Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 15
Zusammenfassung und Interpretation

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Universitätsverlag Göttingen

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

nestor

Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Förderkennzeichen: 01 DL 001 B

Universitätsverlag Göttingen

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

D-GRID GmbH

Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme
hg. v. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig
im Rahmen des Kooperationsverbundes nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Lektorat) und Sonja Neweling (Redaktion) sowie Martina Kerzel und Lajos Herpay (Gestaltung und Montage) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte des Buches stehen auch als Onlineversion
(http://nestor.sub.uni-goettingen.de/bestandsaufnahme/)
sowie über den Göttinger Universitätskatalog
(http://www.sub.uni-goettingen.de) zur Verfügung.
Die URN lautet: http://nbn-resolving.de/urn:nbn:de:0008-2012031401.

Die digitale Version dieses Werkes steht unter einer Creative Commons Namensnennung-Nicht-kommerziell-Weitergabe unter gleichen Bedingungen 3.0 Unported Lizenz.

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.
© Verlag Werner Hülsbusch, Boizenburg, 2012
http://www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen.

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: Kunsthaus Schwanheide
Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-86488-008-7

15 Zusammenfassung und Interpretation

Jens Ludwig

In der Zusammenschau der verschiedenen Disziplinen gewinnt man den Eindruck, dass es sich mit Forschungsdaten verhält wie mit den Tieren in der von Borges beschriebenen, angeblichen chinesischen Enzyklopädie. Sie stelle fest,

> „[...] daß ‚die Tiere sich wie folgt gruppieren: a) Tiere, die dem Kaiser gehören, b) einbalsamierte Tiere, c) gezähmte, d) Milchschweine, e) Sirenen, f) Fabeltiere, g) herrenlose Hunde, h) in diese Gruppierung gehörige, i) die sich wie Tolle gebärden, k) die mit einem ganz feinen Pinsel aus Kamelhaar gezeichnet sind, l) und so weiter, m) die den Wasserkrug zerbrochen haben, n) die von weitem wie Fliegen aussehen'."[1]

Auch unter Forschungsdaten und der dazugehörigen Infrastruktur werden ganz verschiedene Dinge verstanden und ganz verschiedene Dimensionen, um Daten zu beschreiben, und Anforderungen prallen aufeinander. Diese Unterschiede können unnötig kompliziert wirken und wecken vielleicht ein Bedürfnis nach einer einzelnen, klaren Definition für die Wissenschaften insgesamt. Ein Ziel der folgenden Gegenüberstellung wesentlicher Dimensionen von Forschungsdaten, wie sie von den Autoren aus den verschiedenen Disziplinen berichtet werden, ist aber aufzuzeigen, dass diese Vielfalt zumindest überwiegend kein Defizit, Fehler oder mangelnder Entwicklungsgrad der Disziplinen ist, sondern das notwendige Resultat der Ausdifferenzierung der Wissenschaften.

Schon die elf vorgestellten wissenschaftlichen Disziplinen sind in sich komplexe Wissenschaftsbereiche. Mit vielleicht zwei Ausnahmen haben alle darauf hingewiesen, dass es sich bei ihrem Gebiet entweder um interdisziplinäre Unternehmungen handelt, in denen verschiedene Disziplinen zusammen ein Thema untersuchen (z.B. Biodiversität), oder um

1 Foucault (1974), S. 17.

eine stark ausdifferenzierte Disziplin, in der ganz verschiedene Themen untersucht werden (wie z.B. sehr plastisch bei den Geowissenschaften beschrieben). Die zwei Bereiche, bei denen dieser Aspekt nicht so stark im Vordergrund steht, sind die Teilchenphysik und die Astrophysik, die aber schon durch ihren Namen kenntlich machen, dass sie – ohne Anspruch auf wissenschaftshistorische Genauigkeit – als bereits selbstständige Ausdifferenzierungen der gemeinsamen Mutterdisziplin Physik betrachtet werden können.

Eine Schwierigkeit beim Vergleich des Umgangs mit Forschungsdaten in dieser groben Disziplineinteilung ist, dass bereits eine einzelne Disziplin sehr verschiedene Teilbereiche umfasst und es deshalb in ihr nicht nur eine, sondern viele Arten des Datenmanagements gibt. Aber auch eine engere Fassung der Disziplinen würde diese Schwierigkeit nicht beseitigen, denn Forschungsgebiete liegen immer quer zu Aufgaben des Datenmanagements. Es ist geradezu eines der Kennzeichen der Wissenschaft, dass für die Untersuchung ein und desselben Themas die Methoden, Instrumente und Anforderungen an Forschungsdaten vielfältig sind und sich als Teil des wissenschaftlichen Fortschritts verändern. Und so wie der Umgang mit Forschungsdaten in einer fast beliebig weit oder eng gefassten Disziplin vielfältig ist, so wird auch umgekehrt eine Methode und ein Umgang mit Forschungsdaten sich nur sehr selten auf eine Disziplin beschränken, sondern in einer Vielzahl von Disziplinen anzutreffen sein. Disziplinen haben zwar Besonderheiten, aber kein erfolgreicher Forschungsbereich ist so speziell, dass nicht andere die Verfahren für die Behandlung ihrer Themen übernehmen werden (z.B. die medizinische Magnetresonanztomografie zur Visualisierung des Gehirns in der Psycholinguistik) und damit die Besonderheiten der anderen Disziplin relativieren.

