



WissGrid

Deliverable 2.3.1

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Arbeitspaket 2: Blaupausen und Beratung,
Arbeitspaket 3: Langzeitarchivierung

Leitfaden zum Forschungsdaten-Management¹

Version 0.6:
Entwurfsversion zur öffentlichen Kommentierung

¹ This work is created by the WissGrid project. The project is funded by the German Federal Ministry of Education and Research (BMBF).

Herausgegeben von

WissGrid – Grid für die Wissenschaft

Teil des D-Grid Verbundes und der deutschen e-Science Initiative

www.wissgrid.de

BMBF Förderkennzeichen: 01|G09005A-G (Verbundprojekt)

Laufzeit: Mai 2009 - April 2012

Dokumentstatus: Deliverable

Kontakt

Niedersächsische Staats- und Universitätsbibliothek Göttingen

Jens Ludwig

Abteilung Forschung & Entwicklung

Papendiek 14

37073 Göttingen

Leibniz-Institut für Astrophysik Potsdam

Harry Enke

Abteilung e-Science

An der Sternwarte 16

14482 Potsdam

Autoren

Harry Enke, Leibniz-Institut für Astrophysik Potsdam

Norman Fiedler, Institut für Deutsche Sprache Mannheim

Thomas Fischer, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Timo Gnad, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Erik Ketzan, Institut für Deutsche Sprache Mannheim

Jens Ludwig, Niedersächsische Staats- und Universitätsbibliothek Göttingen

Torsten Rathmann, Deutsche Klimarechenzentrum

Gabriel Stöckle, Zentrum für Astronomie der Universität Heidelberg



Der Inhalt dieser Veröffentlichung steht unter einer Creative Commons Namensnennung 3.0 Unported Lizenz (<http://creativecommons.org/licenses/by/3.0/>).

WissGrid, 2011

Inhaltsverzeichnis

Einleitung	4
I Aufgaben im Lebenszyklus von Forschungsdaten	9
1 Planung und Erstellung	9
2 Auswahl und Aufbewahrungsdauer	13
3 Ingest: Einspeisen und Verantwortungsübernahme	16
4 Speicherung und Infrastruktur	19
5 Erhaltungsmaßnahmen und ihre Planung	22
6 Zugriff und Nutzung	23
II Übergreifende Aufgaben des Forschungsdatenmanagements	30
7 Organisation, Management und Policies	30
8 Kosten	33
9 Rechtliche Aspekte von Forschungsdaten	36
10 Metadaten	40
11 Identifikatoren und Informationsobjekte	45
Anhang: Urheberrecht	50

Einleitung

Digitale Forschungsdaten sind wichtige Produkte der wissenschaftlichen Arbeit, in deren Erstellung viel Geld, Zeit und Expertise investiert wird. Gleichzeitig sind durch die modernen wissenschaftlichen Arbeitsinstrumente das Volumen und die Komplexität der Forschungsdaten gestiegen und der sinnvolle Umgang mit Forschungsdaten ist deutlich anspruchsvoller geworden. Vor diesem Hintergrund haben Wissenschaftsorganisationen wiederholt Anforderungen an das Management von Forschungsdaten gestellt, wie z.B. die DFG, die seit 2010 in Anträgen eine Darstellung des Umgangs mit Forschungsdaten verlangt [DFG-Antrag, S. 6]. Dieser Leitfaden und die begleitende Checkliste sollen als Instrument für WissenschaftlerInnen und Service-Einrichtungen (wie z.B. Rechenzentren) dienen, um für ein Vorhaben gemeinsam und systematisch alle wesentlichen Themen des Forschungsdatenmanagements zu untersuchen und einen Plan für das Datenmanagement aufzustellen. Die Kapitel des Leitfadens und die Abschnitte der Checkliste entsprechen sich, sodass die Kapitel als Erklärung der Checklistenabschnitte benutzt werden können.²

Unter Management von Forschungsdaten werden alle Maßnahmen verstanden, die sicherstellen, dass digitale Forschungsdaten nutzbar sind. Was dafür notwendig ist, variiert aber stark mit den verschiedenen Zwecken, für die Forschungsdaten genutzt werden sollen. Es lassen sich vier Arten von Zwecken unterscheiden:

1. Die Nutzung als Arbeitskopie für das wissenschaftliche Arbeiten,
2. die Nachnutzung von Forschungsdaten für spätere Forschung,
3. die Aufbewahrung als Dokumentation des korrekten wissenschaftlichen Arbeitens und
4. die Aufbewahrung, um rechtlichen oder anderen forschungsfremden Anforderungen nachzukommen.

Unabhängig von diesen Nutzungsarten unterliegen digitale Forschungsdaten besonderen Bedingungen, die in den letzten Jahren unter dem Thema der Langzeitarchivierung digitaler Daten behandelt wurden. Langzeitarchivierung umfasst die sogenannte Bitstream Preservation, technische Nachnutzbarkeit und inhaltliche Nachnutzbarkeit. Eine Grundvoraussetzung ist die Bitstream Preservation, d.h. die Erhaltung der Bitfolge, was angesichts der Kurzlebigkeit und Fehleranfälligkeit von digitalen Datenträgern bei großvolumigen Forschungsdaten bereits eine Herausforderung sein kann. Der Erhalt der Bitfolge gewährleistet aber noch nicht, dass Forschungsdaten immer technisch nachnutzbar sind. Die Nutzung von Forschungsdaten kann z.B. durch spezialisierte Dateiformate besondere technische Anforderungen an Software, Hardware oder Infrastruktur stellen, die langfristig schwer erfüllbar sein können. Und selbst wenn die technische Nutzbarkeit gewährleistet ist, erfordert die inhaltliche Nutzung immer Hintergrund- und Kontextwissen, das bei der Erstellung der Forschungsdaten nachvollziehbar dokumentiert werden muss. Zu einem späteren Zeitpunkt können weitere Ergänzungen notwendig werden, wenn sich z.B. in einem wissenschaftlichen Gebiet bisher selbstverständliche

² Es existieren eine Reihe von englischen Checklisten, die vom DataOne-Projekt verglichen werden, siehe <https://www.dataone.org/plans>. Die hier vorgestellte Checkliste ist insbesondere von der DCC Checkliste inspiriert, siehe http://www.dcc.ac.uk/webfm_send/431.

Annahmen, Methoden oder Begriffe ändern, die späteren NutzerInnen explizit erklärt werden müssen. Und schließlich gibt es allgemeine Herausforderungen wie die Fragen der Organisation und Finanzierung: Wer übernimmt die Verantwortung für den Erhalt der Nutzbarkeit, wie aufwändig ist das und wer wird das bezahlen? Für traditionelle Publikationen oder Dokumente öffentlicher Institutionen existiert ein System von Bibliotheken und Archiven, für die es im Bereich der Forschungsdaten nur selten eine Entsprechung gibt.

Der vorliegende Leitfaden behandelt die Aufgabenbereiche, die für das Management von Forschungsdaten bedacht werden müssen, in zwei Teilen. Einige Aufgaben können klar in einem als Lebenszyklus vorgestellten Ablauf von Schritten verortet werden (siehe [Abb. 1](#)). Andere Aufgaben hingegen spielen in jedem Abschnitt des Lebenszyklus eine Rolle (siehe [Abb. 2](#)).

Der Lebenszyklus von Forschungsdaten

Es gibt eine Vielzahl von Lebenszyklus-Modellen für digitale Informationen, von denen das Modell des Digital Curation Centers besonders elaboriert und auf Forschungsdaten ausgerichtet ist.³ Im Folgenden wird ein vereinfachtes Modell benutzt.



Abb. 1: Aufgaben im Lebenszyklus von Forschungsdaten

Planung und Erstellung Um das spätere Management von Forschungsdaten möglichst zu vereinfachen, ist es sinnvoll, die Daten schon entsprechend zu erzeugen. Ein wichtiger Aspekt in dieser Phase ist z.B. die Wahl der richtigen Standards.

Auswahl und Bewertung Nicht alle Forschungsdaten müssen und können auf Dauer aufbewahrt werden. Die Gründe, Methoden und Kriterien der Selektion und die daraus resultierende Dauer der Aufbewahrung von Forschungsdaten müssen geklärt werden.

³ DCC, The DCC Curation Lifecycle Model, <http://www.dcc.ac.uk/resources/curation-lifecycle-model>

Ingest/Übernahme Forschungsdaten, die längerfristig aufbewahrt werden sollen, müssen in eine geeignete Umgebung wie z.B. ein Datenarchiv überführt werden. In dieser Phase werden üblicherweise zusätzliche Checks, Homogenisierungen und Anreicherungen der Daten notwendig, die besonders aufwändig sind.

Speicherung Die langfristige Speicherung von Forschungsdaten mit Verfahren, die die Chancen von Datenverlust minimieren, sollte am besten von erfahrenen Anbietern von Speicherdienstleistungen übernommen werden.

Erhaltungsmaßnahmen Es ist nicht selbstverständlich, dass digitale Forschungsdaten in anderen Umgebungen als der ursprünglichen Erstellungs- und Nutzungsumgebung nutzbar bleiben. Deshalb ist es bereits im Vorfeld sinnvoll zu bedenken und zu dokumentieren, welche Anforderungen an eine technische Umgebung zur Nutzung der Daten gestellt werden und wie mit Veränderungen der Technik umgegangen werden soll.

Zugriff und Nutzung Die besten Daten nutzen wenig, wenn sie nicht gefunden werden. Wie die Daten gefunden werden können, wer autorisiert auf sie zugreifen darf und mit welchen Mitteln, sind daher ebenfalls wichtige Fragen.

Übergreifende Aufgaben

Neben den eindeutig im Lebenszyklus verortbaren Aufgaben gibt es einige Themen, die in jedem Abschnitt des Lebenszyklus wichtig sind. Es handelt sich dabei um Querschnittsthemen, die separat behandelt werden.

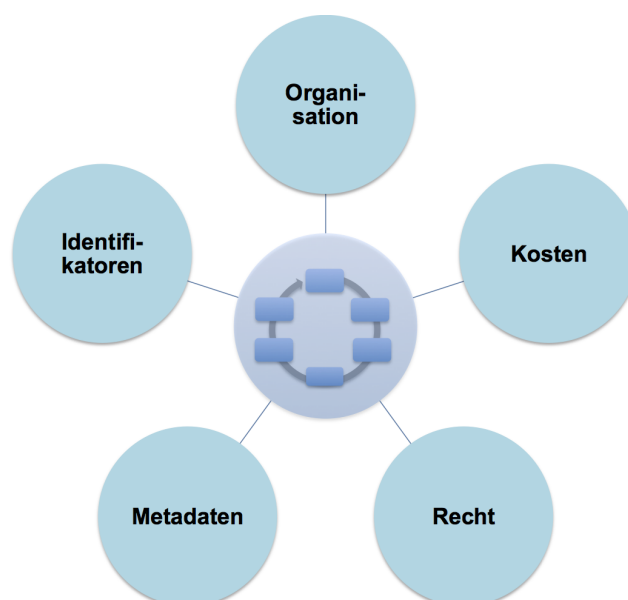


Abb. 2: Übergreifende Aufgaben des Forschungsdatenmanagements

Management, Organisation und Policies Technik allein bewahrt keine Forschungsdaten auf. Es muss immer eine Organisation dafür die Verantwortung übernehmen und mit

definierten Prozessen, die dokumentiert sind, sicherstellen, dass auch langfristig alle notwendigen Maßnahmen erfolgen. Ob es sich bei dieser Organisation dann um ein fachspezifisches Datenarchiv oder Konsortium oder um eine fachübergreifende Einrichtung wie ein universitäres Rechenzentrum handelt, ist nur von nachrangiger Bedeutung.

Recht und Ethik Forschung muss sich an geltendes Recht und ethische Forderungen halten. Für das Management von Forschungsdaten sind u.a. der Schutz personenbezogener Daten, Urheberrechte, Lizenzierung von Forschungsdaten und Vertraulichkeit zu bedenken.

Finanzierung und Förderung Ein limitierender Faktor für die Aufbewahrung von Forschungsdaten sind die damit verbundenen Kosten. Insbesondere die Kostenentwicklung und Gesamtkosten einer dauerhaften, zeitlich unbefristeten Archivierung sind offene Forschungsfragen.

Metadaten Metadaten sind strukturierte Informationen über die vorliegenden Daten und für das Management von Forschungsdaten in jeder Phase des Lebenszyklus unverzichtbar.

Identifikatoren Identifikatoren stellen einen nicht vernachlässigbaren Aspekt des Forschungsdatenmanagements dar. Ein präzises und durchdachtes Konzept zur Identifizierung von Forschungsdaten erfordert auch ein präzises Konzept der Informationsobjekte und klärt dadurch eine Reihe von wichtigen Fragen.

Benutzung der Checkliste

Die in diesem Leitfaden und der begleitenden Checkliste dargestellten zentralen Aufgaben des Forschungsdaten-Managements können weder von den Wissenschaftlern noch von den Service-Einrichtungen allein gelöst werden, sondern erfordern ihre Zusammenarbeit. Entsprechend muss auch die Planung des Datenmanagements gemeinsam erfolgen, z.B. in einer Reihe von Gesprächen oder Workshops. Soll die Checkliste als Gerüst eines Datenmanagementplans dienen, dann sind zudem einige formale Rahmenaspekte zu beachten. Dazu gehört es festzuhalten, wann und von wem der Plan erstellt und abgenommen wurde.

Es ist auch sinnvoll, im Vorfeld zu klären, für welchen Bereich im Modell des „Curation Continuum“ (siehe [Abb. 3](#)) das Datenmanagement geplant wird. Im privaten Arbeitsbereich eines Forschers (private Domäne) werden Daten individuell gesammelt und erstellt, und es muss kein Bedarf für explizite Regelung vorhanden sein; jedoch schon für den Übergang in die Arbeitsgruppe (Gruppendomäne) sind (wenigstens informelle) Regelungen erforderlich, z.B. in Bezug auf standardisierte Metadaten. Für eine längerfristige Nutzung (dauerhafte Domäne) und eine Veröffentlichung (Zugang und Nachnutzung) müssen höhere Maßstäbe angelegt werden, um die Nutzbarkeit in diesem Kontext sicherzustellen. Solche Wechsel von einer Domäne in die nächste sind oftmals kritische Phasen, die nicht nur mit veränderten Anforderungen, sondern auch mit veränderten Verantwortlichkeiten einhergehen.

In allen Fragen des Datenmanagements ist zu bedenken, dass je nach Disziplin und Szenario der Klärungsbedarf und die Antworten auf die Fragen sehr unterschiedlich ausfallen können. Auch wenn die Autoren glauben, alle wesentlichen Punkte für einen Datenmanagementplan

bedacht zu haben, so kann es weitere wichtige Aspekte geben, die hier nicht erwähnt werden. Und umgekehrt wird es oftmals Aufgaben geben, die den Beteiligten für ihren Kontext selbstverständlich erscheinen, die aber in anderen Kontexten expliziter Klärung bedürfen. Für beide Fälle ist es sinnvoll, dass Service-Einrichtungen die vorliegende Checkliste als generisches Instrument verstehen, die sie für bestimmte Zielgruppen erweitern oder einschränken sollten, wenn sie einen Bedarf nach größerer oder geringerer Granularität sehen. Die hier beschriebenen Aufgaben und Fragen sollten deshalb als Hilfsmittel, aber keineswegs als in allen Punkten zwingend oder erschöpfend angesehen werden.

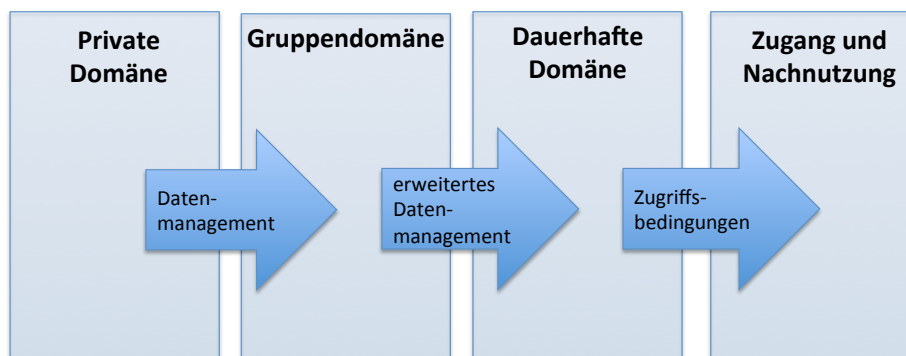


Abb. 3: Das Curation Continuum in Anlehnung an [Treloar and Harboe-Ree \[2008\]](#) und Arbeiten des DFG-Projekts Radieschen

Teil I

Aufgaben im Lebenszyklus von Forschungsdaten

1 Planung und Erstellung

Am Beginn des Lebenszyklus von Daten steht die Projektplanung und die Erstellung bzw. Erfassung sowie erstmalige Speicherung der Daten. Dieses Kapitel fasst die wichtigsten Punkte hierzu zusammen.

Planung

Die Ziele des Projektes bestimmen, welche Daten erfasst und bearbeitet werden. Für deren unmittelbare Bearbeitung und Auswertung sind herkömmliche Arten der Aufbewahrung und des Austauschs oft schon unzureichend. Darüber hinaus besitzen gesammelte Forschungsdaten einen eigenen Wert; ihre Nachnutzung wird zunehmend thematisiert und findet sich in Vorgaben und Erwartungen der Fördergeber.

Die Planung der Daten-Nachnutzung sollte bei Forschungsvorhaben von Beginn an bedacht werden, da dies die Wahrscheinlichkeit erhöht, dass in der weiteren Forschung sinnvolle Datenpflege-Richtlinien beachtet werden. Weiterhin gilt es schon zu Beginn eines Projektes zu beurteilen, wie man geplante und erwartete Datenmanagement-Aufgaben erfüllen kann und ob hierfür genug Personal und Finanzmittel vorgesehen wurden.

Die Frage der Rechte an geistigem Eigentum ist von Anfang zu klären und entsprechend zu dokumentieren; in diesem Zusammenhang ist es unumgänglich, die für eine Veröffentlichung der Daten nötigen Anforderungen und Beschränkungen zu definieren. Abschließend müssen spezifische Rollen und Verantwortlichkeiten so früh als möglich geklärt und dokumentiert werden.

Vorhandene Daten

Es ist wichtig, zu prüfen, in welcher Beziehung neue Datensätze zu bereits vorhandenen stehen. Besonders die Wichtigkeit eines Datensatzes klar darzustellen kann für spätere Nutzer und Archive nützlich bei der Bewertung der Daten sein. In diesem Kontext sollte man Möglichkeiten zur Nachnutzung bestehender Datensätze recherchieren und die technischen und inhaltlichen Bedingungen klären, unter denen neu geschaffene Daten in bereits vorhandene Datenbestände integriert und mit diesen kombiniert werden können. Auch ist zu bedenken, ob ihre weitere Verwendung rechtlich unbedenklich ist (siehe [Kapitel 9 Rechtliche Aspekte von Forschungsdaten](#), S. 36).

Ist Vernetzung und Nachnutzung ein Hauptanliegen des Projektes, so sollte man sich auch Gedanken zur Bereitstellung von Daten machen (siehe [Kapitel 6 Zugriff und Nutzung](#), S. 23).

Wie werden die Daten zur Verfügung stehen, zum Beispiel in öffentlichen Datenbanken? Gibt es Einschränkungen hinsichtlich ihrer Verwendung? Wann werden sie öffentlich zugänglich gemacht?

Arten von Daten

Die hier aufgeführte Aufzählung unterscheidet Daten nach ihrer inhaltlichen Art sowie nach der Art ihrer Erstellung und Erfassung. Eine solche Klassifizierung dient vor allem der Bewertung des Datensatzes und der Entscheidungsfindung bezüglich dessen Nachnutzung und Archivierung. Sind die Daten beispielsweise reproduzierbar, oder handelt es sich um einmalige Messungen? Hier ist es noch unerheblich, um welche technischen Datentypen es sich handelt.

Experimente: Hierbei handelt es sich um Daten, die im Prinzip erneut hergestellt werden können (obwohl dies in der Praxis schwierig oder nicht wirtschaftlich sein kann).

Modelle oder Simulationen: Bei einer Simulation kann es wichtiger sein, das Modell und die dazugehörigen Anfangsbedingungen der Simulation zu erhalten, als die damit berechneten (Roh-)Daten.

Beobachtungen: Spezifische Phänomene zu einem bestimmten Zeitpunkt oder Ort. Diese Daten repräsentieren in der Regel eine einzigartige und nicht wiederholbare Aufzeichnung eines Ereignisses.

Abgeleitete Daten: Durch die Verarbeitung „roher“ und/oder Verbindung verschiedener Daten werden durch spezielle Methoden neue produziert (die Rechte der Eigentümer der Rohdaten sind zu respektieren!). Hier ist eventuell die Provenienz der Daten und die Dokumentation der Verfahren ebenfalls relevant.

Kanonische oder Referenzdaten: Daten, die andere Daten entsprechend gemeinsamer Regeln (lat. canon „Norm, Regel“) beschreiben oder eine „Übersetzung“ dieser Daten in ein Standardformat darstellen. Der Übergang zu Metadaten ist fließend.

Werkzeuge zur Erstellung von Daten, Software

Die Erstellung oder Schaffung unterscheidet sich von der reinen Erfassung von Daten. Um beispielsweise Daten zu reproduzieren, die bei Analysen und Simulationen entstehen, ist es nicht zwangsläufig notwendig, die erzeugten Daten selbst aufzubewahren. Allerdings ist es unumgänglich, die Software und Werkzeuge, mit denen diese Daten geschaffen wurden, zu erhalten und deren Nachnutzung zu thematisieren.

Erfassung und Speicherung

Zum Zeitpunkt der Datenerfassung gilt es ausreichend Informationen über die Daten selbst zu erfassen, um eine weitere Nutzung zu ermöglichen (siehe [Kapitel 10 Metadaten](#), S. 40).

Es ist notwendig, die Kriterien klar zu definieren, unter denen ein Datensatz angelegt wird und welche zugehörigen Metadaten gesammelt werden sollen.

