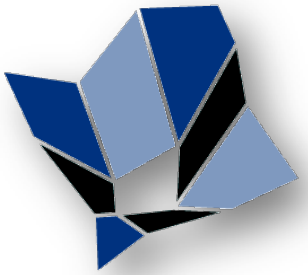

Projekt RADIESCHEN

Rahmenbedingungen einer **di**sziplinübergr**e**ifenden For**sch**ungsdat**en**infrastruktur

Kostenverteilung und Risiken

Entspricht dem Report D4.2 „2. Entwurf LZA-Kosten“ nach Projektantrag

Torsten Rathmann

DFG Deutsche Forschungsgemeinschaft

Inhalt

1. Einleitung

Ziel des Radieschen-Arbeitspaketes 4 (Kosten) ist eine Analyse, welche Kosten bei der Langzeitarchivierung (LZA) und darüber hinaus im gesamten Lebenszyklus von digitalen Forschungsdaten auftreten. Aus Deutschland gibt es zu diesem Thema bisher zwar qualitative, aber kaum quantitative Informationen. Eine der wenigen, mit Zahlen versehenen Studien ist der KoLaWiss-Report zu Kosten der elektronischen Langzeitarchivierung (Dickmann, 2009). Im Ausland sind dagegen zahlreiche Studien erschienen. Einige wichtige Ergebnisse aus diesen Studien sollen an dieser Stelle genannt werden.

Die Kosten für Forschungsdatenarchive sind in den meisten Disziplinen klein verglichen mit denen für die Forschung selbst, d.h. den Kosten für die Produktion/Erhebung der Daten. Die OSI Preservation and Curation Group fand, dass die Kosten für Datenzentren und –dienste im Vereinigten Königreich 1,4-1,5 % der Gesamtkosten einschließlich der Forschungskosten ausmachen (Beagrie, E-Infrastructure strategy for research: final report from the OSI Preservation and Curation Working Group, 2007).

Auf der anderen Seite sind Forschungsdatenarchive deutlich teurer als Archive, die nur Publikationen enthalten. Die Kosten für die zentralen Daten-Repositorien von Cambridge und des King's College London sind um eine Größenordnung höher als die eines typischen Institutions-Repositories, das sich allein auf elektronische Publikationen beschränkt (Beagrie, Chruszcz, & Lavoie, Keeping Research Data Safe, A Cost Model and Guidance for UK Universities, 2008).

Forschungskosten sind jedoch keine Konstante, sondern disziplinabhängig und einer zeitlichen Entwicklung unterworfen. Beispielsweise sind die Kosten für DNA-Sequenzierungen dank neuer Technologien und weitgehender Automatisierung stark gefallen.

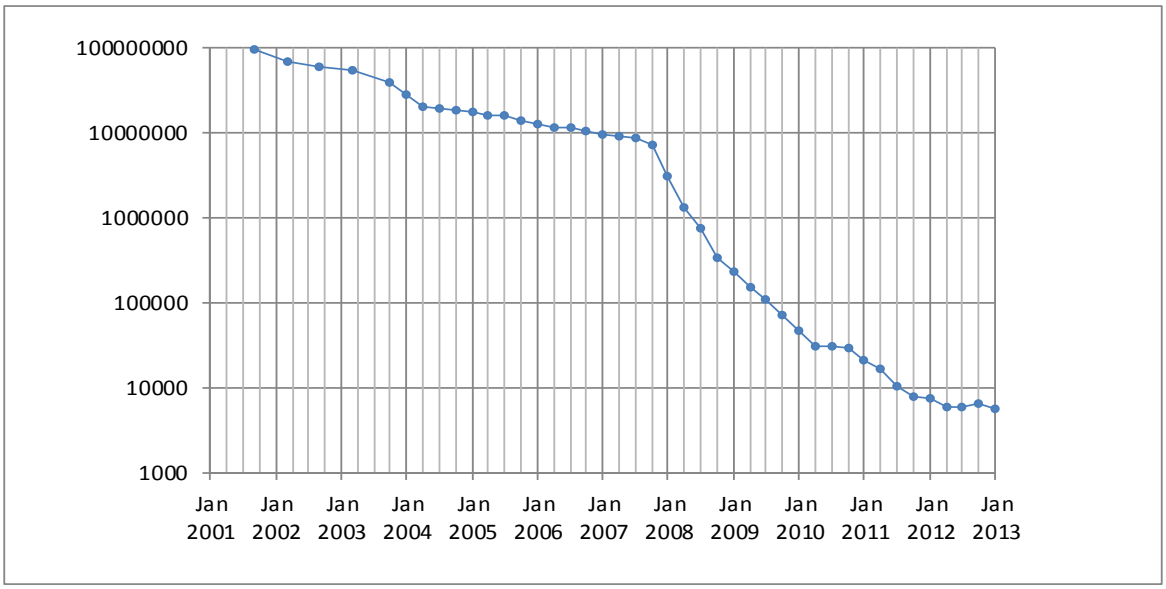

Abbildung 1: Sequenzierungskosten in US-$ für ein menschliches Genom. Daten von (DNA Sequencing Costs, 2013). Die Einführung des Next Generation Sequencing hat 2007/08 einen überexponentiellen Kostenabfall bewirkt.

Wenn der starke Kostenabfall für DNA-Sequenzierungen noch längere Zeit anhält, könnten in diesem Bereich sogar die Archivierungskosten unterschritten werden.

Aber auch Archivierungskosten unterliegen einem Wandel. Personalkosten wachsen leicht an; Hardware und Medien werden — bezogen auf die Leistung — dagegen auf lange Sicht immer günstiger. Die Speicherdichte von Festplatten ist über Jahrzehnte hinweg exponentiell gewachsen. Diese Faustregel wird als Krydersches Gesetz bezeichnet. Da sich der Preis für eine Festplatte nicht wesentlich verändert hat, fallen die Preise pro Datenvolumen entsprechend schnell. Inzwischen sind aber doch Abweichungen vom gleichbleibend schnellen, exponentiellen Wachstum und der Übergang zu einem weniger rasanten Wachstum der Festplatten-Speicherdichte absehbar (Rosenthal, 2013).

Auch die Rechenleistung der Server nimmt ständig weiter zu. Eine dem Kryderschen Gesetz entsprechende Faustregel, das Mooresche Gesetz, sagt aus, dass die Zahl der Transistoren auf einem kostenoptimal hergestellten Chip über Jahrzehnte exponentiell gewachsen ist. Zwar steigt die Rechenleistung nicht ganz linear mit der Zahl der Transistoren, die Leistungszunahme ist aber erheblich.

Die Zahl der unterstützten Formate wirkt sich erheblich auf die Archivierungskosten aus. Weniger häufig verwendete Formate besitzen ein ungünstiges Kosten-Nutzen-Verhältnis. LIFE, ein über ein Jahr laufendes Projekt, das von der British Library und dem University College London 2005/06 zwecks Entwicklung eines LZA-Kostenmanagements durchgeführt worden ist, hat u.a. diese Frage untersucht. In LIFE liegen 85 % der Dokumente in den Formaten PDF, TXT und HTML vor, haben aber nur 7 % der Kosten ausgemacht. Die zwölf am seltensten verwendeten Formate haben nur 0,1 % der Daten umfasst, aber 41 % der Kosten verursacht (Björk, 2007).

Metadaten sollten möglichst gleich bei Erzeugung der Daten geschrieben werden. Einer Schätzung zufolge (Costs of Digital Preservation, 2005) liegen die Kosten für brauchbare Metadaten-Beschreibungen nach zehn Jahren um den Faktor 30 höher.