15.1 Kooperative Strukturen

Wissenschaft lebt vom engen Austausch mit anderen und von Kooperationen. In allen hier vertretenen elf Disziplinen gibt es institutionenübergreifende Kooperationen, wenn auch in unterschiedlichem Maße und aus unterschiedlichen Gründen. Die treibenden Faktoren sind dabei

insbesondere die Untersuchungsinstrumente und die Untersuchungsgegenstände. Am auffälligsten sind vielleicht Teilchenbeschleuniger und Teleskope, die von einzelnen Institutionen weder finanziert noch effizient betrieben werden können. Aber auch wenn keine Instrumente in Gebäudegröße benötigt werden, kann die Datenerhebung so viel Aufwand erfordern, dass er nicht mehr durch einzelne bewältigbar ist und Kooperationen notwendig sind. Dies ist zum Beispiel in den Sozial- oder Erziehungswissenschaften der Fall, deren Großumfragen ein koordiniertes Vorgehen verlangen.

Auch der Untersuchungsgegenstand allein kann ein Grund zur Kooperation sein. Die Instrumente können relativ klein, unspektakulär und von Individuen oder einzelnen Institutionen handhabbar sein, aber die Größe, Entfernung oder Verteilung des Untersuchungsgegenstands kann eine Kooperation erforderlich machen, wie z.B. in der Klimaforschung und in den Altertumswissenschaften. Und schließlich ist für einige Bereiche auch die Interdisziplinarität und Fachausdifferenzierung ein Grund, Kooperationen einzugehen, um die vielen verschiedenen notwendigen Expertisen zusammenzubringen (z.B. erwähnt bei der Biodiversität, Medizin und den Altertumswissenschaften), die nicht mehr von einzelnen Wissenschaftlern allein besessen werden können.

Dies sind primär Kooperationen zu Forschungsfragen, aber auch das Management von Forschungsdaten erfolgt arbeitsteilig. Die Sozialwissenschaften und die Klimawissenschaften sind von den hier vertretenen Disziplinen diejenigen, die diese Aufgabe am stärksten institutionell gebündelt haben. Mit GESIS und dem DKRZ existieren eigenständige Institutionen, zu deren Kernaufgaben es gehört, diese Dienstleistung anzubieten. In den Geowissenschaften, der Pädagogik und zu einem gewissen Grad auch in der Psycholinguistik gibt es mit den World Data Centers, dem DIPF und dem MPI für Psycholinguistik jeweils etablierte Institutionen, die neben ihrer Forschungstätigkeit diese Aufgabe für andere mit übernehmen. In den anderen Disziplinen wird das Forschungsdatenmanagement in Föderationen durchgeführt oder mit Einzellösungen an den Institutionen, an denen die Daten anfallen.

Ob diese Einzellösungen sinnvoll und effizient sind oder nicht, ist nicht einfach zu beurteilen. Häufig wird festgestellt, dass das Datenmana-

gement sowohl fachspezifische als auch generische Aufgaben beinhaltet,[2] wobei erstere nicht von einer übergreifenden Einrichtung ohne Fachkenntnisse kompetent übernommen werden können. Diese fachspezifischen Anteile sprechen eher für Einzellösungen. Eine übergreifende Einrichtung kann aber oftmals für generische Dienstleistungen eine Economy of Scale ausnutzen und verfügt idealerweise über Fachkenntnisse, um auch gewisse fachspezifische Dienstleistungen effizient anzubieten. In welchem Zuschnitt Zentren nun sinnvoll sind – mit welcher Fachgranularität, mit welchen Fachkenntnissen und welchen Aufgaben – oder ob eine Einzellösung besser ist, ist weder pauschal noch einfach zu klären.

Dementsprechend ist auch die Rolle der Institutionen, die Forschungsdaten managen, im Verhältnis zu denen, die sie produzieren oder nutzen, sehr unterschiedlich. Eine klassische, übergreifende Informationseinrichtung wie eine Bibliothek oder ein Archiv ist meist für mehrere andere Institutionen oder Forschergruppen da und betreibt verschiedene Informationssysteme, die wiederum jeweils mehrere Datensammlungen enthalten. In dem Bereich der Forschungsdaten kann dieses Zuordnungsverhältnis von einer Institution zu mehreren Datenbeständen und mehreren Nutzern ganz anders aussehen. Z.B. sind in den Sozialwissenschaften Forschungsdatenzentren oftmals Einrichtungen, die im Wesentlichen die Datensammlung nur einer Quelle wie beispielsweise einer Behörde pflegen. Oder im Fall der Teilchenphysik werden die Daten einer Quelle, des LHC-Beschleunigers, nicht von einer einzelnen Institution, sondern von einer ganzen Föderation aufbewahrt, bereitgestellt und analysiert.

