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

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 10
Klimaforschung

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

10 Klimaforschung

Bernadette Fritzsch

10.1 Einführung in den Forschungsbereich

Im Vergleich zu anderen hier erwähnten Wissenschaften wie der Medizin ist die Klimaforschung ein relativ junger Wissenschaftszweig, der aber hohe gesellschaftliche Relevanz besitzt. Forschungsgegenstand sind die Gesetzmäßigkeiten des Klimas als durchschnittlicher Zustand der Atmosphäre über längere Zeiträume hinweg.[1] Ziel ist es, die Entwicklung des Klimas auf der Erde zu verstehen und vorherzusagen. Dies beinhaltet sowohl die Dokumentation des früheren und jetzigen Zustands (beobachtende Klimatologie) als auch die Beschreibung der zugrundeliegenden Gesetzmäßigkeiten (theoretische Klimatologie).

Das Klimasystem der Erde setzt sich aus mehreren Teilsystemen zusammen, die sich in unterschiedlichen Zeitskalen ändern und in komplexer Weise untereinander Wechselwirkungen eingehen. Neben der Atmosphäre sind vor allem die Hydrosphäre und die Kryosphäre zu nennen. Die Ozeane wirken als Speicher von Wärmeenergie und transportieren sie über ihr Strömungssystem über weite Entfernungen hinweg. Eisschilde greifen in den Energiehaushalt der Erde ein, indem sie die Rückstrahlung von Strahlungsenergie beeinflussen. In ähnlicher Weise wirken auch die Landflächen mittels räumlicher Variation des Absorptions- und Reflexionsverhaltens durch unterschiedliche geografische Gegebenheiten oder Landnutzung. Damit ist die Klimaforschung ein Querschnittsbereich, der Wissenschaftler unterschiedlicher Fachrichtungen vereint: Meteorologen, Ozeanografen, Geografen, Geologen, Physiker etc.

In erdgeschichtlichen Zeiträumen war das Klima bereits sehr großen Schwankungen unterworfen. Diese hohe Klimavariabilität erlaubt es erst

1 Vgl. Lauer; Bendix (2006), S. 10.

nach langen Messreihen, signifikante Trends und systematische Änderungen zu erkennen. Daher haben regelmäßige Beobachtungen über viele Jahrzehnte oder sogar Jahrhunderte hinweg für die Klimaforschung eine große Bedeutung. Die einmal erhobenen Daten besitzen einen hohen Wert: Sie können nicht durch eine einfache Wiederholung der Messung wiederbeschafft werden, da der Zustand an die Zeit gekoppelt ist. Damit ist die langfristige Archivierung der Daten unumgänglich.

Der Mensch greift immer stärker als Akteur in das Klimageschehen ein. Gleichzeitig beeinflussen die Klimaänderungen zunehmend seinen Lebensraum. Daraus ergibt sich die gesellschaftliche Bedeutung der Klimaforschung. Von den Wissenschaftlern werden belastbare Prognosen für die Entwicklung des Klimas in den kommenden Jahren und Jahrzehnten erwartet. Sie werden als Entscheidungsgrundlage benötigt, um daraus Handlungsempfehlungen abzuleiten. Dabei geht es neben Vermeidungsstrategien zunehmend auch um die Anpassung an solche Veränderungen, die nicht vermeidbar sind.

Der einzige Weg, Klimaentwicklungen wissenschaftlich fundiert vorherzusagen, sind Simulationen. Dazu wird das komplexe Klimasystem am Computer modelliert. In Deutschland steht der Klimacommunity neben den allgemeinen nationalen Höchstleistungsrechenzentren auch ein thematisches Höchstleistungsrechenzentrum zur Verfügung: Das Deutsche Klimarechenzentrum DKRZ stellt entsprechend seiner Satzung als „überregionale Serviceeinrichtung Rechenzeit und technische Unterstützung für die Durchführung von Simulationsrechnungen mit aufwändigen numerischen Modellen für die Klimaforschung und verwandte Gebiete“ bereit.[2] Damit können sehr umfangreiche und komplexe Modellierungen durchgeführt werden.

Da Modelle stets gewisse Unsicherheiten in sich tragen, kommt zunehmend der „Multi Model Multi Ensemble“ Ansatz zum Tragen. Darin wird die gleiche Simulation mit unterschiedlichen Modellen und mit jeweils leicht modifizierten Anfangsbedingungen durchgeführt. Die Ergebnisse werden dann mit statistischen Methoden ausgewertet, um eine verlässlichere Aussage zu erhalten. Dieses Verfahren erfordert leistungsfähige Rechner, um Ensemble mit vielen Mitgliedern rechnen zu können. Gleichzeitig erhöht sich mit diesem Ansatz die zu speichernde Datenflut erheblich und stellt neue Anforderungen an das Datenmanagement.

2 Vgl. DKRZ (2011).

Die Verknüpfung der in den einzelnen Datenarchiven gelagerten Informationen ist für den Klimaforscher mit hohem Aufwand verbunden. Dies liegt in der großen Heterogenität der Daten bezüglich ihrer Quellen, Formate und Speichermodalitäten begründet. Das Projekt „Collaborative Climate Community Data and Processing Grid (C3Grid)“ hat sich dieser Herausforderung angenommen und eine Infrastruktur entwickelt und aufgebaut, die einen einheitlichen Zugriff auf die Daten der im Projekt beteiligten Partner erlaubt. Es wurde im Rahmen der D-Grid Initiative vom BMBF, Referat 522 „Internet“ gefördert und gehörte zu den Communityprojekten der ersten Runde.[3] Im Zeitraum von 09/2005 bis 02/2009 wurde prototypisch eine Grid-Infrastruktur für die deutsche Klimacommunity geschaffen. Inzwischen sind die Arbeiten in die zweite Phase getreten. In einem durch das BMBF-Referat „Globaler Wandel“ geförderten Projekt „C3Grid – INAD: Towards an INfrastructure For General Access to climate Data“ sollen die entwickelten Komponenten interoperabel zu den internationalen Infrastrukturen gemacht und in den Produktionsstatus überführt werden.

