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

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 9
Geowissenschaften

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

9 Geowissenschaften

Jens Klump

9.1 Einführung in den Forschungsbereich

Forschungsdaten aus den Geowissenschaften sind so vielfältig wie das Forschungsgebiet selbst. Die Geowissenschaften betrachten die feste Erde und die Prozesse an ihrer Oberfläche. Diese Vorgänge stehen gleichzeitig auch in Wechselwirkung mit Vorgängen in der Biosphäre, Atmosphäre, Hydrosphäre und Kryosphäre. Die räumlichen Dimensionen reichen über mehrere Größenordnungen von 10^{-10} m (Atome) bis 10^{12} m (Sonnensystem), die zeitlichen Dimensionen reichen von 10^{-15} s (molekulare Wechselwirkungen) bis 10^{17} s (Alter des Sonnensystems).

Auch wenn innerhalb einer Teildisziplin eine weniger große Spannweite in den räumlichen und zeitlichen Dimensionen abgedeckt wird, so können durch die langen Zeiträume geologischer Prozesse innerhalb einer Fragestellung durchaus große Spannen innerhalb einer Dimension abgedeckt werden. So lässt sich z.B. die Bewegung der Kontinente (10^{-7} cm/s) in Bezug auf den stellaren Bezugsrahmen messen, in dem man die Verschiebung der Laufzeitunterschiede (10^{-15} s) von Quasar-Signalen aus Radioteleskopen auf unterschiedlichen tektonischen Platten auswertet. Die Auswirkungen der Plattentektonik, die der Bewegung der Kontinente zu Grunde liegt, machen sich jedoch erst nach Millionen von Jahren (10^{13} s) bemerkbar.

Die Plattentektonik wiederum ist der Antrieb für die Auffaltung der über Millionen Jahre abgelegten Gesteinsschichten zu Gebirgen. Die mechanischen Prozesse der Verformung wiederholen sich in ihren räumlichen Dimensionen von der Verformung ganzer Gesteinsschichten bis hinunter zur Verformung des Kristallgitters. Auch die Erforschung des gesamten Planeten, seiner Entstehung, seines Trabanten und seiner Nachbarn erstreckt sich über ein räumlich sehr ausgedehntes Forschungsgebiet. Ein Forschungsgebiet mit solch einem

„universellen“ Anspruch, wie es hier der Fall ist, weist viele Überschneidungen mit anderen Disziplinen auf. Ein Kapitel über Forschungsdaten in den Geowissenschaften kann daher nur exemplarisch einige Bereiche beleuchten.

Forschung in den Geowissenschaften ist in vielen Fällen dadurch gekennzeichnet, dass die untersuchten Phänomene episodischer Natur sind. Das Phänomen, z.B. ein Erdbeben, ist nur in dem einen Moment messbar und dann nie wieder. Oder das „Experiment“ hat in der Natur bereits stattgefunden und muss von den Forschern im Gelände gefunden, untersucht und Proben davon genommen werden. Diese Orte sind oft in schwer zugänglichen Regionen der Erde und müssen durch Expeditionen oder wissenschaftliche Tiefbohrungen erschlossen werden. Der große Aufwand, der notwendig ist um die benötigten Daten zu gewinnen, führt dazu, dass auch diese Daten sehr wertvoll sind, selbst wenn sie prinzipiell ein weiteres mal gewonnen werden könnten. Auf Grund des hohen Aufwands und der komplexen Vernetzung der Fragestellungen und angewandten Methoden werden Forschungsvorhaben in den Geowissenschaften daher oft in Forschungsverbünden durchgeführt.

Der große Aufwand zur Gewinnung der Daten und die oft episodische Natur der beobachteten Phänomene machen die gewonnenen Daten für die gemeinsame Nutzung in Forschungsverbünden und für eine spätere Nachnutzung wertvoll. Schon sehr früh wurde erkannt, dass Strukturen für die Bereitstellung und Langzeiterhaltung der Daten notwendig sind. Bereits für das Internationale Geophysikalische Jahr (IGY, 1957 bis 1958) wurden sog. „World Data Center“ eingerichtet, um die Daten des IGY bereitzustellen und zu erhalten.[1] Auf nationaler Ebene wurden bereits in den 1990er-Jahren die Anforderungen an das Datenmanagement und an kooperative Strukturen beschrieben.[2] In der Zwischenzeit haben sich die Strategien und Techniken für den Umgang mit Forschungsdaten weiterentwickelt. In einem Kapitel ihrer Strategieschrift „Erde im Wandel“ beschreibt die Geokommission der Deutschen Forschungsgemeinschaft (DFG) exemplarisch für die Geowissenschaften eine Strategie für den Umgang mit Forschungsdaten in den Geowissenschaften.[3]

1 Vgl. Pfeiffenberger (2007).
2 Vgl. z.B. Berndt; Fuchs; Vatterott (1997); Diepenbroek; Grobe; Reinke (1997); Lautenschlager; Reinke (1997).
3 Vgl. Wefer et al. (2010).