Auch wenn Förderer und Wissenschaftsorganisationen verstärkt die Beteiligung von traditionellen, disziplinübergreifenden Informationseinrichtungen wie Bibliotheken und Rechenzentren an Kooperationen zum Management von Forschungsdaten fordern,[3] ist nicht von vornherein klar, dass diese Einrichtungen in diesen Kooperationen eine Rolle spielen. Das ist bei Bibliotheken, deren zukünftige Rolle im Bereich der digitalen Information oft als ungewisser betrachtet wird, weniger überraschend als bei Rechenzentren. In den hier vorgestellten Kooperationen zum Forschungsdatenmanagment werden Einrichtungen der Informationsinfra-

2 Vgl. z.B. KII (2011), S. B125.
3 Vgl. z.B. Kapitel 2.1.3; DFG (2009).

struktur ungefähr in der Hälfte der Fälle erwähnt (Geisteswissenschaften, Medizin, Geowissenschaften, Psycholinguistik, Pädagogik, Biodiversität). Rechenzentren und Bibliotheken werden etwa gleich häufig genannt und als Dokumentationszentrum wird das DIMDI erwähnt, das Deutsche Institut für Medizinische Dokumentation und Information. Bei den Bibliotheken handelt es sich um die größeren Einrichtungen, insbesondere die Deutsche Nationalbibliothek (DNB) und die Technische Informationsbibliothek (TIB) Hannover. Die Aufgaben, die die Infrastruktureinrichtungen übernehmen, sind meist das Hosting der Daten oder die Vergabe von persistenten Identifikatoren für Forschungsdaten, z.B. DOIs durch die TIB als Teil des DataCite-Konsortiums, URNs durch die DNB oder Handles durch die Gesellschaft für Wissenschaftliche Datenverarbeitung Göttingen (GWDG). Ob diese grundlegenden Services den disziplinübergreifenden Informationseinrichtungen dauerhaft eine feste Rolle in Kooperationen zum Forschungsdatenmanagement verschaffen und welche weiteren Services wie z.B. Beratungen sie anbieten können, ist wie viele andere organisatorische Fragen noch offen.

15.2 Daten und Metadaten

Auf die Frage, was für Typen von Forschungsdaten in einem Forschungsbereich anfallen, findet sich die eingangs angedeutete volle Vielfalt der Antworten. Die häufigsten Nennungen umfassen Video, Audio, Simulationsdaten, Fotos, quantitative/qualitative Daten, Digitalisate/Scans, Markup/Annotationen, Beobachtungsdaten, Statistiken, Dokumente, Experimentdaten, Zeitreihen und Fernerkundungsdaten. In einer groben Kategorisierung der Nennungen können zwei Hauptarten von Daten unterschieden werden, die als Forschungsdaten betrachtet werden: In ca. 60 Prozent der Nennungen werden die Forschungsdaten extrinsisch bestimmt, d.h. durch ihre Rolle in der Wissenschaft oder durch die Methode, mit der sie erzeugt oder benutzt werden, wie z.B. Simulationsdaten, Beobachtungsdaten, Experimentdaten, Zeitreihen, Interviews etc. Der interne Aufbau und das technische Format dieser unterschiedlichen Datentypen können hingegen sogar identisch sein. In weiteren ca. 30 Prozent der Fälle werden sie stattdessen intrinsisch durch den Medien-

typ charakterisiert wie z.B. Video, Audio, Markup, 3-D-Modelle etc. In diesen Fällen wird der Begriff Forschungsdaten benutzt, um einen Gegensatz zu Dokumenten auszudrücken oder zumindest zu Dokumenten, die nur als Publikation oder Artikel benutzt werden und nicht z.B. zum Protokollieren von Messdaten oder für Interviewtranskripte. Ein gewisser Prozentsatz der Nennungen lässt sich nur schwer zuordnen, wie z.B. Biomaterialdaten, die man vielleicht als durch den Untersuchungsgegenstand bestimmten Datentyp charakterisieren könnte.

Der Grund für die Vielfalt der Antworten ist, dass je nach den Untersuchungsgegenständen und den Methoden des Faches andere Unterscheidungskritieren für Forschungsdaten relevant sind. In den Naturwissenschaften kann der Unterschied zwischen Beobachtungs-, Experiment- und Simulationsdaten hinsichtlich der Kodierung und technischen Anforderungen einzelner Datensätze minimal sein, aber für die Frage der Aufbewahrungswürdigkeit und der notwendigen Hintergrundinformationen ist er entscheidend. Hingegen existieren bei sozialwissenschaftlichen Untersuchungen der Gesellschaft und des Individuums eigentlich nur Beobachtungsdaten. Die grundlegende Entscheidung, welche Aufgaben in Arbeitsabläufen des Forschungsdatenmanagements notwendig sind, wird dort an dem Unterschied zwischen quantitiven und qualitativen Daten getroffen, die z.B. in Bezug auf Datenschutz ganz anders behandelt werden müssen. In geisteswissenschaftlichen Bereichen wiederum, in denen Einzelfalluntersuchungen viel häufiger und wichtiger sind und die Heterogenität der Technologien deutlich größer ist, bilden die Medientypen – ob es sich z.B. um Fotos, Videos oder Dokumente handelt – oft die größte Möglichkeit, Daten zu kategorisieren.