10.2 Kooperative Strukturen

Die Klimawissenschaft basiert ganz grundlegend auf Kollaborationen, die Instituts- und Ländergrenzen überschreiten. Nach außen hin sichtbar wird dies unter anderem durch das Intergovernmental Panel on Climate Change (IPCC; Zwischenstaatlicher Ausschuss für Klimaänderungen). Dieses in Deutschland oft unter dem Namen Weltklimarat bekannte Gremium wurde 1988 vom Umweltprogramm der Vereinten Nationen (UNEP) und der Weltorganisation für Meteorologie (WMO) gegründet und trägt die Ergebnisse der internationalen Forschercommunity zusammen. In regelmäßigen Abständen wird der Stand der Forschung in den sogenannten Wissensstandberichten („IPCC Assessment Reports“, AR) zusammengetragen, um Risiken der globalen Erwärmung zu beurteilen sowie Vermeidungs- und Anpassungsstrategien zu entwickeln.

Ein weiteres international gewichtiges Gremium der Zusammenarbeit ist die World Meteorological Organization (WMO), die eine Sonder-

3 Vgl. BMBF (2004).

organisation der Vereinten Nationen ist. Sie ist nicht nur auf die Klimaforschung beschränkt, sondern will Fachkompetenz auf den Gebieten Wetter, Klima, Hydrologie und Wasserressourcen bereitstellen und internationale Kooperationen auf diesen Gebieten fördern.[4] Zu ihren Zielen gehören Standardisierungsbestrebungen im Bereich meteorologischer und anderer verwandter Beobachtungen.

Die technische Infrastruktur für die Erhebung und Speicherung der Daten ist auf eine Vielzahl von Einrichtungen verteilt. Die für die Beobachtung des Erdsystems notwendigen Großgeräte werden in der Regel in der Verantwortung einer bestimmten Einrichtung betrieben, jedoch von Forschern vieler Institutionen gemeinsam genutzt. So unterhält das Alfred-Wegener-Institut für Polar- und Meeresforschung, Forschungszentrum in der Helmholtz-Gemeinschaft unter anderem den Forschungseisbrecher „Polarstern" und Forschungsstationen in der Arktis und Antarktis für die Community. Wissenschaftler anderer Einrichtungen können sich jeweils mit ihrem Forschungsprogramm bei den Koordinierungsgremien der Forschungsplattformen bewerben. Vielfältige nationale und internationale Kooperationen sorgen für einen regen Wissenstransfer zwischen den Institutionen.

Wegen der großen Datenflut und der teilweise sehr hochvolumigen Daten können die Daten nicht in einem zentralen Archiv gespeichert werden. Vielmehr existiert eine Reihe von Datenarchiven, die sich jeweils auf bestimmte Daten spezialisiert haben und unterschiedlich organisiert sind.

Als ein Beispiel sind die ICSU (International Council for Science) Weltdatenzentren zu nennen. Sie gehen auf international koordinierte Bemühungen zum Internationalen Geophysikalischen Jahr 1957–1958 zurück. Die damals erhobenen Daten sollten dauerhaft für die Nachwelt aufbewahrt und den Wissenschaftlern zugänglich gemacht werden. Inzwischen gehören 52 Einrichtungen in zwölf Ländern diesem Verbund an, die Daten aus vielen unterschiedlichen Bereichen des Erdsystems sammeln. In Deutschland besitzen drei Einrichtungen das Label WDC, mit dem die dauerhafte Verfügbarkeit der dort archivierten Daten gekennzeichnet ist. Die Spannbreite der dort vorhandenen Daten reicht

4 Vgl. WMO (2011).

von Messdaten der Marinen Umweltwissenschaften (WDC MARE) über Satellitendaten (WDC RSAT) bis zu Klimasimulationsdaten (WDC Climate). Wissenschaftler können dort jeweils Daten einstellen, die dann eine genau vorgeschriebene Qualitätskontrolle durchlaufen. Gemeinsam mit dem Datenzentrum des GeoForschungszentrums Potsdam GFZ haben sich diese Weltdatenzentren zum Deutschen WDC Cluster für Erdsystemforschung zusammengeschlossen. Sie haben sich verpflichtet, die internationalen Standards zu beachten und arbeiten gemeinsam an der Weiterentwicklung der Datendienste.

Daneben gibt es mit dem Deutschen Wetterdienst DWD ein Beispiel für ein Datenarchiv, das einer Behörde zugeordnet ist. Dort werden vor allem die Langfristbeobachtungen der meteorologischen Messstationen archiviert. Aber auch weitere meteorologische Daten sind dort verfügbar. Der DWD ist über die WMO (World Meteorological Organizaion) international mit anderen Wetterdiensten verbunden.

Weiterhin gibt es verteilte Datenarchive an diversen universitären und außeruniversitären Forschungseinrichtungen. Sie enthalten oft Daten aus speziellen Messkampagnen, die teilweise nicht ohne weiteres in das Profil der zentralen Datenarchive passen. Außerdem kann durch die dezentrale Archivierung eine enge Kopplung zwischen Daten und zugehöriger fachlicher Expertise erreicht werden.