9.2 Kooperative Strukturen

Die breite Fächerung der Fragestellungen und angewendeten Methoden spiegelt sich auch in der Vielfalt der Strukturen geo- und umweltwissenschaftlicher Projekte wider. Dabei sind die Geowissenschaften, verglichen mit anderen Disziplinen, häufiger in großen nationalen und internationalen Verbundprojekten zusammengeschlossen. Internationale Kooperationen spielen dabei eine wichtige Rolle und gehen, wie eingangs bereits am Beispiel des IGY beschrieben, bis in die 1950er-Jahre zurück. Das daraus entstandene System der World Data Center (WDC) hat sich bis heute erhalten und befindet sich derzeit in einer Modernisierungsphase zum World Data System (WDS). In Deutschland befinden sich drei World Data Center (WDC-MARE[4], WDCC[5], WDC-RSAT[6]).

Im Jahr 1968 begann das Deep Sea Drilling Project (DSDP), zunächst als Programm der Vereinigten Staaten, zur Erkundung des Ozeanbodens durch wissenschaftliche Bohrungen. Auf das DSDP folgte das Ocean Drilling Program (ODP) mit internationaler Beteiligung und seit 2003 das Integrated Ocean Drilling Program (IODP) als gemeinschaftliches Programm, das von den Vereinigten Staaten, Japan und europäischen Staaten durchgeführt wird. Das IODP endet 2013, ein Nachfolgeprogramm ist in Vorbereitung. Für diese Projekte wurde eine gemeinsame Infrastruktur für Forschungsdaten angelegt, die Daten des IODP und seiner Vorgänger können über das Portal SEDIS[7] abgerufen werden. Auch im nationalen Rahmen wurde schon relativ früh damit begonnen, Daten zentral zu sammeln und für die Forschung bereitzustellen.[8] Als Pendant zu den Bohrprogrammen im Ozean wurde 1993 das International Continental Scientific Drilling Program (ICDP)[9] initiiert. Es koordiniert und unterstützt wissenschaftliche Bohrungen an Schlüsselstellen des Erdsystems. Auf nationaler Ebene wurde das Deutsche Forschungsbohrkonsortium GESEP e.V. gegründet, um wissenschaftliche Bohrprojekte zu koordinieren.

---

4 Vgl. WDC for Marine Environmental Sciences (2011).
5 Vgl. WDC for Climate (2011).
6 Vgl. WDC for Remote Sensing of the Atmosphere (2011).
7 Vgl. SEDIS (2011).
8 Vgl. Berndt; Fuchs; Vatterott (1997).
9 Vgl. ICDP (2011).

Langfristige kooperative Strukturen wurden auch in anderen Teilen der Geowissenschaften aufgebaut. Die Seismologie misst mittels eines weltweiten Netzwerks von Seismometern Erdbeben, um daraus neue Erkenntnisse über die Entstehung von Erdbeben zu erlangen, aber auch um durch Erdbebenwellen aus sehr weit entfernten Bebenherden Aufschlüsse über die innere Struktur des Erdkörpers zu bekommen.

Vergleichbar langfristig sind auch die kooperativen Strukturen in der Geodäsie und Geomagnetik. Schon seit Ende des 19. Jahrhunderts werden in vielen Regionen der Welt geophysikalische Observatorien betrieben, in denen langfristige Veränderungen des Erdschwere- und -magnetfeldes gemessen werden. Diese bodengebundenen Stationen wurden in den letzten Jahrzehnten um luft- und satellitengestützte Systeme als „virtuelle Observatorien" ergänzt.[10] Die so gewonnenen Daten werden an mehreren Stellen archiviert und über Portale zur Verfügung gestellt.

Auch wenn kooperative Strukturen in den Geowissenschaften eine wichtige Rolle spielen, so decken die vorhandenen Dateninfrastrukturen längst nicht alle Bereiche dieser Strukturen ab. Außerhalb kooperativer Strukturen gibt es zwar Angebote, die aber bisher nur in geringem Umfang von Forschern zur Sicherung und Veröffentlichung ihrer Daten in Anspruch genommen werden.

