nestor

Langzeitarchivierung von Rohdaten

Thomas Severiens
Eberhard R. Hilf
Institute for Science Networking Oldenburg GmbH
an der Carl von Ossietzky Universität Oldenburg

nestor-materialien 6

This page is intended to be blank.

Studie zum Stand vorhandener Forschungsdaten und Rohdaten aus wissenschaftlichen Tätigkeiten:

Erfordernisse und Eignung zur Archivierung bzw. Zurverfügungstellung in Deutschland (Primärdaten)

Thomas Severiens
Eberhard R. Hilf

Institute for Science Networking Oldenburg GmbH
an der Carl von Ossietzky Universität Oldenburg

nestor – materialien 6

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Herausgegeben von

nestor - Kompetenznetzwerk Langzeitarchivierung und
Langzeitverfügbarkeit Digitaler Ressourcen für Deutschland

nestor - Network of Expertise in Long-Term Storage of Digital Resources

http://www.langzeitarchivierung.de

Projektpartner

Bayerische Staatsbibliothek, München
Bundesarchiv
Computer- und Medienservice / Universitätsbibliothek der Humboldt-Universität zu Berlin
Die Deutsche Bibliothek, Leipzig, Frankfurt am Main, Berlin (Projektleitung)
Generaldirektion der Staatlichen Archive Bayerns, München
Institut für Museumskunde, Berlin
Niedersächsische Staats- und Universitätsbibliothek, Göttingen

Betreuer dieser Veröffentlichung:

Niedersächsische Staats- und Universitätsbibliothek Göttingen
(Heike Neuroth und Stefan Strathmann)

URN: urn:nbn:de:0008-20051114018
http://nbn-resolving.de/urn:nbn:de:0008-20051114018

Die vorliegende Studie wurde von Thomas Severiens und Prof. Dr. Eberhard R. Hilf am Institute for Science Networking (ISN) an der Carl von Ossietzky Universität Oldenburg erstellt. Sie ist im Auftrag von nestor – Kompetenznetzwerk Langzeitarchivierung unter Betreuung durch die Niedersächsische Staats- und Universitätsbibliothek Göttingen entstanden.

Die Expertise will einen Leitfaden an die Hand geben, wie und welche Forschungsdaten/Primärdaten archiviert werden sollten. Sie schätzt die künftig zu erwartenden Datenmengen ab und erläutert, welche Verfügbarkeiten und Zugriffsmodalitäten vorgehalten werden müssen. Dabei stellt sie vorhandene, offensichtlich gewordene Problemfelder dar und zeigt im nationalen und internationalen Kontext implementierte Lösungen auf.

Aus der Expertise - in der Zusammenschau mit der Stellungnahme von Jens Klump vom GeoForschungsZentrum Potsdam - ergeben sich u.a. folgende Brennpunkte künftiger Aktivitäten:

1. Ausbau der Koordination zwischen den an der Archivierung der Forschungsdaten/Primärdaten beteiligten Institutionen.
2. Ausbau der modularen Vernetzung der Datenanbieter im Zuge der Bemühungen um eine angemessene Langzeitarchivierung der relevanten Daten.
3. Integration der Industrie und anderer kommerzieller Datenanbieter in nestor bzw. in die gemeinschaftlichen Bemühungen um die digitale Langzeitarchivierung.
4. Implementierung von beispielhaft gelungenen Umsetzungen (Best Practise) im Bereich der Langzeitarchivierung von Forschungsdaten/Primärdaten. Einen guten Ansatz stellt die Implementierung des im DFG-geförderten Projekt „Zitierfähigkeit wissenschaftlicher Primärdaten“ entwickelten Workflow-Modells dar, das bereits in die Produktion gegangen ist.

Für die Partner des Projekts nestor - Kompetenznetzwerk Langzeitarchivierung

Heike Neuroth und Stefan Strathmann

Niedersächsische Staats- und Universitätsbibliothek Göttingen

Stellungnahme zur Expertise „Studie zum Stand vorhandener Forschungsdaten und Rohdaten aus wissenschaftlicher Tätigkeit: Erfordernisse und Eignung zur Archivierung bzw. Zurverfügungstellung in Deutschland (Primärdaten)“ vom Institute for Science Networking Oldenburg GmbH (ISN)

Jens Klump, GeoForschungsZentrum Potsdam (GFZ), 15.11.2005

Die Studie von Severiens und Hilf zum Stand des Umgangs mit Forschungsdaten in Deutschland in Bezug auf deren Langzeitarchivierung gibt einen Überblick über die Situation in den Datenzentren im Sommer 2004. Ihre Auswertung der Fragebögen und des Workshops zum Thema zeigt die Heterogenität des Feldes. In den anderthalb Jahren, die seit dem vergangen sind, hat sich einiges weiterentwickelt. Es wird zunehmend deutlich, dass die Datenzentren die Anforderungen der Langzeitarchivie-rung erkannt haben, auch jenseits des 10-Jahre-Horizonts der Anforderungen aus den „Empfehlungen für gute wissenschaftliche Praxis“.

In der Studie wird dargestellt, dass es durchaus Ansätze zur Langzeitarchivierung von wissenschaftlichen Daten gibt, diese aber bisher sehr spezifisch auf einzelne Disziplinen zugeschnitten sind. Eine fächerübergreifende Lösung gibt es bisher noch nicht. Ob es dabei jemals einheitliche Datenformate geben wird, darf bezweifelt werden, selbst eine Vereinheitlichung der Metadatenschemata ist nicht absehbar. Vielmehr sind Fortschritte durch eine Verbesserung der Konzepte zur Langzeitarchivierung zu erwarten. Durch eine Vernetzung von Datenzentren untereinander zur besseren Recherche von Daten in Onlinekatalogen und über Webservices wäre eine starke Zentralisierung und eine daraus folgende zu starke Vereinheitlichung – und damit auch unangemessene Vereinfachung – von Metadatenprofilen vermeidbar.

Bemerkenswert ist die Kluft zwischen kleinen wissenschaftlichen Einrichtungen, in denen man sich der Problematik einer Langzeitdatenhaltung sehr bewusst ist, und den Großforschungseinrichtungen, die in Bezug auf Langzeitarchivierung von wissenschaftlichen Daten bereits eigene Lösungen vorantreiben, diese Entwicklungen jedoch bisher

untereinander nicht koordinieren. Leider konnten sich die Forschungsförderungsinstitutionen bislang noch nicht dazu durchringen, Projekte zu fördern, die den Technologietransfer zwischen Großforschungseinrichtungen und kleineren Institutionen zum Inhalt haben. In der industriellen Forschung spielt Langzeitarchivierung von Daten offensichtlich keine Rolle – eine erstaunliche Parallele zur Politik der wissenschaftlichen Verlage. Der Grund liegt vermutlich in der fehlenden Bereitschaft, Daten mit anderen zu teilen, sei es auf kommerzieller Basis oder im offenen Zugang (open access).

Die Studie zeigt auch, dass die Reaktivierung von Datensätzen mit der Zeit stark abnimmt und jenseits der 10-Jahres-Grenze bisher keine Rolle spielte. Insofern bezeichnen die Datenzentren mit der Begriff Langzeitarchivierung einen wesentlich kürzeren Zeitraum als Bibliotheken oder Archive. Ausnahmen intensiver Nachnutzung vorhandener Daten gibt es dort, wo deren Haltung und Verbreitung zentral organisiert werden, wie z.B. in den Weltdatenzentren des International Council of Scientific Unions (ICSU).

Die Studie weist als Beispiel für einen integrierten Ansatz von Archivierung und Zurverfügungstellung von Daten auf das DFG-Projekt „Publikation und Zitierfähigkeit wissenschaftlicher Daten“ (http://www.std-doi.de) hin. In diesem Projekt haben die deutschen Weltdatenzentren, das GeoForschungsZentrum Potsdam und die Technische Informationsbibliothek Hannover ein System aufgebaut, mit dem die Voraussetzung für die wissenschaftliche Publikation und die Zitierbarkeit von elektronischen Objekten geschaffen wurden. Durch die Archivierung der Daten in den Weltdatenzentren, ihren Nachweis in Katalogen und ihre Lokalisierbarkeit über persistente Identifikatoren sind die Voraussetzungen für eine nachhaltige Speicherung wissenschaftlichen Daten und ihrer Nachnutzbarkeit gegeben.

Insgesamt gibt die Studie von Severiens und Hilf einen guten Überblick über die derzeitige Praxis und wertvolle Hinweise auf bestehende Defizite in der Langzeitarchivierung von Daten. Sie kann wesentlich zur weiteren Entwicklung dieses Feldes beitragen.

Für das Projekt nestor *Kompetenznetzwerk Langzeitarchivierung* im Auftrag *Der Deutschen Bibliothek* (DDB) unter aktiver Unterstützung durch die *Staats- und Universitätsbibliothek Göttingen* (SUB) 2004 erstellt.