Eine wesentliche Informationsquelle dieser Studie sind die im Rahmen des Projektes Radieschen durchgeführten Interviews. Da im Interview nicht die Höhe aller Einzelkosten erfragt werden kann, wurde keine Kostenrechnung versucht und kein Kostenmodell aufgestellt. Stattdessen wurden andere Schwerpunkte gesetzt. In verschiedenen Fallstudien wurde gefunden, dass die Personalkosten den größten Anteil der Gesamtkosten der Archivierung ausmachen (Ashley, 1999) (Beagrie, Chruszcz, & Lavoie, Keeping Research Data Safe, A Cost Model and Guidance for UK Universities, 2008) (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010). Die Personalkosten stehen nicht nur im Fokus dieser Studie, weil sie der größte Kostenblock sind, sondern auch weil sich die Zahl der Stellen in einem Interview leichter erfragen lässt als Eurobeträge. Genaue Eurobeträge haben wir in den Interviews erwartungsgemäß nicht erfahren, dafür sind über ein Interview aber Informationen zugänglich, die sich nicht ohne weiteres aus der Buchführung erschließen lassen, z.B. welche Aufgaben das Personal hat und welche Dienste angeboten werden.

Deshalb sind nach der Beschreibung einiger Details der Interviews im nächsten Kapitel Personalaufwand und Datendienste gleich das Thema des übernächsten Kapitels. Andere fachübergreifende Kosten werden in den Kapiteln 4 (Hardwarekosten) und 5 (weitere Kosten) untersucht. Neben den fachübergreifenden gibt es selbstverständlich auch fach- und archivspezifische Faktoren. Auf die wird in Kapitel 6 eingegangen. Eine Risikoanalyse ist Gegenstand von Kapitel 7. Untersucht werden die wirtschaftlichen Risiken der Archivierung aber auch der Nicht-Archivierung. Unter anderem wird anhand von Fallbeispielen untersucht, wie hoch die Folgekosten sind, wenn auf eine Archivierung ganz verzichtet wird.

2. Methode

Im Projekt Radieschen wurden Mitarbeiter von Institutionen und Projekten interviewt, die Forschungsdaten halten oder in Forschungsdatenprojekten arbeiten. Die befragten Institutionen und Projekte kommen aus den Natur-, Bio-, Geistes-, Sozial-, Wirtschafts- und Ingenieurwissenschaften und decken somit ein breites Wissenschaftsspektrum ab. Die Befragten sind in leitender Position an der betreffenden Institution tätig oder direkt in Aufgaben rund um die Haltung von Forschungsdaten involviert oder entwickeln Software hierfür. Die Interviews waren in der Regel auf zwei Stunden angesetzte Präsenz-interviews bei der betreffenden Institution. Zusätzliche Kurzinterviews wurden telefonisch durchgeführt.

Es wurden Fragen zu den Themen Technik, Organisation und Kosten gestellt. Die Fragen zum Thema Kosten sollten darüber Aufschluss geben,

- wie viel Personal vorhanden ist und wie die Arbeitskraft auf Schritte im Daten-lebenszyklus von der Produktion bis zur Bereitstellung verteilt ist
- wie hoch die Hardware-Beschaffungskosten sind
- welche sonstigen Kosten getragen werden müssen (nur qualitativ)
- welche Arbeitsschritte beim Ingest (Aufnahme ins Archiv) typischerweise erfolgen
- wie viele Datensätze üblicherweise in einem Ingest-Vorgang gemeinsam behandelt werden
- welche Datendienste angeboten werden, z.B. wie die Daten bereitgestellt werden

In den Interviews wurden mit Ausnahme der Hardware-Kosten keine Euro-Beträge abgefragt. Wiederholtes Nachschlagen oder Nachrechnen hätte die Gesprächsatmosphäre sicherlich gestört.

Personalkosten werden ausschließlich als Anzahl Vollzeitstellen erfragt. Der Verdienst der Mitarbeiter war nicht Gegenstand dieses Projekts. Außerdem war es für die Interviewten sicherlich einfacher, die Fragen in dieser Form zu beantworten als einen Euro-Betrag zu errechnen.

Von den im Projekt Radieschen befragten Projekten werden in diesem Dokument nur diejenigen betrachtet, die bereits mit Daten arbeiten, denn nur diese konnten Angaben zu Kosten in laufenden Betrieb machen. Projekte in der Planungs- oder Aufbauphase und reine Software-Projekte sind dementsprechend fortgelassen. Insofern unterscheidet sich die Datenbasis dieser Studie von der der anderen Radieschen-Arbeitspakete.

Neben den Interviews wurden die verfügbare Literatur und Erfahrungen am Deutschen Klimarechenzentrum (DKRZ) und Max-Planck-Institut für Psycholinguistik in die Analyse mit einbezogen.

3. Personalaufwand

Ein Schwerpunkt dieses Abschnitts ist die Frage, wie sich die Personalkosten auf die Schritte Auswahl, Ingest, Speicherung, Kuration (Datenpflege) und Bereitstellung (Access) verteilen. In Tabelle 1 sind die bisherigen Ergebnisse zum Personalaufwand zusammengetragen. Selbstverständlich handelt es sich bei den im Interview gemachten Angaben zur Verteilung des Personalaufwandes um Schätzungen.

	Auswahl+Ingest	Speicherung+ Kuration	Bereitstellung [1]
GESIS	10-12 Stellen[2]	2 Stellen + Kern-IT[3]	9-10 Stellen[4]
Transregio TRR 54[5]	~ 2 Stellen		
SILVA	2 Stellen[6]		1 Stelle
megx.net	¼ Stelle[7]	1 Stelle[8]	2 Stellen
PANGAEA	~ 14 Stellen[9]		
WDCC[10]	6 Stellen	4 Stellen	3 Stellen
DoBeS[11]	~ 7 Stellen		
ADS[12]	55 %	15 %	31 %
UK Data Archive[13]	59 %	6 %	35 %
NARA[14]	50 %	33 %	17 %

Tabelle 1: Verteilung des Personalaufwands auf Schritte im Datenlebenszyklus

[1] mit Entwicklung, da Anpassungen von Portalen immer mit Entwicklung verbunden sind

[2] Datenauswahl 1 ständig + Chef + zeitweise andere = 2-4 Stellen; Ingest 8 Stellen, davon 4-5 Verschlagwortung. Für die Berechnung des Prozentwertes nach Formel (1) wurde $z_{11}=11$ gesetzt.

[3] Wie hoch der Personalaufwand im Kern-IT-Bereich ist, lässt sich kaum beziffern, da hier die unterschiedlichsten Aufgaben des IT-Betriebs für das ganze Haus wahrgenommen werden. Für die Berechnung des Prozentwertes nach (1) wurde $z_{12}=2$ gesetzt, da das Datenvolumen eher klein ist.

[4] 3-4 + 6 für Softwarepflege, Beratung und Kumulation. Für die Berechnung des Prozentwertes nach (1) wurde $z_{13}=9{,}5$ gesetzt.

[5] Wachstum und Überleben, Plastizität und zelluläre Interaktivität lymphatischer Neoplasien

[6] Auswahl, Ingest und Speicherung 1 Stelle, Kuration 1 Stelle. Für die Berechnung des Prozentwertes nach (1) wurden die 2 Stellen aufgeteilt in ½ für Auswahl + Ingest und 1½ für Speicherung + Kuration.