LZA-Dienste in den Geowissenschaften

Um Forschungsdaten aus den Geowissenschaften nachnutzbar zu machen, wurden bereits in den 1990er-Jahren, zusätzlich zum eher als Archiv konzipierten WDC-System, zentrale disziplinäre Datenportale aufgebaut. Eine wichtige Rolle in den beiden bereits skizzierten Komplexen des WDS und der Meeresgeologie spielt PANGAEA/WDC-MARE[11], das Mitte der 1990er-Jahre in einer Zusammenarbeit zwischen dem Alfred-Wegener-Institut für Polar- und Meeresforschung und dem Fachbereich Geologie der Universität Bremen aufgebaut wurde. Seit 2001 hat der offen zugängliche Teil des PANGAEA den Status eines ICSU WDC als World Data Center for Marine and Environmental Sciences (WDC-MARE).

10 Vgl. z.B. Mandea; Olsen (2006).
11 Vgl. PANGAEA (2011b).

Auch andere deutsche geowissenschaftliche Großforschungsreinrichtungen, wie zum Beispiel das Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum (GFZ)[12] und das Leibniz-Institut für Meereswissenschaften (IFM-GEOMAR)[13] betreiben Datenarchive und -portale und bieten diese Forschern aus anderen Institutionen zur Nutzung an. Zwischen den deutschen ICSU WDC, dem GFZ und dem IFM-GEOMAR besteht eine enge Zusammenarbeit bei der Entwicklung von Werkzeugen und Diensten für Erfassung, Archivierung, Nachweis und Vertrieb von Forschungsdaten.[14] Dieser Verbund von Datenzentren hat das Potenzial ein künftiger nationaler Dienstleister für die Langzeitarchivierung geowissenschaftlicher Forschungsdaten zu sein.

Zusammenarbeit mit Infrastruktureinrichtungen

Die bisher in den Geowissenschaften etablierten Datenarchive sind aus Projekten in Zusammenarbeit mit Rechenzentren hervorgegangen. In einigen Fällen wurde die Zusammenarbeit mit Bibliotheken gesucht, da der Wert der Bereitstellung von Daten als Veröffentlichung bereits früh erkannt wurde und man diese als neue Form der Veröffentlichung etablieren wollte. Ein Ergebnis aus dieser Zusammenarbeit ist DataCite[15] an der TIB Hannover, das im späteren Verlauf des Kapitels noch vorgestellt wird. Einige Bibliotheken weisen bereits Datenveröffentlichungen in ihren Katalogen nach.[16] Die Entwicklung von gemeinsamen Dienstleistungsangeboten von Bibliotheken und Datenarchiven für Datenproduzenten steht noch ganz am Anfang.

12 Vgl. GFZ (2011).
13 Vgl. IFM-GEOMAR (2011).
14 Vgl. Lautenschlager et al. (2005).
15 Vgl. DataCite (2011).
16 Vgl. z.B. TIBORDER (2011); ALBERT (2011).

9.3 Daten und Metadaten

Typen der Datenherkunft

Grundsätzlich lassen sich in den Geowissenschaften drei Typen von Datenproduktion unterscheiden:

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

Diese drei Bereiche unterscheiden sich in ihren Datenvolumina und -strukturen stark voneinander.

Im Bereich der Großgeräte und Sensorsysteme fallen zum Teil Datenmengen von mehreren Terabyte pro Jahr an. Die Daten werden in automatisierten Abläufen weiterverarbeitet und bereitgestellt. Die Strukturen dieser Datenbestände sind in sich meist homogen mit standardisierten Daten- und Metadatenformaten. Dieser Bereich ähnelt damit anderen Bereichen des *Big Science*, wie z.B. der Klimaforschung. Im Bereich der Großgeräte und Sensorsysteme liegen Forschungsdaten und Metadaten im Allgemeinen in standardisierten Formaten vor. Weit verbreitete Formate sind z.B. SEED in der Seismologie, netCDF in der Fernerkundung und Erdsystemforschung, GeoTIFF in der Fernerkundung, im Zusammenhang mit geografischen Informationssystemen hat sich das Shape-Format als ein *de facto* Standard etabliert. Die semantisch homogenen Datenstrukturen im Bereich der Großgeräte, Sensorsysteme und Modellierung begünstigen die Verwendung von standardisierten Formaten. Wie in anderen Bereichen des Big Science wird auch hier beobachtet, dass die Kapazität zur Erzeugung neuer Daten schneller wächst als die Möglichkeit, diese längerfristig zu speichern.