Für die Speicherung ist es sinnvoll, die erwarteten Datenmengen und Produktionsraten abzuschätzen. Schon bei der Planung der Datenerfassung stellt sich die Frage der Speicherung der Daten auf geeigneten Medien (siehe [Kapitel 4 Speicherung und Infrastruktur](#), S. 19). Ein weiterer Punkt ist die wohldefinierte Identifizierung von Dateien und Inhalten, wobei hier neben einer klaren Zuordnung auch die Sicherung der Qualität der Dateien durch automatische oder gedankliche Überprüfung der Inhalte gemeint ist. Auch gilt es einen sicheren Platz für die erfassten Daten (d.h. ein vertrauenswürdigen Archiv) zu finden und sicherzustellen, dass das Archiv diese Daten aufnehmen kann und wird (siehe [Kapitel 3 Ingest: Einspeisen und Verantwortungsübernahme](#), S. 16).

Kriterien für Dateiformate und Standards zur gemeinsamen Nutzung von Daten

Es muss weiter geklärt werden, in welchen Formaten und gemäß welcher Standards Daten erfasst und gespeichert werden sollen.

Die Auswahl eines geeigneten Formats zur Archivierung der Daten ist eine grundlegende Entscheidung, deren Tragweite man sich bewusst sein sollte. Der Austausch und die Wiederverwendung von Daten erfordert deren Interoperabilität; hierzu müssen Standards für die Datenerfassung, Zitierung, Annotation, Klassifizierung, Integration von neuer Software, Darstellung von Inhalten usw. eingehalten werden. Diese Standards müssen identifiziert – oder, falls nicht vorhanden, entwickelt – werden und sollten dann in einem definierten Format darstellbar sein.

Der Übergang von Dateien zu Daten ist im Rahmen dieser Kriterien fließend. Eine Behandlung von Daten durch Client-basierte Daten-Dienste geht über das Konzept einer Datei hinaus. Dienste, die beispielsweise Daten über ein Webportal anbieten, sollten die Möglichkeit der Verifizierung der Quelle etc. anbieten. Die in diesem Zusammenhang zu beachtenden Kriterien sollten ebenfalls in die Planung des Datenmanagements aufgenommen werden.

Die nachstehend aufgeführten Kriterien für Datenmanagement-Standards wurden sowohl unter Gesichtspunkten der Daten-Langzeitarchivierung(*) als auch im Hinblick auf gemeinsame Datennutzung in kooperativen Forschungsumgebungen(**) gesammelt. Das Domänenmodell aus [Abb. 3](#) auf S. 8 sollte hier Anwendung finden, d.h. die Kriterien sind in den unterschiedlichen Domänen unterschiedlich relevant.

Einfachheit*:** Technische Komplexität erschwert fehlerfreie Entschlüsselung und Nutzung. Je mehr Wissen zur Nutzung notwendig ist, desto eher kann ein Teil des notwendigen Wissens verloren gehen.

Flexibilität:** Große Datenmenge und mehrere Objekte sind in einer Datei speicherbar. Es besteht die Möglichkeit zum Zugriff auf Untereinheiten und zur parallelen Verarbeitung.

Nutzbarkeit:** Verbreitung innerhalb der Community. Verfügbarkeit von Client-Software. Anbindung der Daten an Aufbereitungs-Software. Anwendbarkeit für verschiedenste Szenarien. Stabilität.

Standardisierung^{*,}:** Eine formale Beschreibung/Spezifikation existiert und ist frei verfügbar. Eine Spezifikation ermöglicht es, das Format zu verstehen und eigene Nutzungssoftware zu schreiben.

Referenzierbarkeit/Interoperabilität^{*,}:** Die Daten sind klassifizierbar, kommentierbar, mit anderen Daten verknüpfbar und global referenzierbar.

Datenintegrität^{*}: Verifizierbarkeit der Quelle. Überprüfbarkeit des Inhalts der Daten. Möglichkeit zur automatischen Fehlererkennung.

Provenienz^{*,}:** Möglichkeit, die Bearbeitungshistorie der Daten aufzuzeichnen.

Robustheit^{}:** Hohe Fehlertoleranz bei hardwareseitigen Speicherfehlern

Unabhängigkeit^{*,}:** Datenverarbeitung ist nicht von spezieller Hard- oder Software abhängig. Nutzbarkeit der Daten durch Software unterschiedlicher Versionen.

Schutzmechanismen^{}:** Kopierschutz und Verschlüsselungen sind für die Langzeitarchivierung von Dateien problematisch, da eine Modifikation technisch notwendig werden kann. Gemeinsame Datennutzung macht aber klare Regelungen und Einschränkungen durch Autorisierung und Authentifizierung notwendig.

Selbstdokumentation/Datenbanken: Integration von Metadaten in Daten erleichtert das Verständnis der Daten und verringert Abhängigkeit von externen Datenquellen, andererseits erleichtert getrennte Lagerung von Daten und Metadaten den Zugriff, erhöht aber die Gefahr, dass die Verbindung zwischen Metadaten und Daten langfristig verloren geht. *(Hier kann keine allgemein gültige Richtlinie gegeben werden, es müssen die Vor- und Nachteile der gewählten Vorgehensweise von Fall zu Fall abgewogen werden.)*

Anwendungsfall Klimaforschung

Die Klimaforschung produziert und nutzt unterschiedliche Daten [Overpeck et al., 2011], vor allem Beobachtungsdaten (z.B. Satellitendaten) und Modellrechnungen. Die gespeicherten Daten werden durch die Klimaforschung selbst, aber auch durch die Klimafolgenforschung genutzt. Ein nicht unerheblicher Teil besitzt gesamtgesellschaftliche Relevanz. So sind die Konsortialrechnungen des IPCC (Intergovernmental Panel on Climate Change [IPCC], eine Unterorganisation der UNO) Grundlage für Empfehlungen an die Politik.

In Projekten von besonderer Relevanz wird schon die Datenproduktion gemeinsam von Produzenten, Archiven und Nutzern geplant. Ein solches Projekt ist CMIP5 (Coupled Model Intercomparison Project, Phase 5) [CMIP5]. CMIP ist ein Protokoll zur Analyse von Rechenergebnissen allgemeiner Zirkulationsmodelle (GCM, General Circulation Model) mit gekoppeltem Atmosphären- und Ozeanmodell. CMIP5 liefert eine Infrastruktur für Diagnose, Validierung, Vergleich und Dokumentation solcher Klimamodelle. Es wird erwartet, dass die dabei produzierten Rechenergebnisse in den nächsten Weltklimabericht des IPCC einfließen werden, wie das auch mit den Ergebnissen früherer CMIP-Phasen geschehen ist.

Vor Beginn der Ausführung der Modellrechnungen wurden diese im Detail geplant [CMIP5 Experiment Design], einschließlich deren einheitlicher Benennung für die spätere Speicherung

der Ergebnisse. Dateiformat und Metadatenkatalog sind ebenfalls festgelegt [CMIP5 Datenbeschreibung], und zwar auf NetCDF [NetCDF] und CF [NetCDF CF Metadata Convention]. Eine dreistufige Qualitätskontrolle [CMIP5 Qualitätskontrolle] wird Bestandteil des Ingest sein. Für alle Core-CMIP5-Daten – das sind diejenigen, die in [CMIP5 Experiment Design] angefordert worden sind – ist die Veröffentlichung mit DOI-Vergabe und erneuter Qualitätskontrolle vorgesehen (näheres siehe Kapitel 11, S. 47 ff.). Die Planung sieht die Beteiligung weit entfernter Datenzentren vor. So sollen die Core-CMIP5-Daten nicht nur am WDCC [WDCC] archiviert, sondern zweifach repliziert werden, indem sie auch an das PCMDI (Program for Climate Model Diagnosis and Intercomparison, USA) [PCMDI] und das BADC (British Atmospheric Data Centre) [BADC] geschickt werden. Die bei der Replikation ausgetauschten Datenmengen werden so groß sein, dass diese nicht über das Netz transportiert, sondern auf Festplatten verschickt werden sollen. Die Festplatten sollen wiederverwendet werden.

2 Auswahl und Aufbewahrungsdauer

Gründe zur Aufbewahrung

Wissenschaftliche Daten werden aus ganz unterschiedlichen Gründen aufbewahrt:

- Arbeitskopie: Die Daten werden für die aktive Arbeit während des Projektes gesichert.
- Nachweis der guten wissenschaftlichen Praxis: Die Daten sind Grundlage einer Publikation.
- Nachnutzung: Die Daten sind wichtig für spätere Forschung.
- Auflagen und Selbstverpflichtung
- Dokumentation: Die Daten sind gesellschaftlich relevant, z.B. Grundlage einer politischen Entscheidung.

Bei der *Arbeitskopie* handelt es sich oft um ein Zwischenergebnis oder um Vorläuferversionen der Endfassung. Eine Langzeitarchivierung ist in der Regel nicht erforderlich. Der Vertrag über die Archivierung sollte das Recht einschließen, auch den Lesezugriff auf die eigene Arbeitsgruppe beschränken zu können.

Sind die Daten *Grundlage einer Publikation*, sollten die „Regeln zur Sicherung guter wissenschaftlicher Praxis“ [Max-Planck-Gesellschaft] Anwendung finden. In dem Dokument der Max-Planck-Gesellschaft wird für Primärdaten eine Aufbewahrungsdauer von – wenn möglich – mindestens zehn Jahren gefordert. Darüber hinaus müssen alle wichtigen Schritte einer Forschungsarbeit durch Protokollierung nachvollziehbar gemacht und die Protokolle ebenfalls mindestens zehn Jahre aufbewahrt werden. Die Deutsche Forschungsgemeinschaft hat sich entsprechende Regeln gegeben [Deutsche Forschungsgemeinschaft, 1998]. Gute wissenschaftliche Praxis erhöht die Nachvollziehbarkeit wissenschaftlicher Arbeiten und erleichtert die Aufklärung von Fehlern und Fälschungen. Auch immer mehr Verlage machen eine Veröffentlichung in ihren Medien vom öffentlichen Zugang zu den zugehörigen Forschungsdaten abhängig.

Die Ermöglichung der *Nachnutzung* ist oft ein weiterer Grund für eine Archivierung. Ein Großteil der Forschungsdaten kann nur mit erheblichem Aufwand erstellt werden, und vie-

le Daten können überhaupt nicht oder nicht effizient reproduziert werden. Forschungsdaten sollten als Ressource begriffen und nicht dauerhaft zurückgehalten werden, denn häufig eignen sich einmal produzierte Daten für mehrere Forschungszwecke, oft auch solche, an die bei Erzeugung der Daten noch gar nicht gedacht war. Die mühevollte Arbeit der Datenerstellung sollte belohnt werden, indem das Archiv den Nachnutzer vertraglich verpflichtet, in Veröffentlichungen die Herkunft der Daten in Form eines Zitats kenntlich zu machen. Außerdem sollte Datenerzeugern eine angemessene Sperrfrist für die Erstauswertung eingeräumt werden.

Einer gesetzlichen *Auflage* zur Archivierung muss selbstverständlich nachgekommen werden. So schreibt § 1 der Gentechnikaufzeichnungsverordnung (GenTAufzV) vor, dass bei gentechnischen Arbeiten oder Freisetzungen Aufzeichnungen zu führen und aufzubewahren sind. Die Röntgenverordnung (RöV) legt in § 28 und die Strahlenschutzverordnung (StrlSchV) in § 85 fest, welche Aufzeichnungen bei der Strahlenanwendung am Menschen angefertigt und aufbewahrt werden müssen. Auch die Landesberufsordnungen für Ärzte enthalten bestimmte Dokumentationspflichten. Eine vertragliche Verpflichtung zur Archivierung kann sich ebenso aus den Anforderungen des Projektträgers oder der Mitgliedschaft in einer Forschungsgemeinschaft ergeben, beispielsweise durch Regeln zur Qualitätssicherung. Im kommerziellen Umfeld können auch Geschäftsregeln, Produkthaftung oder Basel II Anlass zur Archivierung sein.

Datenauswahl

Die Selektion, welche Daten aufgehoben werden und welche nicht, muss in transparenter und nachvollziehbarer Weise erfolgen [Whyte and Wilson, 2010]. Dabei sollte vermieden werden, dass allein die Sichtweise einer Person oder Gruppe zum Tragen kommt. Am besten gibt sich das Projekt selbst ein Regelwerk für die Datenauswahl, in dessen Erstellung neben Datenzentren auch Datenerzeuger und Nachnutzer einbezogen werden sollten. Die Selektionsregeln sollten auch vorgeben, wer welchen Teil der Datenbewertung vornimmt.

Bei der Selektion wird darüber entschieden, ob die zur Aufnahme vorgeschlagenen Daten *archivwürdig* und *archivfähig* sind. *Archivwürdig* sind sie, wenn eines der obigen Relevanzkriterien erfüllt ist. Die Archivwürdigkeit kann zusätzlich an bestimmte formale Qualitätskriterien gebunden sein, z.B. daran, dass Folgendes mit den Daten mitgeliefert wird:

- Zitierungen
 - Zitierung der wissenschaftlichen Methode, Normen, Hilfsmittel
 - Zitierung rechtlicher Grundlagen
 - Nachweis von zugehörigen Gegenständen, die in Museen oder Sammelstellen lagern, z.B. Funde, Saatgutproben
- Rohdaten wie z.B. Originalbildmaterial
- Provenienzdaten, in denen die genaue Vorgehensweise protokolliert ist
- Fachgutachten

Archivfähig sind Daten, wenn die technischen Voraussetzungen für die Archivierung erfüllt sind. Digitale Forschungsdatenarchive geben häufig Datenformat und Metadaten-Ausstattung vor. Eine Beschränkung der im Archiv erlaubten Datenformate verringert den Umfang der

zur Pflege benötigten Kenntnisse und Software und damit Aufwand und Kosten. Zur Pflege-Software gehören Standardwerkzeuge zur Formatvalidierung und -konvertierung, deren Zahl direkt von der Zahl der erlaubten Formate abhängt.

Nicht nur neue Daten können einer Selektion unterzogen werden. Es sollte auch in regelmäßigen Abständen überprüft werden, ob Daten, die sich schon lange im Archiv befinden, noch weiter dort aufbewahrt werden sollen; für eine solche Wiedervorlage kann und sollte das Projekt bzw. die Forschungsgemeinschaft verbindliche Regeln aufstellen. Dabei sollte berücksichtigt werden, dass sich die Gründe für die Nachnutzung verändern können und ein anderer als der ursprüngliche Zweck sogar in den Vordergrund des Interesses rücken kann.

Aufbewahrung

Offene Dateninfrastrukturen haben zur Folge, dass sich Datenerzeuger, Manager und Nutzer nicht mehr unbedingt gegenseitig kennen. Vor diesem Hintergrund erscheint es nicht länger angemessen, die Entscheidung über die weitere Aufbewahrung von Daten allein auf fachinterne Kriterien zu stützen. Natürlich sollten fachbezogene Kriterien weiterhin ein starkes Gewicht haben, für die Entscheidung sollte aber weitere hinzugezogen werden. Ein einfach zu ermittelndes Maß für das Nutzerinteresse ist die Zahl der Abrufe, jedoch sollte dies wegen der hiermit verbundenen Manipulationsmöglichkeiten auch nicht das einzige Kriterium sein. Ein geeigneteres Maß kann die Zahl der Zitierungen sein, wenn das Archiv diese erfasst und sammelt. Außerdem können regelmäßige Analysen der Nutzerzusammensetzung helfen, Trends in den Nutzerinteressen frühzeitig zu erkennen. Wenn gesetzlich oder vertraglich ein Löschdatum festgeschrieben ist, hat das Archiv selbstverständlich keine Wahl und muss dem folgen.

Weiterführende Literatur

- Für die Auswahl digitaler Objekte allgemein sowie Auswahlkriterien für Netzpublikationen:
Nestor Handbuch, Version 2.3 ([Neuroth et al., 2010]): Kapitel 3.5, Auswahlkriterien (<http://nbn-resolving.de/urn:nbn:de:0008-2010071949>).

Anwendungsfall Klimaforschung

Die in den Archiven vorgehaltenen Messdaten repräsentieren den Zustand der Umwelt zum jeweiligen Zeitpunkt und können nicht durch Wiederholung der Messung erneut erhoben werden. Ältere Klimamodellrechnungen können nur mit sehr hohem Aufwand wiederholt werden, weil die damalige Hardware nicht mehr existiert und die Programme erst auf die jetzige Hardware portiert werden müssten.

Das World Data Center for Climate (WDCC) [WDCC] ist beschränkt auf verarbeitete Klimadaten. Es handelt sich dabei vorwiegend um Ergebnisse von Modellrechnungen und um solche Beobachtungsdaten, die der Modellvalidierung dienen, z.B. Niederschlagsdaten.

Die Daten dürfen nur in ganz bestimmten Dateiformaten vorliegen, nämlich als ASCII-Textdatei, GRIB [GRIB] oder NetCDF [NetCDF]. Die beiden letzten sind Binärformate, die Header besitzen und in der Klimaforschung üblich sind. Die Daten sollen außerdem den CF-Standard [NetCDF CF Metadata Convention] erfüllen, der z.B. Koordinatensysteme sowie Namen physikalischer und chemischer Größen vorgibt.

Die Daten werden mindestens zehn Jahre aufbewahrt, eine Höchstdauer ist nicht vorgesehen. Ältere Rechenergebnisse sind durchaus gefragt, da diese wichtige Vergleichsdaten für die Entwicklung neuer Klimamodelle sind.

3 Ingest: Einspeisen und Verantwortungsübernahme

Der Begriff *Ingest* bezeichnet den Prozess des Hinzufügens von Daten zu einem Archiv. Zum Ingest gehören alle Vorgänge, die zwischen der Zustimmung für die Aufnahme und dem Ende des Einfüllens ins Archiv liegen. Der Ingest kann gesammelt kurz vor Projektende oder verteilt über die gesamte Projektlaufzeit stattfinden.

Verfahren

In Anlehnung an Digital Curation 101 [Digital Curation 101: Ingest] gehören zum Ingest die folgenden Arbeitsschritte:

- **Transport** der Daten für den Ingest
- **Vorbereitung** der Daten
 - Vergabe eines internen, archivweit eindeutigen **Identifikators**,
 - Test auf **Schadsoftware** in den Daten
 - Übernahme, Extraktion oder Erzeugung der zugehörigen **Metadaten**
 - ggf. **Formatkonvertierung**
 - **Validierung** technischer Details (Daten- und Metadatenformat)
 - Prüfung der Daten und Metadaten auf **Vollständigkeit und Richtigkeit**
 - Aufteilung/Zusammenfassung der Daten in **Containerdateien**
- **Einfüllen** ins Archiv, dabei
 - Erzeugung von **Prüfsummen** zur späteren Prüfung auf ungewollte Veränderungen

Die Arbeitsschritte können je nach Art der Daten sowie Zweck und Aufbau des Archivs unterschiedlich sein. Für *sensible Daten* können z.B. zusätzliche Schritte wie eine Beschränkung des Zugangs und eine Verschlüsselung erforderlich sein. Auch die Reihenfolge der Arbeitsschritte kann variieren. Einige der Arbeitsschritte sollen nun näher beschrieben und diskutiert werden.

Der *Datentransport* kann über das Netzwerk erfolgen oder es werden Datenträger versandt. Wenn die Datenmengen so groß sind, dass eine hohe Beanspruchung des Netzwerks über eine längere Zeit oder sogar ein Abbruch wegen Zeitüberschreitung zu erwarten ist, werden gern Festplatten verschickt. Externe Festplatten besitzen eine hohe Speicherkapazität und lassen sich leicht an bestehende Computersysteme anschließen.

Zur Vorbereitung der Daten gehört die *Zusammenstellung der erforderlichen Metadaten*. Eventuell kann ein Teil der Metadaten maschinell aus den Daten ausgelesen werden (Metadatenextraktion). Die übrigen Metadaten werden meist vom Produzenten geliefert. Das Datenzentrum ergänzt eventuell noch Provenienzinformationen bezüglich des Ingest, d.h. Zusatzinformationen, die den Ablauf des Ingest protokollieren.

Unter *Validierung* soll hier eine Prüfung digitaler Objekte auf technische Funktionsfähigkeit verstanden werden. Dabei wird untersucht, ob die erforderlichen technischen Spezifikationen erfüllt sind. Häufig wird eine Formatvalidierung durchgeführt, d.h. es wird geprüft, ob Daten bzw. Metadaten in einem gültigen Format vorliegen. Bei zusammengesetzten Formaten wie z.B. PDF müssen für eine vollständige Formatvalidierung auch die eingebundenen Objekte auf Gültigkeit ihres Formats geprüft werden.

Eine *Prüfung der Daten auf sachliche Richtigkeit* können nur Fachleute leisten. Eine Möglichkeit ist, die Datenproduzenten diese Prüfung selbst vornehmen zu lassen. Die Produzenten bestätigen nach erfolgter Prüfung die Richtigkeit und Vollständigkeit der Daten gegenüber dem Datenzentrum. Selbstverständlich birgt eine solche Selbstüberprüfung die Gefahr, dass die Produzenten aufgrund der intensiven Beschäftigung mit den Daten eventuell vorhandene durchgängige Fehler bzw. Abweichungen von zuvor festgelegten Anforderungen nicht bemerken. Dieses Risiko kann u.U. durch gezielte Zuordnung von fachkundigen – möglicherweise externen – Gutachtern minimiert werden. Im Falle numerischer Daten ist eventuell eine Plausibilitätskontrolle möglich, welche allerdings auch schon sehr gute Kenntnisse der verwendeten Forschungsmethode und des vorliegenden Datenformats erfordert.

Wissenschaftliche Untersuchungen, Experimente und numerische Rechnungen können nur reproduziert oder rekonstruiert werden, wenn alle wichtigen Schritte nachvollziehbar sind. Beim Ingest sollte deshalb geprüft werden, ob entsprechende Informationen enthalten sind.

Die *Prüfung der Metadaten* auf Vollständigkeit und formale Korrektheit kann anhand eines Katalogs von Pflicht-Metadaten erfolgen, der zuvor in der Community beschlossen worden ist.

Besteht ein Werk oder Dokument aus mehreren zusammengehörigen Dateien, so ist es oft zweckmäßig, diese in eine *Containerdatei* zu packen und als eine Einheit zu archivieren. Containerdateien sollten einen schnellen Zugriff auf die darin befindlichen Objekte ermöglichen, ohne dass der Container erst ganz ausgepackt werden muss.

Verantwortungsübernahme

Schon vor der Übernahme der Verantwortung müssen die wesentlichen rechtlichen Aspekte zwischen Produzent und Datenzentrum geklärt worden sein, um beiden Seiten Rechtssicherheit zu geben. Falls es keine gesetzlich vorgeschriebene Ablieferungspflicht gibt, die für

sich genommen schon die Archivierungstätigkeit regelt, muss zumindest eine Übereinkunft (Lizenzvereinbarung) für den urheberrechtlich wichtigen Bereich getroffen werden [Beinert et al., 2008].