[7] nicht ausreichend, unterfinanziert

[8] Vom Umfang der Arbeit her ist eine Stelle mindestens erforderlich. Die Aufgabe der Speicherung wird von der IT-Abteilung wahrgenommen, die viele Aufgaben für das ganze Haus übernimmt. Der tatsächliche Arbeitsaufwand ist schwer zu ermitteln, da die Abgrenzung schwierig ist.

[9] Kuratoren, die keine PANGAEA-Mitarbeiter sind, sind mitgezählt.

[10] World Data Center for Climate, betrieben vom Deutschen Klimarechenzentrum (DKRZ)

[11] Dokumentation Bedrohter Sprachen: eine Stelle für Pflege der Repository-Software, 2 Stellen für System- und Archivmanagement, 2 Stellen für Anwendungen und Hilfsprogramme, die der Bereitstellung dienen (Wittenburg, Váradi, & Tadić, Budapest Meeting of the Alliance for Permanent Access, 2008). Hinzu kommen 2 Stellen für Digitalisierung und Ingest der Digitalisate. Die Auswahl wird durch den Archivmanager und Senior-Personen vorgenommen, die noch viele andere Aufgaben haben. Der Ingest ist weitgehend automatisiert. (Wittenburg, Privatmitteilung, 2012)

[12] Archaeology Data Service, aus KRDS2 (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010)

[13] berechnet aus den in KRDS2 (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010) tabellierten prozentualen Arbeitszeiten, wobei die in KRDS2 zusätzlich angegebenen Zeiten für Forschung, Entwicklung, Datenmanagement, Verwaltung und gemeinsame Dienste der Vergleichbarkeit wegen hier unberücksichtigt bleiben

[14] US National Archives and Record Administration (NARA). Prozentwerte wurden berechnet aus Zahlen des NARA Performance Report 2002, zitiert in (Costs of Digital Preservation, 2005), wobei die dort zusätzlich angegebenen Zahlen für Datenmanagement und Buchführung der Vergleichbarkeit wegen hier unberücksichtigt bleiben.

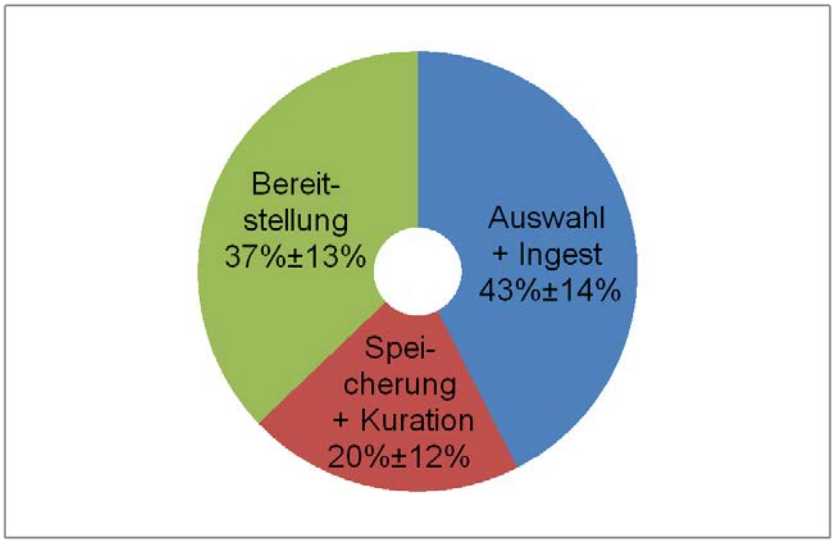

Abbildung 2: Verteilung des Personalaufwandes auf Schritte im Datenlebenszyklus

Abbildung 2 zeigt die Verteilung des Personalaufwandes auf Stationen im Datenlebenszyklus. Datengrundlage sind die Angaben der vier Archive GESIS, SILVA, megx.net und WDCC, die z.B. bei Radieschen-Interviews dazu Informationen liefern konnten, die für eine Auswertung genau genug zu sein scheinen[15]. Die Prozentwerte R_j wurden nach der Formel

$$\boldsymbol{R_j = \frac{\sum_i z_{ij}}{\sum_i \sum_j z_{ij}}} \qquad \textbf{(1)}$$

berechnet, wobei die z_{ij} die Zahl der Vollzeitstellen in Institution i sind, die allein Schritt j zugeordnet sind. Die von den Interviewpartnern genannten Stellenzahlen für einen Schritt, z.B. Bereitstellung, wurden also zunächst über alle befragten Institutionen aufaddiert und dann durch die Gesamtzahl der Stellen dividiert. Institutionen mit vielen Stellen sind dadurch stärker gewichtet als solche mit wenigen.

Primärarchive, d.h. Archive, die direkt Daten vom Produzenten nehmen, haben i.d.R. einen hohen manuellen Aufwand für Kommunikation mit den Datenproduzenten, z.B. um Metadaten zu vervollständigen und in Fragen der Qualitätssicherung. Hinzu kommen eventuell Fragen der Zugriffsrechte und Rechtsfragen im Zusammenhang mit der Übernahme der Verantwortung durch das Archiv. Insgesamt ist der Personalaufwand von Primärarchiven für den Ingest erheblich.

Bei Sekundärarchiven, d.h. Archiven, die ihre Daten nicht direkt vom Datenproduzenten, sondern ausschließlich von anderen Archiven beziehen, ist der manuelle Kommunikationsaufwand wesentlich kleiner. Hier kann der Ingest zu einem weit höheren Anteil automatisiert erfolgen. Bei den von Radieschen befragten sekundären Archiven SILVA und megx.net ist der Ingest nicht wie sonst der aufwändigste, sondern der am wenigsten personalintensive Schritt. An einer direkten Maschine-Maschine-Kommunikation zu anderen Archiven wird in diesen Projekten bereits gearbeitet.

Fast alle von Radieschen befragten Archive prüfen in irgendeiner Weise die Qualität von Daten und Metadaten beim oder vor dem Ingest. In der Regel findet eine teilautomatisierte Qualitätskontrolle statt mit anschließender manueller Bewertung der Ergebnisse durch einen Mitarbeiter. Die meisten Einzelschritte der Qualitätskontrolle sind archiv-

[15] Als maximaler Einzelfehler wurde eine Vollzeitstelle zugelassen. Die in Abbildung 2 angegebenen Fehlergrenzen wurden durch differenzielle Fehlerfortpflanzungsrechnung in der Absolutbetragsnorm (1-Norm) abgeschätzt:

$$|\Delta R_k| \leq \sum_i \sum_j \left|\frac{\partial R_k}{\partial z_{ij}}\right| |\Delta z_{ij}|$$

In der Fehlerfortpflanzungsrechnung wurden ebenfalls maximale Einzelfehler $|\Delta z_{ij}| \leq 1$ angesetzt.

bzw. fachspezifisch. Formatvalidierung und Prüfung auf Vollständigkeit werden ebenso eingesetzt wie statistische Verfahren, Plausibilitätskontrolle oder interne Review-Verfahren.

Auch Speicherung und Kuration variieren von Archiv zu Archiv. Die Daten können kurz- oder langfristig gespeichert werden. Langzeitarchivierung ist nicht das Ziel jedes Archivs. Zur Kuration kann die Vergabe persistenter Identifier (z.B. DOI, Digital Object Identifier) gehören, die Daten zitierfähig werden lässt. Die Anpassung von Metadaten an Veränderungen gehört hier hinein, z.B. die Eintragung von Zitierungen der Daten in neu erschienenen Publikationen. Und selbstverständlich gehört die klassische Kurationsaufgabe dazu, die Erhaltung der Nutzbarkeit, z.B. durch Migration der Daten auf aktuelle Datenformate.