Im Bereich der individuell hergestellten Forschungsdaten fallen nur vergleichsweise geringe Datenmengen an, dafür sind die Stückkosten zur Herstellung der Datensätze sehr hoch. Die Datenstrukturen, Metadaten und Arbeitsabläufe orientieren sich an den individuellen Anforderungen

der Projekte. Standardisierte Datenformate finden daher kaum Anwendung, da die heterogenen Projekte untereinander semantisch inhomogene Strukturen bedingen. Die erwarteten Steigerungsraten sind geringer als die Zunahme der Kapazität der Speichermedien.

Datenveröffentlichung

Im Jahr 2000 wurde im Rahmen des deutschen Beitrags zu CODATA ein Projekt initiiert, das ein Konzept erarbeiten sollte, wie wissenschaftliche Daten publiziert und damit zitierbar gemacht werden können. Im Rahmen des DFG-Projekts „Publikation und Zitierbarkeit wissenschaftlicher Primärdaten" (STD-DOI)[17] wurden ein Konzept und eine technische Infrastruktur aufgebaut, um Daten mittels Digital Object Identifier (DOI) eindeutig identifizierbar und damit auch zitierbar zu machen.[18] Aus diesem Projekt ist 2009 DataCite e.V. hervorgegangen, das als Verbund von Großbibliotheken die Strukturen für die Veröffentlichung und Zitierbarkeit von Forschungsdaten betreibt und weiterentwickelt.

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

Zwischen geowissenschaftlichen Datenzentren, Herausgebern von Fachzeitschriften und Verlagen finden regelmäßig Treffen statt, um sich über die Anforderungen und Vorgehensweise abzustimmen. Auf gemeinsamen Veranstaltungen auf internationalen Konferenzen werden die Konzepte und Angebote zur Veröffentlichung von Forschungsdaten und deren Verknüpfung mit wissenschaftlichen Veröffentlichungen den Fachwissenschaftlern vorgestellt.

17 Vgl. STD-DOI (2011).
18 Vgl. Brase; Klump (2007).

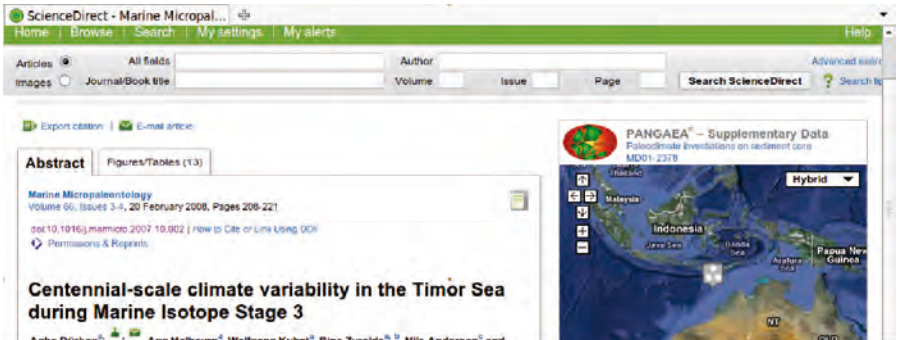

Abb. 1: Datenveröffentlichungen in WDC-MARE/PANGAEA werden in ScienceDirect (Elsevier) zusammen mit dem jeweiligen Artikel angezeigt. Die eingeblendete Landkarte zeigt den Ort der Probennahme und verknüpft den Artikel mit der Datenveröffentlichung.

Datenmanagement und Qualitätssicherung

Mit der Entwicklung eines Systems für die Publikation und Zitierbarkeit von Forschungsdaten stellte sich auch gleich die Frage, wie die Qualität der veröffentlichten Daten geprüft werden kann. Inzwischen haben einige Fachzeitschriften Kriterien für die Bewertung der inhaltlichen Qualität von Daten und für *Peer-Review* Verfahren formuliert.[19] Für einzelne Spezialfälle, in denen einheitliche Konzepte und standardisierte Datenformate existieren, wurden auch schon technische Verfahren zur Prüfung der Konsistenz und Struktur der Daten entwickelt.