Die Vielfalt dieser Charakterisierungen von Forschungsdaten macht auf die Kontextabhängigkeit und Unschärfe des Forschungsdatenbegriffs aufmerksam. Eine inhaltliche Begrenzung, was Forschungsdaten behandeln können oder was ihre Quellen seien können, macht wenig Sinn, da prinzipiell alles Untersuchungsgegenstand der Wissenschaft werden kann. Dass Daten Forschungsdaten darstellen, ist eher eine Aussage über ihre methodische Verwendung in einem bestimmten wissenschaftlichen Kontext. Wenn ein digitalisiertes Buch von einem Literaturwissenschaftler genauso wie sein analoges Pendant gelesen und analysiert wird, dann würde es üblicherweise nicht als Forschungsdatum aufgefasst werden,

nur weil es digital ist. Wenn hingegen dasselbe Buch als Teil eines großen digitalen Textkorpus von Geisteswissenschaftlern nach bestimmten Mustern und Häufigkeiten untersucht wird, dann ähnelt ihre Tätigkeit unter Umständen nicht nur oberflächlich Naturwissenschaftlern, die Messdaten auswerten, sondern sie benutzen gegebenenfalls auch dieselben statistischen Verfahren oder Mustererkennungstechnologien. In diesem Fall ist es naheliegend, das digitalisierte Buch als Teil der Forschungsdaten aufzufassen.[4]

Diese Kontextabhängigkeit hat zusätzlich eine zeitliche Dimension. Daten, die gezielt für die wissenschaftliche Auswertung erzeugt wurden, können von vornherein als Forschungsdaten bestimmt werden. Allerdings können auch Daten, die nicht für die Wissenschaft erzeugt wurden, trotzdem nachträglich zu Forschungsdaten werden, wenn z.B. erst nachträglich ein wissenschaftliches Interesse für sie entsteht und sie erst dann in dieser Rolle in der Wissenschaft genutzt werden.[5] Diese Kontextabhängigkeiten und Abgrenzungsschwierigkeiten von Forschungsdaten haben zur Konsequenz, dass Entscheidungen und Abgrenzungen allein auf der Basis, ob etwas Forschungsdaten darstellt oder nicht, problematisch sein können. Eine (fiktive) Richtlinie einer Universität, dass alle Forschungsdaten durch das Rechenzentrum aufbewahrt werden und alle anderen digitalen Ressourcen durch die Bibliothek, würde sicherlich einige willkürliche Definitionen erfordern und etliche Sonderfälle hervorbringen.

Die Formate der Forschungsdaten existieren in einer Vielzahl und Vielfalt, die kaum sinnvoll wiedergeben werden kann. Alle Fachdiszipli-

---

4 Vgl. z.B. Michel et al. (2011) für eine Verwendung eines digitalisierten Buchbestandes als Forschungsdaten.

5 Wie z.B. Schiffsbücher aus der Zeit des ersten Weltkriegs, die im Old Weather Projekt mit der Beteiligung von Internetnutzern transkribiert wurden, um die darin enthaltenen historischen Wetterbeobachtungen als Wetterdaten für die Klimaforschung nutzbar zu machen (vgl. Old Weather Homepage: http://www.oldweather.org).
Umgekehrt hat auch David Rosenthal wiederholt darauf aufmerksam gemacht, dass interessante Daten nicht gesammelt und archiviert werden, wie z.B. Internetwerbung aufgrund der hohen damit verbundenen technischen und rechtlichen Schwierigkeiten. Die wenigen Institutionen, die Internet-Archivierung betreiben, lassen die Werbung aus, die damit zukünftiger Forschung nicht zur Verfügung stehen wird. Eine ähnliche Situation ergibt sich für zentrale Dokumente des amerikanischen Präsidentschaftswahlkampfs 2008, die u.a. aus Blog-Beiträgen und YouTube-Videos bestanden (vgl. Rosenthal (2011)).

nen scheinen gemeinsam zu haben, dass über die allgemein gängigen Formate hinaus auch fachspezifische Formate existieren. Ein großer Unterschied besteht aber darin, wie die Disziplinen mit der Formatvielfalt und Heterogenität umgehen. Grob können vier Ansätze unterschieden werden: 1. Die Formate sind durch eine Richtlinie eingeschränkt, 2. die Formate sind faktisch eingeschränkt, 3. sie sind faktisch nicht eingeschränkt oder 4. sie können nicht eingeschränkt werden. Im ersten Fall gibt es für die Forschungsdaten explizite Formatvorgaben, die durch die Institution vorgegeben werden. Im zweiten Fall der faktischen Einschränkung gibt es keine Vorgaben durch die datenmanagende Institution, aber die benutzten Forschungsinstrumente liefern nur bestimmte Formate, zum Teil auch weil die Community sich schon auf einen Standard geeinigt hat. Dies ist insbesondere bei Disziplinen der Fall, die auf einzelne Großforschungsinstrumente angewiesen sind, die von einer Vielzahl von Forschern benutzt werden (Teilchenphysik, Astronomie). Wenn hingegen die Formatvielfalt in einer Disziplin nicht eingeschränkt wird, dann kann dies darauf zurückgeführt werden, dass eine Standardisierung einfach noch nicht stattgefunden hat (3.) oder dass sie als prinzipiell nicht durchführbar betrachtet wird. In sehr interdisziplinären Forschungsbereichen, in Forschungsbereichen mit vielen Innovationen und einem dementsprechenden Bedarf an neuen Formaten oder bei der Neuetablierung von Forschungsarchiven, die überhaupt erst einen Bestand aufbauen müssen, können hohe Formatanforderungen so hinderlich sein, dass Standardisierung als prinzipiell nicht durchführbar angesehen wird (4.).