Studie zum Stand vorhandener Forschungsdaten und Rohdaten aus wissenschaftlichen Tätigkeiten: Erfordernisse und Eignung zur Archivierung bzw. Zurverfügungstellung in Deutschland (Primärdaten)

Institute for Science Networking Oldenburg GmbH
an der Carl von Ossietzky Universität Oldenburg
Ammerländer Heerstraße 121,
26129 Oldenburg

Thomas Severiens, *severiens@isn-oldenburg.de*
Eberhard R. Hilf, *www.isn-oldenburg.de/~hilf/*

This page is intended to be blank.

Inhaltsverzeichnis

This page is intended to be blank.

1 Vorwort

Diese Studie soll die Entwicklung einer bundesdeutschen Infrastruktur zur Langzeitarchiverung und Verfügbarhaltung digitaler Rohdaten aus der wissenschaftlichen Forschung unterstützen. Sie wurde im Sommer 2004 vom Institute for Science Networking in Oldenburg unter aktiver Unterstützung durch die Staats- und Universitätsbibliothek Göttingen für das BMBF-Projekt nestor *Kompetenznetzwerk Langzeitarchivierung* entwickelt.

Als selbst nicht in der Geschäftskette und den Entscheidungsebenen der Archivierung und Verfügbarhaltung Beteiligte haben die Autoren dieser Expertise den Vorteil des Blickes von außen. Die Entscheidungen zu politischen Weichenstellungen liegen ebenso wie die Realisierung und Detaillierung der LZA-Infrastruktur (Langzeitarchivierungs-Infrastruktur) beim Staat, den Ländern und den Archivierungsinstitutionen.

Die Expertise will einen Überblick über die vorhandenen Datenquellen, Projekte, Vernetzungen und den Bedarf durch die Wissenschaftler als Datenlieferanten und Datennutzer einer LZA-Infrastruktur für wissenschaftliche Primärdaten bieten. Ziel ist es, die Relevanz primärer Daten bei der LZA zu dokumentieren und eine Abschätzung der Dimension zu archivierender Primärdaten, sowie der von den Wissenschaftlern erwarteten Zugriffsmodalitäten zu bieten.

Vorgabe und Bezug für diese Expertise ist das Exposé des nestor-Projektes in der Fassung vom 3. Dezember 2003.

Thomas Severiens Eberhard R. Hilf

This page is intended to be blank.

2 Zusammenfassung

Wissenschaftliche Primärdaten lassen sich in eine LZA-Infrastruktur ideal aufnehmen. Sie ergänzen vorhandene Volltext-Repositories. Sie sind bereits zu großen Teilen mit Metadaten oder anderen Selbstbeschreibungen inklusive einer technischen Ausleseanweisung versehen, die sie auch für zukünftige Nutzer verständlich beschreiben.

Methodisch resultieren die Empfehlungen und Aussagen aus den Ergebnissen eines Online-Fragebogens und Interviews.

Erwartet wird ein Datenvolumen LZA-relevanter wissenschaftlicher Primärdaten aus der Bundesrepublik Deutschland von mindestens 1.000 bis 2.000 TByte jährlich mit stark steigender Tendenz, wobei diese Zahlen noch mit zahlreichen unbekannten Faktoren behaftet sind und lediglich als Richtgröße, nicht als Planungsgröße zu verstehen sind.

Derzeit werden wissenschaftliche Primärdaten fast ausschließlich in proprietären oder zumindest disziplin-spezifischen binären Dateiformaten vorgehalten, die jedoch im Rahmen der üblichen Publikationsaktivitäten und der Dokumentation der Experimente offen dokumentiert sind. Diese lassen sich daher auf Grund der Dokumentation auch langfristig verstehen und konvertieren. Die Interviewpartner haben bestätigt, dass üblicherweise Checksummen zur Prüfung der Integrität der Daten bereits in die Dateiformate integriert sind. Meist wird aus Gründen der Effizienz im Umgang mit großen Datenmengen auf ein Encoding in XML verzichtet, da dies eine deutliche Vergrößerung des Datenvolumens um etwa den Faktor 10 verursacht, dem wegen der Größe der fachspezifischen Datensammlungen für die Datenerzeuger kein angemessener Struktur-Mehrwert gegenübersteht.

Technisch verfügen die meisten Archive der Primärdatenproduzenten bereits über Schnittstellen für einen effizienten und an die jeweilige Fachdisziplin angepassten externen Zugang zu den Daten und Metadaten.

Als problematisch kann sich bei dem Versuch, alle Primärdatenanbieter in eine LZA-Infrastruktur einzubinden, die Vielfalt der Fachdisziplinen erweisen, die jeweils eigene Datenformate, Metadaten und Schnittstellen verwenden. Hier zu gemeinsamen Standards zu kommen, könnte sehr aufwändig bis unmöglich werden. Machbar hingegen sollte ein modulares System der Vernetzung sein, wie es das WDC-Netzwerk[1] heute bereits verwirklicht ist. Vorbildliche organisatorische und technische Implementierungen einer LZA-Infrastruktur für Primärdaten aus den wissenschaftlichen Disziplinen konnten allenfalls mit dem WDC-Netzwerk ausgemacht werden, ansonsten handelt es sich vorwiegend um archivierende, nicht jedoch langzeitarchivierende Aktivitäten, an die angeknüpft werden kann.

Die Nutzung archivierter Primärdaten erfolgt derzeit zu allermeist nur innerhalb der jeweiligen Fachdisziplin. Aktuell erfasste Primärdaten sind großteils auch heute nur von Wissenschaftlern der aktuellen Fachdisziplin verstehbar und nutzbar. Um diese Primärdaten Wissenschaftlern in der Zukunft zu erschließen, bedarf es der engen Kopplung der Daten an die Dokumentationen der Datenformate, die Methoden der Datenerfassung und deren Dokumentation von Motivation und Zielsetzung.

1 WDC: World-Data-Center, www.ngdc.noaa.gov/wdc/

Die von den Datennutzern erwarteten Zugriffszeiten bewegen sich, je nach Alter der Daten, im Bereich von sofortigem Online-Zugriff bis zur Zusendung binnen einiger Tage. Die Datenanbieter verlangen die Möglichkeit, die Weitergabe der Daten an Dritte sicher zu verhindern. Dieses begründet sich insbesondere aus teilweise enthaltenen personenbezogenen Daten oder der Relevanz der Daten im wirtschaftlichen Wettbewerb. In den bisherigen Archiven werden diese Rechte durch eine Nutzerverwaltung geregelt. Kollegen, die aus wissenschaftlichen Beweggründen Zugriff auf Primärdaten erhalten wollen, erhalten diesen in den allermeisten Fällen. Die Implementation in eine LZA-Infrastruktur setzt die technische Umsetzung dieser bisher menschlichen Zugangskontrolle voraus. Hier steht zu erwarten, dass eine kurzfristige Implementierung problematisch ist. Entsprechend kann auf Modellprojekte zugegriffen werden, die anschließend weitere Archive integrieren können und eine Motivation der Datenerzeuger begründen.

3 Zielsetzung und Definitionen

Diese Studie hat zum Ziel, Hinweise zur Beantwortung der Frage zu liefern: **Wie können wissenschaftlich relevante primäre Forschungsdaten in Deutschland langfristig bereitgestellt und archiviert werden?**

Inhaltlich eng hiermit verwandt sind die Fragestellungen:

1 Welche Daten sind für die Archivierung relevant?

2 Welches Datenvolumen ist zu erwarten?

3 Wie werden Forschungsdaten bisher vorgehalten? Datenträger und Dateiformate?

4 Wie und inwieweit werden Forschungsdaten bereitgestellt bzw. ausgetauscht?

5 Gibt es (international) bereits Ansätze zur Archivierung von Forschungsdaten?

6 Wieso und wie oft wurden Rohdaten aus bisherigen Archiven „reaktiviert"?

7 Welche Verfügbarkeitsarten, Reaktionszeiten und Zugangsmechanismen sind notwendig und werden erwartet?

8 Enthalten Forschungsdaten schon eine „Selbstdokumentation"? Inwieweit ist diese notwendig?

Aufgaben

Aufgaben dieser Studie sind,

- den Umfang vorhandener deutscher Primärdaten bezüglich Volumen und Vorhaltung abzuschätzen,
- die verwendeten Datenformate zu analysieren und insbesondere bezüglich deren Selbstdokumentation und Ihrer Eignung für die LZA zu bewerten,
- die implementierten Verfügbarkeitsmechanismen, Reaktionszeiten und Zugriffsmodalitäten in den vorhandenen Archiven der Datenproduzenten zu analysieren,
- die Einbindung in internationale Archivierungssysteme und LZA-Ansätze für Primärdaten zusammenzustellen.

