hin zu einem in den Forschungsprozess integrierten *data-lifecycle* Management. Die Vision hierbei ist, dass bereits während der Konzeptions- und Durchführungsphase eines Forschungsprojektes die Dokumentation, Archivierung und Veröffentlichung von Forschungsdaten vorbereitet wird, indem beispielsweise die für die Dokumentation erforderlichen Metadaten (manuell, semi- oder vollautomatisch) erfasst und standardkonform aufbereitet werden. Eine eigene Archivierungsphase kann sich dann weitgehend auf die Übergabe der Daten und Datendokumentation an ein Forschungsdatenzentrum beschränken.

Es hat sich immer wieder gezeigt, dass die Auseinandersetzung mit den Erfordernissen der Datenarchivierung und -veröffentlichung nicht erst am Ende eines Forschungsprozesses erfolgen sollte. Nicht selten lassen sich wichtige Aspekte der Datenerhebung nicht mehr rekonstruieren, teilnehmende Forscher sind nicht mehr greifbar oder es wurde auch nur vergessen, sich die für die Veröffentli-

chung der Daten erforderlichen Rechte zu besorgen (z. B. bei Untersuchungen mit Menschen das Einverständnis zur Veröffentlichung der Forschungsdaten).

Die sich abzeichnende zunehmende Verfügbarkeit von Forschungsdaten wird Auswirkungen auf das wissenschaftliche Arbeiten quer über alle Disziplinen haben. Rechenbetonte Methoden (*computational methods*) könnten gleichberechtigt neben die klassische (Hypothesen geleitete) Forschung treten, denn mit Hilfe der enormen verfügbaren Datenmengen können beliebig viele Hypothesen „durchprobiert" werden. Der Begriff „*Data-Driven Science*" wurde als viertes wissenschaftliches Paradigma von Jim Gray 2007 (Hey, Tansley & Tolle 2009) ins Spiel gebracht (die anderen drei sind „Empirie", „Theorie" und „Simulation"). Die großen Herausforderungen des 21. Jh. wie Energie- und Umweltproblematiken sind möglicherweise nur auf diese Weise bewältigbar.

Literaturhinweise

APA (American Psychological Association), 1992. Ethical principles of psychologists and code of conduct. *American Psychologist*, 47(12), S. 1597–1611.

Bessel, F. W., 1823. *Astronomische Beobachtungen auf der Königlichen Universitäts-Sternwarte in Königsberg. 8. Abtheilung vom 1. Januar bis 31. December 1822*. Königsberg, III–VIII.

DataCite, 2011. *Welcome to DataCite*. Online: http://datacite.org [Zugriff am 17.02.2011].

Dehnhard, I. & Weiland, P., 2011. Toolbasierte Datendokumentation in der Psychologie. In: J. Griesbaum, T. Mandl & C. Womser-Hacker, Hrsg. Information und Wissen: global, sozial und frei? *Proceedings des 12. Internationalen Symposiums für Informationswissenschaft.* Boizenburg: Hülsbusch, S. 74–84.

DGPs (Deutsche Gesellschaft für Psychologie), 2004. *Revision der auf die Forschung bezogenen ethischen Richtlinien.* Online: http://www.dgps.de/dgps/aufgaben/ethikrl2004.pdf [Zugriff am 17.02.201].

DFG (Deutsche Forschungsgemeinschaft), 1998. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*. Weinheim: Wiley-VCH.

Donders, F. C., 1868. Die Schnelligkeit psychischer Processe. *Archiv für Anatomie, Physiologie und wissenschaftliche Medizin.* Berlin, S. 657–681.

EROS (Earth Resources Observation and Science), 2007. *Records appraisal tool.* Online: http://eros.usgs.gov/government/ratool/ [Zugriff am 17.02.2011].

Esanua, J. Davidson, J. Ross, S. & Anderson, W., 2004. Selection, appraisal, and retention of digital scientific data: Highlights of an EPRANET/CODATA Workshop. *Data Science Journal.* Online: http://www.jstage.jst.go.jp/article/dsj/3/0/227/_pdf [Zugriff am 17.02.2011].

Haeussler, C., 2010. *Information-sharing in academia and the industry: A comparative study.* (RatSWD Working Paper No. 154) Online: http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_154.pdf [Zugriff am 17.02.2011].

Hey, T. Tansley, T. & Tolle, K., 2009. Jim Gray on eScience: A Transformed Scientific Method. Based on the transcript of a talk given by Jim Gray to the NRC-CSTB in Mountain View, CA, on January 11, 2007. In: A. Hey, St.

Tansley & K.M. Tolle, 2009. *The Fourth Paradigm Data-Intensive Scientific Discovery.* Redmond, Wash.: Microsoft Research, S. xvii-xxxi

Key Perspectives, 2010. *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study.* Edinburgh: Digital Curation Centre. Online: http://www.dcc.ac.uk/scarp [Zugriff am 17.02.2011].

Nature, 2006. A fair share. *Nature,* 444 (7 December 2006), S. 653–654.

NIH (National Institutes of Health), 2003. *NIH data sharing policy and implementation guidance*. (Stand: 5.3.2003) Online: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm [Zugriff am 17.02.2011].

OMICS, 2006. *OMICS A Journal of Integrative Biology,* Vol. 10 (Special Issue on Data Standards).

PARSE.Insight, 2009. *Survey report.* (D3.4) Online: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf [Zugriff am 17.02.2011].

Piwowar, H.A. & Chapman, W.W., 2008. A review of journal policies for sharing research data. In: *Proceedings of the 12th International Conference on Electronic Publishing.* Toronto, Kanada 25.-27. Juni 2008. Online: http://elpub.scix.net/data/works/att/001_elpub2008.content.pdf [Zugriff am 17.02.2011].

PsychData, o.J. *PsychData – Startseite*. Forschungsdaten für die Psychologie. Online: http://www.psychdata.de/ [Zugriff am 17.02.2011].

RIN (Research Information Network), 2008. *To share or not to share: Publication and quality assurance of research data outputs. Annex: detailed findings for the eight research areas.* Research Information Network. Online: http://www.rin.ac.uk/system/files/attachments/To-share-data-outputs-annex.pdf [Zugriff am 17.02.2011].

SOEP (Sozio-oekonomisches Panel), o.J. DIW Berlin. *Startseite SOEP.* Online: http://www.diw.de/soep [Zugriff am 17.02.2011].

Titchener, E. B., 1908. *Lectures on the elementary psychology of feeling and attention*. New York: Macmillan.

UK Data Archive, 2009. *UK Data Archive collections development policy.* Online: http://www.esds.ac.uk/news/publications/UKDACollectionsDevPolicy.pdf [Zugriff am 17.02.2011].

Weichselgartner, E., 2008. Fünf Jahre Primärdatenarchivierung in der Psychologie: Ein Erfahrungsbericht. DGI (Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis), *30. DGI Online-Tagung 2008*. Frankfurt am Main, Deutschland 15.-17. Okt. 2008. Online: http://www.weichselgartner.de/edu/downloads/dgi2008.pdf [Zugriff am 17.02.2011].

Wicherts, J. M. Borsboom, D. Kats, J. & Molenaar, D., 2006. The poor availability of psychological research data for reanalysis. *American Psychologist*, (61), S. 726–728.

Zimmerman, A. S., 2008. New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science Technology Human Values*, (33), S. 631-652.