Ähnlich wie bei den Datenformaten existiert auch bei den Metadatenformaten eine Flut von Standards. Jede Disziplin hat ihre eigenen Metadatenformate und viele basieren auf XML. Eine Besonderheit beim Vergleich der Disziplinen ist, dass gerade in der Biodiversität und Archäologie, in denen ein mangelndes Bewusstsein der Wissenschaftler für die Bedeutung von Standards konstatiert wird, nicht keine, sondern eine Vielzahl von Metadatenformaten existieren. Allerdings sind dies auch Gebiete, die sehr interdisziplinär sind, die eine Vielfalt von Untersuchungstechniken kennen und für deren Beschreibung deshalb eine Vielzahl von Metadatenformaten vielleicht auch notwendig ist.

Eine fehlende Standardisierung von Dateiformaten und Metadatenformaten muss nicht in jedem Fall dramatisch und ein Zeichen eines

Mangels sein. Wie oben angedeutet, kann eine Standardisierung eine Einschränkung bedeuten, die für manche Bereiche prinzipiell oder noch nicht angemessen ist. Schließlich bedeutet Wissenschaft gerade auch neue Sachen zu tun, die kein Standard sind. Für die Erhaltung der Forschungsdaten auf einer basalen Ebene der *bitstream preservation* stellen mangelnde Standards auch kein Problem dar. Die Integrität der Daten kann unabhängig von Datei- und Metadatenformaten sichergestellt werden. Eine Schwierigkeit stellen diese Formate erst auf den Ebenen der technischen und der inhaltlichen Nachnutzbarkeit dar. Eine Begrenzung der Dateiformate und der Metadatenformate ist sinnvoll, um die Anzahl der technischen Umgebungen (wie z.B. Hardware und Software) und die zur Interpretation der Daten notwendigen Hintergrundinformationen zu reduzieren, die Datenarchive beobachten müssen, wenn sie Nachnutzbarkeit in einem anspruchsvollen Sinne anbieten wollen. Dabei geht es aber nicht um das Veralten von Formaten, das fälschlicherweise bzw. in zu undifferenzierter Weise als ein Grundproblem der Langzeitarchivierung angesehen wird. Denn moderne Datei- und Metadatenformate veralten nur selten in dem absoluten Sinne, dass nirgends mehr eine technische Umgebung und Dokumentation für sie zu finden ist. Mit gewissem Aufwand können deshalb auch Daten und Metadaten genutzt werden, die in einem Fachgebiet mit stärkerer Standardisierung nicht vorkämen. Viel bedeutsamer ist das Veralten von Datei- und Metadatenformaten, das relativ zu den technischen und inhaltlichen Anforderungen der Zielgruppen ist. Es mag zwar alte Software und Emulatoren geben, mit denen die Daten im Prinzip, aber nicht mehr entsprechend der Anforderungen der Zielgruppen nutzbar sind und die deshalb ineffizient oder de facto nutzlos sind. Um eine technische und inhaltliche Nachnutzbarkeit von Forschungsdaten entsprechend dieser Anforderungen zu gewährleisten, müssen eine fortlaufende Technologie- und Zielgruppenbeobachtung und daraus resultierende Anpassungen stattfinden. Und genau um diese anspruchsvolle Nachnutzbarkeit effizient zu gewährleisten und die Anzahl der zu beobachtenden und unterstützenden Technologien und Metadaten der Zielgruppen zu reduzieren, sind Standards wichtig.

Entsprechende Anforderungen an die Aufnahme von Forschungsdaten in ein Datenzentrum werden anscheinend nur selten explizit gestellt. Es gibt zwar zum Teil Formatvorgaben (z.B. Klimaforschung) und ein

allgemeines Bewusstsein der Bedeutung von offenen Formaten, aber für viele Bereiche scheinen sich diese Anforderungen dadurch zu erledigen, dass es Standardformate gibt, die ohnehin von den Forschenden verwendet werden und werden müssen, da alle Software und Instrumente darauf basieren. Vereinzelt wird von zusätzlichen Maßnahmen zur Qualitätskontrolle berichtet, wie Plausibilitätstests oder Tests auf Vollständigkeit der Metadaten.

Aufgrund ihrer stärkeren Standardisierung und Homogenität tendieren auch gerade große Datenmengen aus einer Massenproduktion dazu, sehr viel einfacher verarbeitbar zu sein. Problematischer als die Byte-Größe eines einzelnen Datensatzes, die primär eine technische Herausforderung ist, kann die Anzahl der Datensätze/Datenobjekte sein, die als logische Verwaltungseinheiten mehr Aufwand erzeugen. Darüber hinaus lässt sich weder über Forschungsbereiche hinweg, noch für einzelne Forschungsbereiche als Ganzes etwas Allgemeines über das Forschungsdatenvolumen aussagen. Die Volumina sind oft nur für ein Projekt und Instrument angebbar und rangieren dann zum Teil im Tera- bis Petabytebereich pro Jahr.

Dass ältere Forschungsdaten zur Nachnutzung zur Verfügung stehen, ist für alle Disziplinen mit wenigen Einschränkungen von hoher Bedeutung. Dies gilt insbesondere für Disziplinen, in denen die Wissenschaftler z.T. langfristige Beobachtungen durchführen und die Zustände und Veränderungen der Umwelt, des Weltraums oder der Gesellschaft untersuchen. Die Beobachtungsdaten eines Ereignisses und die Messwerte zu einem bestimmten Punkt einer Zeitreihe sind bei Verlust nicht wieder herstellbar. Ein ähnlicher Fall liegt in der Archäologie vor, in der es zwar nicht um die Beobachtung zu einem bestimmten Zeitpunkt geht, dessen Bedingungen danach nicht wiederkehren, aber bei der durch die destruktive bzw. verändernde Untersuchung der Untersuchungsgegenstand danach nicht mehr in seiner ursprünglichen Form existiert.