Allen hier betrachteten Institutionen und Projekten ist gemeinsam, dass Daten nicht nur gespeichert, sondern auch für die Nachnutzung bereitgestellt werden. Ebenfalls allen gemeinsam ist die Bereitstellung über ein Web-Portal oder ein Datenwarenhaus, nur manchmal noch ergänzt durch weitere Angebote wie den Versand von Datenträgern. Die Bereitstellung über ein Web-Portal oder ein Datenwarenhaus kann z.B. einen Download beinhalten. Das ermöglicht Archivnutzern den schnellen Zugriff auf die Daten in einer meist ansprechenden Umgebung und entlastet das Archivpersonal weitgehend von lästigen Routinetätigkeiten wie dem Beschreiben und Versenden von Datenträgern. Über ein Web-Portal oder ein Datenwarenhaus können neben den Standardfunktionen Suche und Download auch Zusatzdienste zur Verfügung gestellt werden wie z.B. grafische Visualisierung oder auf die Community zugeschnittenes Postprocessing. Die angebotenen Zusatzdienste sind von Archiv zu Archiv sehr unterschiedlich.

Die Entwicklung und ständige Anpassung solcher Zusatzdienste ist vom reinen Betrieb häufig kaum zu trennen. In Sachen Bereitstellung sind deshalb in Tabelle 1 und Abbildung 2 Betrieb und Entwicklung zusammengefasst. Der Personalaufwand für die Bereitstellung ist erheblich; er liegt in derselben Größenordnung wie der zusammengefasste Aufwand für Auswahl und Ingest.

4. Hardwarekosten

Hardwarekosten sind in Tabelle 2 zusammengefasst. Einige Institutionen haben absolute Kosten angegeben, andere jährliche. Das Projekt Transregio TRR 54 befand sich zum Zeitpunkt des Interviews noch in seiner Anfangsphase mit weniger als 300 GB Daten. Das WDCC umfasst inzwischen 1,5 PB (Stand September 2012), Sicherungskopien nicht eingerechnet.

GESIS	10-20 k€; Erneuerung alle 4 Jahre
SILVA und megx.net, gemeinsam genutzt	50-70 k€/Jahr
Transregio TRR 54	10 k€ für zwei Mittelklasse-Server
WDCC (World Data Center for Climate)	200 k€/Jahr[16]
DoBeS (Dokumentation Bedrohter Sprachen)	80 k€/Jahr (Wittenburg, WG2-9 Cost Estimations, 2010)

Tabelle 2: Hardwarekosten

[16] Die Hardware wird am DKRZ alle 5 Jahre erneuert. Die Kosten dafür betragen 30 Mill. € pro Zyklus. Hardware für die Datenhaltung macht ein Drittel davon aus. Etwa 10 % der Daten am DKRZ entfallen auf das WDCC. Das ergibt 200000 €/Jahr. Hinzu kommen zwischenzeitliche Hardware-Kosten, z.B. für Reparaturen, von 60000 €/Jahr, wobei hiervon wieder 10 % dem WDCC zugerechnet werden.

Hardwarekosten sind naturgemäß vom zu speichernden Datenvolumen abhängig. Nach unten hin gibt es jedoch eine Grenze, eine Mindestinvestitionssumme, wenn die Hardware selbst betrieben wird. Wegen der Gefahr des Datenverlustes, z.B. durch einen Brand, sollten auch kleine Archive Kopien ihrer Daten an zwei Orten speichern, wobei unterschiedliche Gebäude sicherer sind als zwei Räume im selben Gebäude. Von daher umfasst die Mindestausstattung zwei Server. Ein kleines Archiv würde ausschließlich auf Festplatten speichern, weil sich Tape nur bei größeren Datenmengen lohnt. Wenn die Daten auf zwei redundanten Servern gespeichert werden, die mit RAID 0/1/10-Controller und je vier Wechselrahmen-Festplatten á 600 GB ausgestattet sind, liegt der Anschaffungspreis einschließlich unabhängiger Stromversorgung (USV), verlängerter Garantiezeit und Mehrwertsteuer zurzeit dicht unter jenen 10000 €, die in den Interviews der vom Volumen her kleinsten Archive angegeben wurden.

Mit der Zeit gehen Storage-Kosten bezogen auf das Volumen im Zuge der immer höheren Speicherdichten stark zurück. Gleich bleibenden Preis vorausgesetzt verdoppelte sich bis vor wenigen Jahren das Speichervolumen bei SAN-Festplatten etwa alle zwei Jahre, bei Bandkassetten etwa alle drei Jahre[17]. Wegen der schlimmsten Überschwemmung in Thailand seit einem halben Jahrhundert, bei der auch 40 % der weltweiten Festplatten-Produktionskapazitäten zerstört wurden, haben sich die Preise für Festplatten aber Ende 2011 in kurzer Zeit wieder mehr als verdoppelt (Kunert, 2011).

5. Weitere Kosten

Alle interviewten Partner ermöglichen ihren Mitarbeitern den Besuch wissenschaftlicher Fachtagungen und Konferenzen oder bilden ihre Mitarbeiter fort. Manche Archive suchen darüber hinaus den persönlichen Kontakt zu wichtigen Datenproduzenten und anderen Archiven. Bei allen diesen Aktivitäten fallen Reise- oder Fortbildungskosten an.

Gebäudekosten werden bei allen interviewten Institutionen nicht dem Archiv als Kostenstelle zugeordnet, sondern aus dem allgemeinen Haushalt der Institution bestritten. Zu den Gebäudekosten gehören nicht nur Mieten, Instandhaltungs-, Hausmeister- und Heizkosten, sondern auch die mit der Bewirtschaftung des Arbeitsplatzes verbundenen Betriebskosten für Strom, Telefon und Netzzugang.

Alle Archive, die nicht ausschließlich kostenfreie Software einsetzen, haben Kosten für Software-Lizenzen, vor allem für Datenbank- und Office-Produkte. Hinzu kommen fach- und archivspezifisch die verschiedensten Anwendungs- und Dienstprogramme. Im sozialwissenschaftlichen Bereich ist es z.B. vor allem Statistik-Software, für die Lizenzen benötigt werden.

6. Fachspezifische und archivspezifische Faktoren

Zielsetzung, Archivgröße und die angebotenen Datendienste sind von Fach zu Fach und sogar von Archiv zu Archiv unterschiedlich. Auf diese archivspezifische Seite soll im Folgenden eingegangen werden[18]. Auch die in den Tabellen 1 und 2 zusammengestellten Zahlen müssen selbstverständlich vor dem Hintergrund der unterschiedlichen Ziel-

[17] Nach Erfahrung der GDWG, siehe KoLaWiss-Report (Dickmann, 2009, S. 11)
[18] Mit freundlicher Genehmigung der Interviewpartner wurden Details zu Archiven und teilweise auch zu fachspezifischen Faktoren den Radieschen-Interviews entnommen.

setzungen, angebotenen Datendienste und Archivgrößen gesehen werden. Beispielsweise wäre ein unkommentierter Vergleich der Kosten pro Datenvolumen unangemessen.