3.5 Informationswissenschaftler im Forschungsdatenmanagement

Stephan Büttner [1], Stefanie Rümpel [2], Hans-Christoph Hobohm [1]
[1] Fachhochschule Potsdam
[2] Fachhochschule Düsseldorf

Wissenschaftspolitisch gibt es seit Ende des 20. Jahrhunderts verschiedene Forderungen auf nationaler und internationaler Ebene bez. des verantwortungsvollen Umgangs mit Forschungsdaten (Rümpel, 2010, S. 27). Spätestens mit der Forderung der DFG von 1998 in der Denkschrift „Sicherung guter wissenschaftlicher Praxis“ (DFG, 1998, S. 12) wurde in der Empfehlung 7 angeraten: „Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden.“ Es benötigte jedoch noch viele weitere Jahre, bis diese Empfehlung auch bei der Informationsinfrastruktur ankam. Bei der Informationsinfrastruktur wird traditionell in Deutschland stark zwischen bibliothekarischen, dokumentarischen und archivarischen Tätigkeiten unterschieden, obwohl es allen um Erschließung, Bewertung, Erhaltung, Nutzbarmachung und Vermittlung von Wissen bzw. Bildungs- und Kulturgut auf analogen und digitalen Medien geht. So hat in Deutschland eine integrative Ausbildung (Archivare, Bibliothekare, Dokumentare) immer noch Modellcharakter (Potsdamer Modell an der FH Potsdam).

3.5.1 Rollen für Informationswissenschaftler im Forschungsdatenmanagement

Corrall beschrieb bereits 2008 sogenannte „*Hybrid Information Workers*“ und verdeutlichte, dass drei Arten von Spezialisten zur Umsetzung des Forschungsdatenmanagements benötigt werden: „*Content Specialists*“, „*Conduit Specialists*“ und „*Context Specialists*“ (Corall, 2008, S. 6). Pampel, Bertelmann und Hobohm (2008, S. 166) entwickelten diesen Ansatz weiter und betonten, dass für den Contentbereich Bibliotheks- und Informationsspezialisten, insbesondere für die kooperativen Aufgaben, benötigt werden (s. Abb. 1).

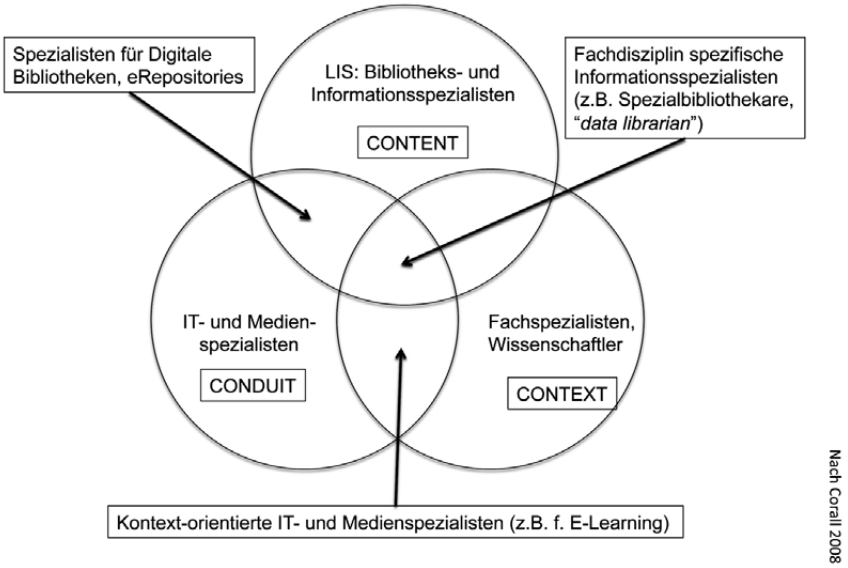

Abb. 1: Rollen für Informationspezialisten in einer hybriden Umgebung (Grafik: Pampel, Bertelmann & Hobohm, 2010, S. 166 nach Corrall, 2008, S. 6)

Seit den ersten nationalen, strategischen Papieren zu Datenmanagement-Fragen in der Wissenschaft in den USA (NSF, 2005), Großbritannien (Lyon, 2007; Swan & Brown, 2008), der EU (Atkinson et al. 2008; OGF22-ET CG, 2008) und vor allem seit den *Guidelines* der OECD (2007) wird international intensiv über die Erfordernisse der Ausbildung von „Datenspezialisten" diskutiert. Grundlegende Überlegungen, die einen gewissen internationalen Konsens gefunden haben, wurden im Auftrag von JISC von Swan und Brown (2008, S. 1) entwickelt (vgl. Hank & Davidson, 2009). Sie schlagen vier Rollen im Forschungsdatenmanagement vor. Neben dem *Data Creator* (datenproduzierender Forscher), dem *Data Scientist* (Wissenschaftler der bspw. bei der Datenanalyse unterstützt) und dem *Data Manager* (verantwortlich für alle technischen Aspekte: Lagerung, Zugriff, Aufbewahrung) wurde auch der *Data Librarian* benannt.

Im Weiteren wird, auch unter Beachtung der Tätigkeiten die im *Data Curation Lifecycle Model* (s. Beitrag von Rümpel. Kap. 1.2) aufgeführt sind, neben dem *Data Manager* und dem *Data Curator* insbesondere auf den *Data Librarian* eingegangen.

Das Rollen-Modell wurde speziell in Großbritannien weiterentwickelt auf einer Reihe von *ResearchData Management Workshops* in Kooperation mit dem *Digital Curation Centre* (DCC). Pryor und Donnelly (2009) schlagen als

Quintessenz einer dieser internationalen Arbeitsgruppen eine Erweiterung und Erläuterung der vier Rollen des JISC Reports (Swan & Brown, 2008) vor, zu dem bereits 2009 auf dem Deutschen Bibliothekartag (vgl. Pampel, Bertelmann & Hobohm, 2010, S. 168) eine Übertragung vorgeschlagen wurde, aus der sich Aufgaben und notwendige Kompetenzen für verschiedene Berufsprofile ergeben:

Data Manager **(Steuerung)** - Juristischer Sachverstand - Nutzungsbedingungen - Notfallplanung / *Risk + Desaster Management* - Sicherheit und Authentifizierung - Prozess-Monitoring (zus. mit *Data Creator*) - Metadaten (zus. mit *Data Creator*) - Bestandserhaltung (zus. mit *Data Librarian*) - Wert von Daten / Wirtschaftsaspekte (zus. mit *Data Librarian*)	***Data Creator*** **(Bearbeitung)** - Dokumentation + Kontext - Prozess-Monitoring (zus. mit *Data Manager*) - Metadaten (zus. mit *Data Manager*) - Datenmodellierung (zus. mit *Data Scientist*)
Data Librarian **(Unterstützung)** - Verhandlungsgeschick - Beschwerdemanagement und Kundenerwartungen - Koordination der Praktiken (Verfahrensregelung) - Bewertung und Bestandsaufbau - Promotion / Marketing / Öffentlichkeitsarbeit - Entwicklung von Standards (zus. mit *Data Scientist*) - Bestandserhaltung (zus. mit *Data Manager*) - Wert von Daten / Wirtschaftsaspekte (zus. mit *Data Manager*)	***Data Scientist*** **(Analyse)** - Informationsmanagement/Wissensmanagement - Datenanalyse / Datenverarbeitung - *Merging* und *Mash-ups* / Integration - Informationsextraktion (aus Datenmodellen und Know How von Personen) - Data Modelling (zus. mit *Data Creator*) - Entwicklung von Standards (zus. mit *Data Librarian*)

Abb. 2: Rollen im Datenmanagement (erweitert in Anlehnung an Donnelly, 2008)

Die Felder in der Tabelle sind als Flächen zu lesen, in denen die Überschriften den Bereich idealtypischer Tätigkeiten umschreiben und die konkreten Kompetenz- bzw. Aufgabenbereiche variabel verorten. Im Zentrum der vier Felder – und damit allen vier Rollen gemeinsam ist die Kompetenz „Anbahnung und Kommunikation“ (*Facilitation / Communication*) angesiedelt. Einige Kompetenzen und Aufgaben sind mehreren Personengruppen zugeschrieben. Aus informationswissenschaftlicher Sicht ergeben sich zwei interessante Dimensionen. In der Horizontalen ist die altbekannte Arbeitsteilung zwischen Infrastruktur und Fachwelt oder zwischen zwei Infrastrukturwelten wie früher zwischen Bibliothek und IT-Abteilung erkennbar. In der Vertikalen jedoch ist zwar links weiterhin die Hierarchie des allgemeinen Qualifikationsrahmens (Bachelor – Master) abgebildet – rechts jedoch ergibt die berufliche Ausdifferenzierung von Rollen keine so klare Niveaueinordnung mehr, denn der *Data Creator* muss nicht unbedingt der ausführende Wissenschaftler, sondern kann durchaus der technische Assistent der Datenerhebung sein. Die alten Strukturen vermischen sich und wir sind geneigt dem, was hier als *Data Librarian* bezeichnet wird, ebenfalls mehr Kompetenzen zuzusprechen als diese Strukturierung es suggeriert. Gerade in kleineren Instituten wird die unterstützende Rolle sich auf die anderen Bereiche ausweiten. Die Rollen im Einzelnen:

3.5.1.1 Data Manager

Der *Data Manager* ist verantwortlich für alle technischen Aspekte, also Lagerung, Zugriff und Aufbewahrung von Daten. Von der Ausbildung her können dies sowohl IT-Spezialisten als auch Informationswissenschaftler (mit IT-Ausrichtung) sein. Sie arbeiten relativ eng mit den Fachwissenschaftlern zusammen, insbesondere was die Auswahl und Zurverfügungstellung der IT-Komponenten, z. B. der virtuellen Wissens- und Forschungsumgebungen betrifft. Manche Datenmanager beschreiben ihre Rolle als „[...] data 'plumber', piping data from one place to another, ensuring data flows operate properly and that valuable data are not lost" (Swan & Brown, 2008, S. 8). Von Bedeutung ist insbesondere die Kommunikation zwischen dem datenerhebenden Wissenschaftler mit seinen Problemen bzw. Bedürfnissen und den Möglichkeiten des Datenmanagers.

3.5.1.2 Data Librarian

Traditionell sind Bibliothekare im Forschungsprozess erst nach der Publikation eingebunden (s. Abb. 3).

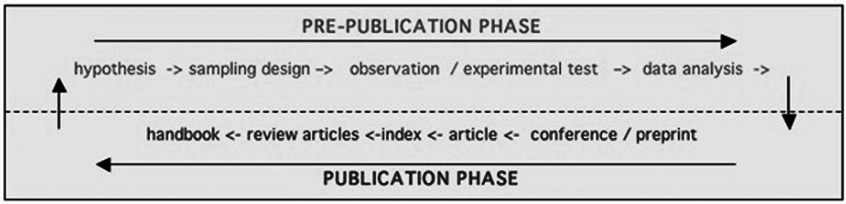

Abb. 3: Data and publication life cycles (Gold, 2007)

Darauf sind im Wesentlichen auch die (bisher) benötigten Fähig- und Fertigkeiten von Bibliothekaren ausgerichtet. Doch als Konsequenz daraus, „[…] dass der Wert der Forschung insbesondere in den Daten steckt und sich daher das Arbeitsspektrum auf das Primärobjekt, die Forschungsdaten, erweitern muss" (Büttner & Rümpel, 2011, S. 108) treten nun mehr Bibliothekare auf den Markt, deren Dienstleistungen sich von der *pre-publication phase* bis zur *publication phase* erstrecken, die „*Data librarians*". Sie brechen ihre traditionelle Rolle auf und erbringen Dienstleistungen im Forschungs*prozess* selbst und nicht „nur" im Verwalten der Forschungs*ergebnisse*. Dies kommt letztlich einem Paradigmenwechsel gleich. Von Dienstleistern der Publikationsphase zu Dienstleistern und Partner des gesamten Forschungsprozesses.

Eine ähnliche Hinwendung zu einer strukturell neuen Positionierung erlebt z.Zt. das Berufsfeld der Archivare, die in der digitalen Welt ebenfalls erkennen, dass ihre Informationstätigkeit nicht erst nach Abschluss des Verwaltungshandelns einsetzen kann, sondern schon die Entstehung von archivierbaren Struktu-

ren und Objekten zum Zeitpunkt ihrer Erstellung steuern müssen im sog. *Records Management*.

Beispiele für die neue Rolle der Bibliothekare im Forschungsdatenmanagement gibt es in Amerika, aber auch im britischen und kanadischen Raum. Rice, eine „*Data Librarian*“ an der University of Edinburgh, teilt ihre Aufgabenbereiche, wie folgt:

> Klassische Aufgabe: Unterstützung des Universitätspersonals und der Studenten bei dem Zugang zu Daten und deren Nutzung
>
> Neue Aufgabe: *Data Curation*, Unterstützung der Bibliothekare beim Verwalten und gemeinsamen Nutzung von Daten (Rice, 2008, S. 13)

Eine Statistik, welche einen Überblick über die Gesamtzahl der *Data Librarians* bietet, existiert nicht (Rümpel, 2010, S. 43). Die Gruppe der *Data Librarians* ist jedoch relativ klein. Eine Analyse von Stellenausschreibungen und Befragungen ergab, dass nur die größeren und größten Hochschulbibliotheken ihr Personal durch diese Stelle erweitern. Diese Position wird i.d.R. auch nur ein Mal je akademischer Organisation besetzt.[1]

> „Eine Ausnahme bildet bspw. die Cornell Universitiy. Dort sind derzeitig drei Data Librarians und zusätzlich noch eine „Datenarchivarin“ angestellt.“ (Rümpel, 2010, S. 42–43)

Nach der Einschätzung von Swan und Brown (2008, S. 1) gibt es in Großbritannien 5 Bibliothekare, die auf Daten spezialisiert sind.

> „In Deutschland ergibt die Recherche nach *Data Librarians* oder nach der deutschen Übersetzung „Datenbibliothekar(e)“ einen einzigen Treffer. Es wird [ein] „Data librarian of PANGAEA“ [benannt]“ (Rümpel, 2010, S. 43)

Die Gründe, weswegen es in Deutschland keine oder kaum *Data Librarians* gibt, sind vielschichtig. Es liegt nahe zu vermuten, dass diese Nachfrage national gar nicht besteht oder die Aufgaben des *Data Librarians* von anderen Berufsständen übernommen werden.

3.5.1.3 Data Curators

Die Rolle eines Datenkurators ergibt sich zwangsweise aus der Forderung nach der Langzeitarchivierung der erhobenen Forschungsdaten. Im *Curation Lifecycle Model*, (s. den Beitrag von Rümpel, im vorliegenden Band Kap. 1.2), werden die Aufgaben des *Curators* besonders deutlich. Das Modell stellt den Wiedergebrauch und die Wiederbelebung der Objekte in den Vordergrund. Beson-

1. Auswertung von Stellenausschreibungen zu *Data Librarians*, siehe Rümpel, 2010.

ders im Fokus liegen die vor dem *Ingest* liegenden Prozesse also die Bildung, Bewertung und Erschließung von Informationsobjekten durch den Produzenten – eine klassische archivarische Tätigkeit (Baumann & Dressel, 2009, S. 1264–1265 zitiert nach Stettler). Weiterhin werden die folgenden Aufgabenfelder aufgeführt: Bewertung, Einarbeitung, Aktivitäten der Bewahrung, Speicherung: Unterstützung bei der Auswahl speicherwürdiger Daten, Betreiben von Repositorien (für Dokumente wie auch Daten), Unterstützung bei der *Curation* und Durchführung von Zufriedenheitsmessungen. Alle Aufgaben, deren alleinige Umsetzung nicht den Produzenten obliegen kann, sondern eine Unterstützung vom *Data Curator* geradezu provoziert. Bisher gibt es jedoch noch keine Untersuchungen/Erhebungen zu diesem Thema in Deutschland.

Zusammenfassend kann konstatiert werden: Informationstechnologie ist nicht der Problemlöser, sondern ein Tool zur Problemlösung. Es wird zunehmend deutlich, dass Datenmanagement eine neue Ausprägung des Wissensmanagements darstellt, mithin ein originäres Thema der Informationswissenschaften ist. Konsequent folgend ist das ein Kern, auf dem eine datenorientierte Ausbildung in den Informationswissenschaften aufbauen kann.