Ganz anders kann dies bei Messdaten aus Laborexperimenten sein. Die Geowissenschaften und die Psycholinguistik konstatieren für solche Daten, dass sie zum Teil schnell veralten, weil sie mit verbesserter Messtechnologie in höherer Präzision wieder erzeugt werden können. Bei diesen und anderen reproduzierbaren Daten werden die Kosten der Aufbewahrung und die Kosten der erneuten Erzeugung zueinander ins Verhältnis gesetzt werden

müssen. Solche Daten werden unter Umständen nur für einen begrenzten Zeitraum für den Nachweis der korrekten wissenschaftlichen Arbeit aufbewahrt. Allerdings kann die Wiederholbarkeit von experimentellen Messungen mit Großinstrumenten eher theoretischer Natur sein. Die Beschleunigerexperimente der Teilchenphysik sind im Prinzip wiederholbar, aber wenn keiner der aktuell betriebenen Teilchenbeschleuniger das Experiment durchführen kann, ist die Wiederholbarkeit aus praktischen und finanziellen Gründen ausgeschlossen. Mit der Verschiebung der Forschungsfragen in der Teilchenphysik ist deshalb auch die Aufbewahrung prinzipiell reproduzierbarer Forschungsdaten wichtig.

Für wen die Forschungsdaten vorgehalten werden, reicht von einer nur forschungsgruppeninternen Verwendung (wie in der Teilchenphysik und Medizin) über eine Bereitstellung für Gruppen, die ein berechtigtes wissenschaftliches Interesse nachweisen können (z.B. Sozialwissenschaften, Pädagogik), bis hin zu überwiegend öffentlichen Datenpublikationen (Geowissenschaften, Klimaforschung). Die Bereitstellung erfolgt über zwei Wege. Zum einen über die Institution oder Föderation, die die Daten auch dauerhaft managt (GESIS/Forschungsdatenzentren, WDC, offene Kooperationen/Föderationen) und sie über Portale und Datenbanken anbietet. Technisch handelt es sich dabei um Einzellösungen und Standards werden abgesehen von der OAI-PMH Metadatenschnittstelle nicht erwähnt. Ein anderer, mehrfach erwähnter Verbreitungsweg sind Verlage und Datenpublikationen.

Die Zielgruppen und Bereitstellungswege sind eng verknüpft mit der Frage der Kontrolle der Daten und der Rechte an ihnen. Die Nachnutzung von Daten unterliegt ungefähr bei der Hälfte der Antworten Einschränkungen. Der häufigste Grund dafür ist, dass es sich um sensible Daten handelt, weil sie dem Datenschutz unterliegen. Dies stellt bei der Medizin, den Sozialwissenschaften und auch der Pädagogik keine Überraschung da, da sie direkt den Menschen untersuchen. Aber auch in Disziplinen wie der Biodiversität, in denen es vielleicht zuerst nicht erwartet wird, gibt es sensible Daten, wie z.B. Brutstätten von gefährdeten Arten. Die Lösungsansätze, die für sensible Daten verfolgt werden, schränken die Zielgruppen mit Authentifizierungs- und Autorisierungsmechanismen ein und benutzen ausgeklügelte Anonymisierungs- und Pseudonymisierungsverfahren (vgl. insbesondere Medizin).

Neben der Sensibilität der Daten sind die Datenproduzenten und ihr Anspruch auf priviligierten Zugang zu den Daten ein Hauptgrund für Nutzungseinschränkungen. Information als digitales Gut hat den Vorteil, dass sie sich nicht durch Nutzung verbraucht und es im Prinzip reicht, wenn eine Institution oder Person Forschungsdaten erzeugt und erhält, damit beliebige weitere Personen die Daten effizient nutzen können. Dies ermöglicht es einerseits, relativ einfach Open Access anzubieten, wie es z.B. bei den Weltdatenzentren (Geowissenschaften, Klimawissenschaften) Teil der Satzung ist. Andererseits hätten Produzenten wenig Anreiz den Aufwand zu übernehmen, wenn sie dadurch keinen Vorteil hätten und es keine Beteiligung oder Anerkennung durch andere Nutzer gibt. Aus diesem Grund haben sich eine Reihe von Regelungen in den verschiedenen Disziplinen etabliert. Erstverwertungsrechte und Moving Walls, die den Produzenten für eine gewisse Zeit die alleinige Nutzung garantieren, werden z.B. in den Geowissenschaften und der Biodiversität eingesetzt. In manchen Disziplinen wie der Psycholinguistik ist je nach Datenbestand auch die explizite Einholung einer Nutzungserlaubnis durch den Produzenten notwendig. Das universelle Mittel, mit dem die Nutzer dem Produzenten die wissenschaftliche Anerkennung für seine Leistungen ausdrücken, ist aber das Zitat. Die Bemühungen, das Zitieren von Forschungsdaten zu etablieren, sind deshalb wichtige Beiträge für ein sinnvolles Management von Forschungsdaten, da sie einen Anreiz dafür schaffen können.