Beim Ingest sollten Produzent und Archiv eine Übernahmevereinbarung abschließen, die die Details der Übernahme der Verantwortung beschreibt und die bisherigen Schritte der Datenvorbereitung dokumentiert. Bestandteile der Übernahmevereinbarung sollten in Anlehnung an nestor [Beinert et al., 2008] insbesondere sein:

- die Liste der zu archivierenden Werke
- die Liste der diese Werke ausmachenden Datenobjekte (z.B. Dateien)
- die zu ihrer Archivierung notwendigen bzw. gewünschten organisatorischen und technischen Rahmenbedingungen (z.B. Aufbewahrungsdauer, Sperrfrist, Zahl der Kopien⁴)
- die erforderlichen Metadaten
- Kostenschätzung
- die rechtlich handelnden Parteien
- Regelungen zu Urheberrecht und Haftung
- Zeitplan für die Durchführung der Informationsübernahme

Alle wichtigen Arbeitsschritte des Ingest sollten protokolliert werden. Dieses Protokoll sollte ebenso dauerhaft erhalten bleiben wie die gespeicherten Inhalte. Das Protokoll sollte eine Liste der aufgenommenen Datenobjekte, den Namen des Produzenten, alle Transformations- und Validierungsschritte einschließlich der Prüfergebnisse und die Zeitstempel aller wichtigen Schritte enthalten [Beinert et al., 2008]. Datenproduzent und Archiv sollten außerdem vereinbaren, wer im Fehlerfall für welche Schritte zur Fehlerbehandlung verantwortlich ist.

Weiterführende Literatur

- nestor materialien 10: Wege ins Archiv. Ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv. Version I ([Beinert et al., 2008]), (<http://nbn-resolving.de/urn:nbn:de:0008-2008103009>).

Anwendungsfall Klimaforschung

Nur im Ausnahmefall werden Daten durch das World Data Center for Climate (WDCC) [WD-CC] vom Produzenten geholt. Im Regelfall stellt das WDCC Platz zur Zwischenspeicherung der Daten zur Verfügung. Der Produzent kopiert die zum Ingest vorgesehenen Daten in diesen zur Verfügung gestellten Bereich. Die Verantwortung für den ordnungsgemäßen Datentransfer liegt beim Produzenten. Das WDCC führt an dieser Stelle keine Prüfsummenkontrolle durch.

Der Produzent ist verpflichtet Metadaten abzuliefern, und zwar auch dann, wenn dieselben Informationen bereits in den Headern der Datenfiles enthalten sind. Die Metadaten können

⁴ Für eine umfassende Liste organisatorischer und technischer Rahmenbedingungen siehe [Bitstream Preservation]

mit Hilfe einer Webapplikation online eingetragen und direkt in eine temporäre Datenbank geschrieben werden. Wenn die neu eingetragenen Metadaten die Qualitätskriterien erfüllen, brauchen sie vom DKRZ nur noch in die Produktionsdatenbank übernommen zu werden. Die Metadaten dürfen aber auch in XML eingebettet als Textdatei abgegeben werden. Eine weitere Möglichkeit ist, XML online mit Hilfe von GeoNetwork [GeoNetwork] zu erzeugen. GeoNetwork ist eine Open-Source-Software des OSGeo-Projektes [OSGeo], die auch von der FAO, WHO, UNEP und vielen anderen Organisationen verwendet wird.

Im Rahmen der technischen Qualitätskontrolle wird überprüft, ob versehentlich leere Dateien und somit überhaupt keine Daten geliefert wurden. Wenn die Daten in den Formaten NetCDF [NetCDF] oder GRIB [GRIB] angeliefert wurden, wird das Format validiert. Je nach Projekt werden weitere Tests durchgeführt, z.B. auf doppelt vorhandene Zeitstempel geprüft: Für ein und dieselbe Zeit zwei Datensätze zu haben, deutet auf einen Fehler hin.

Auch die wissenschaftliche Qualitätskontrolle ist stark vom Projekt abhängig, in dessen Rahmen die Daten produziert wurden. In vielen Fällen ist eine Kontrolle der Daten auf Richtigkeit und Vollständigkeit nicht möglich oder seitens der Produzenten nicht gewünscht. Vom Projekt ist abhängig, ob der Wertebereich von Variablen kontrolliert wird, beispielsweise sollen Daten für die relative Feuchte zwischen null und eins liegen, dies ist in der Praxis aber nicht immer der Fall. Messdaten besitzen eine durch die begrenzte Messgenauigkeit bedingte Toleranz, außerdem kann es auf natürliche Weise zur Übersättigung kommen (Werte > 1). Ergebnisse von Modellrechnungen können aufgrund numerischer Ungenauigkeiten ebenfalls außerhalb des eigentlich zulässigen Intervalls liegen, deshalb muss das Modell jedoch nicht insgesamt schlecht sein. Welche Abweichung toleriert werden kann, können nur Experten entscheiden.

Metadaten werden durch die Ingest-Software auf formale Korrektheit geprüft.

- Metadaten, die aus einer Auswahlliste kommen, müssen einen der erlaubten Werte besitzen. Die erlaubten Werte sind in einer Datenbank abgelegt und beruhen auf Standards.
- Für bestimmte Einträge gibt es Vorgaben bzgl. der Länge.
- Bestimmte Einträge (Datensatznamen) dürfen nicht mehrfach im Archiv vorkommen.
- Ebenfalls wird geprüft, ob das XML valide ist, in dem die Metadaten eingebettet sind.

Am WDCC werden die Daten vor der Einspeisung ins Archiv in Containerdateien gepackt. Die Technik dazu hat das Deutsche Klimarechenzentrum selbst entwickelt, um Lizenzgebühren zu sparen. Der Zugriff auf Daten ist ohne vollständiges Auspacken des Containers möglich.

4 Speicherung und Infrastruktur

Die Speicherung von Forschungsdaten ist eine der grundlegendsten Aufgaben im Datenmanagement. Wenngleich organisatorische Aufgaben und Arbeitsabläufe oftmals die größten Schwierigkeiten darstellen und Speicherkapazitäten in der Wahrnehmung vieler Nutzer immer günstiger, einfacher und zuverlässiger werden, so bleibt doch die Speicherung von Forschungsdaten eine Herausforderung.

Grundlegende Faktoren, welche die Speicherung beeinflussen, sind:

- die Größe der Datensätze,
- die Anzahl der Datensätze und
- wie häufig auf die Datensätze zugegriffen werden soll.

Die derzeitigen Entwicklungstendenzen in Bezug auf Datenmengen, Speicherkapazitäten und Netzwerkkapazitäten lassen dies auch nicht als vorübergehende Schwierigkeiten erscheinen. Zwar wachsen die Kapazitäten von typischen Speichermedien exponentiell (das sogenannte Kryders Law in Anlehnung an Moores Law), aber auch die produzierten Datenmengen wachsen durch verbesserte Hard- und Software (wie z.B. die Verbesserung der Bildsensoren) in einem ähnlichen oder sogar stärkeren Maße. Die Netzwerkkapazität zum Transport der Daten wächst hingegen nicht unbedingt im gleichen Umfang, sodass es zu einer tendenziellen Verlangsamung des Datenzugriffs kommen kann.

Die Speicherung muss nicht immer durch die Institution, die die Verantwortung für das Forschungsdaten-Management übernimmt, selbst durchgeführt werden, sondern sie kann unter Umständen ausgelagert oder durch einen Verbund von Datenzentren/Archiven geleistet werden. Details zu den Erwägungen, wer die Speicherung übernimmt, finden sich im Kapitel „Management, Organisation und Policies“. Unabhängig davon, wer die Speicherung letztendlich durchführt, werden die wesentlichen Anforderungen durch folgende Faktoren vorgegeben:

- Integrität,
- Vertraulichkeit
- Verfügbarkeit und Nutzung

Die Sicherung der Integrität von Forschungsdaten auf der Speicherebene wird Bitstream Preservation genannt. Der prinzipielle Ansatz ist, genügend Kopien der Daten vorzuhalten, die möglichst wenig fehleranfällig und möglichst unabhängig voneinander sind (also z.B. an unterschiedlichen Orten mit unterschiedlichen Technologien gespeichert werden). Wenn die Kopien angemessen häufig auf ihre Integrität überprüft und fehlerhafte Kopien ersetzt werden, kann die Chance eines Datenverlusts sehr niedrig gehalten werden, auch wenn immer eine theoretische Verlustmöglichkeit besteht. Die Schwierigkeit besteht darin, eine sinnvolle Balance zwischen Integritätsanforderungen und Aufwand zu finden.

Die Vertraulichkeit sowie die Verfügbarkeit und Nutzbarkeit sind zwar zuerst Fragen des Zugriffs und werden im nächsten Kapitel detaillierter behandelt, aber sie haben direkte Auswirkungen auf die Speicherung und notwendige Infrastruktur. Es ist wichtig, die Nutzungsszenarien realistisch zu planen, insbesondere wie häufig auf welche Datenbestände zugegriffen wird und zu welchem Zweck. Daten, auf die sehr häufig und in nicht vorhersehbaren Mustern zugegriffen wird und die zudem mit nur geringer Verzögerung bereit stehen müssen, müssen wahrscheinlich eher auf Festplatten („online“) als z.B. auf wesentlich günstigeren Magnetbändern („nearline, offline“) gespeichert werden. Die Netzwerk-Infrastruktur und die Übertragungsprotokolle müssen die vom Nutzungsszenario benötigte Transfargeschwindigkeit unterstützen. Gerade bei großen Datenmengen können für die Verarbeitung und einen schnellen Zugriff Grid- und Cloud-Technologien sehr sinnvoll sein, insbesondere in Szenarien mit einer verteilten Speicher-Infrastruktur. Für spezielle Berechnungen oder Visualisierungen

mit diesen Datensätzen kann es außerdem am effektivsten sein, wenn spezialisierte Hardware auf der Seite des Datenzentrums vorhanden ist.

Weiterführende Literatur

- Jens Ludwig, Torsten Rathmann, Harry Enke, Florian Schintke: Bitstream Preservation: Bewertungskriterien für Speicherdienste. WissGrid Arbeitspaket 3: Langzeitarchivierung von Forschungsdaten, 2011 ([[Bitstream Preservation](#)]), (<http://www.wissgrid.de/workgroups/ap3/2011-03-08--bitstream-preservation.pdf>).

Anwendungsfall Klimaforschung

Um die wachsenden Datenmengen bewältigen zu können, wird am World Data Center for Climate (WDCC) [[WDCC](#)] ein High Performance Storage System (HPSS) betrieben, d.h. die Daten werden zunächst temporär auf Festplatten geschrieben und später automatisch auf Bandkassetten geschoben. Am WDCC hat jede Bandkassette zur Zeit eine Kapazität von einem Terabyte. [Abb. 4](#) zeigt das Innere eines Storage-Containers, der 10000 Bandkassetten fasst. Je drei solcher Storage-Container stehen in zwei getrennten Brandabschnitten. Von allen Daten des WDCC gibt es zwei Kopien: Eine Kopie liegt im ersten, die andere im zweiten Brandabschnitt. Zum Schutz vor Diebstahl ist der Zugang zu den Tape-Libraries auf einige wenige Mitarbeiter beschränkt. Den Schutz vor klimatischen Einflüssen gewährleistet eine Klimaanlage.



Abb. 4: Innenleben eines Storage-Containers. Auf den Schienen laufen die Tape-Roboter, die die Bandkassetten zu den Lesegeräten bringen.

Zur Erkennung von Fehlern werden Checksummen geschrieben, bei Bändern auf Blockebene und auf höheren Ebenen. Dort findet gegebenenfalls auch eine Fehlerkorrektur statt, ein End-to-End-Checksumming ist im Aufbau. Bei einem Umstieg auf neue Medien, welcher alle drei bis fünf Jahre stattfindet, werden die Daten, die in der Regel unverschlüsselt sind, umkopiert.

5 Erhaltungsmaßnahmen und ihre Planung

Um digitale Daten langfristig nutzbar zu halten, sind eine Reihe von Maßnahmen notwendig, die im weiteren Sinne auch alle anderen in diesem Leitfaden behandelten Themen umfassen. Im engeren Sinne sollen in diesem Kapitel die spezifischen Maßnahmen zur Sicherung der technischen und intellektuellen Nachnutzbarkeit behandelt werden.

Erhaltungsmaßnahmen werden notwendig, wenn sich die Anforderungen der Zielgruppe bzw. die Zielgruppe selbst oder die verfügbaren Technologien und Verfahren ändern. Beispiele dafür sind neue Daten- oder Dateiformate, neue Schnittstellen, die von der Zielgruppe für die Arbeit mit neuen Softwareprogrammen oder Arbeitsumgebungen benötigt werden, neue wissenschaftliche Standards oder Arbeitsweisen, die eine Umrechnung in neue Maßeinheiten oder zusätzliche Parameter als Hintergrundinformation erfordern, oder auch die Erweiterung der Zielgruppe auf Laien. Wenn in diesen Zusammenhängen von einem Veralten von z.B. Dateiformaten gesprochen wird, so bedeutet dies in den seltensten Fällen, dass keine funktionierende Nutzungsumgebung mehr verfügbar wäre, sondern dass relativ zu den aktuellen Anforderungen mit den „veralteten“ Mitteln kein effizientes Arbeiten mehr möglich ist. C64-Software und mit ihr erstellte Daten lassen sich üblicherweise gut mit Emulatoren nutzen, aber es entspricht nicht mehr den modernen Erwartungen an Nutzungsumgebungen.

Eine wichtige Aufgabe für die Erhaltung von Daten ist zuerst überhaupt festzustellen, dass es relevante Änderung in der Zielgruppe oder der Technologielandschaft gibt. Dafür ist die regelmäßige Untersuchung der eingesetzten und neuen Technologien (Technology Watch) und der Anforderungen und des Hintergrundwissens der Zielgruppe (Community Watch) z.B. durch ein Datenarchiv notwendig. Bereits vor der Übernahme in ein Datenarchiv sollte eine Dokumentation der ursprünglichen Technologien und Anforderungen erfolgen.

Wie mit diesen Änderungen umzugehen ist, muss in einem gründlichen Planungsprozess entschieden werden. Im PLANETS Projekt wurde zu diesem Zweck ein Verfahren auf Basis der Nutzwertanalyse entwickelt. In einer groben Übersicht sind folgende Schritte durchzuführen:

1. *Anforderungen definieren*: Die Anforderungen an die Erhaltung von Datenbeständen müssen als messbare Eigenschaften definiert werden. Als Leitlinie sollten u.a. Anforderungen an die zu erhaltenden Objekteigenschaften (z.B. Inhalt, Erscheinung, Strukturierung), an die technische Umsetzung der Objekte (z.B. Verbreitung des Dateiformats, einfache Verarbeitbarkeit), an den Erhaltungsprozess (z.B. maximaler Zeitbedarf einer Konvertierung) und an die Infrastruktur (benötigte Hardware, Personal, Kosten etc.) definiert werden.
2. *Alternativen evaluieren*: Die unterschiedlichen Alternativen zur Umsetzung von Erhaltungsmaßnahmen müssen identifiziert und in einem Experiment gemessen werden.
3. *Ergebnisse analysieren*: Anhand der Experimentergebnisse kann gemessen werden, wie gut eine Erhaltungsmaßnahme den verschiedenen Anforderungen entspricht. Durch die Zuordnung von Relevanzfaktoren zu den verschiedenen Anforderungen kann zwischen den Alternativen abgewogen werden.
4. *Erhaltungsplan erstellen*: Schließlich wird aus der Analyse ein Umsetzungsplan generiert.

Typische Maßnahmen zur Sicherung der technischen Nachnutzbarkeit sind die Anpassung der Software-Umgebung (wie z.B. Portierung von Software, Unterstützung weiterer Formate, Emulation) oder die Anpassung der Daten (wie z.B. Formatmigration, Konvertierung) an neue Software-Umgebungen und Anforderungen. Formatmigrationen müssen dabei nicht unmittelbar vorgenommen werden, sondern können auch erst beim erneuten Zugriff auf die Daten erfolgen, wenn entsprechende Konvertierungssoftware vorhanden ist.

Auch die Erhaltung der intellektuellen Nachnutzbarkeit, d.h. dass ein technisch einwandfrei nutzbarer Datensatz auch inhaltlich verstanden werden kann, erfordert Maßnahmen. Diese bestehen oft darin, dass zum Verständnis notwendige Kontextinformationen vor einer Veröffentlichung oder Übernahme in ein Datenarchiv dokumentiert werden und bei Bedarf aktualisiert werden. Das dafür benötigte Hintergrundwissen verlangt meist eine aktive Kooperation mit der wissenschaftlichen Zielgruppe. Der Bedarf für eine Aktualisierung kann z.B. entstehen, wenn sich in einer Disziplin neue Terminologien oder Verfahren etablieren, um Nutzern das Verhältnis zu den älteren Terminologien und Verfahren nachvollziehbar zu machen. Weitere Maßnahmen zur inhaltlichen Nachnutzbarkeit können die Ergänzung oder Korrektur von Datensätzen sein, sofern sie nachvollziehbar versioniert werden und sofern sie nicht als Nachweis des korrekten wissenschaftlichen Arbeitens dienen und deshalb unverändert bleiben müssen.

Der beste Weg zu der Erhaltung von digitalen Daten ist, spätere Erhaltungsmaßnahmen durch eine gute Planung der Erzeugung von Daten und Qualitätskontrollen im Ingest unnötig zu machen (siehe Kapitel 1 und 3).

Weiterführende Literatur

- Christoph Becker et al., Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal for Digital Libraries* ([Becker et al., 2009]), (<http://www.ifs.tuwien.ac.at/~becker/pubs/becker-ijdl2009.pdf>).
- nestor AG Digitale Bestandserhaltung. nestor Materialien 15: Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung ([nestor AG Digitale Bestandserhaltung, 2011]), (<http://nbn-resolving.de/urn:nbn:de:0008-2011101804>).

Anwendungsfall Klimaforschung

Am World Data Center for Climate (WDCC) werden nicht die Daten migriert, sondern die Software-Umgebung an neue Hardware oder Anforderungen der Nutzer angepasst. Da am WDCC nur wenige Datenformate zugelassen sind, ist dies der bequemere Weg.

6 Zugriff und Nutzung

Neben der reinen Erhaltung der Daten gehört die Ermöglichung der Datennutzung zu den vornehmsten Aufgaben der meisten Archive. Ziel des Archivs ist dabei, autorisierten Nut-

zern den Zugang zu den Daten zu verschaffen und zugleich nicht-autorisierten Zugriff zu unterbinden, denn nur wenn letzterer verhindert werden kann, werden Datenerzeuger mit der Aufbewahrung ihrer Daten im Archiv einverstanden sein. Die meisten Archive sind mit den folgenden Aufgaben konfrontiert:

- Nachnutzung muss möglich sein, insbesondere müssen Daten gefunden werden können.
- Nur autorisierte Nutzer dürfen Zugang erhalten.
- Rechtliche Einschränkungen müssen eingehalten werden.
- Interoperabilität soll ermöglicht werden.

Nachnutzung und Suchbarkeit

Wenn Daten durch andere als die Produzenten nutzbar sein sollen und ein Zugriff über das Web realisiert werden soll, bietet sich ein Portal an. Ein Portal ist ein Webservice, unter dem weitere Dienste über eine grafische Oberfläche erreichbar sind. Ein bequemer Zugang zu den Daten kann im Portal mit Suchfunktionen, Orientierungshilfen und Nachbearbeitungsschritten zusammen angeboten werden. In der Praxis wichtige Nachbearbeitungsschritte sind die Visualisierung und die Konvertierung der Daten in ein Format, das von der Software des Nutzers gelesen werden kann. Die Nachbearbeitung kann bis hin zu komplexen Postprocessing-Workflows im Grid gehen. Eine Vielzahl solcher und ähnlicher Dienste kann in einem Portal bereitgestellt werden. Da das Portal gemeinsamer Einstiegspunkt im Web ist, muss die Prüfung der Identität des Nutzers (Authentifizierung) nur einmal pro Sitzung erfolgen, auch wenn mehrere Dienste in einer Sitzung genutzt werden. Als Webservice baut ein Portal auf dem HTTP-Protokoll auf.

Für den Datenzugriff außerhalb des Portals können auch andere Protokolle verwendet werden. Diese werden meist über eine Kommandozeile gesteuert. Häufig werden solche Lösungen zusätzlich zum Portal angeboten.

Für alle autorisierten Nutzer sollten die Daten suchbar sein. Ihre Existenz und der Ort ihrer Aufbewahrung können z.B. über Kataloge und Verzeichnisse herausgefunden werden. Effektiv und schnell ist die Suche in den Metadaten, sofern zweckmäßige, beschreibende Metadaten vorhanden sind. Wenn in den Daten selbst gesucht werden soll, müssen sich diese auf Medien befinden, auf die schnell zugegriffen werden kann, z.B. Festplatten.

Die gleichzeitige Suche in mehreren Archiven kann ermöglicht werden, wenn die für die Suche erforderlichen Metadaten eine einheitliche Struktur besitzen. Bei dem als *Harvesting* bezeichneten Verfahren werden in regelmäßigen Zeitabständen die für die Suche erforderlichen Metadaten bei den einzelnen Archiven eingesammelt und in eine zentrale Datenbank kopiert, in der dann gesucht werden kann [OAI-PMH].

Ist zu Beginn der Suche unklar, welche Archive überhaupt existieren, können *Registries* helfen. In einer Registry können Archive und Dienste gesucht werden, die dort zuvor registriert worden sind.

Offener Zugang versus Zugriffsbeschränkungen

Öffentlichkeit und Politik dürfen erwarten, dass die Mittel für die öffentliche Forschung so effizient wie möglich eingesetzt werden. Die Erzeugung von Forschungsdaten ist fast immer mit hohen Kosten verbunden. In der öffentlichen Forschung muss daher vermieden werden, dass die Produktion von gleichartigen Forschungsdaten mehrfach erfolgt und mehrfach bezahlt wird. Ein offener Zugang zu den Daten würde das gewährleisten, denn dann könnten außer den Datenproduzenten auch andere Forscher auf die Daten zugreifen und bräuchten die Daten nicht neu zu erzeugen.