Ziel ist es, potenziellen Entscheidungsträgern – sowohl auf der Erzeuger-, wie der Archivierungs- und der Nutzerseite – einen Leitfaden an die Hand zu geben, wie Primärdaten in eine LZA-Infrastruktur integriert werden sollten. Dies beinhaltet eine Abschätzung bezüglich anfallender Datenmengen, Verfügbarkeitszeiten und Zugriffsmodalitäten. Die Studie erfolgt zwar für den bundesdeutschen Raum, steht aber immer vor dem Hintergrund der internationalen Vernetzung der LZA.

Definitionen

Zur Klarstellung der Sprache in dieser Studie, seien definiert:

- **Forschungsdaten:** Synonym zu „Primärdaten".

- **LZA:** Langzeitarchivierung, die die Erhaltung des Bit-Streams und der darin enthaltenen Information auf unbegrenzte Zeit garantiert und die Verfügbarkeit und Nutzbarkeit der archivierten Inhalte organisiert und reglementiert.
- **Metadaten:** Oft Synonym zu „Selbstdokumentation", jedoch meist als Beschreibung des Inhaltes mittels kontrolliertem Vokabular verstanden. Da der Begriff „Metadaten" bereits zur Inhaltserschließung textueller Objekte verbreitet ist und die „Selbstdokumentation" von Primärdaten bezüglich der inhaltlichen Beschreibungstiefe in der bisherigen Praxis deutlich darüber hinaus reicht, wird in dieser Studie der Begriff der „Metadaten" soweit möglich vermieden.
- **Primärdaten:** In der Forschung und wissenschaftlichen Arbeit gewonnene Daten aus Messprozessen oder Experimenten. Dies können beispielsweise gemessene Wetterdaten sein, ebenso Messwerte aus Experimenten (Beschleuniger-Experimente etc.) und auch Resultate sozialwissenschaftlicher Forschung (Resultate von Volkszählungen etc.) in einem inhaltlich möglichst vollständigen Zustand, nicht jedoch statistische Zusammenfassungen oder Publikationen über diese Primärdaten.
- **Rohdaten:** Synonym zu „Primärdaten".
- **Selbstdokumentation:** Jede Form von in den Primärdaten enthaltenen oder fest mit diesen gekoppelten Informationen zur Entstehung, Erstellung oder Messung dieser Daten, ihrem Encoding und den verwendeten Verfahren und Apparaten.

4 Vorgehen

Grundlage dieser Studie bilden eine Fragebogen-Aktion, Interviews und ein Workshop. Die Studie versucht, die hierdurch gewonnenen Erkenntnisse, die gesammelten Fakten und offenkundig gewordenen Problembereiche synoptisch zusammenzustellen.

Fragebogen

Der Online-Fragebogen[2] wurde in enger Kooperation mit der SUB Göttingen und dem Wissenschaftsjournalisten Richard Sietmann erstellt und gemäß nestor-interner Absprache vom Medienzentrum der HU Berlin gehostet. Der Fragebogen ist seit dem 17. Mai 2004 auf dem Server abrufbar. In die Auswertung flossen alle Antworten ein, die bis zum 12. Dezember 2004 eingingen.

Der Fragebogen enthielt die folgenden Fragen:

- 1. Does your institute frequently generate research output in the form of primary data? (scientific digital raw data) [yes | no]
- 2. In which format are primary data generated and stored by your institute? (weather observation data, space observation data, accelerator data, social science surveys, epidemiological data, data from clinical studies, gene sequencing data, sound or video data, ...)
- 2.1. Are these raw data edited before their storage?
- 2.2. How are these data stored in your institution? (Tapes, CD-ROM, HDD, ...)
- 2.3. How many Bytes of data are stored per year?
- 2.4. In what file formats are the data stored? (XML, binary, proprietary, ...)
- 2.5. Are the files stored in a self-describing way? (metadata describing the content of each file, condition of measurement included into every file, useable by prospective external or future users ...)
- 2.6. How long will these data be stored in your institution?
- 2.7. Are these data made available for use by scientists from other institutions? On what terms? (on request, restricted access, open access)
- 3. Do you already cooperate with an external agency for long-term preservation of the primary data? If so, which institution? If not, for what reasons?
- 3.1. Do you consider intellectual property or (lack of) Digital Rights Management an impediment to external long-term preservation archiving?
- 3.2. Are there disincentives to external long-term preservation resulting from institutional competition / national security restrictions / intellectual property issues / privacy concerns (clinical studies) / other ?
- 4. Provisions for long-term preservation:

2 http://www2.hu-berlin.de/nestor/questionnaire/q2.php

- 4.1. Which part of the data should be archived for more than 10 years? What are the selection criteria?
- 4.2. What type of primary data should be archived for more than 30 years? What are the selection criteria?
- 4.3. Does your institute have a list of selection criteria for data to be long-term archived? Where is this list available? Who is managing the selection?
- 4.4. Who should pay for long-term preservation of these data?
- 4.5. Who should pay for accessing the data?
- 4.6. Should these data be made available for use by scientists from other institutes?
- 5. Did your institute ever request primary data from other institutes? (for validating published research, or follow-up studies)
- 6. Did your institute ever use data from long-term preservation? Which kind of data? Roughly, how often did you use long-term preserved data?
- 6.1. Did your institute benefit from this usage?
- 6.2. What did you have to pay for the usage of these data?
- 6.3. What would be a fair price?
- 6.4. What would be the maximum response time for getting access to the data? (I would like it all online to access within some seconds, 24 hours would be OK, 7 days)
- 6.5. What should be improved in future services?
- 7. Would you welcome national guidelines, a national framework, or external institutional support for long-term preservation?
- 7.1. In what way would they support the mission of your institute?
- 8. Do you have any other comments on issues that are relevant but not covered in this questionnaire?

Am 17. Mai 2004 (und als Erinnerung an jene, die bis dann noch nicht reagiert hatten, am 5. August 2004 erneut) wurden 327 Personen und Institutionen (davon 276 in Deutschland, 51 im Ausland) per individueller Email angeschrieben. Zielgruppe des Fragebogens und damit Empfänger der Mail waren u.a. alle Max-Planck-Institute, alle Fraunhofer-Institute, alle Institute der Helmholtz-Gemeinschaft, alle Institute der Leibniz-Gemeinschaft, sowie zahlreiche Wissenschaftler, die auf internationalen Tagungen Vorträge zum Themenfeld der Langzeitarchivierung von Messdaten und anderen Primärdaten gehalten hatten. Zusätzlich auch zahlreiche Museen, Archive und Dienstleister mit naturwissenschaftlichem Bezug, sowie einzelne Unternehmen der gewerblichen Wirtschaft, bei denen ein Bezug zur Thematik erkennbar war. Die Universitäten bzw. ihre Arbeitsgruppen waren als Erzeuger und Nutzer von Primärdaten ebenfalls in der Zielgruppe enthalten. Angeschrieben wurden 74 universitäre Empfänger (alle in Deutschland). Die Zielgruppe waren also die Produzenten, die meist gleichzeitig auch die Nutzer von Primärdaten sind, sowie Wissenschaftler als Nutzer von Primärdaten.

Workshop

Am 1. und 2. Juni 2004 veranstalteten die Autoren zusammen mit der SUB in Göttingen einen Workshop[3], auf dem erste Ergebnisse der Fragebogenaktion insbesondere vor dem Hintergrund einer Einbindung der Primärdaten bei Entwicklung einer bundesdeutschen LZA-Policy diskutiert wurden.

Eingeladen wurden alle Adressaten des Fragebogens und zusätzlich die Adressaten des Fragebogens der parallelen Expertise[4] zur Entwicklung eines Leitfadens einer nationalen Policy. Eine Liste der Teilnehmer findet sich auf dem Web-Server der Tagung.

Interviews und vertiefende Gespräche

Mit thematisch relevanten und möglichst repräsentativen Institutionen, die weitestgehend auch den Fragebogen ausgefüllt hatten wurden vertiefende Gespräche geführt, teils am Rande anderer Treffen, teils als Telefoninterviews. Diese Gespräche sollten die Perspektive über den natürlich begrenzten Blick eines Fragebogens um organisationstypische und Aspekte erweitern.

Die Interviews und Gespräche insbesondere mit Vertretern folgender Institutionen flossen in die Auswertungen und Bewertungen ein:

- GSI – Gesellschaft für Schwerionenforschung, Darmstadt
- AWI – Alfred-Wegener-Institut, Bremerhaven
- GKSS Forschungszentrum, Geesthacht
- FH Gelsenkirchen (Prof. Zielesny [Lehrstuhl für Chemo- und Bioinformatik], wegen Kontakten in die chemische Industrie)
- IN2P3 – Institut National de Physique Nucléaire et de Physique des Particules, Lyon (Rechenzentrum)

Interview-Anfragen (teils in Kooperation mit der nestor-Leitung) direkt an die chemische Industrie, die Luftfahrt-Industrie und eine große Versicherungsgruppe führten zu Absagen, die jedoch teilweise relevante Aussagen für diese Studie lieferten und entsprechend einflossen.