Ein zentrales Mittel, um Forschungsdatenzitate zu ermöglichen, sind Persistent Identifier. Bereits nicht-persistente Identifikatoren sind ein wichtiges Instrument für das Datenmanagement. Da die Identifikation stabile und eindeutig abgrenzbare Datenobjekte voraussetzt, werden mit der Entwicklung eines Konzepts für Identifikatoren bereits eine Reihe von wichtigen Themen des Datenmanagements wie z.B. die genaue Definition der Objekte berührt.[6] Persistent Identifier haben zusätzlich den Anspruch langfristig stabil zu sein und damit dauerhaft gültige Zitate zu ermöglichen, auch wenn sich der Speicherort der Daten ändert. Die verschiedenen Fachdisziplinen erwähnen drei Ansätze für Persistent Identifier, von denen zum Teil auch mehrere in derselben Disziplin zum

6 Vgl. PILIN (2008).

Einsatz kommen: DOIs (Klimaforschung, Medizin, Pädagogik, Biodiversität, Sozialwissenschaften), die durch den DataCite-Verbund vergeben werden, der sich besonders um Standards für Forschungsdatenzitate bemüht, Handles (Geisteswissenschaften, Psycholinguistik), die die technische und syntaktische Grundlage von DOIs bilden und durch das EPIC-Konsortium für Forschungsdaten vergeben werden, und URNs (Geisteswissenschaften, Klimaforschung), die eine Untermenge der im Internet verwendeten URIs sind und insbesondere im Bibliotheksbereich verbreitet sind.

15.3 Interne Organisation

Feste Regeln und Prozesse zur Langzeitarchivierung sind nur zum Teil bereits etabliert und oft noch in der Entwicklung. Standardverfahren sind insbesondere in der Medizin vorhanden, deren Umgang mit Daten durch eine Vielzahl von gesetzlichen Vorgaben bestimmt wird. Verallgemeinert kann festgestellt werden, dass Arbeitsabläufe in Einrichtungen etabliert sind, die das Datenmanagement für Disziplinen oder Teilbereiche zentral übernehmen (wie z.B. die Weltdatenzentren in der Klimaforschung). Dies liegt vermutlich auch darin begründet, dass etablierte Prozesse zu haben Teil zu zentralen Organisationen dazugehört.

Eine andere Voraussetzung für ein Datenarchiv ist die Finanzierung. Sie ist nur teilweise durch eine institutionelle Grundfinanzierung abgesichert (Sozialwissenschaft, Klimaforschung, Geowissenschaft, Psycholinguistik) und beruht häufig noch auf Projektbasis (Geisteswissenschaften, Biodiversität, Medizin, Teilchenphysik), wobei es sich aber in einigen Fällen um sehr langfristige Projekte handelt, die eine gewisse Sicherheit schaffen. Projektmittel haben im Finanzierungskonzept der etablierten Datenarchive allerdings auch einen sinnvollen Platz, wo sie wirklich für vorübergehende Aufgaben benutzt werden, wie z.B. um die Weiterentwicklung von Services (Psycholinguistik) oder die einmalige Übernahme von sehr großen oder komplexen Datenbeständen (Geowissenschaften, Klimaforschung) zu finanzieren. Die Kostenhöhe wird von kaum einer Disziplin benannt. Als Personalaufwand werden beispielsweise für das zentrale Datenmanagment des Schwerpunktprogramms „Biodiversitäts-

Exploratorien“ zweieinhalb Vollzeitbeschäftigte angegeben, die Betrieb, Beratung und Weiterentwicklungen gewährleisten, aber keine inhaltliche Qualitätskontrolle realisieren können. Monetäre Kosten werden nur von der Psycholinguistik benannt, wobei knapp hunderttausend Euro jährlich für Technik und rund dreihunderttausend Euro für Personal, das das System betreibt und die Software pflegt, anfallen. Dieses Verhältnis von Technik zu Personalkosten bewegt sich damit durchaus im üblichen Rahmen, der von der Literatur mit einem Personalanteil von 70 Prozent und mehr an den Gesamtkosten benannt wird.[7] Die Gesamtkosten sind recht gering, wenn sie an den typischen Betriebskosten für Datenzentren in Höhe von 3,5 Mio. Euro pro Jahr gemessen werden, die die Kommission „Zukunft der Informationsinfrastruktur“ in ihrem „Gesamtkonzept für die Informationsinfrastruktur in Deutschland“ nennt.[8]

Angesichts des hohen Personalanteils an den Gesamtkosten ist es nicht verwunderlich, dass die personelle Lage der finanziellen ähnelt. Zwar gibt es meist Personal, das sich schwerpunktmäßig mit der Langzeitarchivierung befasst, aber überwiegend wird es aus Projekten bezahlt und befristet beschäftigt. Als eine Ausnahme sind hier die Sozialwissenschaften und die Psycholinguistik zu nennen, in denen zumindest teilweise unbefristetes Personal vorhanden ist. In allen Disziplinen hat das Personal meistens seine Qualifikation praktisch und nicht durch eine Ausbildung erworben.