6.1. Sozialwissenschaftliche Daten

Bei den sozialwissenschaftlichen Daten müssen Einzeldaten und kumulierte Daten (letztere oft in Studien eingebettet) unterschieden werden. Kumulierte Daten sind von hohem öffentlichem Interesse wegen ihrer sozialen, politischen oder wirtschaftlichen Bedeutung. Solche Daten werden keineswegs nur von Fachleuten, sondern von den verschiedensten gesellschaftlichen Gruppen nachgefragt. Entsprechend hoch ist der Auswertungs- und Bereitstellungsaufwand, denn für die jeweiligen Zielgruppen müssen kumulierte Daten in verständlicher und nachnutzbarer Form bereitgestellt werden. Aus den Einzeldaten werden durch Kumulation quantitative Aussagen gewonnen, die einerseits von Interesse für die Zielgruppen sind, andererseits aber keine Rückschlüsse mehr auf einzelne Personen oder Betriebsgeheimnisse mehr ermöglichen.

Einzeldaten dürfen in der Regel nicht weitergegeben werden. Nutzer von Einzeldaten müssen unterschreiben, dass sie das beachten werden.

Die Datenmengen sind meist gering; es gibt nur wenige wirklich große Studien im TB-Bereich. Die Hardware-Kosten halten sich entsprechend in Grenzen.

Bei der oft sehr aufwändigen Qualitätskontrolle kommen häufig statistische Verfahren zum Einsatz. Darüber hinaus wird im Detail überprüft, ob die Daten in die richtigen Felder eingetragen und plausibel sind. Stichproben werden dahingehend überprüft, ob sie repräsentativ sind. Qualitätsmanagement, Bereitstellung für unterschiedliche Zielgruppen und die notwendige Kumulation erfordern viel Personal.

GESIS und ZBW betreiben in Zusammenarbeit mit DataCite auf Grundlage des DOI-Systems die nicht-kommerzielle Registrierungsagentur für Sozial- und Wirtschaftsdaten (da|ra). Viele Daten der befragten Institutionen GESIS und Statistisches Bundesamt gehören nicht zum soziologischen Kernbereich, sondern ins politische oder wirtschaftswissenschaftliche Umfeld.

GESIS hat auch die folgende allgemeine (nicht abschließende Liste) von Kostenkomponenten des Datenmanagements sozialwissenschaftlicher Umfragedaten veröffentlicht (Jensen, 2012):

- Projektmanagement: Verantwortlichkeiten zur Erstellung, Implementierung und Sicherung von Strategien und Verfahren des Forschungsdatenmanagements
- Technische Sachmittel (Software; Hardware) und administrative Maßnahmen zur strukturierten und standardisierten Datenorganisation, -austausch, -speicherung und -sicherheit
- Übersetzung von Fragebögen (Landessprache(n)) und / oder weiteren Studienmaterialien
- Datenaufbereitung (Kontrolle, Bereinigung) und Dokumentation der Datenmodifikationen
- Standardisierte Dokumentation auf Studien- und Variablenebene (Metadaten)
- Nutzung von technischen Dokumentationsstandards (z. B. DDI) und Werkzeugen
- Anforderung an die Datenanonymisierung

- Digitalisierung von Materialien zur Weiterverarbeitung bzw. Vorbereitung der nachhaltigen Sicherung der Studie und der erzeugten Daten (z.B. Messinstrument, Show-Cards)

6.1.1. GESIS

Der Schwerpunkt des Leibniz-Instituts für Sozialwissenschaften GESIS liegt auf bundesweiten und international vergleichenden Daten zu Soziologie und Politik. Die Daten sind für internationale Vergleiche geeignet. GESIS produziert einen Teil der Daten selbst.

Die Daten werden nach Qualitätskriterien selektiert, wobei jedoch immer gegen die vermutete Forschungsrelevanz gewichtet wird. Wenn keine besseren Daten zu einem Forschungsproblem verfügbar sind, die Nachfrage aber hoch ist, wird im Zweifel für die Verfügbarmachung des Datensatzes entschieden. Dabei wird jedoch auf die Dokumentation möglicher methodischer Schwächen der Daten geachtet. Die Daten müssen bevölkerungsrepräsentativ sein. Grundlage sollte eine gute Zufallsstichprobe sein. Weniger gut normierbare Qualitätskriterien gibt es jedoch bei Aggregatdaten (vorwiegend Staaten und Gemeinden).

Der Ingest besteht aus den folgenden Schritten:

- Technische Lesbarkeitsprüfung, Formatprüfung
- In die Datensätze wird hineingeschaut: Ist die Kodierung transparent? Steht in den Datenfeldern das, was dort hineingehört?
- Überprüfung der Metadaten: Sind das diejenigen, die zu den Daten gehören?
- Manuelle Ergänzung der Metadaten:
    - Studienbeschreibung
    - Verschlagwortung
    - Bezugseinheit wie z.B. „national“
    - Primärforscher
    - Verknüpfungen zu anderen Daten
- Ggf. Formatkonvertierung
- Aushandeln einer Vereinbarung für die Übernahme der Verantwortung („Datengebervertrag“; regelt auch, was passiert, wenn es den Geber nicht mehr gibt)

An Speicherdienstleistungen wird angeboten: LZA, DataCite-Datenpublikation (DOI), Verschlagwortung. Hinzu kommt Datenpflege im Zusammenhang mit der ganz unterschiedlichen Aufbereitung.

Die Daten werden über mehrere Portale (derzeit drei) bereitgestellt. Ferner werden Metadaten an internationale Partner geliefert. Die Daten müssen in einer für die Nachnutzung geeigneten Weise kumuliert werden. Wissenschaftler machen nur 60 % der Nutzer aus. Die Daten gehen zu je 20 % an Studenten und an andere. Für die Bereitstellung auf CD werden 25 € in Rechnung gestellt.

Wegen der geringen Datenmenge von etwa 2 TB (ca. 6000 Studien) sind die Hardwarekosten gering. Der Personalbedarf ist nicht nur wegen des Qualitätsmanagements hoch. Die Verschlagwortung, der wegen der Heterogenität der Nutzer hohe Beratungsaufwand und die Kumulation lassen sich nur mit der entsprechenden Zahl von Mitarbeitern bewältigen.

GESIS hat Lizenzgebühren für die Statistik-Software SPSS zu tragen. Insgesamt kostet die Kerninfrastruktur DAS (Datenarchiv für Sozialwissenschaften) 4 Mill. € pro Jahr.

6.1.2. Statistisches Bundesamt

Das Statistische Bundesamt beschäftigt derzeit etwa 2600 Personen in Voll- und Teilzeit. Der überwiegende Teil dürfte unmittelbar Daten produzieren. Gemeinsam mit den statistischen Landesämtern werden pro Jahr zwischen 4000 und 5000 statistische Materialien archiviert. Viele sind kleinere Statistiken von Megabyte-Größe, es gibt aber auch volumenstarke Statistiken von Terabyte-Größe wie den Zensus. Die Daten werden in der Regel den folgenden Bearbeitungsschritten unterzogen:

- Prüfung auf Vollzähligkeit und Vollständigkeit
- Fehlende Daten werden nachgefordert, wenn Auskunftspflicht besteht.
- Plausibilitätsprüfung
- Herstellung der Plausibilität
- Aufbereitung, Auswertung

Derzeit werden auch schon Daten archiviert, aber aufbauend auf ein derzeit laufendes Projekt soll ein einheitliches Verfahren dafür eingeführt werden. Den Kern der Daten, die künftig archiviert werden sollen, bilden Einzelangaben, weil diese für die Zukunft die größten Nachnutzungsmöglichkeiten bieten. Es sind in der Regel plausibilisierte Einzeldaten, d.h. die Daten, die vervollständigt oder über Plausibilitätskontrollen geprüft und korrigiert worden sind. In einzelnen Statistiken werden wohl — weil es entsprechende Anforderungen aus dem Forschungsbereich gibt — auch Rohdaten gespeichert werden. Rohdaten sind so belassen, wie sie eingetroffen sind, also unverändert. Damit kann jederzeit nachvollzogen werden, wie die Daten verändert worden sind, um sie plausibel zu machen. Im Einzelfall werden künftig auch schwach aggregierte Daten archiviert.