Alle Zitate aus den Interviews sind genehmigt. Ein wörtliches Protokoll aller Interviews wurde (bis auf eine Ausnahme aus technischen Gründen) zunächst als Arbeitsgrundlage erstellt und von den Interviewpartnern genehmigt. Ihm wurden die Zitate entnommen. Alle Interviewpartner verweigerten die Freigabe der Gesamtmitschrift. Aufgrund der Vielzahl nebensächlicher Bemerkungen erscheint dies nachvollziehbar und für diese Studie nicht schädlich. Alle Zitate von IN2P3 wurden von Th. Severiens aus dem Französischen übersetzt.

3 http://www.isn-oldenburg.de/nestor-workshop/
4 nestor-materialien 7, http://nbn-resolving.de/urn:nbn:de:0008-20051114021

This page is intended to be blank.

5 Auswertung der Fragebogen-Aktion

Der Rücklauf der 327 versendeten Fragebögen war mit 61 Antworten vergleichsweise hoch. Damit betrug die Rücklaufquote 18%, bei erfreulich vollständig ausgefüllten Fragebögen. Dabei gab es mit einem Rücklauf von 16% der Fragebögen aus dem Ausland keinen signifikanten Unterschied zu den 19% Rücklauf von bundesdeutschen Adressaten. Fachlich waren die Rückläufe wie folgt aufgegliedert: 23% Physik, Hochenergieforschung, physikalische Effekte, etc.; 21% Geowissenschaften, Erdbeobachtung; 16% Astronomie; 16% Medizin, Genforschung etc.; 10% Sprachwissenschaften, Sozialwissenschaften; 9% Fachübergreifend; 5% Regionalwissenschaften und andere.

Unerwartet kamen 41% der Antworten über andere Kanäle als den ausgefüllten Online-Fragebogen: Fax, Email, Telefonanrufe mit Antworten zu einzelnen der Fragen.

Die Autoren sehen die gewonnene Stichprobe als relevant an, da insbesondere jene Institutionen ausführlich geantwortet haben, die als Großproduzenten wissenschaftlicher Primärdaten den Autoren bekannt waren. Auch sind alle Adressatenkreise in der Stichprobe vertreten, wobei sich eine zu feine Aufsplitterung bei 61 Antwortsätzen aus statistischen Gründen verbietet. Leider war der Rücklauf aus den wissenschaftlichen Museen so schwach, dass hier keine Relevanz der Stichprobe erreicht wurde.

Zu einer kritischen Betrachtung der angewandten Methode verweisen wir auf das entsprechende Kapitel am Ende dieser Studie.

Eine Auswertung nach den einzelnen Fragenblöcken zeigt folgendes Ergebnis:

Primärdatenproduktion

Als für die Bewertung der Antworten besonders relevant zeigt sich die Frage nach der Produktion von Primärdaten durch die eigene Institution (**Frage 1.**). 73% der Antworten stammen von Institutionen, die sowohl Primärdaten produzieren, als auch nutzen (hier „Primärdatenproduzenten“ genannt), entsprechend 27% von Institutionen, die keine Primärdaten produzieren (hier „Primärdatennutzer“ genannt). Hierbei ist zwischen deutschen und ausländischen Antworten kein relevanter Unterschied festzustellen.

Speicherformat und Datenhaltung

Dieser Fragenkomplex (**Frage 2.**) wurde ausschließlich den Primärdatenproduzenten gestellt.

Die einleitende Frage nach der Art und dem Genre der erzeugten Primärdaten lieferte eine breite Auflistung (sortiert nach Eingang, thematische Überlappungen zusammengefasst):

- gene sequencing data
- diffraction data

- cosmological simulation data
- images
- movies
- nmr data
- geophysical measurement data
- fusion plasma data
- solar radio spectra
- numerical simulation data
- population health statistical data
- astronomical images
- linguistic data (audio of spoken language)
- results of questionnaires
- marine data
- seismic data
- weather data
- high energy collider data

Frage 2.1 die nach der Bearbeitung der Daten fragt, sollte insbesondere klären, inwiefern es sich tatsächlich um Primärdaten handelt. 45% speichern die „rohen" also unbearbeiteten Daten, 43% bearbeiten die Daten vor der Speicherung, 12% treffen keine eindeutige Aussage. Als Beispiele für die Bearbeitung werden angeführt die Transkription gesprochener Sprache in Text, oder die verlustbehaftete Komprimierung von Bild- und Filmmaterial.

In den Interviews wurde diese Frage weiter vertieft. Insbesondere Großforschungseinrichtungen verwiesen hier geschlossen auf die Richtlinien der DFG und die Förderbedingungen des BMBF, die beide eine Speicherung der unbearbeitet erfassten Daten für 10 Jahre verlangen. Die Pharmaindustrie ist insbesondere bei der Entwicklung und Produktion von Medikamenten mit Mess- und Prozessdaten konfrontiert, sowie mit Messreihen aus dem Zulassungsverfahren von Medikamenten. Diese werden dort noch ausgedruckt bzw. wenigstens solange gespeichert, wie der Patentschutz gilt und das Produkt auf dem Markt ist. Auch hier werden die Daten bewusst unbearbeitet gespeichert.

Frage 2.2 fragt nach den Speichermedien und brachte (gewichtet nach den in Frage 2.3 erfragten Datenvolumina) folgende Medien und Verfahren hervor:

- DLT-Tapes (insbesondere in den hochgradig datenintensiven Disziplinen)
- Datenbank (viele Institutionen lagern ihre Primärdaten komplett in Datenbanken. Entsprechend werden diese mit jedem Wechsel der Datenbank auf aktuelle Hardware migriert und sind jederzeit innerhalb des Intranets abrufbar)

- Opto-chemische Datenträger (CD, DVD, ...)
- DAT-Tapes und andere Tapes (außer DLT)

Frage 2.3 fragt nach dem Umfang der produzierten Primärdaten. Hier lassen sich die antwortenden Institutionen in drei Größenklassen unterscheiden:

- bis 100 GB / a
- bis 1 TB /a
- über 1 TB /a

Zur ersten Gruppe gehören insbesondere die sprachwissenschaftlichen und geisteswissenschaftlichen, sowie soziologischen Institutionen. Diese machen in der Gesamtheit der angeschriebenen Primärdatenproduzenten etwa ein Drittel aus. Hieraus ergibt sich ein vermutliches Datenaufkommen in 2004 von maximal 6,7 TB aus den angeschriebenen Institutionen (100 GB x Anzahl der Institutionen (67) in der Stichprobe der angeschriebenen deutschen 202 Primärdatenproduzenten).

Zur zweiten Gruppe gehören kleinere Institutionen mit naturwissenschaftlichem oder medizinischem Forschungsschwerpunkt. Das arithmetische Mittel der angegebenen Volumina liegt bei 652 GB/a in dieser Gruppe. In diese Gruppe sind etwa die Hälfte der angeschriebenen Institutionen einzugliedern. Damit ergibt sich hier ein vermutliches Datenvolumen in 2004 von 66 TB (652 GB x Anzahl der Institutionen [101]).

Zur dritten Gruppe gehören insbesondere die so genannten Großforschungseinrichtungen, also die Betreiber von Großgeräten, sowie Institutionen, die Videomaterial erzeugen. Das arithmetische Mittel der angegebenen Volumina liegt bei 27 TB/a in dieser Gruppe. In diese Gruppe sind etwa ein Sechstel der angeschriebenen Institutionen einzugliedern. Damit ergibt sich hier ein vermutliches Datenvolumen in 2004 von 909 TB (27 TB x Anzahl der Institutionen [34]). Insbesondere die Interviews zeigten, dass diese Schätzung wohl sehr konservativ ist und sich das Volumen der digitalen Primärdaten inzwischen allein in den bundesdeutschen Institutionen wohl deutlich vergrößert hat, insbesondere wenn man die Resultate internationaler Projekte mit bundesdeutscher Beteiligung hinzunimmt (bspw. CERN). Insgesamt scheint also ein Datenvolumen von 1.000 TByte bis 2.000 TByte jährlich für 2004 eine Richtgröße zu sein, die jedoch mit vielen Unbekannten versehen ist.

Die insgesamt 8 beantworteten Fragebögen aus dem Ausland, von denen 5 von Primärdatenproduzenten stammen, lassen sich mit wissenschaftlich seriösen Methoden nicht weiter auswerten. Insbesondere weil die Rückläufe schwerpunktmäßig von europäischen Großforschungsprojekten stammen, die Primärdaten in Größenordnungen produzieren, die bei Aufnahme in eine deutsche LZA-Infrastruktur diese deutlich majorisieren würden (über 50% der Daten).