3.5.2 Nachfrage

Wie ist nun die Nachfrage nach Informationswissenschaftlern im Forschungsdatenmanagement? Gibt es überhaupt eine? In den Kapiteln 3.5.1.2 / 3.5.1.3 wurde bereits thematisiert, dass gegenwärtig in Deutschland keine oder kaum *Data Librarians* / *Data Curators* auffindbar sind. Eine quantitative Erhebung ist insofern schwierig, da in Deutschland weder die Begrifflichkeit des „*Data Librarian*“ noch „*Data Curators*“ gebräuchlich ist. Datenmanager ist hingegen eine gängige Bezeichnung und Informationswissenschaftler mit einer dokumentarischen Ausbildung sind in diesem Gebiet, zumindest theoretisch, gut positioniert. Aktuelle Verbleibstudien informationswissenschaftlicher Studiengänge in Hamburg, Potsdam, Köln, Darmstadt und Stuttgart belegen dies jedoch nur vage. Die Berufsbezeichnung „*Data Manager*“ tritt nicht auf. Bei den Tätigkeitsmerkmalen wird zwar sehr häufig „Datenpflege“ genannt (Herzberg, 2008, S. 61). Management von Forschungsdaten ist jedoch weit umfänglicher. Grundsätzlich wird national das Aufgabengebiet eines Bibliothekars nicht in seiner Berufsbezeichnung genannt, wie es im amerikanischen Raum gehandhabt wird. Der derzeitige Stand über den Einsatz von Informationswissenschaftlern aller „Sparten“ im Forschungsdatenmanagement ist daher nicht konkret zu ermitteln. Insofern erscheint es als besonders interessant zu fragen, ob überhaupt eine Nachfrage existiert.

Eine Recherche nach Ausschreibungen für Informationswissenschaftler mit Aufgabengebieten im Forschungsdatenmanagement, ergab im Jahr 2010 folgende „Fundstücke“.

- Die Universitätsbibliothek Bielefeld suchte eine/n wissenschaftliche/n Mitarbeiter/in für die Entwicklung des Forschungsdaten-Services. Voraussetzung war der Abschluss eines fachwissenschaftlichen Hochschulstudiums entweder im Bereich Informationswissenschaften oder Informatik. (Ausschreibung lief bis zum 15.02.2010)
- An der Universität Trier wurde jemand für das Zentrum „Informations-, Medien- und Kommunikationstechnologien“ zur Entwicklung und Implementierung einer Dienstleistung zur Langzeitarchivierung von Forschungsdaten gesucht. Erwartet wurde ein abgeschlossenes wissenschaftliches Hochschulstudium (vorzugsweise mit Promotion) der Informatik oder Informationstechnologie oder der Geistes- oder Sozialwissenschaften mit nachgewiesenen Kenntnissen im bibliotheks- oder informationswissenschaftlichen Bereich. (Ausschreibung lief bis zum 4.10.2010)
- Die Hochschulbibliothek der RWTH Aachen suchte eine Wissenschaftliche/n Bibliothekar/in oder Dokumentar/in zur Entwicklung einer virtuellen Forschungsumgebung. (Ausschreibung lief bis zum 30.09.2010)
- Die Stabsstelle Strategische Forschungsentwicklung der Georg-August-Universität Göttingen suchte eine/n Mitarbeiter/in für den Aufbau einer Infrastruktur für die „Sicherung guter wissenschaftlicher Praxis im Umgang mit Forschungsdaten“. Erwartet wurde ein abgeschlossenes Hochschulstudium und die Bewerber sollten über nachweisbare Kenntnisse im Bereich der Informatik/Informationstechnologie oder im bibliotheks- oder informationswissenschaftlichen Bereich verfügen und ein breites Verständnis von verschiedenen wissenschaftlichen Arbeitsprozessen mitbringen. (Ausschreibung lief bis zum 1.10.2010)

Alle Ausschreibungen boten eine Stelle der Entgeltgruppe TV-L 13, erforderten also einen Master-Abschluss. Klassische Dipl.-Bibl. (FH) oder BA sind damit von vornherein ausgeschlossen. Bei zwei Einrichtungen wurden promovierte Wissenschaftler eingestellt. Diese können auf Grund ihrer wissenschaftlichen Tätigkeiten enorme Expertise bez. der Erhebung von Forschungsdaten und den Belangen eines Wissenschaftlers vorweisen.

Eine Nachfrage nach Informationswissenschaftlern ist deutlich vorhanden, jedoch haben diese offenbar starke Konkurrenz aus der fachwissenschaftlichen Praxis. Ebenfalls wird durch die hochgruppierten Stellenausschreibungen betont, welche Art von Bewerbern erwartet werden. Dies lässt den Schluss zu, dass national die Informationswissenschaftler gegenüber Wissenschaftlern aus anderen Fachgebieten für die Aufgaben weniger attraktiv sind. Nach den o.a. Ausführungen zum *Data Librarian* eine zunächst doch eher ernüchternde Erkenntnis.

3.5.3 Mögliche Tätigkeitsfelder für Informationswissenschaftler

Die gegenwärtige Ausbildung von Bibliothekaren ist so konzipiert, dass sie natürlich die Fähigkeiten für die traditionellen bibliothekarischen Aufgaben erwerben: Sammeln, Bewahren, Ordnen, Bereitstellen und Vermitteln. Diese Dienstleistungen beziehen sich auf die Publikationen, welche das Ergebnis der Forschung bilden (Plassmann, Rösch, Seefeldt & Umlauf, 2006, S. 10).

> „[Die erlernten] Aufgabenfelder können ebenfalls auf das Forschungsdatenmanagement übertragen werden. Denn auch Daten werden gesammelt (Aufbau von Datensammlungen im Forschungsdatenrepositorium), bewahrt (*Curation*, *Preservation*), geordnet (mit Metadaten versehen), bereitgestellt (zugänglich in Repositorien) und vermittelt (Beratung, Vermittlung von Datenkompetenz). Der Unterschied zu den Tätigkeitsfeldern mit klassischen Medien ist, dass datenorientierte Bibliothekare nicht Informationsquellen nach der Veröffentlichung vermitteln, sondern diese bereits ab der Entstehung. Doch unabhängig von diesen Differenzen sind es immer Informationsobjekte.“ (Rümpel, 2010, S. 47)

Dass das bibliothekarische Studium aber eine wichtige Grundlage für Positionen im Forschungsdatenmanagement schafft, wurde durch die Sichtung von internationalen Stellenanzeigen für *Data Librarians* erkannt.[2] Grundsätzlich wird dort ein „*Master of Library (and Information) Science (ML(I)S)*“ als Qualifikation erwartet.

Erforderliche Fähig- und Fertigkeiten von *Data Librarians*:

- Kommunikation
- Kundenservice
- Soziale Kompetenz
- Erschließung
- Vermittlung

Diese Kompetenzen werden gegenwärtig auch in der deutschen bibliothekarischen Ausbildung vermittelt.

Hinzu werden spezialisierte Kompetenzen für den Umgang mit Daten erwartet[3]:

- Kenntnis von Datenformaten, Analysesoftware, Analyseverfahren, Urheberrecht

2. Vgl. Rümpel, 2010; sowie Pampel, Bertelmann & Hobohm, 2010, S. 170ff

3. Die Erwartungen wurden in durchgeführten Interviews mit Experten erhoben, siehe Rümpel, 2010

- Bewertung bez. Daten.

Keine neuen Kompetenzfelder, sondern klassische, die ein Stück spezifischer werden.

Als Fazit gilt, dass Bibliothekare auf Grund ihrer Ausbildung große Expertise in traditionellen Tätigkeitsfeldern besitzen, die auf das Forschungsdatenmanagement angewendet werden können. Folglich sind sie per se prädestiniert auch im Forschungsdatenmanagement ihre Dienstleistungen anzubieten. (Rümpel, 2010, S. 46)

3.5.4 Datenorientierte Ausbildungsszenarien für Informationswissenschaftler

3.5.4.1 Grundständige Ausbildung

Die Ausbildung von Informationswissenschaftlern in den angloamerikanischen Ländern und Deutschland unterscheidet sich stark.