Das Statistische Bundesamt hat Kosten für Softwarelizenzen zu tragen. Es werden viele lizenzpflichtige Produkte genutzt, z.B. Oracle und SAS.

6.2. Nukleotidsequenzen

Gemeint ist biologische Erbinformation (DNA- und RNA-Sequenzen), keine technische, künstlich erzeugte DNA. Nukleotidsequenzen können von Forschern aus der ganzen Welt bei der INSDC (International Nucleotide Sequence Database Collaboration) eingereicht werden. Die INSDC garantiert uneingeschränkten und kostenfreien Zugang zu allen vorhandenen Nukleotidsequenzen. In ihr sind die drei Großarchive

- DDBJ (DNA Data Base of Japan)
- ENA (European Nucleotide Archive)
- GenBank (USA) (GenBank Overview)

miteinander verbunden. Die Datenbestände der drei werden täglich synchronisiert. Nach der Annahme wird den Daten eine Accession Number zugeordnet, unter der die Daten wiedergefunden und referenziert werden können.

In biologischen Kernfächern gibt es eine Hinterlegungspflicht für Nukleotidsequenzen, wenn über diese eine Veröffentlichung in einem wissenschaftlichen Journal erfolgen soll. Voraussetzung für eine Veröffentlichung ist die Speicherung der Daten bei der INSDC. Mit der Accession Number weist der Autor dem Journal gegenüber nach, dass er seiner Hinterlegungspflicht nachgekommen ist. Die Accession Number wird in der Veröffentlichung auch als Referenz angegeben.

Die archivierten Datenmengen wachsen schnell. Vor allem haben technische Verbesserungen an den Sequenzierern, insbesondere die Einführung des Next Generation Sequencing, zu dieser Entwicklung geführt aber auch die zunehmende Zahl von Institutionen, die die Sequenzierung im großen Umfang betreiben. Von 2004 bis 2010 hat sich der Sequenzier-Output um den Faktor 100000 erhöht (Kahn, 2011). Wird ein längerer Zeitraum betrachtet, kann sogar ein überexponentielles Wachstum beobachtet werden, ein Wachstum, welches die Fortschritte in der Speichertechnologie repräsentiert durch das Krydersche Gesetz (exponentielles Wachstum der Speicherkapazität von Festplatten) glatt überrundet. Lokale Engpässe bei der Speicherkapazität gibt es bereits. Umgekehrt sind die Kosten für die Sequenzierung, die Erzeugung von Sequenzdaten, dramatisch gefallen, siehe Einleitung.

Die Datenflut hat zu Diskussionen darüber geführt, was eigentlich Rohdaten sind. Viele derzeitige Sequenzierer speichern Bilddaten zu jeder sequenzierten Base, also viel mehr als die Baseninformation selbst.

Zunehmende Probleme erwachsen auch daraus, dass es nicht ausreicht, die Daten nur abzuspeichern. Diese müssen auch verarbeitet werden. Im Interview von SILVA und megx.net wurde dies sehr deutlich gesagt und auch Folgen und Entwicklungstendenzen genannt. Finanzielle Engpässe haben schon dazu geführt, dass GenBank zeitweise keine Daten mehr annehmen kann und Wartezeiten in Kauf genommen werden müssen. Schon heute produziert das BGI (Beijing Genome Institute), das sich inzwischen in Shenzen befindet, so viele Sequenzdaten, dass sie niemand mehr nehmen kann. In China wird bereits für eine Erzeugungsrate von Sequenzdaten im Petabyte/Tag-Bereich geplant, und die chinesischen Planungen schließen entsprechende Datenverarbeitungsfähigkeiten mit ein.

Bei der Qualitätskontrolle spielt die Berechnung von Qualitätsindikatoren eine wesentliche Rolle. Statistische Verfahren kommen zum Einsatz.

6.2.1. Transregio TRR 54: Wachstum und Überleben, Plastizität und zelluläre Interaktivität lymphatischer Neoplasien

Das Archiv nimmt zurzeit Messungen (expression values) von Proben auf einem Mikro-Array (Genchips) auf und in Zukunft auch Ergebnisse von DNA-Sequenzierungsmaschinen. Metadaten werden von den Forschern mitgeliefert, im Wesentlichen als Freitext. Meist wird darin auf eine Veröffentlichung verwiesen.

Die Gesamtmenge ist zurzeit kleiner als 300 GB. Die Hardwarekosten sind daher niedrig. Zurzeit werden ca. 20 GB pro Jahr zum Ingest ausgewählt (etwa 10 Ingest-Vorgänge pro Jahr). Der Ingest besteht im Wesentlichen aus folgenden Schritten:

- Normalisierung des Einzeldatensatzes
- Normalisierung innerhalb eines Experiments
- Berechnung von Qualitätsindikatoren
- Ausfiltern bei offensichtlichen Fehlern in Absprache mit den Erzeugern
- Aggregation auf Gen-Ebene
- Upload

Der webbasierte LymphomExplorer ermöglicht zahlreiche Analysen auf den Daten und besitzt eine Exportfunktion in Standarddatenformate. Dies wird von Forschern genutzt, zunächst von denjenigen, die die Daten erzeugt haben. Eine typische Datenanalyse

vergleicht zwei Datensätze, einmal von einem Gesunden und einmal von einem Kranken, häufig mit Krebs, um dann zu sehen, welche Gene charakteristisch für die Krankheit sind.

Da freie Software verwendet wird, fallen keine Lizenzgebühren an.

6.2.2. SILVA

SILVA ist eine Datenbank für ribosomale RNA-Sequenzen (rRNA). SILVA enthält zurzeit rund 2,8 Millionen Einträge.

Es werden keine Daten direkt von Wissenschaftlern angenommen. Für den Ingest wird stattdessen die komplette EMBL-EBI/ENA-Datenbank (European Molecular Biology Laboratory, European Bioinformatics Institute / European Nucleotide Archive) heruntergeladen. Daraus werden alle Sequenzen extrahiert, die wie die vier Ribosom-Untereinheiten 16S, 18S, 23S und 28S ausschauen. Dies geschieht mit Hilfe von Keyword-Suche aber auch über sequenzbasierte Suche. Dann wird weiter gefiltert, denn für die Aufnahme in SILVA gibt es strenge Qualitätskriterien:

- Mindestanforderung ist die Länge, d.h. die Vollständigkeit der Sequenz.
- Zum Test, ob es sich überhaupt um rRNA-Sequenzen handelt, wird jede einzelne Sequenz mit manuell geprüften Sequenzen aus einem Referenzdatensatz verglichen.
- Wie viele Ambiguities, d.h. nicht aufgelöste Basen, befinden sich in der Sequenz?
- Gibt es eine Vektorkontamination? Beim Prozess der 16S-Generierung wird oftmals kloniert. Dadurch können vorn und hinten an der Sequenz noch Vektorsequenzen hängen. Die sollten eigentlich nicht mehr dort zu finden sein.
- Gibt es Homopolymer-Abschnitte in der Sequenz? Homopolymere bestehen aus nur einer Base. Dass solche Wiederholungen einer Base in der Sequenz natürlichen Ursprungs sind, ist eher unwahrscheinlich. Das sind typische Sequenzierfehler.
- Mittels Chimärentests wird getestet, ob die Sequenz mit hoher Wahrscheinlichkeit von einem Organismus stammt oder ob während der Polymerase-Kettenreaktion oder bei der Probenprozessierung Verunreinigungen hinzugekommen sind.