Frage 2.4 fragte nach dem Dateiformat der archivierten Primärdaten. 70% der Institutionen speichern ihre Daten in einem binären Format, 10% verwenden ein XML-basiertes Format, 20% ASCII oder andere Text-Formate. Rechnet man diese Ergebnisse jedoch auf das Datenvolumen der jeweiligen Institution um, so ergibt sich

ein einheitlicheres Bild: 97,8% der Daten werden in binären Formaten abgespeichert, 0,3% in XML oder Derivaten und 1,9% in Textformaten. Hier gibt es keinen relevanten Unterschied zu internationalen Archiven, wobei zu betonen ist, dass es insbesondere innerhalb der Initiativen und Projekte, die Primärdaten produzieren und archivieren Unterschiede bezüglich des Dateiformates gibt, je nachdem woher die jeweiligen Daten stammen und wozu diese innerhalb des Vorhabens verwendet werden.

Frage 2.5 fragt nach der Selbstbeschreibung der Primärdaten. 63% der Institutionen speichern ihre Primärdaten zusammen mit einer Selbstbeschreibung, die teilweise separat liegt, bei den meisten binären Formaten jedoch Teil der Datei (meist in Form eines Headers) ist. 23% der Institutionen speichern keine beschreibenden Informationen zusammen mit den Daten, 4% äußern sich hierzu nicht. Wiederum umgerechnet auf die Datenvolumina zeigt sich, dass die Großproduzenten von Primärdaten diese öfter mit beschreibenden Informationen koppeln (91% der Daten enthalten beschreibende Informationen). Hierzu fällt auf, dass Daten aus internationalen Projekten praktisch ausschließlich mit Selbstbeschreibungen vorliegen.

Nachfragen zeigten zu diesem Punkt auf, dass innerhalb der Großforschungsinstitutionen detaillierte Richtlinien zu Qualitätsstandards der Metadaten genutzt werden, die innerhalb internationaler Projekte abgestimmt werden. Am weitesten fortgeschritten zu einer übergreifenden, integrierten Standardisierung der zu erfassenden Metadaten ist hier offenkundig das Netzwerk der World-Data-Center. Alle hierzu befragten Institutionen betonten, dass es keine allgemein gültigen Richtlinien geben kann, sondern immer fall- und fachspezifische gibt.

Frage 2.6 fragt nach dem zeitlichen Horizont der hausinternen LZA-Bemühungen, also danach, wie lange die Institution selbst plant die Daten vorzuhalten. Hier zeigt sich ein heterogenes Bild in den Antworten (in Auswahl):

- 3-8 years, several years
- permanent, unlimited
- till deleted
- 10 years
- until the end of the funding period of the projects. Copies may „survive“ elsewhere
- 40 years
- more then 10 years

Die Mehrzahl der Institutionen nennt (unabhängig vom Datenvolumen) den Zeitraum von 10 Jahren, wobei auffällt, dass der genannte Zeitraum in internationalen Projekten größer ist als in rein nationalen Vorhaben.

Frage 2.7 fragt nach der Bereitschaft, die Daten Kollegen außerhalb der Institution zu wissenschaftlichen Zwecken zur Verfügung zu stellen.

45% der Institutionen stellen anderen Wissenschaftlern Primärdaten zur Verfügung.

27% lehnen dies ab.

28% stellen Primärdaten nur unter Vorbehalten oder Bedingungen zur Verfügung:

- erst nach Ablauf von 6 bis 12 Monaten nach der Messung (mehrfach genannt)
- sofern nicht an kommerzielle Datenbank-Anbieter verkauft (mehrfach genannt)
- nur um den Kriterien der DFG zu genügen und dann auch nur in Ausnahmefällen (einmal genannt)

Kooperationen

Der Fragenkomplex zur Kooperation (**Frage 3**) wurde nur von wenigen Institutionen beantwortet.

Frage 3.0 fragt nach vorhandenen Kooperationen mit externen Einrichtungen zur LZA der eigenen Primärdaten. Bis auf eine Institution beantworten alle diese Frage verneinend oder gar nicht. Die einzige Institution, die diese Frage bejaht, berichtet von Aktivitäten in Ungarn, die die Autoren jedoch auch mittels Nachfragen nicht näher spezifizieren konnten.

Frage 3.1 fragt danach, ob die rechtlichen Rahmenbedingungen und DRM als Hindernis für die LZA durch externe Institutionen gesehen wird.

55% der Institutionen verneinen dies.

45% dagegen sehen hierin ein Hindernis. Beklagt wird insbesondere eine rechtliche Grauzone, DRM wird von einigen als Hindernis für den Zugriff auf eigene Daten befürchtet.

In beiden Gruppen sind Primärdaten-Produzenten wie -Nutzer anteilig etwa gleich vertreten.

Frage 3.2 fragt nach Hemmnissen für die externe LZA der Primärdaten, resultierend aus der Natur der Institution oder der Daten.

60% der Institutionen sehen hier keine Hemmnisse.

40% sehen hier Hemmnisse. Insbesondere den Schutz der Daten vor unberechtigtem Zugriff sehen sie als problematisch, sowie die mit der LZA verbundenen Kosten, sofern diese nicht durch eigenes Personal erfolgt.

Auffällig ist, dass ausschließlich Antworten von deutschen Institutionen hier Hemmnisse sehen, während alle ausländischen Antwortenden keine Hemmnisse sehen.

Regelungen

Frage 4.1 fragt nach den Auswahlkriterien für jene Primärdaten, die länger als 10 Jahre gespeichert werden sollen.

33% der Institutionen, hierbei alle Großforschungseinrichtungen, die geantwortet

haben, halten „alle“ Primärdaten für so relevant, dass man diese nicht nach 10 Jahren löschen oder die LZA-Bemühungen einstellen darf.

20% der Institutionen glauben, dass keine ihrer Primärdaten länger als 10 Jahre relevant sind, allen diesen Institutionen ist gemein, im medizinischen oder soziologischen Umfeld zu forschen.

33% der Institutionen halten nur eine Auswahl der Primärdaten für länger als 10 Jahre relevant. Erstaunlich ist, dass alle diese Institutionen keine Auswahlkriterien festgelegt haben (Frage 4.3).

13% der Institutionen haben diese Frage nicht beantwortet.

Frage 4.2 verschärft die Frage 4.1, indem sie nach dem Horizont von 30 Jahren fragt.

20% der Institutionen halten „alle“ Primärdaten für so relevant, dass diese auch nach 30 Jahren noch weiter in einem LZA-System vorgehalten werden sollten. Unter diesen Institutionen finden sich alle World-Data-Center[5], die den Fragebogen beantwortet haben.

27% der Institutionen glauben, dass keine ihrer Primärdaten länger als 30 Jahre relevant sind.

33% der Institutionen halten nur eine Auswahl der Primärdaten für länger als 30 Jahre relevant. Als Kriterien werden genannt:

- Daten resultierten in zitierte Publikationen,
- Daten könnten von allgemeinem kulturellem Interesse sein oder werden,
- Astronomische Daten, da diese auch in der Vergangenheit bereits forschungsrelevant nach mehreren Jahrzehnten genutzt wurden.

20% der Institutionen haben diese Frage nicht beantwortet.

Frage 4.3 fragt nach einer Liste von Auswahlkriterien, welche Daten wie lange verfügbar gehalten werden sollen.

80% der Institutionen haben keine derartige Liste.

Keine Institution hat angegeben, eine derartige Liste entwickelt zu haben oder zu nutzen.

20% der Institutionen haben diese Frage nicht beantwortet.

Frage 4.4 fragt danach, wer die LZA der Primärdaten bezahlen soll.

19% der Institutionen sehen dies als Aufgabe der Daten erzeugenden Einrichtung. Hierbei handelt es sich jeweils um Institutionen, die nur ein kleines Datenvolumen erzeugen.

13% der Institutionen sehen dies als ein staatliche, öffentliche Aufgabe. Teilweise wird gefordert, dies als Projektkosten bereits bei der Bewilligung aufzuführen.

Eine ausländische Institution formuliert hierzu prägnant: „Wir nutzen Steuergelder zur Erzeugung der Daten, gibt es da einen Unterschied, ob wir diese Steuermittel selbst

5 http://www.ngdc.noaa.gov/wdc/

für die Archivierung verwenden oder es der Regierung überlassen, damit eine dritte öffentliche Institution zu beauftragen?“ (ins Deutsche übersetzt von Th. Severiens)

Die Mehrzahl der Institutionen (68%) beantwortet diese Frage nicht.

Frage 4.5 fragt danach, wer für die Nutzung der Primärdaten in einer LZA-Infrastruktur bezahlen soll.