Studierende in den USA, UK etc. absolvieren ein Bachelorstudium in einem frei wählbaren wissenschaftlichen Fach. Auf dieses „Vorstudium“ wird der informationswissenschaftliche Master gesetzt. Folglich verfügen dort die Absolventen immer über die Expertise aus einem wissenschaftlichen Bereich. Die deutsche Ausbildung ist anders konzipiert. Die zukünftigen Bibliothekare absolvieren einen informationswissenschaftlichen Bachelor und setzten darauf den passenden Master.[4]

In einem Beitrag zu Karrieremöglichkeiten im informationswissenschaftlichen Bereich heißt es 2009:

> „The special librarian or information specialist must have a degree in subject specialty as well as in library and information science. In fact, this subject expertise may be so vital to companies and businesses that they may prefer a technician with subject specialization to a professional with a master’s degree in library and information science“ (Taylor, Parish & Roderer, 2009, S. XV)

Dies erscheint ein wichtiger Grund zu sein, weshalb es die Informationswissenschaftler in Deutschland bei der Beteiligung im Forschungsdatenmanagement schwer haben. Hinzu kommt, dass in den existierenden deutschen informationswissenschaftlichen Bachelor-Ausbildungsgängen das Forschungsdatenmanagement z. Z. nicht explizit behandelt wird. Gleichwohl wird auf den Kongressen / Jahrestagungen der Berufsverbände stark darüber diskutiert. Bereits auf dem

4. Eine umfangreiche Aufführung der Masterstudiengänge in Deutschland wurde zusammengestellt von Georgy, 2010, S. 210-216.

Bibliothekartag 2009 konnte konstatiert werden (vgl. Pampel, Bertelmann & Hobohm, 2010, S. 169), dass die meisten der Konzepte der internationalen Curriculum-Diskussion zur Ausbildung von Datenmanagern (vgl. z. B. Abb. 2) ursprünglich aus dem informationswissenschaftlichen Feld stammen, wie z. B. „Wissensmanagement" als Aufgabe für den *Data Scientist* oder „Metadaten" als Anforderung für den *Data Manager* und den *Data Creator*. Einzige für den LIS Bereich zunächst untypische Kompetenzfelder scheinen hierbei „Datenmodellierung" und „Datenanalyse" zu sein – alle anderen sind entweder schon lange Kernaufgaben von Bibliothekaren oder Informationswirten oder beginnen es zu werden, wie die Forderung nach „Verhandlungsgeschick" oder die nach „Wirtschaftlichkeitsüberlegungen zu Informationen". Eine Überprüfung am europäischen Certidoc-Kompetenz Modell der Informationsberufe (vgl. ECIA, 2004, Hobohm, 2005) lässt erkennen, dass *information professionals* schon jetzt genau diese Kompetenzen aufweisen sollten. Auch „Modellierung" und „Analyse" ist eine zunehmend wichtige Kernaufgabe in der Ausbildung von Informationswissenschaftlern, führt man sich vor Augen, welche neuen Anforderungen z. B. das *Semantic Web* an *information professionals* stellt (s. Beitrag von Neher, Ritschel Kap. 3.3 im vorliegenden Band). Einzig das Objekt der Informationsbearbeitung ändert sich, aber ähnliche Arten von Medienwandel haben Informationswissenschaftler und Bibliothekare schon oft gemeistert.

Sind also diese Prioritäten erkannt und in den Curricula umgesetzt, gibt es keine Notwendigkeit für spezielle BA-Studiengänge. So verwundert es nicht, dass es in Deutschland keine BA-Ausbildung zu den konkret benannten Berufsfeldern zum *Data-Information Scientist* / *Data Librarian* angeboten wird. Betrachtet man jedoch den breiteren Kontext und inkludiert die für das Datenmanagement wesentlichen virtuellen *eScience* Konzepte virtueller Forschungsumgebungen, so findet man erste Ansätze (Atkinson, Fergusson & Vander Meer, 2009, S. 5).

> „A curriculum should almost always build up to a project that provides an in-depth experience of applying e-Science methods to achieve research, design, or decision-making goals. For example, if a series of modules has developed expertise in handling distributed heterogeneous data, then a student might pursue a project that applies these skills to a specific goal, such as integrating biochemical and climate data to present potentially predictive information about the onset of algal blooms in a particular sea area or lake." (Atkinson, Fergusson & Vander Meer, 2009, S. 5)

In solchen virtuellen Forschungsumgebungen kommt dem einheitlichen Datenmanagement eine essentielle Bedeutung zu. Betont wird die interdisziplinäre Kommunikation.

> „Research and innovation work best if scientific practitioners in the various disciplines operate in close coordination with technical experts in computer science.“ (Atkinson, Fergusson & Vander Meer, 2009, S. 5)

Eine vorzügliche Möglichkeit für ausdifferenzierte Ausbildungsinhalte für das Datenmanagement bilden demzufolge Masterkurse in den Informationswissenschaften. Aufbauend auf den Grundlagen der BA-Module (informationswissenschaftliche und methodische Kernkompetenzen, IT sowie Projektmanagement) können hier weitere Kernkompetenzen vermittelt werden. Dies sind dann keine klassisch bibliothekarischen, dokumentarischen oder archivarischen Kompetenzen. Vielmehr geht es um methodisch-technische sowie an kognitiven Prozessen orientierte Ansätze. Damit werden die Masterstudierenden befähigt, in einem projektorientierten Berufsfeld, informationswissenschaftliche Aufgabenstellungen auf hohem Verantwortungsniveau in einer fachlich großen Breite durchzuführen (Hobohm, 2010, S. 1). Vorstellbar ist, wie etwa im informationswissenschaftlichen Master der FH Potsdam realisiert, eine Profillinie zum Thema „Wissenstransfer und Projektkoordination“. Eine Ausrichtung dieser Profillinie sind „Virtuelle Wissensumgebungen“. Bereits Atkinson, Fergusson und Vander Meer betonten 2009, „it’s important to provide students with both a conceptual framework that they can use in their disciplines and practical skills that they can use immediately“. (Atkinson, Fergusson & Vander Meer, 2009, S. 5)

Das Lernergebnis ist wie folgt beschrieben:

> „Die Teilnehmer
> - kennen die Anforderungen an die Informationsinfrastruktur in den Stufen des wissenschaftlichen Schaffensprozesses,
> - sind befähigt, die Prozesse, z. B. die Lebenszyklen von Daten (DCC) zu analysieren
> - können ganzheitliche Modellösungen erarbeiten, bei denen es um genuin informationswissenschaftliche Tätigkeiten und Fertigkeiten, wie Wissensmanagement, Fähigkeiten zur Bewertung und Einordnung in Kontexten geht.
>
> Im Ergebnis des Zusammenwirkens von Informationstechnik und der menschlichen Komponente können die Teilnehmer eine Konzeption für einen konkreten Anwendungsfall entwickeln“ (FH Potsdam, 2010, S. 25–26)

Die Lösung unter den gegebenen deutschen Bedingungen ist also nicht die Kombination aus wissenschaftlicher Fachexpertise und informationswissenschaftlicher Kompetenz, sondern die allgemeine Methodenkompetenz der informationswissenschaftlichen Masterausbildung in Kombination mit einer guten IT-Ausbildung und kognitiv-personalen Kompetenzen.