Mit der Berliner Erklärung [[Berliner Erklärung](#)] sind alle wichtigen Organisationen der öffentlichen Forschung Deutschlands eine Selbstverpflichtung eingegangen, den freien Zugang zu Publikationen und allen ergänzenden Materialien zu unterstützen. Zu den Unterzeichnern [[Unterzeichner der Berliner Erklärung](#)] der Berliner Erklärung gehören der Deutsche Bibliotheksverband, die Deutsche Forschungsgemeinschaft, die Fraunhofer-Gesellschaft, die Helmholtz-Gemeinschaft, die Hochschulrektorenkonferenz, die Max-Planck-Gesellschaft, die Leibniz-Gemeinschaft und der Wissenschaftsrat. Dementsprechend wird die Vergabe von Mitteln immer häufiger an die Bedingung geknüpft, Forschungsdaten frei zugänglich zu machen.

So wünschenswert der freie Zugang auch ist: Daten, die einer Geheimhaltungspflicht unterliegen, dürfen nicht öffentlich zugänglich gemacht werden. Dies trifft für personenbezogene Daten zu, aber z.B. auch für Einzeldaten statistischer Ämter, die zu Forschungszwecken weitergegeben wurden (§ 16 Bundesstatistikgesetz).

Die Notwendigkeit, den Zugang zu beschränken, ergibt sich nicht unbedingt nur aus gesetzlichen Vorschriften. Diejenigen, bei denen Daten zu Forschungszwecken erhoben worden sind, verlangen häufig eine Beschränkung des Zugangs, beispielsweise wenn die Daten Betriebs- oder Geschäftsgeheimnisse enthalten. In solchen Fällen werden die Daten möglicherweise erst zur Verfügung gestellt, wenn die Zugriffsbeschränkung in einem Vertrag verankert ist.

Selbst für Daten, die für die Weitergabe vorgesehen sind, können Beschränkungen gefordert werden. Ein Beispiel hierfür sind kommerziell verwertbare Daten. Hier verlangt der Datenerzeuger häufig eine Gebühr oder eine Erklärung, dass die Daten nur für wissenschaftliche oder nicht-kommerzielle Zwecke verwendet werden.

Ein ganz anderer Grund Daten zurückzuhalten ist Unfertigkeit. Daten, an denen noch gearbeitet wird (Arbeitskopie), sollten nur Mitgliedern der Arbeitsgruppe zugänglich sein.

Sensible Daten können durchaus gleichzeitig durch Beschränkung des Zugangs geschützt und gemeinschaftlich für Forschungszwecke genutzt werden. Ist die Nutzung eingeschränkt, müssen Nutzer ihr Einverständnis erklären, bestimmte Nutzungsbedingungen zu beachten, bevor sie Zugang zu den Daten erhalten. Die Bedingungen können die Nutzung auf einen bestimmten Zweck einschränken, z.B. fachbezogene Forschung, oder bestimmte Formen der Nutzung ausschließen, z.B. kommerzielle Nutzung oder die Rückgängigmachung von Pseudonymisierung bzw. Anonymisierung der Daten.

Die Erklärung, die Nutzungsbedingungen zu beachten, kann je nach Erfordernissen elektronisch oder schriftlich bei der Registrierung als Nutzer abgefordert werden. Es können auch strengere Zugangsregeln für sensible Daten festgelegt sein, beispielsweise:

- Einschränkung des Zugangs auf bestimmte Gruppen oder Personen
- Forderung einer Erlaubnis des Eigentümers der Daten
- Gesicherter Zugang, über den die Daten nur analysiert, nicht aber kopiert werden können
- Öffnung des Zugangs erst nach Ablauf einer Sperrfrist, d.h. zu einem Termin, ab dem die Vertraulichkeit nicht länger fortbesteht. Über eine Sperrfrist kann Datenerzeugern auch eine angemessene Zeit für die Erstnutzung der Daten eingeräumt werden.

Nutzer müssen sich authentifizieren, bevor ihnen Zugang zu zugriffsbeschränkten Daten gewährt wird. In Computernetzwerken geschieht die Authentifizierung durch Versenden von Zeichenketten, die eindeutig oder nahezu eindeutig einem Nutzer zugeordnet werden können. Sehr häufig werden Benutzername/Passwort, Public-Key-Zertifikate oder OpenID verwendet, seltener Hardware wie Smartcards oder biometrische Daten wie Fingerabdrücke.

Public-Key-Zertifikate beruhen auf einem Schlüsselpaar bestehend aus einem öffentlichen und einem geheimen Schlüssel. Das eigentliche Zertifikat setzt sich aus dem Namen, einigen weiteren Daten und dem öffentlichen Schlüssel zusammen und wird zur Authentifizierung verschickt. Der Nutzer erhält sein persönliches Zertifikat von einer Certification Authority (CA), nachdem er seinen Personalausweis bei einer Registration Authority (RA) vorgelegt hat, die für die CA die Ausweiskontrolle vornimmt.

Eine OpenID ist eine dem Nutzer zugeordnete Webadresse, die von einem OpenID-Server vergeben wird. Nutzer melden sich einmal pro Sitzung bei einem OpenID-Server mit ihrem Benutzernamen und Passwort an und können sich danach mit ihrer OpenID bei allen das System unterstützenden Webservices authentifizieren.

Interoperabilität

Immer häufiger wird die Bereitstellung von Diensten erwartet, die Interoperabilität ermöglichen. Interoperabilität ist unverzichtbar für gemeinschaftlich betriebene Informationsarchitekturen. Wortbedeutung und praktische Verwirklichung dieses Konzeptes sind jedoch außerordentlich unterschiedlich [Gradmann, 2008]: Interoperabilität kann aus einer objektbezogenen oder einer funktionalen Perspektive gesehen werden, aus der Sicht der Institution oder der des Nutzers. Darüber hinaus wurde Interoperabilität auf verschiedenen Abstraktionsebenen konzipiert, siehe [Abb. 5](#).

Von der Funktion her betrachtet können interoperable Dienste einfach digitale Inhalte austauschen. Die Funktionalität kann aber auch weit darüber hinausgehen. Beispielsweise können digitale Objekte zu einer gemeinsamen Inhaltsschicht verknüpft werden. Auch kann über Interoperabilität eine gemeinsame Dienstarchitektur etabliert werden.

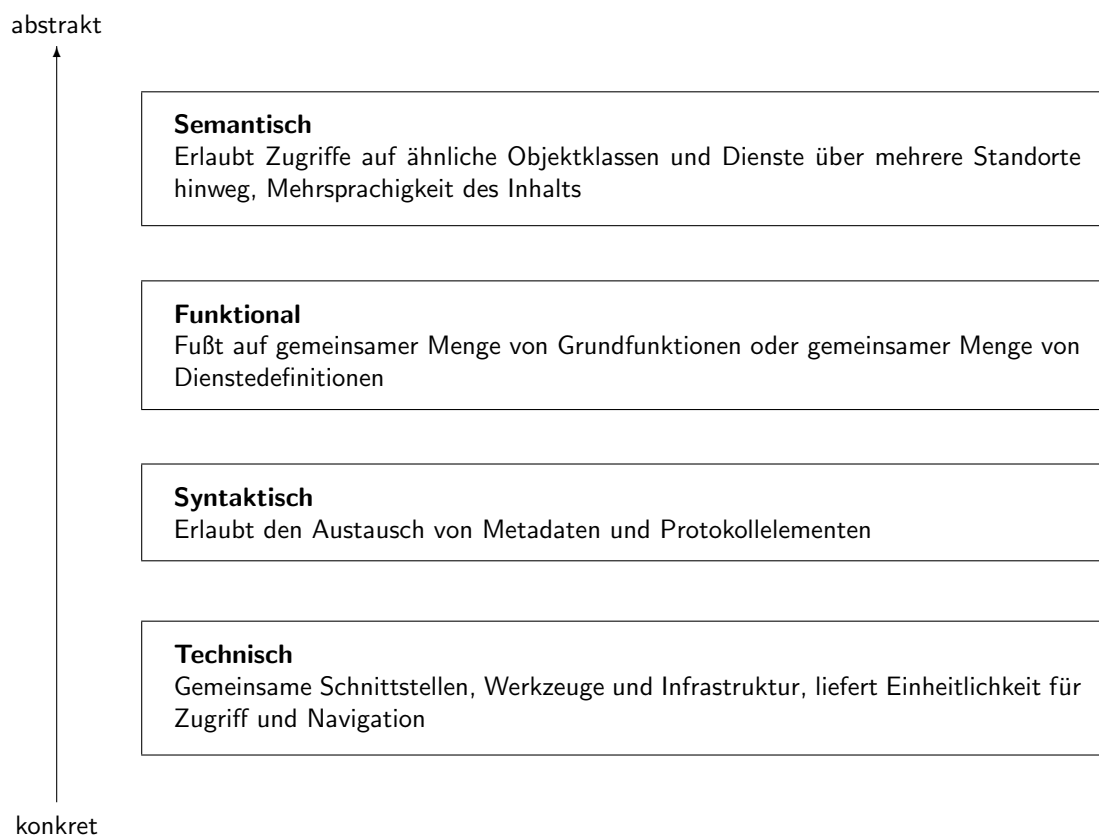


Abb. 5: Abstraktionsstufen der Interoperabilität nach [Gradmann, 2008]

Die zusammenwirkenden Elemente können sowohl traditionelle Institutionen (Archive, Bibliotheken, Museen) oder digitale Repositories, eScience- und eLearning-Plattformen oder einfach nur Webservices sein. Interoperabilität ist also keineswegs auf Institutionen beschränkt, sondern schließt auch Fragen wie die folgende mit ein: Wie können IT-gestützte Tools geeignete Informationen finden, um mit fachfremden oder ungewöhnlich strukturierten Daten arbeiten zu können?

Weiterführende Literatur

- Nestor Handbuch, Version 2.3 ([Neuroth et al., 2010]): Kapitel 9.3, Retrieval, (<http://nbn-resolving.de/urn:nbn:de:0008-2010071949>).
- Stefan Gradmann: Interoperability: A key concept for large scale, persistent digital libraries. DigitalPreservationEurope (DPE) Briefing Paper, 2008 ([Gradmann, 2008]), (<http://www.digitalpreservationeurope.eu/publications/briefs/interoperability.pdf>).

Anwendungsfall Klimaforschung

Der Zugang zum World Data Center for Climate (WDCC) [WDCC] kann online über grafische Benutzerschnittstellen auf zwei verschiedenen Wegen erfolgen.

Der eine Weg führt über das CERA-Portal [CERA]. CERA steht für „Climate and Environment data Retrieval and Archiving system“. Das CERA-Portal wird wie das WDCC vom Deutschen Klimarechenzentrum betrieben. Über das CERA-Portal sind alle Datensätze des WDCC erreichbar. Gesucht werden kann in den Metadaten, und zwar nach folgenden Strategien:

- Über eine Liste der Experimente⁵
- Über Begriffe, die aus einer Liste ausgewählt werden können
- Über den Namen der Modell-Software (Code-Suche)
- Volltextsuche
- Hierarchische Suche in einer Baumstruktur aus Oberbegriffen und Begriffen
- Suche in einer Tabelle mit Autorennamen, Titel und DOI

Nach erfolgreicher Suche kann auf die Daten per Download zugegriffen werden. Für die Nachbearbeitung stehen Zeit- und Datenformatkonvertierer sowie fachspezifische Berechnungswerkzeuge zur Verfügung.

Der zweite Weg zu Daten des WDCC führt über das C3Grid-Portal [C3Grid-Portal]. Über das in Abb. 6 auf S. 29 gezeigte Webformular kann in den Metadaten gesucht werden. Nach erfolgreicher Suche können die Daten per Grid-Job geholt werden. Dabei kommt im C3Grid die Eigenentwicklung GNDMS (Generation N Data Management System) [GNDMS] zum Einsatz, die auch Postprocessing-Workflows steuert. Solche Workflows zur rechnerischen Weiterverarbeitung der Daten können ebenfalls über das C3Grid-Portal gestartet werden. Hierfür sucht sich der Nutzer einen der vorgegebenen, fachspezifischen Workflows aus einer Liste aus und startet den Grid-Job über das Portal. Die Workflow-Software kann nicht über das Portal verändert werden. Zusätzlich benötigte Parameter können aber über ein Webformular mitgegeben werden.

Die Nutzungsbedingungen der Archive werden durchgesetzt, indem der Nutzer zusätzlich zum Benutzerkonto für das C3Grid-Portal künftig auch die Nutzungsberechtigungen für die gewünschten Archive haben muss. Am WDCC muss der Nutzer bei erstmaliger Anmeldung die Nutzungsbedingungen akzeptieren und bekommt bei Aufnahme in den Nutzerkreis ein persönliches Nutzerkonto. Nutzerkonten für Gruppen gibt es nicht. Der Zugang zu Daten, vor deren Nutzung der Nutzer eine Erklärung unterschreiben muss, wird gesondert freigeschaltet. Dies geschieht, indem der Nutzer in die entsprechende Gruppe der Nutzungsberechtigten für diese speziellen Daten eingetragen wird. Technisch werden die Zugriffsbeschränkungen durch Abfrage einer Oracle-Datenbank durchgesetzt. Dort werden die Tabellen mit den Berechtigengruppen gehalten.

⁵ „Experiment“ bedeutet hier Modellrechnung. Modellrechnungen sind sozusagen numerische Experimente.

Manche Dateneigentümer wollen informiert werden, welche Nutzer ihre Daten heruntergeladen haben. Je nach Wunsch des Dateneigentümers geht diese Information sofort nach dem Download an den Dateneigentümer oder in regelmäßigen Zeitabständen in Form einer zusammenfassenden Liste.



Welcome , Torsten Rathmann [Profile](#) [Home](#) [Logout](#)

C3 Status My C3Grid **Search & Download** Workflows Test Suite

Data Retrieval

Data Retrieval

Free Search Advanced Search Browse Catalogs My Stored Queries My Stored Downloads

reset start search

Free Search ?

Time Constraints ?

activate this box

Start Date after 1900-01-01 12:00

End Date before 2100-01-01 12:00

Vertical Constraints ?

activate this box

Min Vertical 1 level

Max Vertical 1 level

Vertical Type Column Density

Geographical Constraints ?

activate this box

Min Lat -90.00 World

Max Lat 90.00

Min Lon 0.00

Max Lon 360.00

within

Content Constraints ?

- 2m temperature
- FAO data set (soil data flags)#nonCF
- air_pressure_at_convective_cloud_top (maximum)
- air_pressure_at_sea_level
- air_temperature
- air_temperature (mean per month of four values a day)
- air_temperature (mean per month)
- air_temperature-at2m
- air_temperature-at2m (maximum per ...)
- air_temperature-at2m (mean per month of maximum)
- air_temperature-at2m (mean per month of minimum)
- air_temperature-at2m (mean per month)

Keywords ?

- "climate simulation"
- "scenario run"
- Arctic River Discharge
- CMIP
- Climate Modeling
- ECHAM4
- ECHAM5
- ECHO-G
- Eurasia
- HOPE
- Hydrological Cycle
- IPCC AR4

Abb. 6: Formular zur Datensuche im C3Grid-Portal

Inhaltlich ist das Datenangebot der Portale verschieden. Über das CERA-Portal sind alle WDCC-Daten und nur diese zugänglich. Über das C3Grid-Portal ist eine Teilmenge der Daten aller am C3Grid beteiligten Institutionen zugänglich. Eine Verpflichtung, den Zugriff auf Daten über das C3Grid zu ermöglichen, besteht nicht. Welche Daten über das C3Grid zur Verfügung gestellt werden, entscheidet jede Institution für sich.

Teil II

Übergreifende Aufgaben des Forschungsdatenmanagements

Die in jeder Station des Lebenszyklus der Daten vorkommenden Aufgaben (siehe [Abb. 2, S. 6](#)) sind im Wesentlichen keine IT-bezogenen Fragen, sondern beziehen sich auf die Herstellung von geeigneten Rahmenbedingungen, unter denen das Datenmanagement bzw. die Langzeitarchivierung gesichert werden kann.

7 Organisation, Management und Policies

Die Forschungsdaten werden von einzelnen Forschern, Forschungsgruppen innerhalb einer Einrichtung, überinstitutionellen Gruppen sowie Kollaborationen auf nationaler oder internationaler Ebene erzeugt.

Organisation

Es muss geklärt werden, an welcher Stelle die Daten abgelegt werden können und sollen, wer (welche Personen, Abteilung, Organisation) für die Aufbewahrung zuständig ist und nach welchen Kriterien die Weitergabe der Daten erfolgen soll.

Ein Datenarchiv ist die Organisationseinheit, die die Aufgabe des Datenmanagements in einem festgelegten Bezugsrahmen verantwortlich übernimmt.

Es gibt verschiedene Formen eines Repository (Datenarchiv), die sich in ihrem Bezugsrahmen unterscheiden und damit auch in der Festlegung ihrer Managementstrukturen und Policies unterschiedlich sind.

Institutionelles Repository: Als Beispiel kann eine Universität dienen, die neben den Diplomarbeiten/Dissertationen/Habilitationen auch die Daten zu diesen Arbeiten speichert. Die Konzeption eines solchen Repository bedarf einer Policy der Universität, in welcher die Form und die Verpflichtung der Abgabe dieser Daten sowie deren Nutzungsbedingungen geregelt sind. Eine solche Policy regelt konzeptuell das Datenmanagement eines Teils der von der Institution erzeugten Daten. Der Datenmanagement-Plan muss darüber hinaus die Implementierung des Repository und die Sicherstellung des Betriebs beinhalten. Die Universität hat hierfür organisatorische Voraussetzungen zu schaffen: Bereitstellung von IT-Infrastruktur, Bereitstellung von Services für den Ingest der Daten und deren Publikation. (Diese Aufgabe ist nicht spezifisch für eine einzelne Universität; die Entwicklung von standardisierten Metadaten und Verfahren hierfür wäre sinnvoll.)

Kollaborations-Repository: Hierfür lässt sich als Beispiel eine (internationale) Kollaboration zur Nutzung eines Instruments nennen. Solche Kollaborationen entstehen oft aus der Notwendigkeit, knappe Ressourcen optimal auszunutzen. Die Nutzung der Daten

wird von den Beteiligten in erster Linie als Eigennutzung konzipiert, und unter diesem Gesichtspunkt werden die Vorgaben für die Datenspeicherung und Nutzung vor allem von den individuellen Interessen der Beteiligten gelenkt.

Die Fördergeber verbinden mit der Vergabe der Mittel jedoch mittlerweile eine – zeitgemäße – Auflage zur Veröffentlichung der Daten. Zudem sind die Größenordnungen dieser Daten auch für eine Auswertung/Speicherung durch die einzelnen Beteiligten meist nicht mehr geeignet. Daher werden in größeren Vorhaben explizit Mittel für ein gemeinsames Datenmanagement bereitgestellt, und eine oder mehrere der beteiligten Institutionen stellen die technische Infrastruktur sowie Know-How und Dienste für das Management der Daten des Instruments bereit.

Für die Policy eines solchen Repository sind in erster Linie die Kollaborationsvereinbarungen bindend. Die langfristige Sicherstellung der Verfügbarkeit der Daten ist in diesen Fällen oft von den Institutionen abhängig, die das Repository betreiben. Charakteristisch sind zeitlich befristete Zugangsbeschränkungen und partielle Publikation von Daten; eine Strukturierung der Daten unter Einbeziehung der Gesichtspunkte der späteren Publikation ist daher sinnvoll.

Thematisches Repository: In diese Kategorie fallen z.B. Einrichtungen wie das Institut für Deutsche Sprache, das Deutsche Institut für Wirtschaft oder auch eine Gen-Datenbank. Während der Bezugsrahmen durch das Thema vorgegeben ist, werden hier vor allem Datenschutz, Vertraulichkeit und auch wirtschaftliche Interessenkonstellationen einen großen Einfluss haben. Auch hier ist eine Policy erforderlich, welche die thematische Abgrenzung der Daten, die Aufnahme der Daten und deren Publikation bzw. Nutzung regelt.

Die Policies dieser thematischen Repositories sind stark beeinflusst von den Vorgaben der Fördermittelgeber oder der „Datenprovider“ (beispielsweise Statistische Ämter, Sozial- oder Arbeitsämter oder Wirtschaftsverbände). Die Bereitstellung der Daten und deren Pflege ist Daseinszweck dieser Art von Repositories, und die Festlegung von Policies für den Ingest, den Zugriff und die Leistungen des Archivs ist ein wesentliches, konstituierendes Element.

Management

Die Strukturierung der Daten nach bestimmten (meist fachlichen) Kriterien sowie die Auswahl von zu verwendenden Metadaten zur Beschreibung von Inhalt und Kontext sind Komponenten des Datenmanagements. Zwar hat jede Domäne (siehe [Abb. 3, S. 8](#)) eigene Notwendigkeiten, jedoch ist es sinnvoll klare Regelungen zu haben, welche die Aufbewahrung und Weitergabe der Daten betreffen. Die Handhabung der Daten kann durch die Regelungen der Fachgemeinschaft, der Institution, der Kollaboration oder auch des Fördergebers beeinflusst oder vorgegeben sein.

Der Datenmanagementplan regelt die Implementierung und die Sicherstellung des Betriebs des Datenarchivs. Zur Implementierung des Datenmanagements ist eine Analyse des Workflows von der Erzeugung der Daten bis zu deren Nutzung notwendig. Eine Abgrenzung von Daten, die im Repository vorzuhalten sind, und intermediären Daten muss anhand dieser

Analyse vorgenommen werden. Zudem sind Regelungen über die zeitliche Dauer der Datenvorhaltung notwendig. Ein weiterer Bestandteil des Datenmanagementplans ist die Qualitätskontrolle, welche den Gehalt der Daten und deren technische Integrität bewertet. Beim Übergang von der privaten zur Gruppendomain ist eine Selektion und Qualitätskontrolle relativ leicht zu realisieren, z.B. in einer Kollaboration, in der die Forscher „live“ mit den Daten arbeiten. Beim Übergang von der Gruppen- in die Public Domain sind adäquate Formen der Qualitätskontrolle und Selektion bisher nur in Ansätzen vorhanden. Eine dieser Formen ist das Data-Release für Kollaborations-Repositories, welches eine wissenschaftliche Bewertung der publizierten Daten beinhaltet. Für thematische oder institutionelle Repositories existieren solche Prozesse derzeit nur bedingt.