47% der Institutionen halten es für angemessen, dass kommerzielle Nutzer die tatsächlichen Kosten zahlen, während wissenschaftliche, universitäre Nutzer einen symbolischen (22 Prozentpunkte) oder gar keinen (25 Prozentpunkte) Beitrag zahlen.

19% der Institutionen halten es für wichtig, dass die Daten kostenfrei nutzbar sind. Lassen allerdings offen, wer die Kosten tragen soll.

35% der Institutionen beantworten diese Frage nicht.

Frage 4.6 fragt danach, wer auf die Daten in einer LZA-Infrastruktur nach 30 Jahren zugreifen darf.

68% der Institutionen halten die vollkommene Freigabe der Primärdaten für das Ziel der LZA und favorisieren diese.

Lediglich eine Institution hält ein geschlossenes Archiv auch nach 30 Jahren für geboten mit Verweis auf den Datenschutz (medizinisch pharmazeutischer Forschungsschwerpunkt).

Die verbleibenden Institutionen beantworten diese Frage nicht.

Nutzung von Primärdaten

Die **Frage 5** erfragt, ob in der Institution Erfahrungen mit der Nutzung von Primärdaten aus anderen Institutionen bestehen (also nicht notwendig alten Daten).

60% der Institutionen berichten, dass regelmäßig Primärdaten mit anderen Institutionen ausgetauscht werden.

27% der Institutionen berichten, dass die Nutzung externer Primärdaten nicht bekannt sei.

13% der Institutionen beantworten diese Frage nicht.

Erfahrungen mit alten Primärdaten

Hier sollten die Erfahrungen mit alten Primärdaten (nicht notwendig digitaler Natur) und die Erwartungen an eine LZA-Infrastruktur erfragt werden.

Frage 6 fragt, ob in der Institution bereits Erfahrungen mit der Nutzung alter Primärdaten bestehen und um welche Daten es sich handelt.

54% der Institutionen berichten von Erfahrungen mit der Nutzung alter Primärdaten. Alle beziehen sich auf Daten, die älter als 10 Jahre sind. Sie berichten von seltenen aber regelmäßigen Nutzungen. Als Grund für die Nutzung wird aufgeführt:

- Vergleich mit aktuellen Daten

- Neue Auswertung der alten Daten

34% der Institutionen berichten, keine Erfahrungen mit der Nutzung alter Primärdaten zu haben.

13% der Institutionen haben diese Frage nicht beantwortet.

Frage 6.1 fragt, ob die Institution von der Möglichkeit dieser Nutzung von Primärdaten profitiert hat.

75% der Institutionen, die alte Primärdaten genutzt haben, berichten, davon profitiert zu haben.

25% dieser Institutionen beantworten diese Frage nicht.

88% der Institutionen, die alte Primärdaten genutzt haben berichten in **Frage 6.2**, dass sie bisher nie für die Nutzung alter Primärdaten bezahlen mussten.

12% dieser Institutionen beantworten diese Frage nicht.

Nur Institutionen, die Erfahrungen mit Nutzung alter Primärdaten haben, haben die Frage nach dem Preis, den sie als „fair“ hierfür empfinden würden (**Frage 6.3**) beantwortet.

37% dieser Institutionen halten nur den kostenfreien Zugriff für „fair“.

12% dieser Institutionen halten einen Beitrag zu den Kosten der Archivierung für „fair“.

52% dieser Institutionen beantworten diese Frage nicht.

Frage 6.4 nach der maximal akzeptierten Reaktionszeit einer professionellen LZA-Infrastruktur auf die Anfrage nach Primärdaten wurde entsprechend der vorgegebenen Antwortmöglichkeiten wie folgt beantwortet:

- 7%: sofortiger Online-Zugriff
- 9%: Zugriff binnen 24 Stunden
- 29%: Zugriff binnen 7 Tagen
- 13%: Je älter die Daten sind, desto länger darf es dauern
- 42%: Keine Antwort

Frage 6.5 nach Ideen und Verbesserungen für zukünftige Dienste zum Zugriff auf alte Primärdaten wurde lediglich von einer Institution beantwortet, die darin die weitere Standardisierung der Dateiformate anmahnt.

Guidelines und weitere Kommentare

Frage 7 erfragt, ob die Institution nationale Richtlinien und Hilfestellungen zur LZA begrüßen würde.

54% der Institutionen bejahen diese Frage, wobei einige anmerken, dass es sinnvoller sein könnte, nicht als Nation zu agieren, sondern eher nach Fachdisziplinen.

20% der Institutionen lehnen Richtlinien und Hilfestellungen ab.

26% der Institutionen beantworten diese Frage nicht.

In **Frage 7.1** wird erfragt, zu welchen Aspekten der LZA Richtlinien und Hilfestellungen besonders erwünscht sind.

Genannt werden (in Reihenfolge des Eintrages, teilweise zusammengefasst):

- Datenformate (mehrfach genannt)
- Dokumentationsrichtlinien
- Nutzerschnittstellen
- Interne Datenverwaltung (mehrfach genannt)
- Persistent Identifier
- Rechtliche Beratung (mehrfach genannt)

Frage 8 bietet den Antwortenden die Gelegenheit, zu diversen Punkten, die sie für relevant halten, Kommentare abzugeben.

Hier die Kommentare (teilweise gekürzt), sofern diese für diese Studie relevant sein können.

- Mehrfach wird angemerkt, dass es in verschiedenen Disziplinen abweichende Definitionen von Primärdaten gibt. Insbesondere die Regionalwissenschaften verstehen hierunter vorwiegend Publikationen.
- „Websites should be archived. A lot of data can be found there but people move and with it valuable resources."
- „Metadata should contain also software package and version information related to the primary data, otherwise the data is lost as well."

This page is intended to be blank.

6 Schlussfolgerungen unter Einbeziehung der Interviews und des Workshops

Die Schlussfolgerungen aus dem Fragebogen, den Interviews und dem Workshop, sollen hier in der kompakten Form eines Leitfadens für die LZA wissenschaftlicher Primärdaten dargelegt werden, um im Rahmen der Entwicklung einer LZA-Infrastruktur schnell zur Hand zu sein.

Welche Primärdaten sollten archiviert werden?

Die fachliche Vielfalt der Institutionen (und Personen), die den Fragebogen beantwortet haben spiegelt wieder, dass Primärdaten in vielen Formaten und Genres relevant sein werden.

Die Antworten auf Frage 4.1 und 4.2 zeigen, dass die Selbsteinschätzung der (öffentlich finanzierten) Großforschungseinrichtungen ist, möglichst alle Daten so lange wie technisch möglich zu erhalten, weil diese eben zukünftig von derzeit unbekanntem Wert sein können. AWI: „Wenn man heute entscheiden würde, bestimmte Daten nicht mehr weiter zu pflegen, weiß man ja nie, ob diese in 10 Jahren vielleicht dringend gebraucht würden." IN2P3: „Wir schreiben alle Daten von CERN auf Tapes, denn niemand weiß, welche Informationen darin verborgen sind und vielleicht erst in 30 Jahren gefunden werden, während das aktuelle Experiment längst abgebaut wurde."

Dies entspricht auch dem von der DFG in ihren „Grundsätzen zur Sicherung guter wissenschaftlicher Praxis"[6] empfohlenen Verhalten. Darin wird empfohlen, dass alle Daten (inkl. ihres Kontexts), die Grundlage einer Veröffentlichung waren, für zehn Jahre auf sicheren Speichermedien archiviert werden sollen. Wobei zu betonen ist, dass hiermit eben eine Archivierungsaufgabe verknüpft ist, die einen möglichen Ansatzpunkt für die ausstehende Langzeitarchivierungs-Aufgabe sein kann.

Leider zeigt Frage 2.6, dass die Mehrzahl der Primärdatenproduzenten noch nicht die Relevanz der eigenen Daten in der Zukunft, nach Ablauf des Zehn-Jahres-Horizontes sieht, bzw. sich selbst bezüglich deren Erhaltung nicht als zuständig empfindet. Hier sehen die Autoren dieser Studie dringenden Handlungsbedarf. Auch im Sinne des offenen Zugangs zu Daten muss in den wissenschaftlichen Communities noch einiges an Überzeugungsarbeit über den Wert von Daten und Zugang zu selbigen geleistet werden.