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

19 Vgl. z.B. Pfeiffenberger; Carlson (2011); weitere Artikel in dieser Ausgabe des D-Lib Magazine.

nehmern über den Umgang mit im Projekt gewonnenen Daten. In der Mehrzahl der Projekte mit kleinen Datenmengen gibt es kein systematisches Datenmanagement.

Rechtliche Rahmenbedingungen und Lizenzen

Nutzungsbeschränkungen auf Daten aus öffentlich geförderter Forschung, z.B. aus Gründen des Datenschutzes oder anderer rechtlicher Vorgaben, spielen in den Geowissenschaften kaum eine Rolle. In Projekten mit Beteiligung von Partnern in der Industrie können Daten allerdings durchaus Beschränkungen unterliegen. Auch Daten aus der Fernerkundung (Luft- und Satellitenbilder) sowie Kartenmaterial und Daten aus Umweltmessnetzen (Wetterdaten, Flusspegel, u.ä.) unterliegen oft Nutzungsbeschränkungen. Diese sind meist durch die Nutzungsvereinbarungen mit den Datenproduzenten bestimmt, nicht durch gesetzliche Vorgaben.

In den meisten Verbundprojekten werden inzwischen unter den Projektpartnern Vereinbarungen über den Umgang mit Daten getroffen.[20] Dabei wird anerkannt, dass es unter Forschern einen starken *sense of ownership* in Bezug auf Forschungsdaten gibt, auch wenn dieser urheberrechtlich strittig ist. Aus diesem Grund wird den Forschern meist eine Frist von bis zu zwei Jahren nach Projektende für ausschließliche Nutzung der Daten eingeräumt. Unabhängig von der unklaren Urheberrechtssituation für Forschungsdaten besteht die Möglichkeit diese Daten mit einer Lizenz zu versehen.[21] Einige Projekte geben ihre Daten bereits schon zur Laufzeit des Projekts zur Nutzung durch Dritte frei.[22]

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

20 Vgl. z.B. IODP (2011).
21 Vgl. Ball (2011).
22 Vgl. Pfeiffenberger; Klump (2006).

Nachnutzung älterer Datenbestände und Materialien

Wie bereits eingangs geschildert, sind vielen Daten in den Geowissenschaften episodischer Natur oder der Aufwand, der betrieben werden muss, um sie zu gewinnen, macht sie zu Daten einmaliger und unwiederbringlicher Beobachtungen. Aus diesem Grund besteht für viele Daten ein großes Interesse an einer späteren Nutzung.[23] Die hohen Kosten der Datenerhebung in der Geophysik machen auch hier manchen älteren Datenbestand für die Neuprozessierung interessant. Dies ist allerdings nicht allgemein gültig. Insbesondere im Bereich geochemischer Analytik sind die Fortschritte so immens, dass hier ältere Datenbestände nur noch wenig nachgenutzt werden. Dieses Muster findet sich auch in anderen Disziplinen wieder.[24] Ältere, bisher unerschlossene Datenbestände werden nur in seltenen Fällen für die Nachnutzung erschlossen, da der hiermit verbundene Aufwand sehr hoch ist.

An Stelle der Daten werden in der Geochemie die Probenstücke selber nachgenutzt, um sie mit neuen oder verbesserten Methoden erneut zu bearbeiten. Umso wichtiger sind hier die Sammlungen der Institute, Museen und staatlichen geologischen Dienste, in denen die originalen Proben als Referenzmaterialien aufbewahrt werden. Ähnlich wie bei Daten wurde auch bei Probenstücken festgestellt, dass eine eindeutige Identifizierbarkeit der Stücke notwendig ist. An einzelnen besonders wertvollen Probenstücken werden über Jahre hinweg immer neue Analysen gemacht und veröffentlicht. Bisher gab es in den meisten Fällen keine international gültige Namenskonvention. Zudem wurde in vielen Projekten eine bereits verabschiedete Namenskonvention später nicht durchgehend eingehalten. Dies führte teilweise zu einer Verwirrung bei den Probenbezeichnungen, so dass sich Daten, die in der Literatur zu bestimmten Stücken veröffentlicht wurden, nicht mehr eindeutig zu Proben zuordnen und damit auch nicht in größer angelegten Studien integrieren lassen.

Um das Problem der Benennung der Proben zu lösen wurde vorgeschlagen, eine International Geo Sample Number (IGSN) einzuführen.[25] Auf einem Workshop an der University of California San Diego wurden

23 Vgl. Pfeiffenberger (2007).
24 Vgl. Severiens; Hilf (2006).
25 Vgl. Lehnert et al. (2006); IGSN (2011).

im Februar 2011 die konzeptionellen und organisatorischen Grundlagen für die Einführung der IGSN nach dem Vorbild von DataCite gelegt.