Der Bereich Organisation bietet insgesamt ein schwaches Bild. Auch die größeren Ansätze und Entwicklungen, die organisatorische Aspekte behandeln, wie z.B. die Kriterienkataloge für vertrauenswürdige digitale Archive[9], das Data Seal of Approval[10] oder der nestor Ingest-Leitfaden[11] werden sehr selten oder gar nicht erwähnt. Zu der Kosten- und Finanzierungsfrage sind ebenfalls eine Reihe von Studien veröffentlicht worden, die die methodischen Grundlagen und eine Vielzahl von Fallstudien liefern. Vor diesem Hintergrund scheint sich zu zeigen, dass die Schwierig-

7 Vgl. KRDS (2010), S. 14.
8 Vgl. KII (2011), S. B122.
9 Vgl. RLG-NARA (2007); nestor (2008a).
10 Vgl. Data Seal of Approval (2011b).
11 Vgl. nestor (2008b).

keit, die organisatorischen Aspekte und die Kosten für das Management von Forschungsdaten zu klären, nicht aus mangelndem theoretischem und methodischem Wissen resultiert. Dieses Wissen und die Technik sind bedeutsamer für den Bereich Daten und Metadaten. Die Schwierigkeiten der Organisation und Kosten liegen hingegen in der praktischen Durchführung in der jeweiligen konkreten Situation.

15.4 Perspektiven und Visionen

Es ist bemerkenswert, dass die Bedeutung von Forschungsdaten in allen Fachdisziplinen sehr betont wird, aber sehr viele Fragen noch offen sind, von denen die technischen die geringsten sind. Unter den besonderen Herausforderungen, mit denen sich die Disziplinen konfrontiert sehen, gibt es grob drei gemeinsame Themengruppen. Die erste Gruppe von Herausforderungen betrifft die Vermittlung der Bedeutung und des Nutzens des Forschungsdatenmanagements. Viele Disziplinen sehen sich bei den Wissenschaftlern mit einem geringen Bewusstsein für den Wert des Archivierens und Teilens von Forschungsdaten (z.B. Geistes- und Sozialwissenschaften) konfrontiert und wollen die Nachnutzung der vorhandenen Bestände verbessern. So sollen Daten der Klimaforschung auch für die Wirtschaft (z.B. Tourismusbrache) nutzbar gemacht werden. Komplementär zum Bewusstsein für die Aufgabe sehen die Disziplinen in der Wahrnehmung der Archive eine Reihe weiterer Herausforderungen: Um von den Wissenschaftlern als verlässliche und wertvolle Institution wahrgenommen zu werden, müssen Fragen des Umgangs der Archive mit den Forschungsdaten geklärt und kommuniziert werden. Sind die Archive vertrauenswürdig? Werden die intellektuellen Leistungen und Rechte der Wissenschaftler von den Archiven berücksichtigt (Medizin)? Ist der notwendige Persönlichkeitsschutz gewährleistet und im Rahmen von Zertifizierungsverfahren geprüft (Pädagogik)? Als dritte Gruppe von Aufgaben wird schließlich die qualitative und quantitative Verbesserung des Datenbestands benannt, wie z.B. die Ausweitung der Langzeitarchivierung auf die zur Datennutzung notwendige Verarbeitungssoftware (Teilchenphysik) oder die Weiterentwicklung von Metadatenstandards und Datenqualität (Pädagogik, Archäologie, Biodiversität).

Die praktischen Möglichkeiten, die gesehen werden, die Archivierung und Nachnutzung von Forschungsdaten zu fördern, können in einem weiteren Sinne disziplinübergreifend als eine stärkere Integration des Forschungsdatenmanagements in die wissenschaftlichen Arbeitsabläufe beschrieben werden. Dies reicht von der personellen Unterstützung der Forscher durch Datenmanagementspezialisten (Sozialwissenschaften), über die technische Einbindung einzelner Datendienste wie z.B. automatische Qualitätskontrollen und Datenrepositorien in Arbeitsabläufe (Geisteswissenschaften, Biodiversität) bis hin zur Schaffung virtueller Forschungsumgebung (Geowissenschaften). Eine langfristigere Möglichkeit und Notwendigkeit, die gesehen wird, ist es, das Thema Langzeitarchivierung in die wissenschaftliche Ausbildung mit einzubringen und schon dort das Bewusstsein dafür zu schaffen (Klimaforschung, Sozialwissenschaft, Teilchenphysik).

Über das letztendliche Ideal für das Management von Forschungsdaten herrscht angesichts der heterogenen Ausgangssituation erstaunlich große Einigkeit und Klarheit. Als Ideal wird die Existenz von fachlichen Kompetenzzentren für Forschungsdaten betrachtet, die zentral oder in einem dezentralen Verbund die Langzeitarchivierung von Forschungsdaten, Entwicklung von Standards und Beratung von Wissenschaftlern übernehmen (Sozialwissenschaften, Geisteswissenschaften, Teilchenphysik, Altertumswissenschaften, Psycholinguistik, Pädagogik, Biodiversität). Wie oben ausgeführt, wird es keine einfache Aufgabe sein, den Zuschnitt solcher Zentren zu bestimmen und zu entscheiden, wie viel von diesem disziplinübergreifenden Ideal disziplinspezifischer Zentren mit einer disziplinübergreifenden Infrastruktur realisiert werden kann. Es gibt aber keinen Zweifel, dass Zentren als der Weg angesehen werden, um die Verfügbarkeit und effiziente Nutzung von Forschungsdaten zu verbessern.