Aus wissenschaftlicher Sicht ist eine Verknüpfung dieser unterschiedlichen Datenquellen unabdingbar. Sie wird jedoch behindert durch technische Hürden: Der Wissenschaftler muss zunächst einmal überhaupt wissen, welche Daten sich in welchem Datenarchiv befinden. Dann muss er sich über die spezifischen Zugangsmodalitäten informieren. Viele Daten sind bereits über die Webportale der Archive zugänglich. Es fehlen aber übergreifende Strukturen für einen einheitlichen Datenzugriff. Dies wurde im bereits erwähnten C3Grid in Angriff genommen. Mithilfe der Gridtechnologie sollte die Heterogenität für den Nutzer transparent gemacht werden.

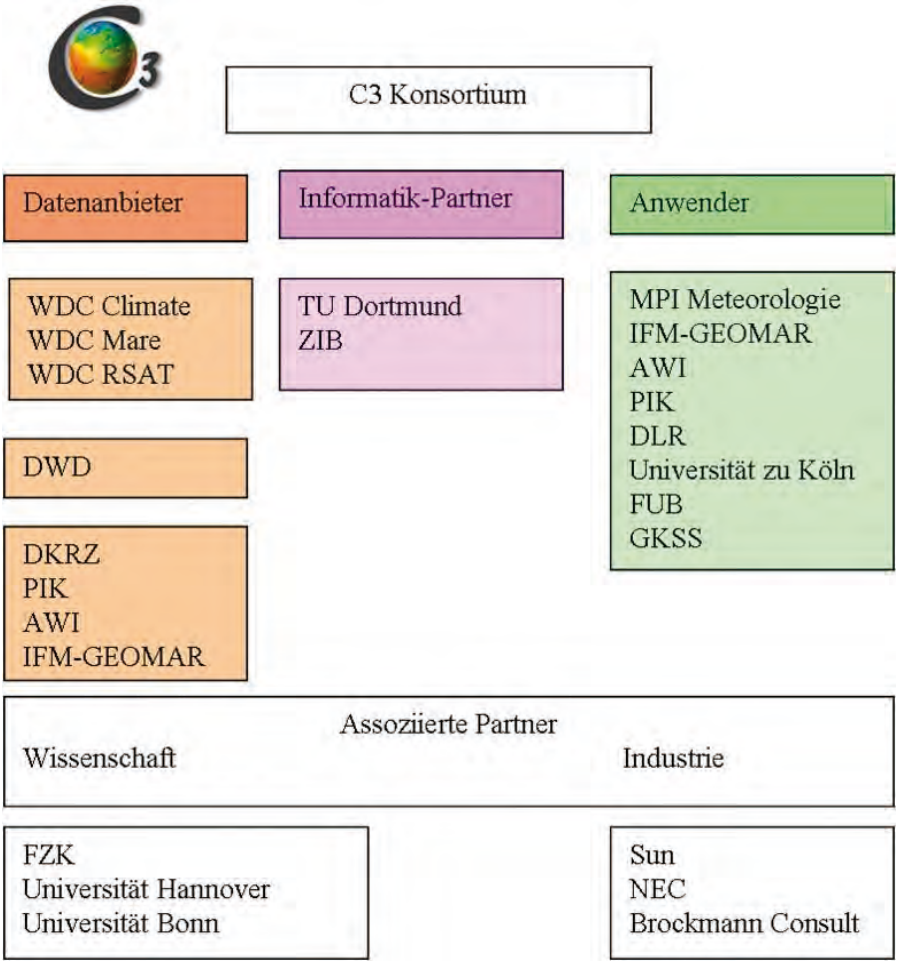

Abb. 1: Beteiligte Partner im C3Grid

Um die Entwicklung auf eine breite Basis zu stellen, wurden in C3Grid wesentliche Einrichtungen der deutschen Klimaforschung im Konsortium versammelt. Ihm gehörten zum einen die großen Datenprovider an, zu denen die Weltdatenzentren WDC Climate, WDC MARE und WDC RSAT, der Deutsche Wetterdienst DWD sowie die vorhandenen Filearchive weiterer Institutionen wie DKRZ, AWI, Potsdam Institut für Klimafolgenforschung (PIK), Leibniz-Institut für Meeresforschung IFM-

GEOMAR und Universität zu Köln zählen. Um die Werkzeuge für die Analyse der Daten zu entwickeln, waren Wissenschaftler wichtiger Forschungseinrichtungen für die Spezifikation der Anforderungen und die Einbringung von Workflows zuständig. Hier sind als zusätzliche Partner zu nennen: das Max-Planck-Institut für Meteorologie, das Meteorologische Institut der Freien Universität Berlin und das Helmholtzzentrum GKSS. Zur Umsetzung der Anforderungen wurden Informatikpartner vom ZIB und der TU Dortmund eingebunden, die bereits über Grid-Erfahrungen verfügten. Als assoziierte Partner waren weitere Forschungseinrichtungen aus dem Klima-Umfeld beteiligt sowie als Industriepartner zwei Rechnerhersteller und ein mittelständisches Unternehmen aus dem Bereich der Umweltinformatik.