Umgang mit Metadaten

Das Thema Metadaten löst, wie in andern Disziplinen auch, bei Forschern keine Begeisterungsstürme aus. Im Rahmen der Forschung in Messnetzwerken war allerdings die Notwendigkeit der Standardisierung der Formate von Daten und Metadaten von Anfang an offensichtlich. Einen wichtigen Impuls erhielt die Standardisierung von Metadaten durch das zeitgleiche Entstehen des Internets und der verstärkten Erforschung des globalen Klimawandels. Mit dem Global Change Master Directory (GCMD), einem Nachweissystem für Daten mit Relevanz für die Klimaforschung, setzte die US Weltraumbehörde NASA einen Maßstab für Datenportale und für die Beschreibung von Daten in den Geowissenschaften durch Metadaten. Der von der NASA entwickelte Metadatenstandard Directory Interchange Format (DIF) wurde dadurch weltweit verbreitet und wird stetig weiterentwickelt.[26]

Mit dem Ziel einer internationalen Standardisierung wurde 1999–2002 ein ISO-Standard für Metadaten zu Daten mit Raumbezug entwickelt, ISO 19115:2003 Geographic Information – Metadata. Träger der Entwicklung neuer Standards ist das Open Geospatial Consortium (OGC). Weitere wichtige Entwicklungen im Rahmen des OGC sind Standards für Webservices der Geodateninfrastruktur und für die Vernetzung von Sensoren in Umweltmessnetzen (Sensor Web Enablement, SWE). Konzeptionell wegweisend für die Beschreibung von Umwelt- und Labormessdaten war die Entwicklung des „Observations and Measurements“ Modells (O&M)[27]. Speziell für geowissenschaftliche Daten wurde GeoSciML entwickelt.[28]

Die Infrastructure for Spatial Information in the European Community (INSPIRE) ist eine Initiative der Europäischen Kommission mit dem Ziel, ein europäisches Netzwerk von integrierten raumbezogenen Infor-

26 Vgl. DIF (2011).
27 Vgl. OGC (2011).
28 Vgl. Sen; Duffy (2005).

mationsdiensten zu schaffen.[29] Die öffentlichen Einrichtungen beginnen auf Grund der in Kraft getretenen INSPIRE-Richtlinie, ihre Geodaten INSPIRE-kompatibel aufzubereiten, wobei der Zeitplan zunächst die Erzeugung einheitlicher Metadaten vorsieht. Die INSPIRE-konformen Webdienste bauen auf den Normen der ISO 191xx Familie auf.

Bei den hier vorgestellten Metadatenstandards handelt es sich fast ausschließlich um XML-kodierte Standards für den Nachweis von Daten in Datenportalen und in standardisierten Webschnittstellen. Darüber hinaus existieren auch noch unzählige disziplinspezifische Metadatenstandards für die fachliche Beschreibung, z.B. QuakeML für seismische Ereignisse, Darwin Core für Biodiversitätsdaten, ThermoML für Daten zur Thermodynamik und viele weitere.[30] Einige der Metadatenschemata wurden immer komplexer, da immer neue Aspekte der Modellierung der beschriebenen Daten mit aufgenommen wurden.[31] Dies führte zu Problemen bei der Akzeptanz der Standards in der Community, bei der Anwendung der Standards und bei der Verarbeitung der Daten bei der Übermittlung. Gerade bei der Übermittlung und Speicherung kleiner Datenobjekte wird deshalb auch erwogen, auf XML zu verzichten, da hier das Volumen der umgebenden XML-Struktur gegenüber dem eigentlichen Datenvolumen übermäßig groß ist.[32;33]

Gerade bei der Verknüpfung von Diensten kommt es nicht nur auf die Syntax der Metadaten an, sondern auch auf die Semantik. Um eine semantische Interoperabilität der Dienste zu ermöglichen werden neben kontrollierten Vokabularen auch zunehmend Identifikatoren (PID oder URI) eingesetzt um Begriffe eindeutig und maschinenlesbar zu benennen, ein Ansatz, der in der Biodiversitätsinformatik bereits mit Erfolg angewandt wird.[34]

29 Vgl. Schilcher et al. (2007).
30 Eine Übersicht über Metadatenstandards in den Erd- und Umweltwissenschaften ist auf den Seiten der Marine Metadata Initiative zu finden (vgl. Marine Metadata Initiative (2011)).
31 Vgl. Klump; Huber (2011).
32 Vgl. Barkstrom (2011).
33 Vgl. Riede et al. (2010).
34 Vgl. Page (2008).