Als kritisch wird bewertet, dass es bisher keinerlei Kriterien gibt welche Primärdaten mit welchem Aufwand erhalten werden sollen. Alle Interviewpartner äußerten sich zu diesem Problem in ähnlicher Weise: „Derzeit wächst der für das gleiche Geld erhältliche Speicherplatz schneller als die Menge an Daten, warum sollten wir also eine Auswahl betreiben?" Aber einfach alles zu speichern, erhöht die Anforderungen an die Strukturierung mittels Metadaten, um die relevanten und gesuchten Daten zu finden. In der freien Wirtschaft ist es derzeit üblich, dass jedes Projekt für seine eigenen Daten entscheidet, welche wie lange verfügbar sein sollen. Dabei wird

6 http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

bewusst die Auseinandersetzung mit dem Widerspruch zwischen der Nachhaltigkeit und der Machbarkeit vielfach wiederholt eingegangen. Generelle, übergreifende Kriterien könnten die Effizienz hier deutlich steigern und somit die Wettbewerbsfähigkeit verbessern.

Welches Datenvolumen ist zu erwarten?

Das erwartete Volumen wissenschaftlicher Primärdaten in bundesdeutschen LZA-Infrastrukturen hängt wesentlich davon ab, wieweit hier auch die Primärdaten internationaler Großprojekte aufgenommen werden sollen. CERN produziert je nach Ablauf und Art der Experimente mehrere hundert TByte Daten jeden Monat (Quelle IN2P3). Die derzeit laufend erfassten Geobasisdaten sprengen leicht jedes Archivierungskonzept: „Allein der Datenzuwachs im Bereich von Orthofotos (Luftbildern) der staatlichen Vermessungsämter ist größer als der zu erwartende Zuwachs an Rechenkapazität um die Bilder für eine Archivierung vorzubereiten.“ (Zitat J. Klump, GFZ Potsdam)

Die konservative Schätzung der in dieser Studie befragten Stichprobe zeigt eine untere Schranke von 1.000 TByte pro Jahr (Frage 2.3) der produzierten Primärdaten. (202 Produzenten a 100 GByte plus 101 Produzenten a 652 GByte plus 34 Institutionen a 27.000 GByte ≈ 1.004 TByte). Diese Abschätzung ist jedoch nur als absolut unterste Schranke zu verstehen, da wie oben geschrieben, einige Datenproduzenten nicht eingeflossen sind oder das Wachstum des Datenaufkommens so stark variiert, dass eine realistische Abschätzung wissenschaftlich nicht möglich ist, lediglich eine Prognose der unteren Schranke.

Hierbei sollte beachtet werden, dass die Menge produzierter Primärdaten mit jedem Jahr zunimmt. Beispielhaft sei die Errichtung laufend neuer Wettermessstationen genannt.

Welche Datenträger und Dateiformate werden bisher zur Archivierung von Primärdaten verwendet?

Die Mehrzahl der Primärdaten wird entweder direkt auf Tapes (derzeit DLT-Tapes) geschrieben oder zunächst in Datenbanken gespielt und hier zur Nutzung vorgehalten (Frage 2.2). Opto-chemische Datenträger spielen derzeit eine untergeordnete Rolle bei der Speicherung wissenschaftlicher Primärdaten.

Insbesondere die Welt-Daten-Zentren halten die Primärdaten zur Nutzung in Datenbanken bereit und stellen deren Backup auf Tapes sicher.

Die Daten selbst werden fast ausschließlich binär kodiert gespeichert (Frage 2.4). XML ist hier (mit 0,3% des Datenvolumens) nicht relevant. Die Primärdaten enthalten überwiegend (91%) (Frage 2.5) eine Selbstbeschreibung.

Inwieweit werden Forschungsdaten (Primärdaten) ausgetauscht?

Frage 5 zeigt, dass 60% der Institutionen aus der Stichprobe regelmäßig Primärdaten mit anderen Institutionen austauschen. Die Interviews zeigten, dass insbesondere die

Großforschungseinrichtungen aktiv im Rahmen von Verbundprojekten Primärdaten austauschen.

Das Interview mit Prof. Zielesny zeigte jedoch auch, dass in der chemischen Industrie nicht unbedingt der Wunsch nach einem Austausch von Primärdaten besteht. Diese enthalten Informationen, die die Prozesse zur Produktion von Substanzen beschreiben, die die Wirkung von Medikamenten belegen oder eben auch nicht belegen. Viele dieser Informationen sind relevant zur Durchsetzung von Patenten.

Ähnlich äußerte sich auch die DASA auf eine Interview-Anfrage. Hier sieht man Techniken und Inhalte der Archivierung von Prozessdaten als Wettbewerbsfaktor und war nicht bereit, über diesen Aspekt der Produktion von Luftfahrzeugen zu reden. Rechtlich ist hier die Dokumentation jedes Arbeitsschrittes vorgeschrieben, ebenso wie die Archivierung über mindestens 30 Jahre. Ein Austausch der Daten erfolgt nur mit der zuständigen Aufsichtsbehörde.

Zusammenfassend kann festgestellt werden, dass derzeit noch nicht der maximal mögliche und notwendige Mehrwert des Datasharings erschlossen ist. Langfristig wird dieses Konzept nur dann etablierbar sein, wenn es der Reputation der Wissenschaftler dienlich ist.

Gibt es Ansätze zur Archivierung von Primärdaten?

Diese Frage lässt sich eindeutig bejahen. Mit dem 1958 im Rahmen des internationalen geophysikalischen Jahres gegründeten und unter der Rigide der ICSU[7] stehenden Verbundes der Welt-Daten-Zentren gibt es ein Netzwerk von Institutionen, die seit 47 Jahren Daten elektronisch verfügbar halten.

Dies ist, so berichten die deutschen WDCs übereinstimmend, nur möglich, wenn die Daten „lebendig“ bleiben. Sie werden in Datenbanken vorgehalten und regelmäßig auf aktuelle Systeme übertragen. Die Daten werden in Formaten vorgehalten, die sie fest „mit ihren Metadaten verdrahten“ (Zitat AWI). Aktuell befindet sich das System der WDCs „im Umbruch und neue Policies werden [...] formuliert. Ein Kerngedanke ist dabei, den Status eines WDC an die Existenz eines Online-Katalogs des WDC-Bestände zu koppeln.“ (Zitat J. Klump, GFZ Potsdam)

CERN betreibt ein System von Backup-Rechenzentren, die gleichzeitig für die Erstellung von Archiv-Tapes zuständig sind. „Die Dokumentation der Daten, die wir auf die Tapes schreiben, erfolgt entweder innerhalb der Daten selbst oder meist als Publikation des Experimentes vorweg.“ (Zitat IN2P3)

Wie oft werden Primärdaten aus bisherigen Archiven „reaktiviert“?

Hierzu gibt der Fragebogen keine wirkliche Auskunft. Keine der befragten Institutionen konnte hierzu gesicherte Zahlen vorlegen, so dass hier nur ein Schätzwert geliefert werden kann. Nach Schätzungen der für die Archive Zuständigen in den kontaktierten Institutionen liegt derzeit die Quote jener digitalen Daten, die älter als 10 Jahre sind und noch einmal oder mehrmals genutzt werden, bei klar unter

7 International Council of Science – www.icsu.org

1%. Hieraus folgt aber keine Aussage über den Stellenwert dieser Daten. Frage 6 und 6.1 belegen vielmehr, dass diese Daten mit hoher Wahrscheinlichkeit an eine wissenschaftlich erfolgreiche Publikation angekoppelt sind.

„Eine Verbesserung der Nachnutzung könnte erreicht werden, wenn Datenveröffentlichungen selbst den Rang zitierfähiger Publikationen bekämen." (Zitat J. Klump, GFZ Potsdam) Das DFG-Projekt „Publication and citation of Scientific Primary Data"[8] beschreitet diesen Weg. Durchgeführt von der TIB Hannover in Kooperation mit den Weltdatenzentren „Climate" in Hamburg, „Mare" in Bremen und dem GFZ in Potsdam, werden hier Persistent Identifier (DOI und URN) für Primärdaten vergeben. Diese Daten sind damit in bibliothekarischen Nachweisdiensten zusammen mit der Literatur recherchierbar (hier: TIBORDER). Die Aufgaben der Archivierung und Qualtiätssicherung verbleiben bei den erzeugenden Institutionen (WDC). Diese erfolgversprechenden Ansätze lassen international hoffen und sollten weiterhin beobachtet und vorangetrieben werden. Der Anteil der bisher hier nachgewiesenen Daten mit insgesamt etwa 250.000 Datensätzen (Ende 2005) ist jedoch nur prototypisch, so dass die Frage einer technischen und organisatorischen Skalierbarkeit hier noch zu klären bleibt.

Notwendige und erwartete Verfügbarkeiten und Zugangsmechanismen

Die Antworten auf die Fragen 3.2, 4.6 und 6.4 zeigen, dass die Datenlieferanten erwarten, dass eine LZA-Infrastruktur den Anforderungen des Datenschutzes genügt. Die Datenlieferanten wollen beispielsweise in der chemischen Industrie die Kontrolle über den Zugriff auf ihre Daten behalten, obwohl sie umgekehrt den Service einer LZA-Infrastruktur gerne nutzen wollen, so das Interview mit Prof. Zielesny.