Werden Forschungsdaten durch ein internationales Konsortium erzeugt, so wird die Verwendung von Standards für diese Daten geradezu eine unumgängliche Notwendigkeit. Insbesondere um diese Daten für IT-basiertes Data-Mining zugänglich zu machen, werden die z.B. vom Virtual Observatory in der Astronomie oder dem Open Geospatial Consortium vorgeschlagenen Standards und Tools ein wichtiger Teil der vom Datenmanagement zu berücksichtigenden Komponenten.

Ein wesentliches Charakteristikum des modernen Datenarchivs ist, dass Aufbau und Betrieb kaum mehr als Nebenaufgaben eines oder mehrerer beteiligter Wissenschaftler zu realisieren sind. Dementsprechend gehört auch die Bereitstellung von ausreichenden Personalreserven mit den erforderlichen Qualifikationen zu den Komponenten eines Datenmanagementplans.

Policies

Policies regeln in Bezug auf das Datenarchiv die grundlegenden Verfahren zur Aufnahme, zur Bereitstellung und zum Zugriff auf die Daten. Solche Policies werden innerhalb eines organisatorischen Rahmens festgelegt und gelten nur innerhalb desselben. Bei überinstitutionellen Zusammenschlüssen (Kollaborationen, Arbeitsgruppen) ist es erforderlich, dass der Betrieb des Repository von bestehenden Organisationen/Institutionen verpflichtend übernommen wird, und dass Regelungen für die Zeit nach Beendigung der Zusammenarbeit getroffen werden. Darüber hinaus wird durch Policies festgelegt, welche Nutzung der Daten vom Archiv als zulässig gewertet wird, ob und in welcher Form Lizenzen, Gebühren etc. erhoben werden, welche Nutzergruppen autorisiert sind usw.

Anwendungsfall Klimaforschung

Das C3Grid (Collaborative Climate Community Data and Processing Grid) [[C3Grid Website](#)] ist ein echtes Datengrid, in dem die Daten über verschiedene Standorte verteilt liegen. Drei ICSU-Weltdatenzentren (WDCC [[WDCC](#)], WDC-MARE [[WDC-MARE](#)], WDC-RSAT [[WDC-RSAT](#)]), der Deutsche Wetterdienst (DWD) [[Deutscher Wetterdienst](#)] und eine Reihe weiterer Archive kooperieren im C3Grid, um eine einheitliche Plattform für den Zugriff auf Klimadaten anzubieten. Die Entscheidung, welche Daten über das C3Grid-Portal [[C3Grid-Portal](#)] zugänglich gemacht werden, liegt bei den einzelnen Datenzentren. Generell nicht über das Grid zugänglich sind Daten, vor deren Nutzung eine schriftliche Erklärung unterzeichnet werden muss.

Wie wird die Einhaltung der Planung überprüft oder nachgewiesen? Im C3Grid erfolgen Überprüfung und Nachweis der Einhaltung der Planung durch Projektberichte. Am WDCC gibt es drei Nachweis- bzw. Prüfverfahren:

- regelmäßige Reviews (Metadaten und Datenkonsistenz)
- halbjährliche Überprüfung auf nationaler Ebene durch den Wissenschaftlichen Lenkungsausschuss (Verfügbarkeit der WDCC-Dienste für die Community)
- alle 5 Jahre internationale Begutachtung im Rahmen der Rechnerbeschaffung (Ausstattung)

8 Kosten

Mit den fallenden Preisen für Speichermedien scheint oftmals der Eindruck entstanden zu sein, dass das Aufbewahren von Daten kaum Kosten und Aufwand verursachen kann. Aber eine der größten Gefahren für den langfristigen Erhalt von Forschungsdaten sind fehlende finanzielle und personelle Ressourcen. Forschungsinstitutionen leiden fast immer unter Mittelknappheit, und es ist notwendig, die teuren und aufwändigen Aufgaben zu erkennen. Bei Entscheidungen und Abschätzungen der benötigten finanziellen und personellen Ressourcen sollten einige grundlegende Faktoren bedacht werden:

Teil der Nutzungskosten Das Aufbewahren und die Pflege von Forschungsdatenbeständen wird nicht um der Tätigkeit selbst willen durchgeführt, sondern weil es möglich sein soll, die Forschungsdaten zu nutzen. Entsprechend wichtig ist es, den Nutzen eines Forschungsdatenbestandes herauszuarbeiten und die Kosten des Datenmanagements nicht als zusätzliche und optionale Kosten, sondern als genauso notwendig für die Nutzung aufzufassen wie die Produktionskosten selbst.⁶ Die Kosten des Datenmanagements sind somit ein notwendiger Teil der gesamten Nutzungskosten.

Höchste Kosten am Anfang Digitale Forschungsdaten haben oftmals einen langfristigen Wert, aber genauso langfristig müssen auch Aufwände und Investitionen erfolgen, um diesen Wert zu erhalten. Bei einer groben Einteilung der Datenerhaltung in die Übernahme-, Speicher- und Zugriffsphase lassen sich ungefähr die Hälfte der Kosten der Übernahme in das Archiv zuordnen. Die zweitaufwändigste Phase ist der Zugriff, die Speicherphase ist am günstigsten.⁷

⁶ Um den Nutzen bei Bedarf genauer auszuarbeiten, bieten sich die im Rahmen der *Keeping Research Data Safe*-Projekte entwickelten Instrumente an, siehe [KRDS-I2S2-Tools]. Einige grundlegende ökonomische Eigenschaften der Bewahrung digitaler Daten wurden im Abschlussbericht der Blue Ribbon Task Force erläutert, siehe [Blue Ribbon Task Force, 2010].

⁷ Für den Archaeology Data Service wird eine als typisch angesehene Kostenstaffelung von circa 55 Prozent in Outreach/Acquisition/Ingest, circa 31 Prozent in Access und ungefähr 15 Prozent bei Archival Storage und Preservation angegeben, vgl. [Beagrie et al., 2010, S. 79].

Warten und Nichtstun ist teuer Diese am Anfang entstehenden hohen Kosten können kaum aufgeschoben werden, weil das spätere Nacharbeiten noch mehr Kosten verursacht. Ein Beispiel sind die Aufwände für Qualitätskontrolle und Metadaten, die zudem eine höhere Effizienz bei den restlichen Archivabläufen ermöglichen.⁸

Personalkosten Den größten Anteil an den Kosten für ein verlässliches Datenmanagement stellt ausreichend qualifiziertes Personal dar und nicht Hard- oder Software. KRDS gibt eine Größenordnung von 70 Prozent und mehr für Personalkosten an [Beagrie et al., 2011, S. 14].

Sinkende jährliche Kosten Aufgrund der hohen einmaligen Anfangskosten und der zunehmenden Effizienz der Technologien sinken die jährlichen Kosten für die Aufbewahrung eines Datenbestand. Dies hat z.B. zur Entwicklung eines simplen Geschäftsmodells „Pay Once, Store Forever“⁹ an der Princeton University geführt, das aber auch nur einen sehr begrenzten Service vorsieht.

Anreize Nicht nur unklare Vorstellungen vom langfristigen Nutzen und mangelnde Mittel können notwendige Maßnahmen verhindern, sondern auch fehlende Anreize. Anders als viele analoge Güter werden Informationsgüter nicht durch die Nutzung aufgebraucht. Es reicht daher im Prinzip eine Partei aus, die den Aufwand des Forschungsdaten-Managements treibt, damit beliebig viele andere ohne Aufwand die Daten nutzen können.¹⁰ Gesamtwirtschaftlich betrachtet kann das sehr sinnvoll sein, aber einzelne Wissenschaftler oder Institutionen könnten demotiviert werden, diesen Aufwand zu erbringen, wenn Dritte den Nutzen ohne Gegenleistung erhalten. Aus diesem Grund sind Erstverwertungsrechte oder das Zitieren von Forschungsdaten wichtige Faktoren im Forschungsdaten-Management, um die notwendige Anerkennung zu gewährleisten.

Zur eigentlichen Bestimmung der Kosten der Bewahrung digitaler Daten haben zwei mehrphasige Projekte Modelle und Fallstudien entwickelt: Keeping Research Data Safe und das LIFE-Projekt. In beiden Fällen orientieren sich die Berechnungsgrundlagen am DCC Curation Lifecycle Model (siehe [Abb. 1 Aufgaben im Lebenszyklus von Forschungsdaten](#), S. 5). Eine Kostenplanung kann eine Abschätzung der Kosten an den einzelnen Phasen orientiert durchführen. Im KRDS-Modell gliedern sich die Kostenkategorien, in welche die einzelnen Phasen einsortiert werden, grob wie folgt [Beagrie et al., 2010, S. 14 – 26]:

- Vorarchiv-Phase: Hier fallen neben den Kosten für die Erschaffung der Daten auch Kosten an für Beratung, Schulung und die Planung des Datenmanagements selbst.
- Archiv-Phase: Die wesentlichen Kostenkategorien in diesem Abschnitt sind die Kosten für alle einzelnen Lebenszyklusphasen, die in dem Modell dieses Leitfadens von Auswahl und Bewertung bis Zugriff und Nutzung reichen. Zusätzlich werden insbesondere Innovationskosten für die Entwicklung von neuen Werkzeugen, Standards etc. darunter verbucht.

⁸ In Studien wurde abgeschätzt, dass es durchaus eine Größenordnung teurer sein kann, nachträglich Metadaten zu erzeugen, vgl. [Archief, 2005, S. 15].

⁹ siehe [Goldstein and Ratliff, 2010, S. 1]

¹⁰ „Digital assets are nonrival in consumption and create a free-rider potential.“ [Blue Ribbon Task Force, 2010, S. 24]

- Unterstützungsdienste: Kosten für die Verwaltung aller Aktivitäten und die allgemeine IT-Basisinfrastruktur.
- Gebäude: Aufwände im Zusammenhang mit benötigten Räumen und Gebäuden.

Weiterführende Literatur

- Charles Beagrie Ltd, JISC (Hrsg.): Keeping Research Data Safe Factsheet, 2010 ([Charles Beagrie Ltd and JISC, 2010]), (http://www.beagrie.com/KRDS_Factsheet_0910.pdf).
- Neil Beagrie et al.: User Guide for Keeping Research Data Safe, 2010 ([Beagrie et al., 2011]), (http://www.beagrie.com/KeepingResearchDataSafe_UserGuide_v2.pdf).
- Ayris, P. et al.: The LIFE2 final project report. LIFE Project, London, UK. 2008 ([Ayris et al., 2008]), (<http://eprints.ucl.ac.uk/11758/1/11758.pdf>).
- Blue Ribbon Task Force, Sustainable Economics for a Digital Planet, 2010 ([Blue Ribbon Task Force, 2010]), (http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf).

Anwendungsfall Klimaforschung

C3Grid [C3Grid Website] hat von September 2005 bis August 2008 eine Projektförderung vom Bundesministerium für Bildung und Forschung (BMBF), Referat „Internet“, bekommen. Das Nachfolgeprojekt C3-INAD¹¹ wird jetzt vom Fachreferat „Globaler Wandel“ des BMBF gefördert. Der Förderungszeitraum reicht von Oktober 2010 bis September 2013.

Das World Data Center for Climate (WDCC) [WDCC] wird am Deutschen Klimarechenzentrum (DKRZ) [Deutsches Klimarechenzentrum] betrieben, dessen Betrieb wiederum von den vier Gesellschaftern (Helmholtz-Zentrum Geesthacht, Max-Planck-Gesellschaft, Freie und Hansestadt Hamburg, Alfred-Wegener-Institut für Polar- und Meeresforschung Bremerhaven) finanziert wird.

Die Archivierung von Daten ist am WDCC für externe Auftraggeber kostenpflichtig. Die Kosten für die Archivierung müssen in der Regel von demjenigen getragen werden, der den Auftrag für den Ingest gegeben hat. Das Abrufen von Daten über das Web ist dagegen kostenfrei, sofern die Daten nur für wissenschaftliche Zwecke verwendet werden. Im Falle von Sonderwünschen, z.B. Bereitstellung von Daten auf CD, werden die Kosten für das Kopieren und Verschicken in Rechnung gestellt.

Bisher wurden hauptsächlich Klimadaten von Institutionen archiviert, die zugleich DKRZ-Nutzer sind. In diesem Fall werden die Kosten mit den bewilligten Kontingenten verrechnet und die Speicherung ist de facto kostenfrei. In letzter Zeit sind die Archivierungsdienste des WDCC anscheinend auch für externe Kunden interessant geworden, z.B. für Forschungsinstitute, die sich den Aufbau eines eigenen Langzeitarchivs ersparen wollen. Von daher zeichnet

¹¹ INAD steht für „Towards an Infrastructure for General Access to Climate Data“.

sich eine zusätzliche, wenn auch eng begrenzte Einnahmequelle für das Archiv ab. Kunden müssen lediglich einen Selbstkostenbeitrag leisten. Gewinne darf das DKRZ als gemeinnützige GmbH nicht erzielen. Ein höherer als der Selbstkostenpreis widerspricht auch den Regeln [WDC-Principles] für ICSU-Weltdatenzentren [International Council for Science]. Solange es keine umfassenden Erfahrungen mit den tatsächlich anfallenden Kosten gibt, stellt das DKRZ den Einstellern Schätzkosten nach Tab. 1 in Rechnung. Auch von den zahlenden Kunden werden nur Daten angenommen, die den Regeln für eine Archivierung im WDCC entsprechen¹².

	1 Experiment, 1 TB	10 Exp. ähnl. Struktur, 15 TB
Summe Daten-/MetadatenSpeicherung	0,30 PM	0,76 PM
LZA (10 Jahre) inkl. Pflege	0,35 PM	1,35 PM
Medien und Betriebskosten für 10 Jahre inkl. 2 Kassetten-Upgrades	400 Euro	6000 Euro
Kosten für DOI-Registrierung (optional)	0,15 PM	0,55 PM

PM = Personenmonate, Experiment steht für Modellrechnung (numerisches Experiment)

Tab. 1: Kosten für die Datenspeicherung am DKRZ

9 Rechtliche Aspekte von Forschungsdaten

Im wesentlichen lassen sich für Forschungsdaten zwei Bereiche unterscheiden, in denen rechtliche Fragen Bedeutung erlangen. Zum einen ist dies das Feld des Datenschutzes, typischerweise der Schutz personenbezogener Daten vor unbefugtem Zugriff, und zum anderen Fragen des Urheberrechtes. Bei letzterem sind wiederum zwei verschiedene Aspekte zu betrachten, nämlich einerseits die Be- oder Nachnutzung von urheberrechtlich geschützten Werken (z.B. Datenbanken oder Programmen), zum anderen die Erstellung neuer urheberrechtlich geschützter Objekte bzw. die Kontrolle über die bei der Erstellung sich ergebenden Rechte.

Datenschutz - personenbezogene Daten

Forschungsdaten unterliegen einer Reihe von deutschen und internationalen Gesetzen. Das Bundesdatenschutzgesetz (BDSG) stellt die primäre Referenz der gesetzlichen Bestimmungen über die Erhebung, Verarbeitung und Nutzung personenbezogener Daten dar und betrifft sowohl Behörden als auch nicht-staatliche Institutionen, neben dem Bundesrecht weist jedes der 16 Bundesländer eine eigene Gesetzgebung auf. Personenbezogene Daten dürfen nur erhoben werden, wenn dies gesetzlich zulässig ist (z.B. mittels gerichtlicher Anordnung) oder die betroffene Person eingewilligt hat; in der Regel muss diese Einwilligung schriftlich gegeben werden.

Zweck des BDSGs ist es, Einzelne vor Missbrauch ihrer personenbezogenen Daten zu schützen. Solche Daten umfassen Namen, Geburtsdatum, Patientendaten und andere vertrauliche Informationen. Personen, die mit der Erhebung oder Verarbeitung personenbezogener

¹² Regeln siehe Abschnitte 2 und 3

Daten beschäftigt sind, ist es untersagt, diese Daten ohne Genehmigung zu beschaffen, zu nutzen oder weiter zu verarbeiten, und sie sind verpflichtet, das Datengeheimnis auch nach Ende ihrer Tätigkeit zu wahren.

Das BDSG enthält besondere Bestimmungen zur Verwendung personenbezogener Daten zum Zwecke der Forschung. Es besagt, dass personenbezogene Daten, die zu wissenschaftlichen Zwecken erhoben oder gespeichert wurden, nur im Rahmen dieser Zwecke verarbeitet oder genutzt werden dürfen. Die so erhobenen Daten müssen anonymisiert werden, sobald es der Forschungsprozess erlaubt. Bis zu diesem Zeitpunkt müssen Daten, die einer bestimmten Person zugeordnet sind oder werden können, gesondert aufbewahrt werden und dürfen nur soweit mit anderen Daten kombiniert werden, wie es der Forschungszweck erfordert. Wissenschaftliche Institutionen dürfen personenbezogene Daten nur veröffentlichen, wenn das Einverständnis der entsprechenden Person vorliegt oder die Veröffentlichung der Daten für die Präsentation von Forschungsergebnissen zu aktuellen Ereignissen unverzichtbar ist.

Urheberrecht

Prinzipiell unterliegen Forschungsdaten in Deutschland dem Urheberrecht. Der Urheberschutz greift allerdings erst dann, wenn ein Werk die Erfordernisse des persönlichen Schaffens, der wahrnehmbaren Formgestaltung, des geistigen Gehalts und der eigenpersönlichen Prägung erfüllt.¹³ Auf Forschungsdaten trifft dies jedoch nicht immer zu (z.B. bei unstrukturierten Messdaten); sie sind daher oft nicht urheberrechtlich geschützt. Anders verhält es sich, wenn für die Erstellung, Darstellung oder Auswertung der Forschungsdaten eigene Programme entwickelt oder die Daten in einer Datenbank gesammelt werden. Dies kann durchaus einen eigenen Urheberschutz begründen und sollte daher von Projektseite bedacht und gestaltet werden, insbesondere wenn diese Werke in die Speicherung der Forschungsdaten einbezogen werden.

Wenn das Werk vom deutschen Urheberrecht geschützt ist, sind nur bestimmte Nutzungsarten ohne Zustimmung des Urhebers zulässig. Dazu gehören:

- Vervielfältigung von bloßen Fakten im Rahmen eigener Interpretation oder Wortwahl,
- Vervielfältigungen zum privaten Gebrauch,
- Vervielfältigung, Verbreitung und öffentliche Wiedergabe im Rahmen eines Zitats,

Urheberrecht bei fremden Daten

Grundsätzlich müssen bei wissenschaftlicher Forschung die Rechte an geistigem Eigentum berücksichtigt werden. Ist ein Werk urheberrechtlich geschützt, ist die Einwilligung der Urheber zu dessen Vervielfältigung oder Weiterverbreitung unabdingbar. In diesem Zusammenhang ist zu berücksichtigen, dass Datenbanken¹⁴ nach deutschem Recht einem spezifischen Schutz

¹³ siehe Peter Lutz: Grundriss des Urheberrechts. C.F. Müller, Heidelberg 2009, Rn. 37–86d, zitiert nach Wikipedia, http://de.wikipedia.org/wiki/Deutsches_Urheberrecht

¹⁴ Gemeint sind hier die in Datenbanken gespeicherten Daten, also nicht die Architektur oder Verwaltungssoftware einer Datenbank.

unterliegen, der den Erstellern der Datenbank das alleinige Recht zu ihrer Verbreitung und Vervielfältigung gewährt. Lediglich die Verwendung eines *unwesentlichen* Teils der Datenbank ist ohne Zustimmung des Datenbankherstellers erlaubt; für die Vervielfältigung eines *wesentlichen* Teils einer Datenbank ist eine Einwilligung außerdem dann nicht erforderlich, wenn diese Vervielfältigung zum privaten Gebrauch oder zur Veranschaulichung im Unterricht erfolgt.

Darüber hinaus sind Kopien – sowohl von Datenbanken als auch allgemein von urheberrechtlich geschützten Werken – zum persönlichen wissenschaftlichen Gebrauch¹⁵ zulässig, wenn die Vervielfältigung zu diesem Zweck geboten¹⁶ ist und keinen gewerblichen Zwecken dient. Die Speicherung von urheberrechtlich geschützten Daten durch Forschungseinrichtungen fällt hingegen normalerweise nicht unter diese Regel, da dieser Vorgang üblicherweise auf eine Verfügbarmachung für mehr als eine Person und den Austausch mit anderen Forschern abzielt.

Insofern ist bei der Verwaltung von Forschungsdaten zu bedenken, welche Fremddaten und -programme benutzt wurden und welche Einschränkungen mit deren Verwendung verbunden sind. Insbesondere stellt sich die Frage, ob diese Daten und Programme mit in die Archivierung einbezogen werden dürfen. In Zweifelsfällen sollte eine Klärung mit den Rechteinhabern angestrebt werden, die gegebenenfalls in die Form eines rechtsverbindlichen Vertrages münden sollte.

Urheberrecht bei eigenen Daten

Nicht nur für den rechtlichen Schutz der verwendeten fremden Daten sollten Überlegungen angestellt werden, sondern auch bzgl. der Rechte an den im Rahmen des Projektes erstellten eigenen Daten, und wie die Einhaltung dieser Rechte kontrolliert werden kann. Hierbei sollten auch die in [Kapitel 8](#) erwähnten Anreize und deren rechtliche Umsetzung beachtet werden.

Zur Festlegung von Nutzungsrechten existiert mittlerweile ein breites Spektrum von möglichen Lizenzierungsmodellen, deren Erläuterung den Rahmen dieses Leitfadens sprengen würde. Wenn keine oder nur spezifische Restriktionen gewünscht werden, so können diese wohl am einfachsten mit einer geeigneten offenen Lizenz (z.B. GPL oder Creative Commons) versehen werden. Zu beachten sind jedoch Fälle, bei denen in die Erstellung von eigenen Programmen fremde Software eingegangen ist: Hier können die dafür bestehenden Lizenzen die Wahl eigener Lizenzierungsmöglichkeiten einschränken (z.B. durch das „Copyleft-Prinzip“).

Weiterhin zu beachten sind Fälle mit bestimmten Datenarten, in denen das Patentrecht greift. Wenn verwendete oder erzeugte Daten von wissenschaftlichen, technischen, oder methodischen Patenten geschützt sind, so sollte ein erfahrener Patentanwalt konsultiert werden, um einen Lizenzvertrags auszuarbeiten oder um Probleme, welche die Daten während deren gesamter Lebensdauer beeinträchtigen können, zu überprüfen.