Ganz am Ende steht die Taxonomie. Jede Sequenz wird nach den Qualitätsprüfungen in den Baum eingerechnet. Damit wird die Klassifikation der Sequenz praktisch noch einmal manuell überprüft. Berücksichtigt wird dabei, dass Organismen in der Zwischenzeit evtl. umbenannt worden sind. Die neuen Namen müssen eingepflegt werden. Speziell in der Mikrobiologie werden immer wieder Organismen und sogar ganze Gruppen umbenannt. Bei diesem Schritt ist ein noch relativ großer manueller Kurationsprozess erforderlich. Die Kuration schlägt allein mit einer Stelle zu Buche.

Der Zugang zu Daten und etlichen Analysediensten ist über ein Portal möglich. Die Daten werden außer von Wissenschaftlern auch von Firmen genutzt. Kommerzielle Nutzung ist kostenpflichtig, wobei sich die Gebühren an den Selbstkosten orientieren.

6.2.3. megx.net

Das Ziel von megx.net ist die Integration von Umweltdaten und Sequenzdaten. Wie SILVA wird megx.net vom Max-Planck-Institut für Marine Mikrobiologie in Bremen betrieben. Beide Datenbanken sind sekundäre, d.h. die Daten werden aus anderen Archiven übernommen.

In megx.net werden nur georeferenzierte Daten aufgenommen, die aus nicht passwort-geschützten Bereichen heruntergeladen werden können. Der Ingest besteht aus den folgenden Schritten:

- Daten herunterladen
- Formatkonvertierung
- Einladen
- Semantische Umformung in das megx.net-Datenmodell

Zu den Speicherdienstleistungen gehören auch Kuration und Taxonomie. Die Daten werden nur von Wissenschaftlern genutzt. Es stehen mehrere Portale zur Auswahl. SILVA und megx.net arbeiten mit freier Software und Eigenentwicklungen. Nach den Personalkosten sind Reisekosten und Hardwarekosten die wichtigsten Posten.

6.3. Umweltdaten

Typisch für Umweltdaten ist deren Bindung an Ort und Zeit. Ohne Orts- und Zeitangabe sind die Daten praktisch wertlos. Homogene Datenstrukturen im Bereich der Modellierung, Großgeräte und Sensorsysteme begünstigen die Verwendung standardisierter Formate. In diesen Bereichen wächst die Kapazität zur Erzeugung neuer Daten schneller als die Speicherkapazität (Klump, 2012). Die Datenmengen sind häufig schon jetzt sehr groß, z.B. bei Satellitendaten und Klimamodellrechnungen. Andererseits gibt es auch in Handarbeit aufgenommene Kleinserien, z.B. Mikroskopaufnahmen. Hier ist die Vielfalt der vorhandenen Datenformate dann doch wieder groß.

Die Qualitätskontrolle von Umweltdaten ist wegen der Vielzahl möglicher Koordinatensysteme und physikalischer und chemischer Parameter oft eine Herausforderung. Häufig kann eine inhaltliche Qualitätskontrolle mit vertretbarem Aufwand nur von den Datenproduzenten selbst vorgenommen werden. Einige Größen, wie z.B. Temperaturen, können durch Vergleich mit Maximal- oder Minimalwerten auf Messfehler bzw. numerische Fehler geprüft werden. Welche Schwellenwerte zur Anwendung kommen, können nur Experten entscheiden, denn die Vorgehensweise ist nicht unproblematisch, wurden doch z.B. vom Satelliten aus gemessene, niedrige Ozonkonzentrationen im Ozonloch über der Antarktis von einem automatisch arbeitenden Auswerteprogramm fälschlicherweise als Messfehler eingestuft.

6.3.1. PANGAEA

PANGAEA ist ein Umweltdatenarchiv für georeferenzierte Daten und wird durch das AWI Bremerhaven und MARUM Bremen betrieben. PANGAEA nimmt Daten direkt vom Wissenschaftler. Ein Teil der Daten wird von Messgeräten an Observatorien direkt abgegriffen. Der Ingest besteht aus folgenden Schritten:

- Kurator vervollständigt in Absprache mit dem Wissenschaftler die Metadaten
- Qualitätskontrolle
    - Statistische Verfahren oder Plausibilitätskontrolle der Maximal- bzw. Minimalwerte je nach Art der Daten
    - Sind Lagekoordinaten vernünftig?
    - Sind Metadaten vollständig?
- Endkontrolle durch den Wissenschaftler selbst
- Evtl. Formatkonvertierung

- Einfüllen

Da die eingelieferten Daten sehr unterschiedlicher Natur sind, ist die Kommunikation mit den Wissenschaftlern aufwändig. Entsprechend hoch ist der Personalbedarf. Den Kontakt zum Wissenschaftler halten Datenkuratoren. Neben den Kuratoren, die bei PANGAEA angestellt sind, kommen einige auch aus den Projekten oder Instituten, mit denen PANGAEA zusammenarbeitet.

An Datendiensten im Bereich Speicherung werden LZA, Kuration und DOI-Vergabe angeboten. Über ein Portal können PANGAEA-Daten heruntergeladen werden. Der Zugang zu den Metadaten ist durch umfangreiche Katalogdienste gewährleistet. Außerdem besteht die Möglichkeit, in einem Datenwarenhaus die Daten selbst zuzuschneiden und so für das Postprocessing vorzubereiten.

PANGAEA hat Hardwarekosten, auch wenn diese nicht in Tabelle 2 quantifiziert sind. Außerdem hat PANGAEA Lizenzgebühren für die Datenbanksysteme zu tragen. Reisekosten müssen eingeplant werden, auch um Kontakte zu pflegen.

6.3.2. megx.net

siehe Abschnitt 6.2.3

6.3.3. Deutsches Klimarechenzentrum (DKRZ)

Am DKRZ befindet sich das World Data Center for Climate (WDCC), das mit zum WDS (ICSU World Data System) gehört. Im WDCC liegen überwiegend Ergebnisse von Klimamodellrechnungen. Daneben gibt es solche Beobachtungsdaten, die der Modell-validierung dienen, z.B. Niederschlagsdaten. Daten werden direkt vom Wissenschaftler angenommen. Einen Teil der Daten produziert das DKRZ selbst. Der Ingest besteht aus folgenden Schritten:

- Vervollständigung der Metadaten gemeinsam mit dem Wissenschaftler
- Technische Qualitätskontrolle
    - Prüfung auf leere Dateien
    - Formatvalidierung
    - Projektabhängig weitere Tests, z.B. auf doppelt vorhandene Zeitstempel
- Projektabhängige wissenschaftliche Qualitätskontrolle
- Prüfung der Metadaten auf formale Korrektheit
- Packen der Daten in Containerdateien und Einfüllen ins Archiv

Am WDCC wird die Vergabe der persistenten Identifier DOI (Digital Object Identifier) und URN (Uniform Resource Name) angeboten. Der Zugang zu den Daten ist über das CERA-Portal realisiert. Für die Nachbearbeitung stehen dort Zeit- und Datenformatkonvertierer sowie fachspezifische Berechnungswerkzeuge zur Verfügung. Zusätzlich werden WDCC-Daten künftig in den Datenverbünden C3Grid (Collaborative Climate Community Data and Processing Grid), ENES (European Network for Earth System Modelling) und ESGF (Earth System Grid Federation) verfügbar sein.