Waren dies wegen der Einbettung des Projekts in die D-Grid Initiative ausschließlich deutsche Partner, so ist das Folgeprojekt C3-INAD der internationalen Einbettung gewidmet. Damit soll ein für die Wissenschaftler effektives Werkzeug implementiert werden, das den Anforderungen im Zusammenhang mit dem kommenden Sachstandsbericht IPCC AR5 genügt. Weltweit beteiligen sich ca. 1.000 Wissenschaftler an den Rechnungen, wesentlich mehr werden die Ergebnisse für weitere Arbeiten nutzen.

Die Datenflut für dieses Dokument wird alle bisherigen Vorhaben übersteigen. Allein für den Teil der globalen Simulationen werden ca. zehn PB an Resultaten erwartet, so dass nur eine verteilte Datenhaltung in Frage kommt. Weltweit verteilte Datenknoten speichern die Ergebnisse der Rechnungen. Eine definierte Teilmenge von ca. 10% (Core data) wird in drei Daten Gateways gesammelt und verfügbar gemacht, die ihre Daten jeweils untereinander spiegeln. Eines dieser Gateways entsteht gerade am WDC Climate. Es bildet mit seinem Datenbestand auch die Schnittmenge zwischen C3Grid und der IPCC Datenföderation. Die im Rahmen des IPCC AR5 bereitgestellten Daten werden auch im C3Grid bekannt gemacht, indem die Metadaten dem dort gültigen Metadatenprofil angepasst und dann publiziert werden. Da C3Grid neben den Simulationsdaten aber auch weitere Mess- und Beobachtungsdaten zur Verfügung stellt, kann der C3Grid-Nutzer durch Verknüpfung unterschiedlicher Datenquellen neue Erkenntnisse für seine wissenschaftliche Arbeit gewinnen.

10.3 Daten und Metadaten

Die Daten in der Klimaforschung zeigen eine sehr große Heterogenität. Da sehr unterschiedliche Wissenschaftsdisziplinen zusammenarbeiten, gibt es auch bei den Daten eine große Bandbreite bezüglich der Quellen, der Formate und Speichermodalitäten. Im Folgenden sollen einige Beispiele die Vielfalt schlaglichtartig kurz beleuchten.

Eine wichtige Datenquelle für die Klimaforscher sind die Beobachtungsdaten aus den Messreihen des Deutschen Wetterdienstes. Dem DWD als teilrechtsfähiger Anstalt des öffentlichen Rechts im Geschäftsbereich des Bundesministeriums für Verkehr, Bau und Stadtentwicklung obliegt per Gesetz der Betrieb der Mess- und Beobachtungssysteme und die Bereithaltung, Archivierung und Dokumentierung meteorologischer Daten und Produkte. Er unterhält ein dichtes Netz aus Wetter- und Niederschlagstationen und Beobachtungsstellen, die kontinuierlich in einem definierten Verfahren Daten erfassen.

Mit der Nutzung von Satelliten für das Monitoring der Erde wurde seit den 1970er-Jahren eine weitere wichtige Datenquelle erschlossen, deren Datenströme seitdem immer weiter zunehmen. Sie liefern durch ihre großräumige Abdeckung wertvolle Informationen. Ausgestattet mit einer Vielzahl von Sensoren werden unterschiedlichste Phänomene beobachtet. Als ein Beispiel sei hier die Bestimmung der Meereisbedeckung genannt. Dazu nutzt man die Tatsache, dass die von der Erde emittierte Strahlung im Mikrowellenbereich wesentlich durch die Oberflächenbeschaffenheit bestimmt wird. Daher kann man eisbedeckte von eisfreien Flächen unterscheiden und kann Rückschlüsse auf die Eisverteilung und sogar den Eistyp ziehen. Durch eine qualifizierte Prozessierung der Satellitendaten entstehen so tagesaktuelle Karten zur Eisbedeckung.[5]

Seit längerem dagegen werden in der Meteorologie Ballons für die Messungen der Atmosphäre genutzt. Sie sind mit unterschiedlichen Messgeräten versehen. Weltweit existieren mehrere hundert Stationen, von denen aus regelmäßig Ballonaufstiege durchgeführt werden. Damit werden Profile der Atmosphäre bis ca. 20–30 km Höhe aufgenommen.[6]

---

5 Vgl. Heygster; Spreen; Kaleschke (2007).
6 Vgl. z.B. Lanconelli et al. (2010).

Ein wichtiges Instrument zur Bestimmung des Klimazustands der Erde in vergangenen Jahrhunderten sind Eiskernbohrungen. Dabei werden tiefe Bohrungen in die Landeisschilde eingebracht und die gewonnenen Bohrkerne analysiert.[7] Ähnlich wie bei den Jahresringen der Bäume erlauben die übereinanderliegenden Eisschichten (Jahresschichten) eine Tiefen-Zeit-Zuordnung und damit Aussagen zum Klima der Vergangenheit. Je nach Länge des Eiskerns und den lokalen Akkumulationsraten ergeben sich Zeitserien im Bereich von bis zu einigen Hundertausend Jahren.

Eine eigene Klasse von Daten stellen Simulationen dar. Sie beruhen auf der Modellierung des Klimageschehens im Rechner. Während Beobachtungsdaten meist durch kleine Datenvolumina, aber große Heterogenität charakterisiert sind, besitzen die Klimamodelldaten große Volumina (Archive im Petabyte-Bereich sind keine Seltenheit), aber eine weitgehend homogene Struktur. Die Erhebung, Validierung und Bereitstellung von qualitativ hochwertigen und dokumentierten Daten ist mit einem hohen Aufwand verbunden. Um diese Arbeit angemessen würdigen zu können, gibt es in der Klimacommunity vielfältige Bestrebungen zur Publikation von Forschungsdaten. Dies ist auch im Hinblick auf die Nachnutzung von Daten von Bedeutung. Die Datenzentren sind nur attraktiv für Nutzer, solange immer wieder Wissenschaftler ihre Daten in den jeweiligen Archiven zur Verfügung stellen.