¹⁵ Der „persönliche wissenschaftliche Gebrauch“ umfasst das Kopieren innerhalb einer unzugänglichen Umgebung und den Ausschluss einer Weitergabe an Dritte.

¹⁶ Als „geboten“ kann die Vervielfältigung dann bezeichnet werden, wenn sie die wissenschaftliche Forschung erfordert und der Kauf einer Kopie nicht zumutbar oder problematisch ist.

Schranken des Urheberrechts

Geschützte (Text-)Daten unterliegen einer Frist, nach der ihre Autoren und andere Rechteinhaber ihre Exklusivrechte verlieren und die Daten gemeinfrei werden (z.B. 70 Jahre nach deren Tod). Es ist möglich, dass diese Zeitspanne während des Aufbewahrungszeitraumes endet oder sich ändert.

Nach Ablauf dieser Frist unterliegen Forschungsdaten nicht mehr dem urheberrechtlichen Schutz und eine weitergehende (oder uneingeschränkte) Nutzung der Daten ist möglich. Zu beachten ist dabei, dass die gesetzlichen Fristen im Urheberrecht in und außerhalb der EU im Fluss sind¹⁷.

Weitere Aspekte

Um rechtliche Unklarheiten beim Management Ihrer Forschungsdaten zu vermeiden, sollten Sie Ihre Vertrags- und Lizenzentwürfe von einem Anwalt absichern lassen. Insbesondere die z.T. sehr unterschiedlichen Datenschutz- und Urheberrechtsbestimmungen in internationalen Kontexten und die sich daraus ergebende Komplexität der Rechtsansprüche sollte nicht unterschätzt werden. Aufgrund der noch relativ neuen und sich stetig ändernden Rechtslage kann auch ein System zur regelmäßigen Rechtsberatung in Betracht gezogen werden um festzustellen, wie sich zukünftige Gesetzesänderungen auf Ihre Datenhaltung auswirken.

Weiterführende Literatur

- Madeleine de Cock Buning, Barbara van Dinther, Christina G. Jeppesen de Boer, Allard Ringnalda: Report on the Legal Status of Research Data in the Knowledge Exchange partner countries. Centre for Intellectual Property Law (CIER), The Netherlands, 2011 ([de Cock Buning et al., 2011]), (<http://www.knowledge-exchange.info/Default.aspx?ID=461>).
- Gerald Spindler und Tobias Hillegeist: KoLaWiss Project: Arbeitspaket 4 - Recht. Göttingen, 2009, ([Spindler and Hillegeist, 2009]), (http://kolawiss.uni-goettingen.de/projektergebnisse/AP4_Report.pdf).

Anwendungsfall Klimaforschung

Klimadaten unterliegen keiner gesetzlichen Archivierungspflicht. Das World Data Center for Climate (WDCC) [WDCC] bekennt sich aber zu den Regeln zur Sicherung guter wissenschaftlicher Praxis [Max-Planck-Gesellschaft], aus denen sich eine Selbstverpflichtung ergibt, Primärdaten mindestens zehn Jahre lang aufzuheben. Eine Verpflichtung zur Langzeitarchivierung ergibt sich auch aus den Prinzipien [WDC-Principles], denen die ICSU-Weltdatenzentren unterliegen. Hier ist z.B. vorgeschrieben, dass die Daten im Falle der

¹⁷ In der EU: http://en.wikipedia.org/wiki/Copyright_law_of_the_European_Union#Duration_of_protection – In anderen Ländern: http://en.wikipedia.org/wiki/List_of_countries'_copyright_length

Schließung eines Welt Datenzentrums an ein anderes Welt Datenzentrum weitergegeben werden müssen, außerdem garantieren die WDC-Prinzipien den (fast) kostenfreien Zugang zu den Daten für Wissenschaftler aller Länder.

Dies spiegelt sich auch in den Nutzungsbedingungen [WDCC-Nutzungsbedingungen] wider: für Forschungszwecke dürfen die Daten des WDCC frei genutzt werden, fließen sie in eine Veröffentlichung ein, muss eine entsprechende Referenzierung erfolgen. Weitergehende Vorschriften der Datenbesitzer müssen allerdings beachtet werden, so wird manchmal von Datenbesitzern die Unterzeichnung einer schriftlichen Erklärung gefordert, dass die Daten nur für Forschungszwecke genutzt werden.

10 Metadaten

An jedes wissenschaftliche Projekt wird die Forderung gestellt, die Erstellung und Verarbeitung von Forschungsdaten umfassend so zu dokumentieren, dass die Entstehung der Daten selbst wie auch die daran geknüpften Interpretationen jederzeit inhaltlich nachvollzogen werden können; dazu dienen typischerweise „Metadaten“.

Metadaten sind „(mehr oder weniger) strukturierte Informationen, die das Erstellen, Verwalten und Nutzen von Datensätzen dauerhaft in und zwischen den Bereichen, in denen sie erzeugt wurden, ermöglichen. [Sie] können benutzt werden, um Datensätze und die Personen, Vorgänge und Systeme, die sie erzeugen, verwalten, unterhalten und nutzen, zu identifizieren, zu beglaubigen und in einen Zusammenhang zu stellen.“ (Paraphrasiert nach „recordkeeping metadata“ in [Wallace, 2001], S. 255, zitiert nach [Day, 2006], S.8.)

Metadaten können aus verschiedenen Perspektiven betrachtet werden, hier als verschiedene „Dimensionen“ beschrieben, obwohl diese nicht ganz voneinander unabhängig sind. Trotzdem müssen alle diese Aspekte (letztlich gleichzeitig!) in Betracht gezogen werden, wenn es an den Aufbau eines effektiven und nützlichen Metadaten systems geht. Mit „Metadaten system“ bezeichnen wir hier die Einheit aus den Begriffs- und/oder Felddefinitionen, die für das Projekt relevant sind (das „Datenmodell“), und seiner Verwaltungsinfrastruktur, z.B. einer Datenbank.

Zweck

Zentrale Frage vor der Einrichtung eines Metadaten systems ist die Frage nach dem Zweck: wozu sollen die Metadaten dienen oder benutzt werden? Je nach Arbeitsbereich und Zielgruppe kann die Erstellung und Bereitstellung von Metadaten sehr unterschiedlichen Zwecken dienen, die letztlich bestimmen, wie sie aussehen und was mit ihnen geschieht. Dies kann unterschiedlich klassifiziert und strukturiert werden, hier wird die folgende Gliederung vorgeschlagen¹⁸:

Datenbestand sichtbar machen: Viele Metadaten dienen dazu, Datenbestände oder Dokumente verfügbar zu machen. Beispiele sind alle Arten von Katalogen (z.B. Biblio-

¹⁸ Day [Day, 2006] zitiert eine Auflistung von Haynes [Haynes, 2004]: „Metadaten dienen der Beschreibung, dem Auffinden, der Verwaltung, der Beschreibung von Eigentum und Echtheit sowie der Interoperabilität von Daten.“

theyskataloge), aber auch Zeitschriften oder Repositories können ihren Inhalt per Metadaten strukturiert anbieten. Typischerweise werden hier Informationen bereitgestellt, die potentiellen NutzerInnen Rückschlüsse auf den Inhalt der beschriebenen Ressourcen ermöglichen.

Daten interpretierbar machen: Oft können Daten nur verstanden werden, wenn die Rahmenbedingungen ihrer Erhebung (Ort, Zeit, Messinstrumente etc.) einerseits und die Bedeutung der Daten (Kategoriensystem, Skalierung etc.) berücksichtigt werden.

Daten austauschen: Metadaten können dem Austausch von Daten zwischen mehr oder weniger eng miteinander assoziierten Partnern dienen, z.B. um einen gemeinsamen virtuellen Korpus zu erzeugen, Daten an bestimmten Punkten zu aggregieren oder zu synchronisieren oder die für den Austausch nötigen Informationen bereitzustellen (z.B. Datumsangaben).

Verwaltung und Pflege: Von zentraler Bedeutung sind stabile Kennzeichen der Daten zu ihrer Identifikation, außerdem Informationen über Zugriffsrechte, etwaige Formataktualisierungen und die (Sicherung der) Echtheit von Daten.

Präsentation: Um Daten (effektiv) zu nutzen sind oft zusätzliche Informationen nötig, z.B. um Objekte (Bilder, Seiten) zur Darstellung zusammenzuführen, sie mit anderen Informationen zu verknüpfen oder die Darstellung an die NutzerInnen anzupassen (z.B. Sprache).

Arten der Information

Aus dem Zweck der Metadaten ergibt sich weitgehend, welche *Arten* von Informationen erhoben und verwaltet werden müssen. „Klassisch“ ist die Einteilung in deskriptive, administrative und strukturelle Metadaten, wie sie auch im METS¹⁹ repräsentiert wird.

Auch hier ist das Projektziel von entscheidender Bedeutung. Z.B. spielt in der Dokumentation wissenschaftlicher Forschung die Provenienz von Daten eine große Rolle, da sie nachvollziehbar macht, wie Daten erhoben und verarbeitet wurden und so die Überprüfbarkeit der Forschungsergebnisse gewährleisten soll. In anderen Zusammenhängen wird der Kontext der Daten und Dokumente betont, der insbesondere bei Archiven eine zentrale Rolle spielt.

Unabhängig vom jeweiligen Projekt erscheinen die folgenden Informationen von allgemeiner Bedeutung:

Objekte: Beschreibung und Identifikation der Objekte, die im Arbeitsprozess entstehen bzw. be- oder verarbeitet werden. Ohne eindeutige und dauerhafte Identifikation der beschriebenen Objekte sind alle weiteren Informationen nur noch schwer zuzuordnen und zu nutzen.

Akteure: Einerseits Personen, Gruppen und/oder Organisationen, die an der Entstehung und Bearbeitung von Daten beteiligt waren; andererseits können auch technische Systeme als Akteure gefasst werden, wenn sie Aktionen anstoßen oder durchführen. Dies ist vor allem für Abrechnungs- und Sicherheitsanforderungen von Bedeutung.

¹⁹ „Metadata Exchange and Transmission Standard“, siehe <http://www.loc.gov/standards/mets/>

Quellen: Dokumentation des Entstehungskontexts von Daten, von Ort, Zeit und Umständen, unter denen sie erhoben wurden und welche Akteure (Personen oder technische Systeme) daran beteiligt waren.

Vorgänge: Dokumentation der Bearbeitungsschritte, denen die Daten unterworfen wurden, von welchen Akteuren welche Aktionen angestoßen wurden, welche Programme und Systeme dabei eingesetzt wurden, und welcher zeitliche Verlauf sich dabei ergab.

Ergebnisse: Die Ergebnisdaten, die für die direkte Nutzung vorgesehen und daher von primärer Relevanz für die Nachnutzung sind, müssen durch Metadaten auffindbar gemacht und eventuell für die Präsentation angereichert werden.

Je nach Projekt und geforderter Vollständigkeit kann das Erheben und Verwalten dieser Daten einen nicht unerheblichen Aufwand fordern – der Umfang der Daten kann durchaus die Masse der Primärdaten erreichen oder überschreiten. Wo irgend möglich sollten sie automatisch erhoben und gespeichert werden; wenn darüber hinaus menschliche Intervention nötig ist, so muss die korrekte und vollständige Erhebung gesichert werden. In jedem Fall muss eine sorgfältige Abwägung (unter Berücksichtigung der notwendig zu erfüllenden Ansprüche) zwischen dem erwarteten Nutzen und dem dafür notwendigen Aufwand an Ressourcen stattfinden.

Semantik

Um Metadaten „verstehen“ zu können, muss man die Bedeutung der verwendeten Terme kennen. Ein simples Beispiel ist die Frage, ob unter „Titel“ eine Überschrift oder ein Namenszusatz zu verstehen ist. In historisch gewachsenen Bereichen wie den Bibliothekswissenschaften oder dem Archivwesen haben sich komplexe Begriffssysteme herausgebildet, die den dortigen speziellen Anforderungen entsprechen, z.B. MARC²⁰ oder ISO 15489 für die Schriftgutverwaltung; ebenso für einige fachspezifische Informationsfelder, z.B. die INSPIRE-Richtlinien zur Geo-Information²¹. Diese sind aber oft nicht einfach auf neuere (digitale) oder andere Objekte übertragbar. Im Gegenzug hat sich mit dem „Dublin Core Metadata Element Set“ (DCMES, [DCMES]) ein Minimalstandard etabliert, der mit 15 Elementen eine elementare und mit den 55 „DCMI Metadata Terms“ [DCTerms] eine erweiterte Beschreibung ermöglicht²². Es zeigt sich darüber hinaus, dass spezielle Wissensbereiche vor allem zur inhaltlichen Beschreibung zusätzlicher Klassifikationen oder Ontologien bedürfen, die entweder einer Allgemeinklassifikation (z.B. der „Dewey Decimal Classification“ [Dewey Decimal Classification]) oder einer fachlichen Spezialklassifikation (z.B. der „Mathematical Subject Classification“ [MSC 2010]) entnommen sind.

Insbesondere für Datenaustausch und -zusammenführung können die unterschiedlichen benutzten Begriffs- und Klassifikationssysteme fast unüberwindliche Hindernisse bilden, so dass oft das Zurückfallen auf einen Minimalstandard wie das DCMES die einzige Möglichkeit

²⁰ <http://www.loc.gov/marc/>

²¹ <http://inspire.jrc.ec.europa.eu/>

²² zur Anwendung siehe <http://dublincore.org/specifications/>

bildet, eine zumindest minimale Interoperabilität zu erreichen. Daher wird das Dublin Core Model auch als „Pidgin language for the digital tourist“²³ bezeichnet.

Neben fachspezifischen Begriffen, Formaten und Ansprüchen gibt es einen gemeinsamen Kern, der durch das DCMES abgedeckt und noch informativ genug ist, um einen elementaren Datenaustausch zu ermöglichen. Typische Elemente und ihre Zwecke sind dabei:

- Der **Identifikator** (das Kennzeichen), der erlaubt, auf das Objekt zuzugreifen.
- Eine **Bezeichnung** (Titel, Label, Etikett), damit man das Objekt bezeichnen und darüber reden kann.
- Das **Datenformat** (evtl. eine Referenz auf eine Formatedatenbank), damit das Objekt präsentiert bzw. gelesen werden kann.
- **Beteiligte Akteure** (Personen, Körperschaften, Programme, Prozesse), damit klar ist, wer Aktionen angestoßen hat bzw. sie verantwortet.
- **Thematische Informationen** (Schlagwörter, Klassifikation, Ontologie), damit potentielle NutzerInnen die Relevanz für sich beurteilen können.

Über diesen Kern hinaus wird es vielfältige verschiedene Zusatzinformationen geben, die jenseits des gegebenen Faches oder Kontextes aber nicht genutzt oder interpretiert werden können.²⁴

Syntax

Metadaten werden typischerweise innerhalb eines Datei- oder Datenbanksystems gespeichert und verwaltet, dazu wird ein Datenmodell benötigt, das reichhaltig genug ist, um alle gewünschten Informationen aufzunehmen. Dies wird von den einzelnen Projekten auf mehr oder weniger individuelle Weise gelöst und ist für sich genommen unproblematisch. Wichtig ist aber die Frage, wie der Austausch von Daten ermöglicht wird, dazu muss natürlich zunächst die oben erwähnte Semantik der Partner zueinander passen. Der *Inhalt* von Datensätzen kann auf verschiedene Weise „verpackt“ werden, z.B. regelt ISO 19139 eine XML-Verpackung der Geoinformation nach ISO 19115 oder MARC XML ein „framework for working with MARC data in a XML environment“ [MARC-XML]. Das Ziel ist hierbei immer, die Daten so transportabel zu machen, dass möglichst keine Informationen verloren gehen. Gleichzeitig sollen die entstehenden Datenpakete möglichst universell und ohne großen technischen Aufwand interpretiert werden können.

Innerhalb enger Partnerschaften kann man natürlich ein beliebiges Austauschformat wählen, z.B. auch Datenbankformate (bibliographisch oder relational). Wegen möglicher technischer Komplikationen bei der Übertragung werden aber üblicherweise textbasierte Formate bevorzugt, welche die Daten mehr (z.B. XML) oder weniger (z.B. CSV) strukturiert bereitstellen.

²³ Tom Baker, Ricky Erway; siehe „digital tourist“ auf <http://dublincore.org/documents/usageguide/glossary.shtml#D>

²⁴ Vgl. dazu auch das Schalenmodell in [Krause, 1996], das auf einer elementareren Ebene die Qualitäten von Metadaten differenziert.

Darüber hinaus gibt es mit RDF die Möglichkeit, komplexere Sachverhalte zu beschreiben. In Kurzform lassen sich die genannten Formate grob wie folgt zusammenfassen:²⁵

- Interne (Datenbank-)Formate: effektiv, aber auf engen Nutzerkreis beschränkt,
- CSV: einfaches Textformat, eher für einfache Strukturen geeignet, recht kompakt,
- XML: vielfältige Möglichkeiten, auch komplexe Strukturen abzubilden, relativ großer Überhang an Text,
- RDF: Möglichkeit, komplexe Zusammenhänge abzubilden, Standard für „Semantic Web“,
- Graphen: stärker visuell orientiert, noch wenig eingesetzt.

Weiterführende Literatur

- Murtha Baca (Editor): Introduction to Metadata Version 3.0, Getty Publications, Los Angeles, 2008 ([Baca, 2008]).
Digitale Ausgabe: http://www.getty.edu/research/publications/electronic_publications/intrometadata/
- M. Day: Digital Curation Manual: Instalment on “Metadata”, 2006 ([Day, 2006]), (<http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/metadata/metadata.pdf>).
- PREMIS (Preservation Metadata: Implementation Strategies) Editorial Committee: PREMIS Data Dictionary for Preservation Metadata, version 2.1, 2011 ([Committee, 2011]), (<http://www.loc.gov/standards/premis/v2/premis-2-1.pdf>).
- A. Powell and P. Johnston: Metadata Guidelines for the Resource Discovery Taskforce, 2011 ([Powell and Johnston, 2011]), (<http://rdtfmetadata.jiscpress.org/>).

Anwendungsfall Klimaforschung

Für Metadaten gibt es im World Data Center for Climate (WDCC) [WDCC] Regeln, die im CERA-2-Datenmodell [CERA2-Datenmodell] zusammengefasst sind. Das CERA-2-Datenmodell ist konform mit dem internationalen Metadaten-Standard ISO 19115/19139.

Es gibt zwei Kategorien deskriptiver Metadaten entsprechend der Struktur des WDCC, das aus den folgenden drei Schichten besteht:

1. Metadaten des Experimentes (Zusammenfassung von Datensätzen): Hier finden sich alle Metadaten zum Experiment, z.B. Titel und Autoren.
2. Metadaten der einzelnen Datensätze
3. Datensätze selbst

²⁵ Mehr zur Präsentation von RDF findet sich z.B. zu JSON unter <http://www.json.org/>, zu Turtle unter <http://www.w3.org/TeamSubmission/turtle/>, zu Graphen in http://portal.acm.org/ft_gateway.cfm?id=1060835&type=pdf.

Von großer Bedeutung sind die Standardnamen, die in Metadaten und Daten verwendet werden und physikalische bzw. chemische Größen benennen, z.B. „air_temperature“. Diese erleichtern die Arbeit mit den Daten und entsprechen der in der Klimaforschung international anerkannten CF-Namensliste²⁶.

11 Identifikatoren und Informationsobjekte

Identifikatoren sind ein wichtiges Querschnittsthema im Datenmanagement. Sie sind ein fundamentales Instrument, um sich auf Informationsobjekte zu beziehen. Es ist zwar möglich, sich auf Informationsobjekte durch relative Angaben und Eigenschaften zu beziehen, wie z.B. „Der Datensatz, der unter Adresse X gespeichert ist“ oder „der Datensatz, der am Datum Y erzeugt wurde“. Dies wird aber schnell unpraktisch und für eine vereinfachte und robustere Verarbeitung ist es notwendig, Daten eigene und eindeutige Identifikatoren zuzuweisen.

Prinzipiell kann man alles identifizieren, seien es einzelne Dateien, eine bestimmte Menge von Dateien, die durch die Dateien repräsentierten abstrakten Inhalte, Teile von digitalen Objekten, dynamische Objekte, Funktionalitäten von Webservices, analoge Gegenstände, Personen oder beliebige andere Objekte. Es muss deshalb explizit definiert werden, was das Informationsobjekt ist, das ein Identifikator bezeichnen soll, und wie der Identifikator aufgelöst werden soll. Diese Entscheidung hängt von dem jeweiligen Anwendungsfall und den Nutzungsszenarien ab. Es kann Fälle geben, in denen Daten in einer bestimmten technischen Version mit immer derselben Bitfolge benötigt werden, z.B. um die Integrität automatisch zu prüfen. Ein entsprechend technisch definierter Identifikator verweist immer auf dieselbe Bitfolge. In anderen Fällen können hingegen nur die Inhalte relevant sein, unabhängig davon ob sie durch eine CSV-, eine EXCEL-Datei oder ein anderes Dateiformat ausgedrückt werden. Ein entsprechend inhaltlich definierter Identifikator könnte dieselben Inhalte in den jeweils benötigten, unterschiedlichen Datenformaten liefern. Beide Identifikatoren können unter Umständen sogar die gleiche Datei liefern, obwohl sie unterschiedliche Informationsobjekte als identisch betrachten.

Das PILIN-Projekt hat eine Reihe von Leitfragen formuliert, um auf Basis eines Informationsmodells bei der Entscheidung zu helfen, wie Identifikatoren in einem Forschungsprojekt vergeben werden sollten [[PILIN Transition project](#)].

Was für Dinge existieren? Der erste Schritt beinhaltet sich darüber klar zu werden, welche Dinge im Forschungsbereich existieren, d.h. welche analogen oder digitalen Objekte, welche Akteure, Funktionalitäten, Relationen etc. Hierfür gibt es zum Teil domänenspezifische Standards oder Ontologien.

Welche Dinge basieren auf anderen? Einige Objekte sind von anderen abgeleitet, und je nach Anwendungsfall können die Ursprungsobjekte oder die resultierenden Elemente wichtiger sein. Z.B. kann es sich um Ausschnitte aus einem größeren Datensatz, inhaltlich überarbeitete Versionen oder in einem anderen Format gespeicherte Präsentations-

²⁶ CF steht für „Climate and Forecast“ und ist eine Konkretisierung des Standard-Datenformates NetCDF [[NetCDF](#)] für die Bedürfnisse der Klimaforschung. Die Namensliste ist ein Anhang zur CF-Konvention [[NetCDF CF Metadata Convention](#)].

oder Arbeitsversionen handeln.