9.4 Interne Organisation

LZA-Strategien und Verfahren

Auf Grund der Heterogenität der Geowissenschaften sind LZA-Strategien und -Verfahren bisher nur in einigen Teilbereichen etabliert. Insbesondere die Seismologie ist durch große Datenmengen und internationale Messnetzwerke bestimmt, gleichzeitig beschäftigt sie sich mit Daten aus episodischen Ereignissen. Diese Umstände führten schon früh zu einer technischen Vereinheitlichung von Formaten und Metadaten, sowie einem internationalen Austausch von Daten und deren Archivierung in zentralen Datenarchiven.

Je weiter man sich jedoch von der Vorgabe der Vernetzung von homogen strukturierten Datenbeständen entfernt, desto seltener werden Vereinbarungen oder gemeinsame Strategien für die Archivierung und Erhalt der Nachnutzbarkeit von Forschungsdaten getroffen. Ausgehend von den geowissenschaftlichen Großforschungseinrichtungen hat jedoch ein kultureller Wandel begonnen, der dazu führt, dass durch die Zitierbarkeit von Daten diese deutlich aufgewertet werden. Damit steigt auch das Interesse für die Langzeitarchivierung von Daten.

Finanzierung und Organisation

Die Finanzierung der WDC und verwandter Datenarchive ist im Allgemeinen gegliedert in datenkuratorische Aufgaben einerseits (OAIS Ingest) und den Betrieb des Archivs (OAIS Archiving) und des Vertriebs der Daten andererseits (OAIS Dissemination). Betrieb des Archivs und der Dissemination erfolgt meist aus der Grundfinanzierung der Institution, die das Datenarchiv betreibt. Die Finanzierung der datenkuratorischen Aufgaben wird nach Möglichkeit von den Projekten eingeworben, die das Datenarchiv nutzen. Da der kuratorische gegenüber dem technischen Aufwand überwiegt[35], werden von großen Projekten nach Möglichkeit Datenkuratoren eingestellt, die für die Verwaltung und Sicherung der im Projekt erzeugten Forschungsdaten verantwortlich sind. Für

35 Vgl. Beagrie; Lavoie; Woollard (2010).

die Daten kleinerer Projekte werden meist individuelle Lösungen gefunden, ein ständiges Dienstleistungsangebot fehlt jedoch noch.

Die Kosten für die Entwicklung und den Betrieb eines Datenarchivs sind nur schwer abschätzbar. Entscheidend für die Kosten ist, ob in der Entwicklung bereits bestehende Komponenten, auch aus externen Quellen, nachgenutzt werden können und ob im Betrieb Skalierungseffekte genutzt werden können, um die Stückkosten zu senken. In diesen Punkten unterscheiden sich Datenarchive in den Geowissenschaften nicht von anderen wissenschaftlichen Datenarchiven.[36] Da kleineren Gruppen diese Skalierungseffekte fehlen, werden zunehmend Dienste Dritter für die vertrauenswürdige Langzeitarchivierung von Forschungsdaten in Anspruch genommen.

Personal in der LZA

Bisher gibt es nur in wenigen geowissenschaftlichen Forschungseinrichtungen Personal, das mit der Langzeitarchivierung von Forschungsdaten beauftragt ist. Der Schwerpunkt der Arbeit liegt im Allgemeinen im Datenmanagement im Projekt und unmittelbar vor der Archivierung der Daten. Hier hat es sich bewährt, die Mitarbeiter aus der Informatik und aus dem jeweiligen Fachgebiet zu rekrutieren, da dies die Verständigungsprobleme zwischen den Forschern und den Archivmitarbeitern verringert. Die hier eingesetzten Datenmanager sind meist Wissenschaftler, die im Rahmen eines Projektes befristet eingestellt werden. Auch hier setzt ein Umdenken ein, es wird zunehmend erkannt, dass für den Betrieb eines digitalen Langzeitarchivs ein fester Stamm von Mitarbeitern notwendig ist, um eine organisatorische und technologische Kontinuität zu ermöglichen.