In den deutschen Weltdatenzentren werden die Datenobjekte beim Ingest im Rahmen eines Publikationsprozesses mit persistenten Identifiern versehen. Damit werden die Daten zitierfähig in wissenschaftlichen Arbeiten und es wird die Arbeit der veröffentlichenden Wissenschaftler anerkannt.

Daneben gibt es seit 2008 mit dem „Earth System Science Data – The Data Publishing Journal“ den Versuch, die Datenpublikation in einen ähnlichen Kontext zu stellen wie die Veröffentlichung von wissenschaftlichen Artikeln.[8] Mit einem Review-Prozess analog zu dem in üblichen wissenschaftlichen Journalen soll gesichert werden, dass die Daten qualitativ hochwertig, frei zugänglich (open access) und mit Standardmetadaten beschrieben sind.

7 Vgl. Udisti et al. (2008).
8 Vgl. Pfeiffenberger; Carlson (2011).

Die einzelnen Datenzentren haben ihre spezifischen Anforderungen bei der Aufnahme von Forschungsdaten in ihre Archive[9],[10]. Es existieren definierte Prozeduren, nach denen die Daten einer Qualitätskontrolle unterzogen werden. Diese Prozeduren sind in den einzelnen Datenzentren leicht unterschiedlich, da die verschiedenen Datenquellen jeweils angepasste Verfahren erfordern. Sie zeigen jedoch einige Gemeinsamkeiten.

Wesentlich ist zunächst einmal die Definition von Verantwortlichkeiten im Prozess der Langzeitarchivierung. Beim Dateneintrag liefert der Autor der Daten bzw. der Verantwortliche des zugehörigen Projekts (principal investigator) die Daten an das Archiv. Er erstellt die notwendigen Metadaten zur Beschreibung des Datensatzes, wobei das Metadatenprofil jeweils vom Archiv vorgegeben ist. In der Klimacommunity hat sich dabei das durch den internationalen Standard[11] ISO 19 115/139 definierte Metadatenprofil als Basis durchgesetzt, das dann jeweils von den einzelnen Archiven an die jeweiligen Bedürfnisse angepasst wird. Der Datenkurator prüft dann die Vollständigkeit der Metadaten und die Datenkonsistenz und sorgt für den Import der Daten in die jeweilige Dateninfrastruktur. Nach einem ersten Test auf Lesbarkeit der Daten und Verfügbarkeit durch den Kurator wird eine technische Qualitätskontrolle vorgenommen und dem Datensatz ein Digital Object Identifier (DOI) zugeteilt. Danach muss der Autor ebenfalls die importierten Daten und Metadaten prüfen und deren Vollständigkeit und Korrektheit bestätigen. Er ist für die wissenschaftliche Qualität seiner Daten verantwortlich. Erst nach seiner Bestätigung werden die Daten veröffentlicht und können über die zugeteilte DOI referenziert werden.

Die in den Datenzentren gespeicherten Informationen wachsen stetig. Dies lässt sich zum einen am Datenvolumen ausmachen. So stieg im WDC Climate das Datenaufkommen von vier TB in 2004 auf 420 TB in 2010. Für 2011 wird das Erreichen der PB-Grenze erwartet. Jedoch liefert eine Fixierung auf das Volumen allein kein aussagekräftiges Bild über die Dynamik der Datenhaltung in der Klimacommunity. In einigen Bereichen ist es vor allem die große Anzahl der Datenobjekte, die eine Heraus-

9 Vgl. z.B. Grobe (2005).

10 Die Data Policy des WDC Climate sind in den Terms of Use festgehalten und folgen den Open Access Prinzipien (http://cera-www.dkrz.de/WDCC/ui/docs/TermsOfUse.html).

11 Vgl. ISO (2011).

forderung für die Betreiber des jeweiligen Datenarchivs bedeutet. In der PANGAEA Datenbank des WDC MARE befinden sich derzeit über 5,5 Milliarden Datenobjekte mit jährlichen Steigerungsraten von 5–10%.

Die Beschränkung auf einzelne Formate in einem Datenarchiv bringt eine Reihe von Vorteilen für die Handhabung der Daten im Bereich der Langzeitarchivierung. Jedoch gibt es nicht ein einzelnes Datenformat, das für alle Informationen geeignet wäre, die in der Klimaforschung relevant sind. Je nach Spezialisierung kann daher in den Datenarchiven nur teilweise eine Beschränkung auf Datenformate erfolgen. Das WDCC ist durch die Fokussierung auf Klimadaten mit Modellierungsbezug ein Beispiel dafür, während dies beim WDC MARE nicht möglich ist, da dort gerade die Vielzahl von unterschiedlichen Beobachtungsdaten gesammelt wird. Diese beiden Datenzentren sollen im Folgenden auch immer wieder als Beispiele diskutiert werden, um einerseits die Bandbreite des Datenmanagements in der Klimaforschung zu illustrieren, andererseits aber auch zu dokumentieren, dass die Community sich bereits auf einige Standards in der Datenhaltung einigen konnte.