3.5.4.2 Weiterbildung

Eine viel größere Bedeutung kommt in diesem Kontext jedoch der Weiterbildung zu. Da die Nachfrage nicht ausreichend für grundständige Studiengänge ist, wird die Qualifizierung stärker über verschiedene Formen der Weiterbildung realisiert. So z. B. über Workshops wie „Providing Social Science Data Services: Strategies for Design and Operation" in den USA (Jacobs, 2010). Partner sind dabei universitäre Ausbildungseinrichtungen, sowie Forschungseinrichtungen. Im Idealfall basiert ein solches Weiterbildungskonzept auf zwei Säulen. Säule 1 (Grundlagen) besteht aus periodisch stattfindenden *Sommerschools*/ Ringvorlesungen zu aktuellen Themen. Die Ausbildungseinrichtungen bringen dabei das didaktische Konzept und die Infrastruktur ein, die Forschungsinstitute entsprechende, aus der Praxis kommende, Inhaltsideen und Experten. Als Themen eignen sich z. B.:

- Systeme & technische Infrastruktur
- Data Service Design
- Metadaten & Standards für Forschungsdaten
- Datenqualität
- Datenvisualisierung und -analyse
- Wissensmanagement
- *Data Curation Practices / Digital Curation*
- Langzeitverfügbarkeit und Zugang zu Forschungsdaten
- *Data Librarians*: Erfahrungen aus der Praxis

Die *Summerschools* können von mehreren Tagen bis zu max. 1 Woche dauern. Wichtig ist das abschließende Zertifikat, z. B. nach dem *Graduate Certificate in Digital Information Management*[5]. Beispiel in Deutschland sind die Nestor *Summerschools*[6].

Die Themen sollten in Blöcken zusammengefasst werden, dass eine getrennte Buchung und Teilnahme, z. B. nach Zielgruppen, ermöglicht wird. In der Praxis bewährt, hat sich folgende Organisation:

- Wahlmöglichkeit aus insgesamt mehreren eintägigen Veranstaltungen (Module)
- Jeder Tag / Modul wird in Sessions mit Vortrag und Diskussion zerlegt, so dass pro Modul 3–4 Einzelthemen (mit unterschiedlichen Referenten) bearbeitet werden können.

5. Informationen zum *Graduate* Certificate in *Digital Information Management* unter: http://digin.arizona.edu/ [Zugriff am 18.08.2011].

6. siehe Webauftritt der *Summerschools* von NESTOR unter: http://nestor.sub.uni-goettingen.de/education/index.php [Zugriff am 18.08.2011].

Die zweite Säule dient der Verstetigung der Grundlagen und beinhaltet eintägige Seminaren zu Spezialthemen, zielgruppenspezifisch nach Bedarf zu den Themenfeldern der Summerschools. Insbesondere aus der Praxis kommt zusätzlich der dringende Wunsch nach Online-Kursen bzw. Selbstlernangeboten.

Die Ausbildungseinrichtungen bringen wieder das didaktische Konzept und die Infrastruktur ein, die Forschungsinstitute die praxisrelevanten Themen. In der FH Potsdam wurden damit gute Erfahrungen gemacht. Bereits bearbeitete Themen in einer Kooperation zwischen der FH Potsdam und dem GFZ (Geoforschungszentrum) waren[7]:

- Wissenschaftskommunikation im Wandel
- Open Access / Open Data

Zur Zeit wird an der FH Potsdam im Rahmen des brandenburgischen Forschungs- und Innovationsförderungsprogramms an einem Projekt für Selbstlernangebote im Datenmanagement gearbeitet (MWfK-Projekt: Werkzeuge und Selbstlernangebote für den nachhaltigen Einsatz in der Klimaforschung).

7. Projektseminar der FH Potsdam im WS 2010/2011 unter der Leitung von Prof. Büttner und Dipl. Bibl. Heinz Pampel.

Literaturhinweise

Atkinson, M. et.al., 2008. *Education and Training Task Force. Long Report.* Lugano, 2008. (Version 6, final, 09.06.2008) Online: http://www.e-irg.eu/images/stories/publ/task_force_reports/ettf_long_report_final_july08.pdf [Zugriff 17.08.2011].

Atkinson, M. Fergusson, D. & Vander Meer, E., 2009. Curricula Development for e-Science: Meeting the Challenges. *IEEE Computing Now Online Only,* March 2009. Online: http://www.computer.org/portal/c/document_library/get_file?uuid=c2c272c7-bd01-4c70-9faf-f34bf86b6e12&groupId=53319 [Zugriff am 17.08.2011].

Büttner, S. & Rümpel, S., 2011. Bibliotheken und Bibliothekare im Datenmanagement. In: S. Schomburg et. al., Hrsg. 2010. *Digitale Wissenschaft. Stand der Entwicklung digital vernetzter Forschung in Deutschland*. Konferenzband. Köln: hbz, S. 107–114. Online: http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf [Zugriff am 17.08.2011].

Bailey, J. & Ball, R., 2008. Die Einbindung von Bibliotheken in das integrative Wissenschaftskonzept. E-Science und Bibliotheken. *B.I.T. online*, 11(1), S. 15–24.

Baumann, F. & Dressel, B., 2009. NESTOR Spring School 2009 – ein Teilnehmerbericht. *Bibliotheksdienst*, 43(12). Online: http://www.zlb.de/aktivitaeten/bd_neu/heftinhalte2009/Technik011209BD.pdf [Zugriff am 17.08.2011].

Corrall, S., 2008. Research Data Management: Professional Education and Training Perspectives. Vortrag. *2nd DCC / RIN Research Data Management Forum. Roles and Responsibilities for Data Curation.* Manchester, Großbritannien 26.-27. Nov. 2008. Online: http://www.dcc.ac.uk/sites/default/files/documents/RDMF/RDMF2/07%20Corrall.pdf [Zugriff am 17.08.2011].

DFG (Deutsche Forschungsgemeinschaft), 1998. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*. Weinheim: Wiley-VCH. Online: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf [Zugriff am 17.08.2011].

Donnelly, M., 2008. RDMF2. Core Skills Diagram. Research Data Management Forum, 17.12.2008. Online: http://data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html [Zugriff am 17.08.2011]

FH (Fachhochschule) Potsdam, 2010. *Konsekutiver Masterstudiengang Informationswissenschaften. Modulbeschreibungen*. (Stand: 27.09.2010) Online: http://iw.fh-potsdam.de/fileadmin/FB5/Dokumente/Master_IW/Modulbeschreibungen-MA-I.pdf [Zugriff am 17.08.2011].

ECIA (European Council of Information Associations ECIA) Ed., 2004. *Euroguide. Handbuch für Informationskompetenz (BID*). Frankfurt. Online: http://www.certidoc.net [Zugriff am 17.08.2011]

Georgy, U., 2010. Übersicht Master-Studiengänge. *Information. Wissenschaft und Praxis*, 61(3), S. 210–216.

Gold, A., 2007. Cyberinfrastructure, Data, and Libraries, Part 1. A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine*, 13(9/10). Online: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html [Zugriff am 17.08.2011].

Hank, C.; Davidson, J. 2009: International Data Curation Education Action (IDEA) Working Group. A Report from the Second Workshop of the IDEA. *D-Lib Magazine* 15(3/4). DOI:10.1045/march2009-hank.

Herzberg, E., 2008. *Der Dokumentar in der Informationsgesellschaft*. Diplomarbeit, Fachhochschule Potsdam.

Hobohm, H.-C., 2005. Der Bibliotheks-Bachelor. Oder was ist wirklich neu am neuen Berufsbild des Bibliothekars? In: E. Kolding, E. et al., Hrsg. 2005. *Die innovative Bibliothek.* Elmar Mittler zum 65. Geburtstag. München, S. 275–285.

Hobohm, H.-C., 2010. *Master of Arts – Informationswissenschaften. Spezialisierung „Wissenstransfer / Projektkoordination“* Internes Arbeitspapier FH Potsdam.

Jacobs, J., 2010. *IASSIST Workshop: Providing Social Science Data Services: Strategies for Design and Operation.* Online: http://www.iassistdata.org/blog/workshop-providing-social-science-data-services-strategies-design-and-operation [Zugriff am 17.08.2011].

Lyon, L. 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships.* Consultancy Report. Online: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf [Zugriff am 17.08.2011].