Welche Dinge sind wichtig? Manches muss mit Identifikatoren versehen werden, anderes nicht. Dinge, die außerhalb des Entstehungskontext referenziert werden sollen, wie z.B. ein publizierter Datensatz, sollten üblicherweise Identifikatoren erhalten. Hingegen müssen einzelne Teile von schon mit Identifikatoren versehenen Objekten oftmals nicht neue Identifikatoren bekommen, da sie über relative Angaben adressiert werden können, wie z.B. Zeitabschnitte in Videos oder Seiten in Dokument.

Wie werden Identifikatoren aufgelöst? Es ist nicht von vornherein klar, was mit einem Identifikator gemacht werden kann. Häufig ist die Erwartung, dass das Objekt herunterladbar ist. Wenn aber z.B. abstrakte Objekte identifiziert werden, die in verschiedenen Formaten oder Versionen vorliegen, ist festzulegen, welches Format oder welche Version ausgeliefert wird. Unter Umständen wird auch nicht das Herunterladen angeboten, sondern es sind nur andere Funktionalitäten über den Identifikator ansprechbar, wie z.B. die Darstellung von Metadaten, Service-Operationen etc. Diese sind insbesondere bei nicht als Dateien repräsentierten Objekten wie z.B. Webservices oder analogen Objekten zu definieren.

Wann werden Dinge identifiziert? Der Zeitpunkt der Identifikation muss festgelegt werden. Je nach Szenario kann es sinnvoll sein, sie gleich bei der Erzeugung, erst nach einer Qualitätskontrolle oder erst bei der Publikation zu vergeben.

Es gibt eine Reihe von Schemata für Identifikatoren für unterschiedlichste Zwecke und viele Domänen haben ihre eigenen Standards. An dieser Stelle wird nur auf die sogenannten *Persistent Identifier* eingegangen, die eine besonders dauerhafte Identifizierung ermöglichen sollen, wie sie zum Beispiel zur Zitation benötigt wird. Dies sind u.a. DOI, Handle, URN und PURL. Persistent Identifier werden generiert und in einem Verzeichnisdienst (Resolver) mit Informationen wie der Zugriffsadresse (z.B. URL) gespeichert. Die Identifikatoren selbst werden nicht mehr geändert und bei Veränderungen der Zugriffsadresse oder weiterer Informationen werden nur die Daten im Resolver aktualisiert. Dadurch dass für einen Zugriff zuerst vom Resolver die aktuellen Zugriffsinformationen abgefragt werden, können der Identifikator konstant gehalten und für den Zugriff wechselnde Speicherorte benutzt werden. Zitate und Referenzen, die den persistenten Identifikator anstelle der URL benutzen, werden bei Veränderungen der Speicheradresse nicht ungültig.

Bei Persistent Identifier ist es wichtig zu bedenken, dass sie nicht automatisch persistent sind. Dass die Verknüpfung zwischen dem Identifikator und dem identifizierten Informationsobjekt persistent ist, hängt von einer fortwährenden Pflege der Informationen im Resolver und des Informationsobjekts ab. Persistent Identifier sind nur ein Hilfsmittel. Auch ein mit einem DOI versehener Datensatz kann verloren gehen oder nicht mehr zugreifbar sein, weil er z.B. nur auf einem Arbeitsplatzrechner ohne professionelles Backup gespeichert war oder es versäumt wurde, die Zugriffsadressen im Resolver nachzutragen.

Weiterführende Literatur

- Einführungen zu Persistent Identifier bietet der Australian National Data Service (ANDS) auf verschiedenen Niveaus an (2009):

- Persistent Identifiers Guide Awareness Level ([Australian National Data Service, 2009a]),
(<http://ands.org.au/guides/persistent-identifiers-awareness.pdf>),
Persistent Identifiers Guide Working Level ([Australian National Data Service, 2009b]),
(<http://ands.org.au/guides/persistent-identifiers-working.pdf>),
Persistent Identifiers Guide Expert Level ([Australian National Data Service, 2009c]),
(<http://ands.org.au/guides/persistent-identifiers-expert.pdf>).
- PILIN, Information Modelling Guide for Identifiers in e-research. University of Southern Queensland, 2008 ([PILIN Transition project]),
(<http://resolver.net.au/hdl/102.100.272/6R22YGTRH>).
 - John Kunze, Towards Electronic Persistence Using ARK Identifiers, 2003 ([PILIN Transition project]), (<https://confluence.ucop.edu/download/attachments/16744455/arkcdl.pdf>).
Ein einflussreicher Artikel, der herausarbeitet, dass die Persistenz von Identifikatoren nicht technisch durch Resolver lösbar, sondern eine Selbstverpflichtung zur Pflege der Daten ist.

Anwendungsfall Klimaforschung

Das GeoForschungsZentrum Potsdam [GeoForschungsZentrum] und die drei ICSU-Weltdatenzentren WDCC [WDCC], WDC-MARE [WDC-MARE] und WDC-RSAT [WDC-RSAT] bieten an, in deren Archiven vorhandene Daten mit den Persistent Identifiern DOI (Digital Object Identifier) und URN (Uniform Resource Name) zu versehen. Damit verbunden ist ein Eintrag im GetInfo-Katalog [GetInfo-Katalog], welcher der Suche in den Beständen der Technischen Informationsbibliothek, der Deutschen Zentralbibliotheken für Medizin und Wirtschaftswissenschaften sowie der Suche in weiteren Fachdatenbanken dient.²⁷ Den Eintrag im GetInfo-Katalog nimmt die Technische Informationsbibliothek (TIB) [Technische Informationsbibliothek] mit Sitz in Hannover vor. Die TIB gehört zu einem weltweiten Netz von Registrierungsagenturen, die der Internationalen DOI-Foundation (IDF) [Internationale DOI-Foundation] angeschlossen sind.

Wie alle Persistent Identifier erleichtern DOI und URN die Zitierung, indem der Identifikator einfach als Referenz benutzt wird. Beispiel für ein Zitat mit URN:

Zahn, M (2010), Climate Simulation with CLM, Scenario A1B run no.1, North Atlantic region, WDCC, urn:nbn:de:tib-10.1594/WDCC/CLM_A1B_ZS9

Der zugehörige DOI hat die folgende Gestalt: 10.1594/WDCC/CLM_A1B_ZS

Am WDCC wird der DOI zusammengesetzt aus

- dem Zahlencode 10., der den String als DOI kennzeichnet,
- einer vom IDF für WDCC-Daten fest vergebenen Nummer (1594),
- der Zeichenkette /WDCC und
- einem vom WDCC selbst für die spezielle Veröffentlichung vergebenen Teil.

²⁷ Nach DOI bzw. URN gesucht werden kann unter <http://nbn-resolving.de/>

Der URN ist mit einem Vorspann und am Ende zusätzlich mit einer Prüzfiffer versehen. Das `nb:n` im Vorspann steht für „National Bibliography Number“ (RFC 3188), ein international verwalteter Namensraum der Nationalbibliotheken [[Neuroth et al., 2010](#)].

Am WDCC ist die DOI/URN-Vergabe in einen Qualitätssicherungs-Workflow eingebunden [[WDCC-Qualitätssicherung](#)]. Der Datenproduzent ist für die wissenschaftliche Qualitätskontrolle allein und für die technische Qualitätskontrolle gemeinsam mit dem WDCC verantwortlich. Von Seiten des WDCC werden die folgenden technischen Eigenschaften überprüft:

- Die Zahl der Datensätze ist korrekt und nicht 0.
- Die Größe eines jeden Datensatzes ist nicht 0.
- Die zu den Datensätzen gehörigen Metadaten sind vorhanden und zugreifbar.
- Der Gesamtumfang der Daten ist korrekt.
- Die in den Metadaten angegebenen Zeiten (Start- und Stopdatum, Zeitschritt) sind mit den Daten konsistent. Im Falle von Beobachtungsdaten können einzelne Zeitstempel jedoch fehlen, wenn zu den betreffenden Zeiten nicht gemessen werden konnte.
- Das Datenformat ist valide.
- Variablenbeschreibungen und Daten sind konsistent.

Der Workflow für die eigentliche Publikation ist in [Abb. 7](#) dargestellt. Der Datenproduzent liefert Metadaten wie `Creator` und `Title`, die dann vom Publication Agent durch weitere Metadaten ergänzt werden. Zu den Ergänzungen gehört der Publisher (WDCC) und der DOI und URN selbst.

Die in XML überführten Metadaten²⁸ werden an die TIB geschickt und dort in den GetInfo-Bibliothekskatalog [[GetInfo-Katalog](#)] integriert. DOI/URN und die zugehörige URL werden von der TIB an die Resolver-Anbieter weitergeleitet, welche diese in ihre Resolver aufnehmen.

Nach der DOI-Vergabe dürfen die Daten und diejenigen Metadaten, die Teil des Zitates sind, nicht mehr verändert werden. Im Gegensatz dazu kann die zugehörige URL aber jederzeit durch eine neue ersetzt werden. Dies muss sogar geschehen, wenn sich die Webadresse geändert hat und die alte URL nicht mehr länger besteht.

²⁸ Die erforderlichen und optionalen Metadaten sind in [[DataCite-Metadatenchema](#)] aufgeführt und erfüllen den Standard ISO 690-2 für bibliografische Referenzen. Die Metadaten können auf Dublin Core [[DCMI](#)] abgebildet werden, sowohl auf das Simple Dublin Core Metadata Element Set (DCMES) als auch auf Qualified Dublin Core.

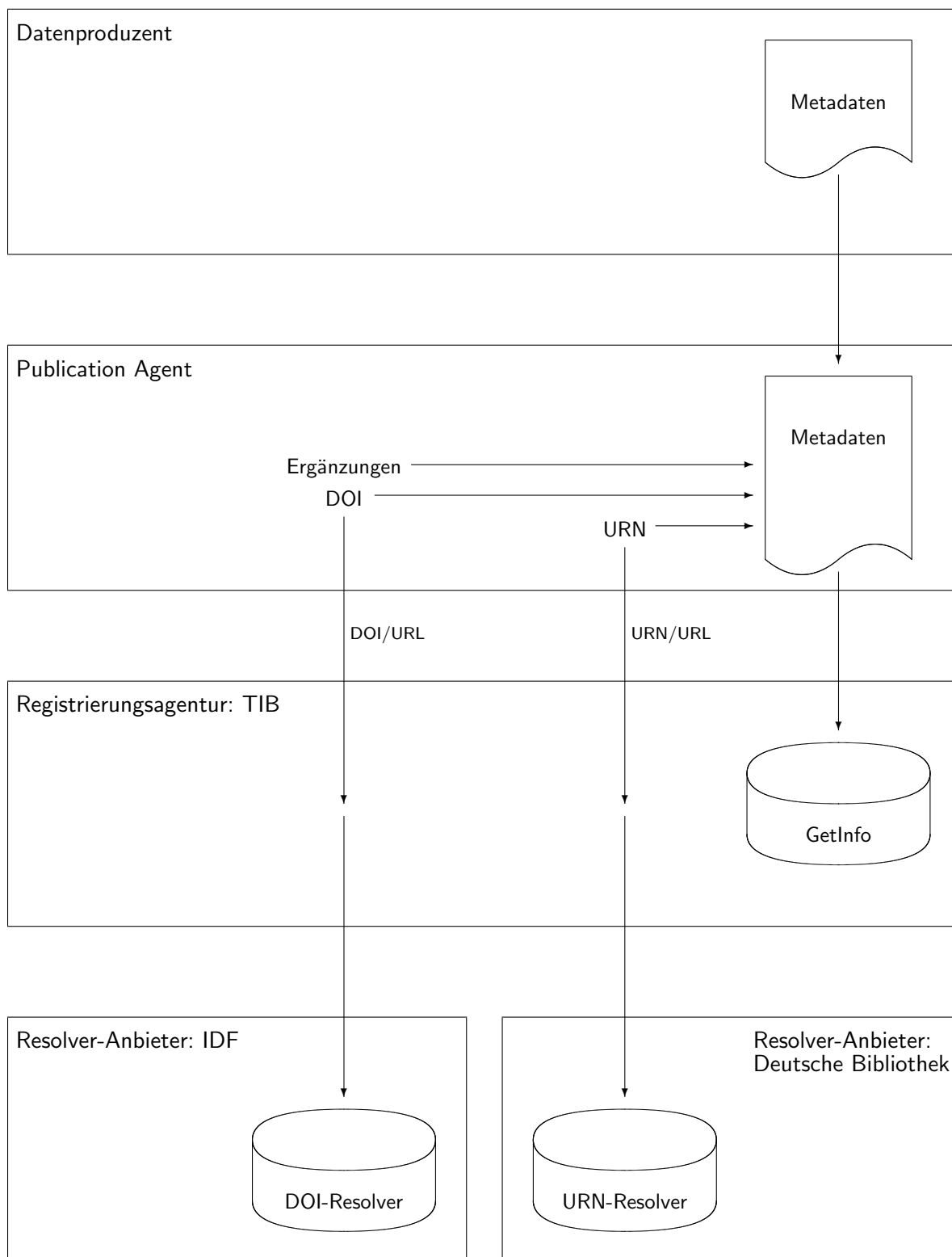


Abb. 7: Workflow zur DOI- und URN-Vergabe (Qualitätssicherung nicht mit dargestellt)

Anhang: Urheberrecht

Einführung

Mit der zunehmenden Verbreitung digitaler Medien ergeben sich umfangreiche neue Möglichkeiten der Verbreitung und Nutzung von Informationen kultureller Äußerungen aller Art. Datenspeicherung geht immer mit rechtlichen Fragestellungen einher. So hat der Gesetzgeber für den Abruf von gespeicherten Daten zwar bereits einige Maßnahmen getroffen (§52b UrhG Elektronische Leseplätze und §137I UrhG neue Nutzungsarten und Retrospektive Digitalisierung). Doch bleiben einige Grauzonen im Bereich langfristiger Datenhaltung bislang ungeklärt.

Um aber auch in Zukunft ihrem gesetzlichen Auftrag nachkommen zu können, sind Gedächtnisorganisationen bei der Sammlung und Erhaltung digital aufgezeichneter kultureller Äußerungen in immer stärkerem Maße auf Langzeitarchivierungsstrategien angewiesen. Um digitale Daten über lange Zeiträume für eine Nutzung zu erschließen und persistent lesbar zu erhalten bedarf es jedoch neben dem gängigen Vorgang des Datensammelns auch Maßnahmen der Datenbearbeitung. Hierbei werden Archive, Museen und Bibliotheken jedoch vor einige juristische Schwierigkeiten gestellt, sobald sie in der Erfüllung ihres Auftrags mit dem Urheberrecht konfrontiert werden. Vom Blickwinkel des Urheberrechtsgesetzes (UrhG) aus gesehen unterscheidet sich die Haltung digitaler Daten signifikant von analogen Daten wie dem klassischen Buch — stellt doch jede elektronische Aktivierung wie Download, Ausdruck etc. im Gegensatz zum bloßen Aufschlagen eines Buches bereits ein Akt der Vervielfältigung dar.

Im Gegensatz zu klassischen Verfahren der Buchrestauration erfolgt die Erhaltung digitaler Daten im Rahmen der üblichen Sicherungsstrategien ausschließlich über die Herstellung von Kopien (Backups), die als Eingriff in die urheberrechtlich geschützte Integrität des Datensatzes interpretiert werden kann und der Zustimmung des Rechteinhabers (Urhebers) bedarf. Es stellt sich demnach die Frage, inwiefern die geltenden Bestimmungen auf den Umgang mit digitalen Daten angewandt werden können.

Rechtliche Situation

Grundsätzlich generiert das UrhG keinen Automatismus, der sämtliche existierenden Daten unter Schutz stellt sowie Zugriff und Verbreitung reglementiert. Vielmehr bedarf es ggf. der Einzelfallprüfung, ob betroffene Datensätze durch ihre Schöpfungshöhe einen geistigen Wert darstellen, der gesetzlichen Schutz genießen kann. So stellt eine Ansammlung gänzlich ungeordneter wissenschaftlicher Rohdaten (z.B. Klimamessdaten) nicht zwangsläufig ein schutzwürdiges Werk *sui generis* dar. Desgleichen greift der urheberrechtliche Schutz nicht bei gemeinfreien Werken wie amtlichen Bekanntmachungen (§5 UrhG) oder Texten, deren Schutzfrist abgelaufen ist – üblicherweise ist dies 70 Jahre nach dem Tod des Urhebers der Fall.

Allerdings bewegt sich jede Gedächtnisorganisation im Umgang mit verwaisten Werken ("orphan works") bereits in einer rechtlichen Grauzone: Wie sind Werke zu bewerten, die nicht

mehr aufgelegt oder vergriffen sind und deren urheberrechtliche Situation somit nicht mehr zu ermitteln ist? Unterliegen bestimmte Werke aber dem bisher geltenden Urheberrecht, werfen sich sogleich Fragen des Bestandaufbaus, der Datensicherung, Datenbearbeitung und Haftung auf.

Bestandsaufbau und Datensammlung

Zur Erweiterung ihres Bestandes sammeln Archive und Bibliotheken Daten, um sie Nutzern zugänglich zu machen. Bei der Aufnahme und Erschließung neuer Werke im Bestand ist die Rechtslage eindeutig: Eine Erschließung ohne vorherige Einwilligung ist ausgeschlossen. Da die Zahl der im Internet frei zugänglichen Daten – insbesondere im Bereich wissenschaftlicher Publikationen – in den letzten Jahren stark zugenommen hat, werden Gedächtnisorganisationen zusehends auch auf diese Bestände zugreifen müssen, um ein möglichst breites Spektrum anbieten zu können. Ein gängiges Verfahren zur Datengewinnung stellt hierbei das Web-Harvesting dar.

Auch im Zeitalter digitaler Netzwerke und Web-Publishing bedeutet die Publikation im Internet noch keinen Blankoscheck für eine ungehinderte Weiterverbreitung oder Vervielfältigung der frei zugänglichen Inhalte. Dies wird vor allem dann relevant, wenn, wie oft zu beobachten, Anbieter der Daten und Urheber nicht identisch sind. Automatisierte Abfragen von Web-Inhalten stoßen an Ausschließlichkeitsrechte von Urhebern, die auch nicht durch Schrankenregelungen des Urheberrechts (§53 Abs. 2 UrhG s.u.) gedeckt sind.

Daher unterliegen Internetpublikationen sehr häufig speziellen Lizenzvereinbarungen wie GNU, GPL oder Creative Commons, die nicht nur das Verhältnis zwischen Anbieter und Urheber regeln, sondern auch eindeutige Passagen zu ihrer Vervielfältigung enthalten. Diese Verträge haben in der Regel Vorrang vor anderen urheberrechtlichen Bestimmungen. Anders als in den Vereinigten Staaten, wo die Auswertung von Internetangeboten bereits als rechtmäßig gilt, wenn der Urheber eine Nachnutzung nicht ausgeschlossen hat oder ihr nachträglich widerspricht, bedarf nach deutschem Recht eine Auswertung digital publizierter Daten oder eine Vervielfältigung durch Harvesting gewonnener Daten grundsätzlich immer der Zustimmung des Urhebers. Seltene Ausnahmen können sich bisweilen für bestimmte Gedächtnisinstitutionen wie Pflichtexemplarbibliotheken (Deutsche Nationalbibliothek oder Bundesarchiv) ergeben, die der Spezialrechtsprechung des Bundes-/ Landesplichtexemplar- und Archivrechts unterliegen.

Ebenfalls unproblematisch nimmt sich die Übernahme gemeinfreier Daten aus, die digitalen Quellen entnommen wurden. Dennoch ergibt sich hier das Dilemma, dass es selbst Pflichtexemplarbibliotheken nicht gestattet ist, jenseits ihres gesetzlichen Auftrages selbstständig zu sammeln und beispielsweise verwaiste Werke in ihren Bestand aufnehmen. Auch die vom Gesetzgeber angestrebten Änderungen (aktueller Gesetzentwurf zum UrhWahrnG für die Nutzung verwaister und vergriffener Werke siehe: BTDS 17/3991 vom 30.11.2010) betreffen lediglich entsprechend der Archivschranke des §53 Abs. 2 UrhG verwaiste Werke, die sich bereits im Bestand des Archivs befinden. Frei verfügbare, etwa im Internet publizierte Daten, deren urheberrechtliche Situation unklar ist, können so nicht erfasst werden. Den Archiven fehlen somit treffsichere Schrankenregelungen, die eine Erfüllung ihres gesetzlichen Sammlungsauftrages gewährleisten.

Bestanderhaltung und Kopien

Vergleichsweise komplexer ist die Situation bei der Handhabung bereits im Archiv befindlicher Bestände. Zur Erhaltung des Bestandes an digitalen Daten einer Gedächtnisorganisation ist die regelmäßige Anfertigung von Kopien unabdingbar, will man nicht Gefahr laufen, die Daten aufgrund eines veralteten, nicht mehr lesbaren Dateiformats unwiederbringlich zu verlieren. Prinzipiell liegt die Entscheidung über die Vervielfältigung von geschützten Werken stets bei deren Urheber.

Urheberrechtlich vergleichsweise unproblematisch gestaltet sich die Anfertigung von Archivkopien – Vervielfältigungen also, die an ohnehin bereits im Bestand einer Gedächtnisorganisation befindlichen Primärwerken zum Zwecke der Bestandswahrung vorgenommen werden: Die Archivschränke des §53 Abs. 2 Satz 1 UrhG erlaubt die Übernahme von bereits vorhandenen Archivkopien (analog/digital) in eben dasselbe Archiv zum Zweck der Sicherung oder der internen Nutzung, sofern sie nicht zur Erweiterung des eigenen Bestandes vorgenommen wurde. Grundvoraussetzung ist das Vorhandensein einer originalen und rechtmäßig erworbenen Vorlage eines eigenen Werkstücks im Besitz des Archivs. Einschränkend gilt diese Regelung jedoch nur für Archive, deren Tätigkeit in öffentlichen Interesse liegt.