Bei den modellierenden Klimaforschern haben sich in den letzten Jahre vor allem zwei Formate duchgesetzt: netCDF[12] und grib[13]. Beiden gemeinsam ist, dass die Daten in einem definierten Gitter abgelegt sind. Dies spiegelt das Vorgehen bei der Simulation wider, die auf der Diskretisierung der dem Klimageschehen zugrundeliegenden Gleichungen basiert. Daher akzeptiert das WDCC mit seiner Spezialisierung auf Daten mit Modellierungsbezug auch nur Daten in diesen beiden Formaten und ASCII. Damit kann der Aufwand für die Datenpflege minimiert werden, da nur noch wenige Werkzeuge für den Zugriff auf die Daten vorgehalten werden müssen. Eine weitere Beschränkung erfolgt bezüglich des Inhalts der Dateien, indem die abgespeicherten Variablen der Climate Forecast Convention[14] genügen müssen. Dadurch werden für die wichtigsten Größen eindeutige Bezeichnungen und Maßeinheiten vorgegeben, die dann die Suche erleichtern und die Vergleichbarkeit von Datensätzen ermöglichen.

12 Vgl. Rew; Davis (1990); NetCDF (2011).

13 Vgl. GRIB (1994).

14 Vgl. Gregory (2003) und die Liste der aktuellen CF Namen unter http://cf-pcmdi.llnl.gov/documents/cf-standard-names.

Dieses Vorgehen ist jedoch nicht immer möglich. So ist es gerade ein wesentliches Anliegen von PANGAEA im WDC MARE, möglichst viele unterschiedliche Beobachtungsdaten zu sammeln, um dem Wissenschaftler einen breiten Blick zu ermöglichen. Dies ist sowohl für die umfassende Dokumentation des Klimas wichtig als auch für die Weiterentwicklung der Modelle. Deren Qualität wird entscheidend daran gemessen, dass sie die beobachteten Daten möglichst korrekt wiedergeben. Nur dann kann erwartet werden, dass die mit dem Modell berechneten Prognosen für die Zukunft ebenso verlässlich sind.

Zur Sammlung des WDC MARE gehören Messungen klimarelevanter Größen, die sich in dem bevorzugten ASCII Format darstellen lassen, ebenso wie beispielsweise Bilder von Unterwasserregionen[15], Videos[16] und Audio-Dateien[17], seismische Daten[18] oder Daten, die von Echosoundern aufgenommen und in proprietären Formaten vorliegen.[19]

Die meisten Daten in den Archiven sind frei verfügbar. Die WDCs akzeptieren in ihren Policies die Regeln des Open Access, wie sie in den Empfehlungen der Budapest Open Access Initiative[20] und der Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen[21] dargelegt werden. Als Weltdatenzentren sind sie auch an die Prinzipien und Verantwortlichkeiten der ICSU Datenzentren[22] gebunden. Häufig wird dem Datenproduzenten für eine gewisse Zeit das Recht zugestanden, den Zugriff auf die von ihm erhobenen Daten zu reglementieren, so dass z.B. nur registrierte Nutzer Zugang zu den Daten für weitere Auswertung haben. Damit soll gewährleistet werden, dass die Datenproduzenten in diesem Zeitraum ein bevorzugtes Recht auf Nutzung der Daten haben und sie als Grundlage weiterer Publikationen verwenden können, bevor die Daten dann für alle freigegeben werden.

---

15 Vgl. z.B. die Daten in Soltwedel et al. (2003a) und den zugehörigen Artikel Soltwedel et al. (2003b).

16 Vgl. z.B. Schönberg (2007); Schönberg; Shields (2008).

17 Ein Beispiel von „singenden Eisbergen“ findet sich bei Müller et al. (2005).

18 Vgl. z.B. Jokat (2008).

19 Die verschiedenen Hersteller von Sonar/Echolot-Systemen haben jeweils eigene Formate, in denen die gemessenen Daten vorliegen. Eine Standardisierung ist hier noch nicht absehbar. Ein Beispiel für solche Daten ist Pätzold (2009).

20 Vgl. Budapest Open Access Initiative (2011).

21 Vgl. Berliner Erklärung (2003).

22 Vgl. ICSU (1987).

In einigen Fällen wird beim Zugriff auf die Daten unterschieden zwischen der wissenschaftlichen und der kommerziellen Nutzung. Daher ist eine Authentifizierung der Nutzer in den meisten Archiven notwendig.

Die Informationen in den Datenarchiven repräsentieren vielfach den Zustand des Klimas zu einem bestimmten Zeitpunkt. Daher sind auch ältere Forschungsdaten sehr wichtig, weil sie Zeugnis geben von den seither erfolgten Veränderungen. Diese alten Daten sind damit von unwiederbringlichem Wert und ihre sichere Aufbewahrung unbedingt notwendig.

Zur Beschreibung der Daten sind bereits seit Jahren vielfältige internationale Bemühungen zu verzeichnen. Ausdruck dessen sind die in den ISO Standards festgeschriebenen Vorgaben. So definiert der ISO 19115 die für die Beschreibung von geografischen Informationen notwendigen Angaben. Er ist sehr umfangreich und beinhaltet die räumliche und zeitliche Zuordnung sowie vielfältige Informationen zu den Daten. Von den insgesamt über 400 Metadatenelementen sind etwa 20 Elemente verpflichtend. Die Norm ist wichtige Grundlage beim Aufbau von Geodateninfrastrukturen (GDI). Während in ISO 19115 das Metadatenschema an sich beschrieben wird, definiert ISO 19139 die XML-Implementierung des Schemas. Beide Standards bilden damit die Basis für die Metadatenschemata, die in den Datenarchiven der Klimaforschung verwendet werden.