9.5 Perspektiven und Visionen

Eine Gruppe von Geowissenschaftlern plant ein Projekt zur Entstehung des heute längst abgetragenen Gebirges im Gariep-Gürtel an der südafrikanisch-namibischen Grenze. Zur Entschlüsselung des Geschehens vor

36 Vgl. ebenda.

540 Millionen Jahren werden geologische Karten erstellt und ausgewertet, sowie umfangreiche geochemische Analysen an Gesteinen und einzelnen Mineralen vorgenommen. In der Vorbereitung des Antrags werden nicht nur die wissenschaftlichen Vorarbeiten ausgewertet und die Logistik der Geländearbeit geplant, das Team wird auch von einem Datenzentrum beraten, welche Werkzeuge ihm zur Verfügung stehen und es wird ein Datenmanagementplan entworfen.

Für das Projekt wird eine virtuelle Forschungsumgebung entworfen, die den verteilt arbeitenden Forschern erlaubt ohne großen Aufwand Daten für ihre Kollegen bereitzustellen. Daten aus analytischen Instrumenten, wie Massenspektrometern oder Röntgenfluoreszenspektrometern, werden durch die Angaben in den Konfigurationsdateien der Experimente weitgehend automatisch dokumentiert und können durch eine Blog-Funktion annotiert und diskutiert werden.

Die virtuelle Forschungsumgebung dient auch der Aufbereitung von Daten für die Veröffentlichung und Archivierung. Datensätze können als abgeschlossen gekennzeichnet werden und werden damit an das Langzeitarchiv und an den Publikationsagenten übergeben. Wenn diese Daten Grundlagen einer Veröffentlichung sind, können sie mit dieser verknüpft werden und sind über das Datenportal der Forschungseinrichtung zugänglich.

Diese Vision ist *work in progress*, aber noch ist sie nicht umgesetzt. Gemessen an anderen Disziplinen ist der Umgang mit Forschungsdaten in Teilen der Geowissenschaften bereits weit entwickelt. Dennoch wird auch hier erst ein kleiner Teil der Daten in langfristige Strukturen überführt. Denn auch wenn sich die Publikation von Daten allmählich als anerkanntes Verfahren durchsetzt, so scheuen viele Wissenschaftler noch den vermuteten Aufwand, Daten für eine Langzeitarchivierung, und gegebenenfalls Veröffentlichung, aufzubereiten. Die bereits vollzogenen Änderungen in der Förderpolitik der DFG und der Europäischen Kommission haben mit dazu beigetragen, ein Umdenken über den Wert von Daten einzuleiten, auch wenn die Verpflichtung zum Datenmanagement als Last wahrgenommen wird.[37]

Hemmnisse sind bei der Umsetzung einer Strategie zur Langzeitarchivierung von Forschungsdaten meist die fehlenden organisatorischen

37 Vgl. Feijen (2011).

und technischen Strukturen. Es fehlen Ansprechpartner in den Institutionen und Werkzeuge, mit denen der datenkuratorische Prozess unterstützt werden kann. Insbesondere bei den Werkzeugen müssen weiter Werkzeuge und Konzepte entwickelt werden, die eine nahtlose Integration der datenkuratorischen Aufgaben in die Arbeitsabläufe der Forschung ermöglichen.

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

> „Wenn aus Projektmitteln systematisch (Mess-) Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch – sofern vorhanden – die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien.“[38]

Um diesen Anspruch zu erfüllen, müssen in den nächsten Jahren auch von den Datenarchiven neue Angebote entwickelt werden. Diese müssen von Geschäftsmodellen flankiert sein, die es erlauben, diese Dienstleistungen mit den Projekten abzurechnen. Aktuell fällt es den Datenzentren noch schwer, die Kosten der Langzeitarchivierung von Forschungsdaten zu beziffern. Das heißt auch, dass an den Zentren, die LZA-Dienstleistungen anbieten, entsprechende organisatorische Strukturen geschaffen werden müssen.

Auf der Seite der Datenzentren besteht das Problem, dass diese meist immer noch als „Silo“ angelegt sind, d.h. der Inhalt der Systeme ist nicht über automatisierte Verfahren zugänglich. Dies ist insbesondere bei der Überführung von Forschungsdaten in ein Datenarchiv problematisch, weil Medienbrüche stets eine Hürde im Lebenszyklus der Daten darstellen, an denen das Risiko besonders hoch ist, dass die Kette der Bearbeitungsschritte im Datenlebenszyklus abreißt. Integriertes Datenmanagement und die Langzeitarchivierung von Forschungsdaten müssen erst noch Bestandteil des wissenschaftlichen Alltags werden. Wir arbeiten daran.

38 DFG (2010c).