NSF (National Science Foundation)/ National Science Board, 2005. *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century.* Arlington. Online: http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf [Zugriff 17.08.2011]

OGF22-ET CG 2008. Open Grid Forum – Education and Training Community Group. Meeting. 22: *Curricula for Undergraduate and Masters Level Courses in e-Science*. Report from the ICEAGE Curricula Development Workshop, Brüssel, Belgien 14.-15. Feb. 2008.

OECD (Organisation for Economic Co-operation and Development), 2007. *OECD's Principles and Guidelines for Access to Research data from Public Funding*. Paris. Online: http://www.oecd.org/dataoecd/9/61/38500813.pdf [Zugriff 17.08.2011]

Pampel, H. Bertelmann, R. & Hobohm, H.-C., 2010. „Data Librarianship" – Rollen, Aufgaben, Kompetenzen. In: C. Schmiedeknecht & U. Hohoff, Ed. 2010. Ein neuer Blick auf Bibliotheken. *98. Deutscher Bibliothekartag*. Erfurt, Deutschland 2.-5. Juni 2009. Deutscher Bibliothekartag: Kongressbände. Hildesheim: Olms. S. 169–176. (auch als Working Paper des RatSWD: Pampel, H. Bertelmann, R. & Hobohm, H.-C., 2010. *„Data Librarianship" – Rollen, Aufgaben, Kompetenzen*. Berlin: Rat für Sozial- und Wirtschaftsdaten / BMBF (Working paper series des RatSWD, 144).

Plassmann, E. et. al., 2006. *Bibliotheken und Informationsgesellschaft in Deutschland. Eine Einführung*. Wiesbaden: Harrassowitz.

Pryor, G. & Donnelly, M. 2009. Skilling up to Do Data. Whose Role, Whose Responsibility, Whose Career? *The International Journal of Digital Curation*, (4), S. 158–170.

Rice, R., 2008. *Roles & Responsobilites for Data Curation. The Data Librarian. 2nd DCC / RIN Research Data Management Forum.Roles and Responsibilities for Data Curation.* Manchester, Großbritannien 26.-27. Nov. 2008. Online: http://www.dcc.ac.uk/sites/default/files/documents/RDMF/RDMF2/05%20Rice.pdf [Zugriff am 17.08.2011].

Rümpel, S., 2010. *Data Librarianship – Anforderungen an Bibliothekare im Forschungsdatenmanagement*. Diplomarbeit, Fachhochschule Potsdam. Online http://opus.kobv.de/fhpotsdam/volltexte/2010/163/pdf/10600.pdf [Zugriff am 17.08.2011].

Swan, A. & Brown, S., 2008. *Skills, Role & Career Structure of Data Scientists & Curators. Assessment of Current Practice & Future Needs*. Report to the JISC. Online: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf [Zugriff am 17.08.2011].

Taylor, T. A. Parish, J. R. & Roderer, N., 2009. *Career opportunities in library and information science*. New York: Ferguson.

Die Autoren

Vorwort – Dr. Stefan Winkler-Nees

Stefan Winkler-Nees ist Meeresgeologe. Nach zahlreichen Forschungsprojekten war er für ein Software Unternehmen tätig. Seit 2007 arbeitet er für die DFG, zunächst im Fachreferat, dann in der Gruppe Wissenschaftliche Literaturversorgungs- und Informationssysteme. Sein Aufgabenfeld beinhaltet die Entwicklung und Umsetzung von Maßnahmen zur Verbesserung des Umgangs mit Forschungsdaten.

Kapitel 1.1 – Prof. Dr. Stephan Büttner, Prof. Dr. Hans-Christoph Hobohm, Lars Müller

Stephan Büttner ist seit 2004 Professor für Digitale Medien an der FH Potsdam. Nach dem Studium der Physik promovierte er auf dem Gebiet der Informationswissenschaft. Zu seinen Lehr- und Forschungsschwerpunkten zählen: Theorie und Praxis digitaler Medien, Projektmanagement, Digital Rights Management, Forschungsdatenmanagement.

Hans-Christoph Hobohm Bibliothekswissenschaftler, ehem. Dekan des Fachbereichs Informationswissenschaften und Forschungsprofessor im Innovationskolleg der Fachhochschule Potsdam, ursprünglich von der empirisch arbeitenden, historischen Sozialforschung kommend.

Lars Müller ist akademischer Mitarbeiter am Fachbereich Informationswissenschaften der FH Potsdam, er arbeitet in FuE-Projekten zu Forschungsdaten. Ausbildung: Europäische Ethnologie und Geschichte M.A., 1998; MA LIS, 2008. Schwerpunkte: Forschungsdaten, elektronische Fachinformation, Informationskompetenz.

Kapitel 1.2 – Stefanie Rümpel

Stefanie Rümpel erhielt ihren Abschluss als Diplom-Bibliothekarin an der FH Potsdam mit ihrer Arbeit zum Thema Data Librarians. Nach Mitarbeit an dem Projekt „WernEr“ (Semantisches Datenmanagement – Werkzeuge und Selbstlernangebote für den nachhaltigen Einsatz in der Klimaforschung) an der FH Potsdam, ist sie nun an der Hochschulbibliothek der FH Düsseldorf tätig.

Kapitel 1.3 – Prof. Dr. Gert G. Wagner, Prof. Dr. Notburga Ott, Denis Huschka, MA, Claudia Oellers

Gert G. Wagner, geb. 1953, ist Vorsitzender des Rats für Sozial- und Wirtschaftsdaten (RatSWD). Hauptberuflich ist er Vorstandsvorsitzender des Deutschen Instituts für Wirtschaftsforschung (DIW Berlin), Professor für Volkswirt-

schaftslehre an der TU Berlin und Max Planck Fellow am MPI für Bildungsforschung in Berlin.

Notburga Ott, Dipl. Volkswirtin, Professorin für Sozialpolitik und Institutionenökonomik an der Ruhr-Universität Bochum. Stellv. Vorsitzende des Rates für Sozial- und Wirtschaftsdaten und des wissenschaftlichen Beirats für Familienfragen beim BMFSFJ. Arbeitsschwerpunkte: Sozialpolitik, insbes. Verteilungsfragen, Familien- und Gesundheitspolitik.

Denis Huschka, geb. 1975, studierte Soziologie und Politikwissenschaft in Berlin. Er ist seit 2007 Geschäftsführer des Rates für Sozial- und Wirtschaftsdaten (RatSWD) und seit 2011 Executive Director der der International Society for Quality of Life Studies (ISQOLS). Er war forschend am Wissenschaftszentrum für Sozialforschung Berlin, der Rhodes University Grahamstown/Südafrika, an der FU Berlin und dem DIW Berlin tätig und gründete 2010 die Gesellschaft für Wissenschaftspolitik und Infrastrukturentwicklung (GWI Berlin).

Claudia Oellers, Diplompolitologin, ist seit 2007 wissenschaftliche Mitarbeiterin der Geschäftsstelle des Rates für Sozial- und Wirtschaftsdaten.

Kapitel 2.1 – Heinz Pampel, Roland Bertelmann

Heinz Pampel arbeitet im Koordinationsbüro des Open-Access-Projektes der Helmholtz-Gemeinschaft. Das Koordinationsbüro ist am Deutschen GeoForschungsZentrum – GFZ und am Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) angesiedelt. Er ist Mitglied in verschiedenen Gremien rund um die dauerhafte Zugänglichkeit von Forschungsdaten.

Roland Bertelmann ist Leiter der Bibliothek des Wissenschaftsparks Albert Einstein, eine gemeinsame Bibliothek des Deutschen GeoForschungsZentrums GFZ, des Potsdam-Instituts für Klimafolgenforschung, der Forschungsstelle Potsdam des Alfred-Wegener-Instituts für Polar- und Meeresforschung und des IASS Potsdam – Institute for Advanced Sustainability Studies

Kapitel 2.2 – Prof. Dr. Gerald Spindler, Tobias Hillegeist

Gerald Spindler ist Direktor des Instituts für Wirtschaftsrecht, Lehrstuhl für Bürgerliches Recht und u.a. Multimedia- und Telekommunikationsrecht; ordentliches Mitglieder der Akademie der Wissenschaften; Habilitation Universität Frankfurt 1996, Rufe an die Universitäten zu Köln, Bielefeld, Frankfurt und die ETH Zürich.