Im europäischen Rahmen bemüht sich die Initiative Infrastructure for Spatial Information in the European Community (INSPIRE)[23] um eine Harmonisierung der Dateninfrastrukturen für Raumdaten, zu denen auch die Klimadaten gehören. Damit soll der Austausch von Umweltdaten zwischen unterschiedlichen öffentlichen Einrichtungen erleichtert und der Zugang zu Raumdaten in Europa vereinfacht werden. Ein Teil dieser Initiative ist auch die Definition von einheitlichen Metadaten[24], die als Grundlage die oben bereits diskutierten ISO Metadaten haben.

Bereits seit einigen Jahren werden persistente Identifier in den Datenarchiven der Klimacommunity genutzt, um eine dauerhafte Identifizierung der Datenobjekte zu gewährleisten. In dem von der DFG geförderten Projekt „Publication and Citation of Scientific Primary Data“

23 Vgl. INSPIRE (2007).
24 Vgl. INSPIRE (2010).

(STD-DOI) haben die Datenarchive WDC Climate, WDC MARE, WDC RSAT und das Datenarchiv des GFZ Potsdam gemeinsam mit der TIB Hannover ein System aufgebaut, das die Zitierbarkeit von Daten gewährleistet (DataCite[25]). Die in diesen Archiven veröffentlichten Daten werden mit persistenten Identifikatoren (DOI und URN) versehen und machen die veröffentlichten Datensätze dadurch dauerhaft findbar und zugänglich. Damit ist eine Grundvoraussetzung für deren Zitierbarkeit erfüllt.[26]

10.4 Interne Organisation

Für die Weltdatenzentren existieren wohldefinierte Workflows für den Dateningest und die Behandlung der Daten in ihrem Lebenszyklus. Sie ergeben sich teilweise bereits durch die Zugehörigkeit der Einrichtungen zu ICSU. Damit wird gesichert, dass die Daten langfristig verfügbar sind.

Die Finanzierung der Datenarchive ist in der Klimacommunity unterschiedlich geregelt. Die Weltdatenzentren werden in der Regel über den Grundhaushalt der Einrichtungen finanziert, an denen das Datenarchiv betrieben wird.

Doch nicht alle Vorhaben können mit dieser Grundfinanzierung abgefangen werden. Immer wieder gibt es Projekte, die ein hohes zusätzliches Datenaufkommen generieren. Da heutzutage elektronische Speichermedien auch mit großen Volumina relativ preiswert zu haben sind, erkennen nicht alle Wissenschaftler die Problematik, die sich beim Übergang der Daten in die Langzeitarchivierung ergibt. Hier muss es noch zu einer stärkeren vorherigen Absprache mit den Datenzentren kommen.

Allerdings wird in diesem Zusammenhang ein weiteres Problem deutlich. Es gibt derzeit noch keine verlässlichen Kostenmodelle für die Langzeitarchivierung von Daten. Die teilweise bereits erhobenen Datenpauschalen stellen nur einen ersten Versuch dar, bei hohem zusätzlichen Datenaufkommen angemessen reagieren zu können. Hier bedarf es noch weiterer Evaluierung, die umso verlässlicher gestaltet werden kann, je besser die Prozesse bei der Langzeitarchivierung in einer Einrichtung definiert

25 Vgl. DataCite (2011).
26 Vgl. Klump et al. (2006).

sind. Außerdem müssen oftmals noch Service Level Agreements formuliert und ausgehandelt werden. Darin sollten Art und Umfang der zugesicherten Leistungen des Datenarchivs festgelegt werden, damit Nutzer die Angebote unterschiedlicher Anbieter vergleichen können.

Um eine Vorstellung von den notwendigen finanziellen Ressourcen für eine Virtuelle Forschungsumgebung in einem datenintensiven Wissenschaftszweig zu bekommen, sollen hier einmal schlaglichtartig einige Investitionen im Umfeld des C3Grid aufgelistet werden. Im Rahmen der D-Grid-Initiative förderte das BMBF neben den Projekten in insgesamt drei Calls auch durch Sonderinvestitionen den Aufbau der Gridinfrastruktur in Deutschland. Dabei wurden in C3Grid Archivkomponenten für ein verteiltes C3Grid Community Repository Center (C3RC) für Langzeitarchivierung zur Unterstützung der C3-Community beschafft. Diese Speichersysteme wurden dabei in die bereits bei den Partnern vorhandene Storage-Infrastruktur eingebunden, um die vorhandenen Erfahrungen vor Ort zu nutzen und eine hohe Kosteneffizienz zu erreichen. Konkret wurden für insgesamt 2,25 Mio. € fünf Archivsysteme mit entsprechenden Medien angeschafft, die insgesamt um die fünf PB an Daten in den beteiligten Zentren (AWI, DKRZ, Universität zu Köln, Zuse-Institut Berlin, DWD) aufnehmen können.

Weitere größere Investitionen wurden in Vorbereitung auf den bereits erwähnten 5. Sachstandsbericht des IPCC getätigt. Das DKRZ wird als eines von drei Daten-Gateways eine zentrale Rolle in der international verteilten Datenhaltung spielen. Dafür wurden 1,1 Mio. € vom BMFB zur Verfügung gestellt, für die neben den für das Datenprozessing notwendigen Servern vor allem Laufwerke und Tapes mit einer Kapazität von 1,6 PB und für einen schnellen Zugriff 700 TB Platten angeschafft und in die bereits vorhandene Dateninfrastruktur des DKRZ integriert wurden.