Dass dennoch eine LZA-Infrastruktur nach dem Willen der Datenlieferanten nicht ein „closed cage" werden soll, zeigen die Antworten auf Frage 4.6, dass die Mehrzahl der Institutionen eine vollkommene Freigabe der Informationen nach Ablauf einer gewissen Schutzfrist unter Wahrung eventueller Persönlichkeitsrechte begrüßt. Auf Nachfrage im Rahmen der Interviews wurde hierzu meist vorgeschlagen, eine Schutzfrist von nur 6 Monaten einzuführen, die aber von den Institutionen beliebig verändert werden könne, bis eben zum vom UrhG maximal garantierten Zeitpunkt. Die Mehrzahl der Primärdaten steht ohnehin für wissenschaftliche Zwecke bereits heute offen auf Anfrage zur Verfügung (Frage 2.7). „Wer uns fragt und einen Grund für den Zugriff hat, bekommt natürlich alle Unterstützung von uns" (GSI).

Die Antworten auf Frage 6.4 zeigen klar, dass die Datennutzer bereit sind, auf den Zugriff um so länger zu warten, je älter die Daten sind. Online-Access wird für Daten, die älter als 10 Jahre sind, nur noch selten erwartet.

8 http://www.std-doi.de

Enthalten Primärdaten heute schon eine Selbstdokumentation? Ist diese notwendig?

Primärdaten, die nicht detailliert beschreiben, wo sie wann mit welcher Motivation mit welchen Geräten wie und von wem erfasst wurden, sind wissenschaftlich kaum nutzbar. „Daten ohne Metadaten kann man gleich löschen." (AWI).

Problemfelder

Die Auswertung zeigt eine Reihe offener Problemfelder, die bisher großteils nur pragmatisch umgangen werden oder als zukünftige Arbeitsfelder definiert werden. Hierzu zählen:

Kriterien für die Auswahl zukünftig relevanter Daten: Derzeit werden in den meisten Institutionen alle Primärdaten so lange gespeichert, bis diese irgendwann schleichend verloren gehen. Dies resultiert insbesondere daraus, dass die Langzeitarchivierung bisher noch fast immer mit der Archivierung gleich gesetzt und mit eigenen Personalkapazitäten betrieben wird. Einen Ansatz, trotz der begrenzten Ressourcen, Daten langfristig zu erhalten und zu selektieren, bietet das pragmatische Vorgehen der WDC.

Uneinheitliche Datenstruktur: Die verwendeten Dateiformate sind bisher weitestgehend projekt-spezifisch, teilweise zusammenfassend innerhalb von Fachdisziplinen einheitlich definiert. Die Mehrzahl der Daten ist nur mittels spezifischer Software intellektuell erschließbar. Eine Möglichkeit zur Konvertierung in XML ist bisher nur für wenige Datenformate implementiert, aufgrund der durchgehenden Dokumentation der Datensätze jedoch meistens möglich (Ausnahmen wurden weder in den Fragebögen noch in den Interviews offensichtlich). Jedoch wurde diese Konvertierung bisher als nicht notwendig und sinnvoll angesehen. Ansätze zum Übergang zu LZA-kompatibleren Datenstrukturen liefern hier eventuell Vorhaben, im IN2P3, die Anzahl der Primärdatenstrukturen zu verringern.

Datensicherheit: Als Haupthindernis für das Outsourcing bzw. Offshoring der LZA-Aufgabe wurden in den Fragebögen und Interviews immer wieder fehlende Regelungen des Datenschutzes bzw. ein mangelndes Vertrauen in die Sicherheit der eigenen Daten vor Fremdzugriff genannt. Lösungen hierzu gibt es ansatzweise in der Industrie, die bereits seit langem Teile der Archivierung beauftragt, jedoch wird auch hier bisher großer Wert darauf gelegt, dass die beauftragten Firmen keinesfalls Aufträge der Konkurrenz annehmen. Inwiefern ein staatliches Archiv hier eine Vertrauensstellung schaffen kann und den Zugriff sicher authentifiziert, bleibt ein relevantes Kriterium für den Erfolg der Bemühungen zum Aufbau einer LZA-Infrastruktur.

This page is intended to be blank.

7 Kritische Betrachtung der gewählten Methode

Die gewählte Methode zur Erstellung dieser Studie soll abschließend kritisch beleuchtet werden, um Rückschlüsse für zukünftige Studien zu bieten und dem Leser eine Abschätzung der Relevanz der Ergebnisse zu ermöglichen.

Zur Erfassung der Daten wurde die Methode einer Einladung zur Teilnahme an einer Online-Umfrage gewählt. Hierbei vertrauten die Autoren auf den inzwischen verbreiteten guten Namen des Projektes *nestor* und die Akzeptanz, die sich aus der Nennung des Mittelgebers, des BMBF ergibt. Der Fragebogen wurde in englischer Sprache verfasst, um internationale Experten und Institutionen einzubinden. Auf die Erstellung einer deutschen Version wurde dabei bewusst verzichtet, da die Autoren voraussetzten, dass in wissenschaftlichen Institutionen arbeitende Personen, die mit dem wichtigen Thema der LZA beauftragt sind, Englisch nicht als Hemmnis oder Hürde empfinden.

Die Resultate des Fragebogens wurden in einem zweiten Schritt im Rahmen eines Workshops diskutiert und dann mit teilweise erweiterter Fragestellung mittels direkter Interviews möglichst im persönlichen, notfalls im telefonischen Gespräch vertieft.

Die Fragebogen-Aktion verlief aus Sicht der Autoren erfolgreich, es zeigte sich jedoch eine gewisse Abhängigkeit der Rücklauf-Wahrscheinlichkeit von der Größe der angeschriebenen Institution. Je kleiner die Institution war, desto wahrscheinlicher war es, dass der Fragebogen ausgefüllt wurde. Genauer wurde diese Beobachtung jedoch nicht untersucht. Hieraus resultiert jedoch eine gewisse geringere Repräsentanz der Großforschungseinrichtungen in den Resultaten, die die Autoren durch die bevorzugte Auswahl großer Institutionen bei den Interviews auszugleichen versuchten. Die Zielgruppen der Umfrage sind statistisch bis auf die wissenschaftlichen Museen hinreichend abgedeckt worden.

Von zwei Max-Planck-Instituten gab es telefonische Anfragen nach einer deutschen Version des Fragebogens. Die Autoren haben daraufhin stützend telefonisch die Beantwortung des Fragebogens begleitet.

Als wenig glücklich stellte sich die technische Implementation des Fragebogens, bei einem der nestor-Partner, heraus. Neben einem Crash des Datenbankservers genau zu dem Zeitpunkt, zu dem vermutlich die Mehrzahl der Antworten eingegangen wären (und einige wohl auch gelöscht wurden), beobachteten die Autoren auch im Regelbetrieb das sporadische Löschen ganzer Antwortsätze. Vermutliche Ursache kann der fehlende Passwortschutz der Administrator-Oberfläche sein, so dass die Löschfunktion offen zugänglich war. Der regelmäßige Ausdruck der Antwortsätze taugte als Sicherung dabei nur bedingt. Dies ist ein gutes Beispiel, dass die LZA der Rohdaten bereits die Implementierung geeigneter Verfahren bei der Datenerfassung erfordert und nur gelingen kann, wenn der gesamte Workflow stimmig und nachhaltig angelegt wird.

Der Workshop stand unter dem Schwerpunkt der parallelen Expertise zur Policy der LZA. Hier wurden die Aspekte dieser Studie nur peripher diskutiert, auch da der Kreis der Primärdatenerzeuger und -nutzer dort kaum vertreten war, sondern hier

vorwiegend die sammelnden Institutionen vertreten waren. Hilfreich war der Workshop, um die Erfahrungen und Anforderungen der wissenschaftlichen Museen zu erfassen, die hier mit zwei Vorträgen vertreten waren.

Die Interviews waren eine hilfreiche und notwendige Erweiterung des durch den Fragebogen natürlich begrenzten Themenkreises und boten die Möglichkeit, den Kreis der eingebundenen Institutionen gezielt zu erweitern, gingen jedoch hinsichtlich des notwendigen Aufwandes zur Erfassung und Auswertung der Aussagen deutlich über die Fragebögen hinaus.

Die Einbindung der gewerblichen Wirtschaft als Produzenten und Nutzer von Primärdaten gestaltete sich sehr schwierig, da anders als ursprünglich erwartet, nur in wenigen Branchen Interesse und Notwendigkeit an Langzeitarchivierung digitaler Primärdaten besteht. Diese Branchen (chemisch-pharmazeutische und Luftfahrt-Industrie bspw.) behandeln ihre LZA-Methoden jedoch vertraulich, so dass hier nur indirekte Aussagen einfließen konnten.