Tobias Hillegeist ist Doktorand am Lehrstuhl von Prof. Dr. Gerald Spindler, Uni Göttingen und Referendar am Landgericht Lüneburg.

Kapitel 2.3 – Uwe Jensen

Uwe Jensen, Dipl.-Psych., ist seit 1996 wissenschaftlicher Mitarbeiter im GESIS Datenarchiv und in mehreren EU Projekten europäischer Datenarchive tätig gewesen (u. a. CESSDA PPP). Aktuell befasst er sich mit sozialwissenschaftlichen Metadatenstandards sowie Archivstandards und -prozessen im Rahmen des langfristigen Managements von Forschungsdaten.

Kapitel 2.4 – Uwe Jensen, Dr. Alexia Katsanidou, Wolfgang Zenk-Möltgen

Uwe Jensen siehe 2.3

Alexia Katsanidou ist Politologin mit Schwerpunkt vergleichende Wahlforschung. Sie leitet seit 2010 das Team Internationale Dateninfrastrukturen der GESIS (Leibniz Institut der Sozialwissenschaften) mit dem Ziel, die Kooperation zu Fragen des Datenmanagements zwischen europäischen Datenarchiven und Forschungsprojekten zu fördern.

Wolfgang Zenk-Möltgen, M.A. ist seit 1996 wissenschaftlicher Mitarbeiter im GESIS Datenarchiv und in Projekten zur Entwicklung von Archivtools und Standards tätig (u. a. DDI, DataCite, da|ra Datenregistrierung, STARDAT). Schwerpunktmäßig beschäftigt sich Herr Zenk-Möltgen mit Datenbank- und Anwendungsentwicklung zur Datendokumentation und -archivierung.

Kapitel 2.5 – Dr. Andreas Aschenbrenner, Dr. Heike Neuroth

Andreas Aschenbrenner hat langjährige Erfahrungen im Aufbau von institutionellen und internationalen digitalen Infrastrukturen. Seine Kernaufgaben liegen dabei in der Entwicklung von Daten- und Prozessmodellen, sowie System-Architekturen.

Heike Neuroth leitet die Abteilung Forschung und Entwicklung an der Niedersächsischen Staat- und Universitätsbibliothek (SUB) Göttingen. Ihre Schwerpunkte sind die Langzeitarchivierung digitaler Objekte, Forschungsdaten, Virtuelle Forschungsumgebungen und Forschungsinfrastrukturen.

Kapitel 2.6 – Dr. Jens Klump

Jens Klump koordiniert eScience Projekte am Deutschen GeoForschungsZentrum in Potsdam. Als Geologe beteiligt er sich aktiv in Forschungsprojekten, um herauszufinden, welche Werkzeuge für den Umgang mit Forschungsdaten in den Projekten und für deren Veröffentlichung und Archivierung gebraucht werden.

Kapitel 2.7 – Matthias Razum

Matthias Razum ist Abteilungsleiter e-Publishing und eScience am FIZ Karlsruhe. Nach seinem Studium der Wirtschaftsinformatik befasste er sich mit webbasierten wissenschaftlichen Informationssystemen, virtuellen Forschungsumgebungen, Forschungsdatenmanagement und *Digital Preservation*.

Kapitel 2.8 – Prof. Dr. Bettina Berendt, Dr. Joaquin Vanschoren, Bo Gao

Bettina Berendt ist Professorin am Fachbereich Informatik der K.U. Leuven. Sie habilitierte sich in Wirtschaftsinformatik an der Humboldt-Universität zu Berlin und promovierte in Informatik/Kognitionswissenschaft an der Universität Hamburg. Sie forscht u.a. in den Bereichen Web Mining, Privacy, Informationsvisualisierung mit interdisziplinären Zugängen zu diesen Themen.

Joaquin Vanschoren ist Postdoktorand an der Universität Leiden. Er promovierte an der K.U. Leuven über die Verbindung von maschinellem Lernen, Meta-Lernen und Datenbanken. Seine Forschungsinteressen umfassen des Weiteren Zeitreihenanalyse, Semantische (Web-)Technologien, eSciences, und Data Mining und seine Anwendungen auf großen Datenmengen.

Bo Gao ist Doktorand am Fachbereich Informatik der K.U. Leuven. Zuvor erwarb er einen Master in Künstlicher Intelligenz an der K.U. Leuven. Derzeit arbeitet er im SPION-Projekt zu Sicherheit und Privacy im Social Web. Seine Forschungsinteressen umfassen Informationsvisualisierung, Data Mining, maschinelles Lernen und webbasierte Privacy-Anwendungen.

Kapitel 3.1 – Dr. Michael Lautenschlager

Michael Lautenschlager leitet die Abteilung Datenmanagement am Deutschen Klimarechenzentrum und ist Direktor des am DKRZ angesiedelten ICSU World Data Center Climate. Er studierte Physik in Hamburg und beschäftigte sich nach der Promotion einige Jahre mit Klimamodellierung am Max-Planck-Institut für Meteorologie in Hamburg bevor er ins DKRZ wechselte. Seit mehr als 15 Jahren gestaltet er das wissenschaftliche Datenmanagement am DKRZ maßgeblich mit.

Kapitel 3.2 – Sünje Dallmeier-Tiessen

Sünje Dallmeier-Tiessen ist Geowissenschaftlerin und war in der klima- und geowissenschaftlichen Forschung tätig, bevor sie sich dem Themenfeld wissenschaftliches Publizieren zugewandt hat – zu Beginn in einem Verlag, dann als wissenschaftliche Mitarbeiterin der Helmholtz Gemeinschaft mit einem Schwerpunkt auf Open Access und dem Zugang zu und der Publikation von Forschungsdaten. Diesen Forschungsschwerpunkt verfolgt sie seit Ende 2009 nun weiter am CERN und an der HU, gefördert durch ein Stipendium des BMBF (Gentner).

Kapitel 3.3 – Prof. Dr. Günther Neher, Bernd Ritschel

Günther Neher ist Physiker, Promotion in physikalischer Chemie an der FU-Berlin. Freiberufliche Tätigkeit als Fachreferent für FIZ-Technik, Frankfurt/ Main. Selbständigkeit im Bereich Konzeption und Entwicklung web-basierter, datenbank-gestützter Informationssysteme. Seit 2005 Professor für Webtechnologie und Semantic Web-Anwendungen am Fachbereich Informationswissenschaften der FH Potsdam.

Bernd Ritschel ist Physiker, Informationswissenschaftler und Projektleiter im Bereich wissenschaftliche Infrastruktur: Aufbau des GFZ Informationssystems und Datenzentrums (GFZ ISDC) für geowissenschaftliche Daten von Satellitenmissionen und in-situ Observatorien, Design der ISDC-Ontologie für die Semantik-basierte Präsentation, Verknüpfung und Recherche geowissenschaftlicher Datenprodukte.

Kapitel 3.4 – Dr. Erich Weichselgartner

Erich Weichselgartner ist seit 2000 stellvertretender wissenschaftlicher Leiter des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID) in Trier und Leiter der Bereiche IT und Entwicklung. Vorherige Stationen waren die Fraunhofer-Gesellschaft in München, die Universität Regensburg (Habilitation) und die New York University (Promotion).

Kapitel 3.5 – Prof. Dr. Stephan Büttner, Prof. Dr. Hans-Christoph Hobohm, Stefanie Rümpel

Stephan Büttner siehe 1.1

Hans-Christoph Hobohm siehe 1.1

Stefanie Rümpel siehe 1.2