Für den Umgang mit digitalen Daten im Archiv kennt das Gesetz indes einige Spezialfälle, die das Anfertigen von Kopien im Einzelfall regeln. So unterliegen Datenbankwerke nach §87a Abs. 1 UrhG oder Computerprogramme (§69d UrhG; Ausnahme: Fehlerbeseitigung, Sicherungskopie des Besitzers) strengen Restriktionen und dürfen grundsätzlich nicht kopiert werden (§53, Abs. 5 UrhG). Hierunter fallen somit auch komplexe Webseiten sowie systematisierte oder elektronisch erschlossene Datensammlungen. Besteht weiterhin ein technischer Kopierschutz an einem digitale Daten bewahrenden Speichermedium (CD-ROM etc.), so darf dieser keinesfalls überwunden oder umgangen werden (§95a UrhG). Allerdings ist der Urheber des so gesicherten Materials gesetzlich verpflichtet, einem Archiv, dem er die Anfertigung einer Archivkopie zur Vermehrung des Bestandes eingeräumt hat, die Mittel zur Beseitigung an die Hand zu geben (§95b UrhG).

Bei der Sicherung digitaler Daten ist es, ganz ähnlich dem natürlichen Verschleiß von Bucheinbänden, grundsätzlich unvermeidbar, dass Speichermedien und -formate dem technischen Verfall unterliegen und mit der Zeit veralten. Eine überzeugende LZA-Strategie schließt daher notwendigerweise auch Eingriffe in die technische Integrität der digitalen Bestände ein, um eine nachhaltige Lesbarkeit des Datenmaterials zu gewährleisten. Gängige Verfahren hierzu wie Emulation, Migration und Konversion stellen stets Interpretationsprozesse unlesbarer Datensätze dar, wobei der informationshaltige Kernbestand der Daten weitgehend unberührt bleiben sollte.

Digitale Datenträger unterscheiden sich aber von analogen Speichermitteln in ihrer deutlich kürzeren Haltbarkeit. Technische Eingriffe in digitale Datenspeicher fallen demnach u.U. wesentlich früher an als die Restauration eines Buches. Juristisch ist dies insofern von Belang, als Eingriffe in den inhaltlichen Bestand eines Datensatzes einer urheberrechtlichen Schutzfrist unterliegen. Da diese erst 70 Jahre nach dem Tode des Urhebers verstreicht, die meisten technischen Eingriffe in Speichermedien aber deutlich früher anstehen, ist es unabdingbar die nach wie vor umstrittene Frage eindeutig zu regeln, ob in einem solchen Verfahren juristisch eine Veränderung im Sinne von §23 UrhG zu sehen ist oder eine bloße Vervielfältigung nach §16 UrhG vorliegt.

Verfechter der Hypothese, eine Migration stelle einen Akt der Vervielfältigung dar, führen das Argument ins Feld, dass derartige Maßnahmen zwar den technischen Träger von Daten manipulieren, als solche aber als rein mechanischer Akt zu verstehen wären und die dem eigentlichen Schaffensprozess des Urhebers entspringenden Daten davon nicht betroffen seien. Als Vergleich wird der Austausch eines Bucheinbandes zu Restaurationszwecken herangezogen, der die äußere Hülle der Schrift intakt lässt und die den Informationsgehalt tragenden Seiten weder in Zahl, Reihenfolge noch Aussagekraft manipuliert.

Trifft dies zu, bleiben Verfahren wie Migration und Emulation – sofern §53 Abs. 2 UrhG eintritt, der die Parameter für die Anfertigung einer Archivkopie definiert – durch das Urheberrecht abgedeckt. Folgt man dieser Sichtweise, bedürfen lediglich Eingriffe in den originalen Informationsbestand der Daten des Einverständnisses ihres jeweiligen Urhebers. Weiterhin ungeklärt bleibt selbst bei dieser Interpretation der Sachverhalt bei ausgesprochen sensiblen Daten, wie etwa dem Aktenbestand medizinischer Archive über die Behandlung von Patienten.

Kritiker dieser Sichtweise halten die Schrankenregelung des UrhG für digitale Daten für gänzlich unzureichend. Demnach reichen Änderungen an der digitale Daten erschließende Software über eine bloße Restauration des rasch veraltenden Datenträgers hinaus – stellen doch Anpassungen am Format des Informationsgehalts bereits einen hinreichenden Eingriff in die Integrität des Datensatzes dar. Fallen solche Maßnahmen fernerhin noch in die bereits angesprochene Schutzfrist des UrhG, wären sie demzufolge durch die bereits existierende Archivschränke nicht gedeckt, wodurch sich zumindest eine zweideutige Rechtslage ergibt.

Zusammenfassend lässt sich feststellen, dass sich insbesondere Forschungsinstitutionen angesichts dieser ungeklärten Rechtslage zusehends mit einem Kostenaufwand für Lizenzen urheberrechtlich geschützter Forschungsdaten konfrontiert sehen, die zum Zwecke der Verifizierbarkeit, Replizierbarkeit und Nachnutzung wissenschaftlicher Arbeit gehalten werden. Dass das UrhG ausschließlich natürliche Personen als Urheber ausweist, kommt für juristische Personen wie Forschungsinstitutionen erschwerend hinzu. Folglich können urheberrechtliche Ansprüche von institutionalisierten wie nicht-institutionalisierten Forschungsverbänden und -einrichtungen auf unter ihrer Ägide entstandene Daten nicht geltend gemacht werden und müssen ggf. durch den verantwortlichen Einzelwissenschaftler nachträglich eingeräumt werden.

Zugriffsrechte

Rein urheberrechtlich gesehen definiert §53 Abs. 2 UrhG Archive als Gedächtnisorganisationen, denen als zentrale Aufgabe das Sammeln, Bewahren und Sichern ihrer Bestände obliegt. Da die Erschließung von Beständen für Nutzer als wesentlicher Teil im Selbstverständnis der meisten Archive liegt, sind sie von den Privilegierungen des Urheberrechts streng genommen ausgeschlossen, was eine nachhaltige LZA verunmöglicht, sobald sie die Einsichtnahme in den gespeicherten Datenbestand umfasst. Zu den bisher vorgesehenen drei Nutzungsvarianten zählen daher ausschließlich die interne, die eingeschränkte und die offene Nutzung.

Die interne Nutzung sieht lediglich den Zugriff auf Daten durch Mitarbeiter der Gedächtnisorganisation vor, die zu Archivzwecken (Metadatenanreicherung, Katalogisierung, Sicherung

etc.) Einsicht in die jeweiligen Inhalte nehmen müssen. Sobald aber, wie im Falle digitaler Daten oft notwendig, der Zugriff mittels Download oder Computerausdruck erfolgt, muss das Archiv gewährleisten, dass die Einsichtnahme nur zu wissenschaftlichen Zwecken geschieht (§53 Abs. 2 S. 1 Nr. 1 UrhG).

Um die zuvor angesprochene Problematik der fehlenden Regelungen zur Einsichtnahme in digitale Daten zu kompensieren hat der Gesetzgeber mit §52b UrhG eine Möglichkeit geschaffen, digitale Inhalte in eingeschränkter Nutzung über öffentliche Bildschirmleseplätze einzusehen, die sich in den Räumlichkeiten des Archivs befinden. Gedeckt wird auch die Darstellung von Digitalisaten zuvor analog gespeicherter Medien. Obgleich diese Regelung eine deutliche Erleichterung für Gedächtnisorganisationen darstellt, greift diese Privilegierung auch in diesem Falle nur für Archive, wenn ihre Sammelaktivitäten in öffentlichem Interesse liegen. Schul- oder universitäre Institutsbibliotheken sowie kommerzielle Archive bleiben von dieser Verbesserung ausgeschlossen.

Erfolgt die Zugriffnahme an den öffentlich zugängigen Stellen wiederum über Download oder Ausdruck, muss auch hier das Kriterium der Nutzung zu wissenschaftlichen Zwecken nachweislich erfüllt sein. Sollten aber bestimmte Umstände dazu führen, dass §53b UrhG nicht greift, kann eine Gedächtnisorganisation ihre Bestände ausschließlich auszugsweise für einen relevanten und eng begrenzten Personenkreis erschließen, welcher die Einsichtnahme zu wissenschaftlichen Zwecken vornimmt (§52a UrhG).

Die offene Nutzung digitaler Daten hängt hingegen gänzlich von der Zustimmung des Urhebers ab: Solange ein Urheber die von ihm geschaffenen Daten nicht freigibt, dürfen sie von keiner Gedächtnisorganisation ortsungebunden publiziert werden.

Haftung

Neben rein urheberrechtlichen Fragestellungen unterliegt die langfristige Speicherung digitaler Daten weiteren juristischen Kriterien. So ist von jeder Gedächtnisorganisation darauf zu achten, dass keine Daten mit volksverhetzenden, pornografischen, ehrverletzenden oder gegen Bestimmungen des Patentrechts verstoßenden Inhalten gespeichert werden. Die §§7-10 des Telemediengesetzes (TMG) unterscheiden in Fragen der Haftung von Archiven zwischen eigenen und fremden Inhalten. Für die Zuweisung zu der einen oder anderen Kategorie ist die Nutzersicht ausschlaggebend: Sollte ein Nutzer bei der Einsicht in einen Datensatz auf rechtswidrige Inhalte stoßen und diese urheberrechtlich eindeutig als Eigentum der jeweiligen Gedächtnisorganisation erkennen (Archivzeitschriften, Kataloge etc.), werden die fraglichen Daten als dem Archiv eigen klassifiziert. Dies hat zur Folge, dass das Archiv für evtl. resultierende Rechtsverletzungen haftbar gemacht werden kann.

Um in solchen Fällen eine eindeutige Rechtslage herzustellen, ist es jeder Gedächtnisorganisation angeraten, eigene Inhalte deutlich als solche auszuweisen und einen Haftungsausschluss für Daten aus externen Quellen zu formulieren. Sollten dennoch gesetzeswidrige Inhalte in Archiven entdeckt werden, unterliegt jede Gedächtnisorganisation einer Sorgfaltspflicht (§7 Abs. 2 TMG), die ihr eine Sperrung von Daten evident rechtswidrigen Inhalts auferlegt. Hierbei genügt es, ohne eine endgültige Löschung vorzunehmen, die Nutzung und Zugänglichkeit solcher Daten zu verwehren, um einem Haftungsanspruch zu entgehen.

Fazit

Angesichts der beschriebenen Komplexität der juristischen Situation sind Gedächtnisorganisationen mit dem aktuellen juristischen Handwerkszeug zwar imstande, ihrem Auftrag zur Sammlung analoger Daten nachzukommen, sehen sich jedoch kaum in der Lage, eine rechtlich abgesicherte LZA-Strategie für digitale Daten zu verfolgen. Zur Klärung der Situation warten sie vielmehr auf einen Satz an griffigen Schrankenbestimmungen, welche die digitale Langzeitarchivierung jenseits vervielfältigender Maßnahmen zur Bestandserhaltung um Möglichkeiten der Bearbeitung und Umgestaltung von Dateiformaten und des Bestandsaufbaus ergänzen (Web-Harvesting). Es bleibt festzuhalten, dass das Ausbleiben solcher Instrumente, die neben den bisher privilegierten Pflichtexemplarbibliotheken dem ganzen Spektrum an öffentlichen Gedächtnisorganisationen zugute kommen, eine umfassende und zuverlässige Dokumentation digital fixierter kultureller Äußerungen auf absehbare Zeit verhindern könnte.

Literatur

- Nationaal Archief. Costs of Digital Preservation, 2005. URL <http://www.nationaalarchief.nl/sites/default/files/docs/kennisbank/codpv1.pdf>.
- Australian National Data Service. Persistent Identifiers Guide Awareness Level, 2009a. URL <http://ands.org.au/guides/persistent-identifiers-awareness.pdf>.
- Australian National Data Service. Persistent Identifiers Guide Working Level, 2009b. URL <http://ands.org.au/guides/persistent-identifiers-working.pdf>.
- Australian National Data Service. Persistent Identifiers Guide Expert Level, 2009c. URL <http://ands.org.au/guides/persistent-identifiers-expert.pdf>.
- Paul Ayriss, Richard Davies, Rory McLeod, Rui Miao, Helen Shenton, and Paul Wheatley. The LIFE2 Final Project Report, 2008. URL <http://eprints.ucl.ac.uk/11758/1/11758.pdf>.
- Murtha Baca, editor. *Introduction to Metadata Version 3.0*, 2008. Getty Publications. URL http://www.getty.edu/research/publications/electronic_publications/intrometadata/.
- BADC: British Atmospheric Data Centre (BADC). URL <http://badc.nerc.ac.uk/>.
- Neil Beagrie, Brian Lavoie, and Matthew Woollard. Keeping Research Data Safe 2, 2010. URL <http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf>.
- Neil Beagrie, Anna Palaiologk, Daphne Charles, Rachel Beagrie, Rob Beagrie, and Brian Lavoie. User Guide for Keeping Research Data Safe, 2011. URL http://www.beagrie.com/KeepingResearchDataSafe_UserGuide_v2.pdf.
- Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, and Hans Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, 10, 2009. doi: DOI10.1007/s00799-009-0057-1. URL <http://www.ifs.tuwien.ac.at/~becker/pubs/becker-ijdl2009.pdf>.
- Tobias Beinert, Georg Büchler, Sabine Graf, Karsten Huth, Christian Keitel, Jens Ludwig, Peter Rödiger, and Tobias Steinke. *nestor-materialien 10: Wege ins Archiv / Ein Leitfaden für die Informationsübernahme in das digitale Langzeitarchiv, Version I*. nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland, Koblenz, 2008. URL <http://nbn-resolving.de/urn:nbn:de:0008-2008103009>.
- Berliner Erklärung. URL <http://oa.mpg.de/lang/de/berlin-prozess/berliner-erklarung/>.
- Bitstream Preservation. Bitstream Preservation: Bewertungskriterien für Speicherdienste, 2009. URL <http://www.wissgrid.de/workgroups/ap3/2011-03-08--bitstream-preservation.pdf>.

Blue Ribbon Task Force. Sustainable Economics for a Digital Planet, 2010.

URL http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

C3Grid-Portal. URL <http://www.c3grid.de/portal/>.

C3Grid Website. Collaborative Climate Community Data and Processing Grid Webseite. URL <http://www.c3grid.de/C3-Grid>.

CERA: Climate and Environment data Retrieval and Archiving system (CERA).

URL <http://cera-www.dkrz.de/CERA/>.

CERA2-Datenmodell.

URL <http://www.mad.zmaw.de/wdc-for-climate/cera-data-model>.

Charles Beagrie Ltd and JISC, editors. *Keeping Research Data Safe Factsheet*. Charles Beagrie Ltd, 2010. URL http://www.beagrie.com/KRDS_Factsheet_0910.pdf.

CMIP5: Coupled Model Intercomparison Project, Phase 5.

URL <http://cmip-pcmdi.llnl.gov/cmip5/>.

CMIP5 Datenbeschreibung.

URL http://cmip-pcmdi.llnl.gov/cmip5/data_description.html.

CMIP5 Experiment Design.

URL http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf.

CMIP5 Qualitätskontrolle. URL <http://purl.org/org/cmip5/qc>.

DataCite-Metadatenchema.

URL http://datacite.org/schema/DataCite-MetadatenKernel_v2.0.pdf.

Michael Day. Digital Curation Manual. Instalment on "Metadata". Version 1.1, 2006. URL <http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/metadata, geladenam24.11.11>.

DCMES: The Dublin Core Metadata Element Set, Version 1.1.

URL <http://dublincore.org/documents/dces/>.

DCMI: The Dublin Core Metadata Initiative. URL <http://dublincore.org/>.

DCTerms: The DCMI Metadata Terms.

URL <http://dublincore.org/documents/dcmi-terms/>.

Madeleine de Cock Buning, Barbara van Dinther, Christina G. Jeppesen de Boer, and Allard Ringnalda. Report on the Legal Status of Research Data in the four partner countries, 2011. URL <http://www.knowledge-exchange.info/Default.aspx?ID=461>.

Deutsche Forschungsgemeinschaft, editor. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*. Wiley-VCH, 1998.

Deutscher Wetterdienst (DWD). URL <http://www.dwd.de/>.

- Deutsches Klimarechenzentrum (DKRZ). URL <http://www.dkrz.de/>.
- Dewey Decimal Classification: The Dewey Decimal Classification (DDC) system.
URL <http://www.oclc.org/dewey/>.
- DFG-Antrag: Leitfaden für die Antragstellung, 2011.
URL http://www.dfg.de/formulare/54_01/54_01_de.pdf.
- Digital Curation 101: Ingest.
URL <http://www.dcc.ac.uk/sites/default/files/DC%20101%20Ingest.pdf>.
- GeoForschungsZentrum Potsdam. URL <http://www.gfz-potsdam.de/>.
- GeoNetwork: GeoNetwork opensource: a standards based Geographic Data and Information Management System for the web. URL <http://www.osgeo.org/geonetwork/>.
- GetInfo-Katalog: GetInfo-Katalog für die Suche in den Beständen der Technischen Informationsbibliothek, der Deutschen Zentralbibliotheken für Medizin und Wirtschaftswissenschaften sowie in weiteren Fachdatenbanken. URL <https://getinfo.de/>.
- GNDMS: Generation N Data Management System. URL <http://gndms.zib.de>.
- Serge J. Goldstein and Mark Ratliff. DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data, 2010.
URL <http://arks.princeton.edu/ark:/88435/dsp01w6634361k>.
- Stefan Gradmann. Interoperability. A key concept for large scale, persistent digital libraries. In *Digital Preservation Europe (DPE) Briefing Paper*. Digital Preservation Europe, 2008. URL <http://www.digitalpreservationeurope.eu/publications/briefs/interoperability.pdf>.
- GRIB: GRIdded Binary (GRIB).
URL http://www.cpc.ncep.noaa.gov/products/wesley/reading_grib.html.
- D. Haynes. *Metadata for information management and retrieval*. Facet, 2004.
- International Council for Science. URL <http://www.icsu.org/>.
- Internationale DOI-Foundation (IDF). URL <http://www.doi.org/>.
- IPCC: Intergovernmental Panel on Climate Change (IPCC). URL <http://www.ipcc.ch/>.
- Jürgen Krause. Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung – Schalenmodell. IZ-Arbeitsbericht Nr. 6, IZ Sozialwissenschaften, Bonn. 1996. URL http://www.thesis.org/fileadmin/upload/forschung/publikationen/thesis_reihen/iz_arbeitsberichte/ab6.pdf.
- KRDS-I2S2-Tools: KRDS/I2S2 Digital Preservation Benefit Analysis Tools, 2011. URL <http://beagrie.com/krds-i2s2.php>.
- MARC-XML: MARC in XML. URL <http://www.loc.gov/marc/marcxml.html>.
- Max-Planck-Gesellschaft, editor. *Verantwortliches Handeln in der Wissenschaft*. Max-Planck-Gesellschaft.

MSC 2010: The Mathematical Subject Classification 2010.

URL <http://www.zentralblatt-math.org/msc/data/msc2010.pdf>.

nestor AG Digitale Bestandserhaltung. nestor Materialien 15: Leitfaden zur digitalen Bestandserhaltung. Vorgehensmodell und Umsetzung, 2011.

URL <http://nbn-resolving.de/urn:nbn:de:0008-2011101804>.

NetCDF: Network Common Data Form.

URL <http://www.unidata.ucar.edu/software/netcdf/>.

NetCDF CF Metadata Convention: NetCDF Climate and Forecast (CF) Metadata Convention. URL <http://cf-pcmdi.llnl.gov/>.

Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann, and Karsten Huth, editors. *nestor-Handbuch, Version 2.3*, 2010.

URL <http://nestor.sub.uni-goettingen.de/handbuch/>.

OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

URL <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

OSGeo: The Open Source Geospatial Foundation. URL <http://www.osgeo.org/>.

Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. Climate Data Challenges in the 21st Century. *Science*, 331:700–702, 2011.

URL <http://www.sciencemag.org/content/331/6018/700.full>.

PCMDI: Program of Climate Model Diagnosis and Intercomparison (PCMDI).

URL <http://www-pcmdi.llnl.gov/>.

PILIN Transition project. Information Modelling Guide for Identifiers in e-research, 2008. URL

<http://www.linkaffiliates.net.au/pilin2/files/infomodellingresearch.pdf>.

Andy Powell and Pete Johnston. Metadata guidelines for the resource discovery taskforce, 2011. URL <http://rdtfmetadata.jiscpress.org/>.

PREMIS (Preservation Metadata: Implementation Strategies) Editorial Committee, editor. *PREMIS Data Dictionary for Preservation Metadata, version 2.1*, 2011.

URL <http://www.loc.gov/standards/premis/v2/premis-2-1.pdf>.

Gerald Spindler and Tobias Hillegeist. KoLaWiss Project: Arbeitspaket 4 - Recht, 2009. URL

http://kolawiss.uni-goettingen.de/projektergebnisse/AP4_Report.pdf.

Technische Informationsbibliothek, Hannover. URL <http://www.tib-hannover.de/>.

Andrew Treloar and Cathrine Harboe-Ree. Data management and the curation continuum: how the Monash experience is informing repository relationships, 2008. URL

http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf.

Unterzeichner der Berliner Erklärung.

URL <http://oa.mpg.de/lang/de/berlin-prozess/signatoren/>.

D. Wallace. Archiving metadata forum: report from the Recordkeeping Metadata Working Meeting, June 2000. *Archival Science*, 1:253–269, 2001.

WDC-MARE: World Data Center for Marine Environmental Sciences.

URL <http://www.wdc-mare.org/>.

WDC-Principles: Principles and Responsibilities of ICSU World Data Centers.

URL <http://www.ngdc.noaa.gov/wdc/guide/gdsystema.html>.

WDC-RSAT: World Data Center for Remote Sensing of the Atmosphere.

URL <http://wdc.dlr.de/>.

WDCC: World Data Center for Climate (WDCC).

URL <http://www.mad.zmaw.de/wdc-for-climate/>.

WDCC-Nutzungsbedingungen: Nutzungsbedingungen für WDCC-Daten.

URL <http://cera-www.dkrz.de/WDCC/ui/docs/TermsOfUse.html>.

WDCC-Qualitätssicherung: Qualitätssicherungs-Workflow bei DOI/URN-Vergabe am WDCC. URL [http://umwelt.wikidora.com/wikidora/wiki/Standard%20Procedure%20\(Objectives\)](http://umwelt.wikidora.com/wikidora/wiki/Standard%20Procedure%20(Objectives)).

Angus Whyte and Andrew Wilson. How to Appraise & Select Research Data for Curation, 2010. URL

<http://www.dcc.ac.uk/resources/how-guides/appraise-select-research-data>.