Datenbibliotheken mit dem erwähnten Umfang können nur dann erfolgreich betrieben werden, wenn dazu speziell geschultes Personal verfügbar ist. Die technische Grundlage schaffen die IT-Experten. Die enge Anbindung an Rechenzentren ist dabei von Vorteil: Die technische IT-Infrastruktur des WDC MARE wird durch das Rechenzentrum des AWI betrieben, das WDC Climate ist Teil des Deutschen Klimarechenzentrums DKRZ. An beiden Einrichtungen gibt es Mitarbeiter für die Langzeitarchivierung der Forschungsdaten. So existiert am RZ des AWI eine Abteilung „Forschungsplattformen“,

die sich um die Datenströme aus den vom Institut betriebenen Forschungsplattformen (Forschungsschiff Polarstern, Antarktisstation Neumayer III,…) kümmert. Dazu ist die Expertise sowohl zu den datenproduzierenden Messeinrichtungen als auch zu den zugehörigen Formaten, Metadaten notwendig. Für die Datenakquise aus verschiedenen Experimenten sind Datenkuratoren zuständig, die jeweils den Datenimport in das Archiv begleiten. In ähnlicher Weise gibt es am WDC Climate Mitarbeiter, die sich der großen Datenlieferanten annehmen. Exemplarisch seien hier die Aktivitäten zur Datenhaltung im Zusammenhang mit CMIP5 (Coupled Model Intercomparison Project) genannt, in dem mit verschiedenen Modellen die globalen Simulationen für den IPCC AR5 erstellt werden. Für das am DKRZ/WDC Climate angesiedelte Daten-Gateway sind zwei Projektmitarbeiter zum Aufbau des IPCC-Datenarchivs angestellt worden, die den Datenfluss aus den Hochleistungsrechnern in die Datenarchive begleiten.

Nur ein Teil der beschriebenen Mitarbeiter im Zuständigkeitsbereich des Datenmanagements ist fest bei den Einrichtungen angestellt. Vielmehr werden viele von ihnen über Drittmittelprojekte finanziert. Dies mag sinnvoll sein, um ein zeitlich begrenztes überdurchschnittlich hohes Datenaufkommen aufzufangen, wie es das CMIP5-Projekt ist. Jedoch ist damit die Gefahr verbunden, dass durch die personellen Fluktuationen wichtige Expertise in der Datenhaltung verlorengeht.

10.5 Perspektiven und Visionen

Die Klimacommunity hat schon frühzeitig die Bedeutung einer koordinierten Datenarchivierung erkannt, um die einmal gewonnenen Informationen für breite Kreise nachnutzbar zu machen[27]. Die im Laufe der Jahre aufgebauten Infrastrukturen sind für die Anforderungen gut aufgestellt. Sie sollten parallel zu dem weiter steigenden Datenaufkommen weiter ausgebaut und verknüpft werden.

Die erhobenen Daten haben ihren Wert nicht nur für die Erstauswertung in einer Publikation. Vielmehr werden oft durch Wiederverwendung und Verbindung mit anderen Datenquellen weitere Erkenntnisse gewon-

27 Vgl. Dittert; Diepenbroek; Grobe (2001), S. 393.

nen. Daher muss bei den beteiligten Wissenschaftlern ein Bewusstsein geschaffen werden, wie wichtig eine ordnungsgemäße Archivierung von Daten ist. Sie sollte nicht einfach nur den Schlusspunkt in der Arbeit bilden, der manchmal etwas stiefmütterlich abgearbeitet wird, sondern integraler Bestandteil des wissenschaftlichen Arbeitsprozesses sein, der kontinuierlich durchlaufen wird. Hier haben die Universitäten die Aufgabe, heranwachsenden Wissenschaftlern bereits in der Ausbildung die Verantwortung für die Daten nahezubringen und Grundlagen der Langzeitarchivierung zu vermitteln. Dies wird insbesondere immer wichtiger, da die Klimaforschung durch die Klimamodellierung und das globale Monitoring zunehmend ein Wissenschaftsbereich aus der Klasse der data intensive sciences wird.

Eine weitere Herausforderung für die Datenarchive der Klimaforschung liegt darin, die dort gesammelte Information nicht nur für die Community selbst bereitzustellen, sondern sie auch für weitere Wissenschaftsdisziplinen nutzbar zu machen. Wenn Strategien für die Anpassung an die Veränderungen des Klimas entwickelt werden sollen, dann müssen die Daten zum Klimawandel aufbereitet werden, damit sie auch von Fachfremden richtig genutzt werden können. So erwarten Mediziner belastbare Aussagen zu Änderungen in der Pollenflugsaison, um Asthmatikern eine angemessene Medikation zukommen zu lassen. Die Tourismusbranche möchte vor größeren Investitionen in Wintersportgebieten wissen, ob dort auch in Zukunft mit genügend Schneefall zu rechnen ist. Die Straßenämter müssen eventuell alternative Fahrbahnbeläge ausbringen, um den mittelfristig sich ändernden Witterungsbedingungen im Winter adäquat zu begegnen. Diese beliebig ausgewählten kleinen Beispiele machen deutlich, dass ein weiter Bedarf an Datenprodukten besteht, die von den Datenzentren bereitgestellt werden sollten. Hier eröffnet sich ein weiteres Betätigungsfeld, dessen Bedeutung aber zukünftig sicher noch steigen wird.

Danksagung

Hiermit möchte ich Michael Lautenschlager (DKRZ) und Michael Diepenbroek (MARUM) recht herzlich für die Informationen zu den Weltdatenzentren WDC Climate und WDC MARE danken.