Das DKRZ hat Lizenzkosten zu tragen für

- HPSS (High Performance Storage System, eine hierarchische Storage-Lösung des HPSS-Konsortiums, Vertrieb und Support über IBM)
- Oracle ACSLS zur Steuerung der Tape-Library

- Oracle-Datenbanksoftware
- GHI (GPFS-HPSS-Interface, eine Software, die das Archiv wie ein Filesystem aussehen lässt)
- IBM AIX
- Software auf den Arbeitsplatzrechnern, z.B. MS Office

6.4. Geisteswissenschaftliche Daten

Auch in den Geisteswissenschaften werden immer mehr digitale Daten produziert. Wegen der teilweisen Gegenständlichkeit des kulturellen Erbes spielen Digitalisierungen eine große Rolle. Es gibt aber auch — wie in allen Wissenschaften — immer mehr Daten, die bereits ursprünglich in digitaler Form erzeugt wurden, darunter viele Bilder, Video- und Audioaufzeichnungen. Inhalte und Datenformate sind in den Geisteswissenschaften sehr unterschiedlich.

6.4.1. DoBeS

Am Max-Planck-Institut für Psycholinguistik befindet sich das digitale Zentralarchiv für DoBeS (Dokumentation Bedrohter Sprachen). Das Archiv enthält (teilweise annotierte) Audio- und Videoaufnahmen, Texte und Fotos. Im Jahr 2010 waren das zusammen etwa 50 TB, die in einem hierarchischen Speichersystem liegen und an vier andere Institute repliziert werden (Wittenburg, WG2-9 Cost Estimations, 2010).

Beim Ingest wird geprüft, ob alle Objekte unter Verwendung einer IMDI-Metadatenbeschreibung assoziiert sind. Weiter wird geprüft, ob die Daten in einem akzeptierten Format vorliegen (MPEGx, mJPEG2000, XML, linear PCM,…) und dieses valide ist. An dieser Stelle müssen manchmal Kompromisse geschlossen und wichtige Daten in einem Format angenommen werden, für das es keine Prüfmethoden gibt (Wittenburg, Drude, & Broeder, Psycholinguistik, 2012).

	k€/Jahr	Kommentar
Grundlegende IT-Infrastruktur	80	Erneuerungszyklus 4-8 Jahre
Digitalisierung und Workflow	10	neue Geräte
Kopien (30 TB) bei 4 Datenzentren	<5	zurzeit nicht abgeführt
Systemmanagement	60	für andere Aktivitäten mitgenutzt
Archivmanagement	80	Kuration, Konsistenz
Betreuung Reposity-Software	60	ohne neue Funktionalität
Betreuung Anwendungs- u. Hilfssoftware	>120	weites Spektrum an Tools
Insgesamt	~ 415	

Tabelle 3: Kosten für das DoBeS-Archiv **(Wittenburg, WG2-9 Cost Estimations, 2010)**

Tabelle 3 gibt eine Kostenübersicht. Gebäude-, Energie- und weitere Kosten sind in der Aufstellung nicht enthalten. Diese werden vom Institut übernommen. Ebenfalls nicht enthalten sind die Wartungskosten für die umfangreiche Software LAT (Language Archiving Technology), die für Management und Bereitstellung der Daten genutzt wird.

7. Risiken und Folgekosten, Nutzen der Archivierung

In diesem Kapitel werden wirtschaftliche Aspekte der Risiken von Archivierung und Nicht-Archivierung untersucht. Technische Fragen wie z.B. Datenverlustwahrscheinlichkeiten bei Speicherung und Übertragung werden in dieser Arbeit nicht thematisiert. Die Zahlenbeispiele sind aus der Literatur zusammengetragen, mit Ausnahme der Kosten für die Wiederherstellung von Klimamodelldaten, die aus dem DKRZ stammen.

7.1. Risiken bei Verzicht auf Archivierung

Zunächst sollen die Risiken untersucht werden, die mit einem Verzicht auf Archivierung verbunden sind. Bei Nicht-Archivierung wird das Risiko in Kauf genommen, dass die Daten noch einmal benötigt werden, aber nicht mehr zur Verfügung stehen. Sollen die Daten dennoch genutzt werden, müssen sie erst wiederhergestellt werden, wenn das überhaupt möglich ist. Die Wiederherstellung kostet Geld, die Wiederherstellungskosten, und die Nutzer auch Zeit, die Zeit des Wartens auf die Wiederherstellung. Werden die Daten nicht wiederhergestellt, fällt der Nutzen durch jegliche Nachnutzung aus. Vor der Wiederherstellung und wenn die Daten nicht wiederhergestellt werden, besteht auch das Risiko der Vergeudung von Arbeitszeit durch vergebliche Suche nach den nicht mehr existenten Daten. Für die Ölexploration hat BP-Manager Simon Hendry in einer aufgezeichneten Expertendiskussion (Hawtin & Lecore, 2011) den Anteil der Datensuche mit 20-40 % der Arbeitszeit beziffert.

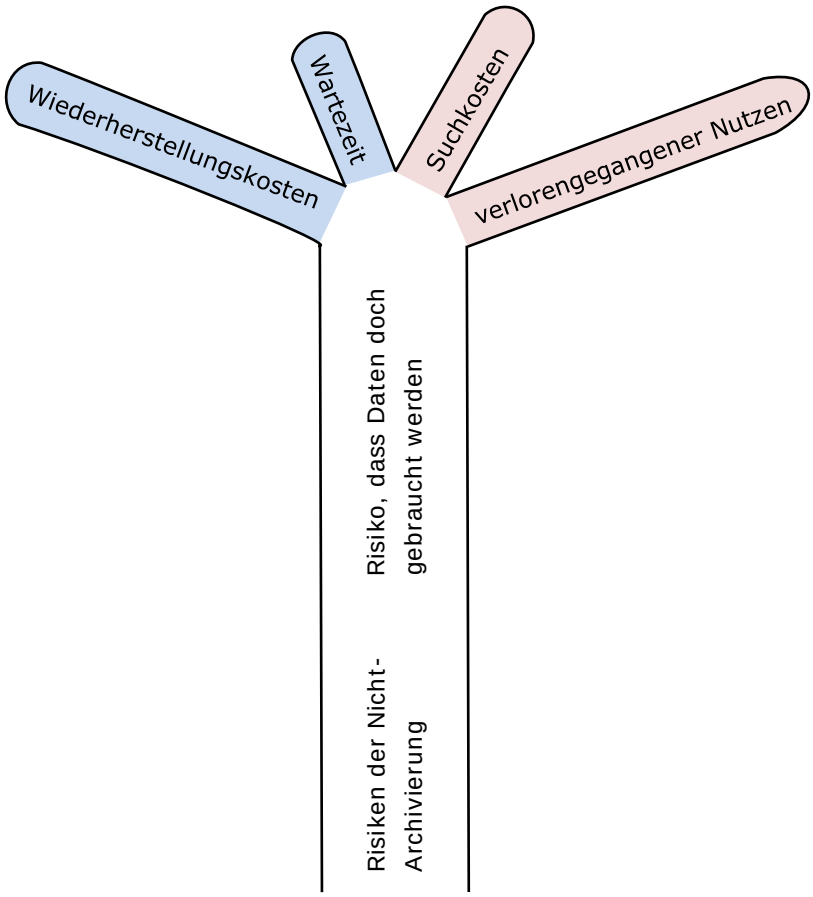

Abbildung 3: Risiken der Nichtarchivierung als Baum. Werden die Daten wiederhergestellt, kostet das Geld und Wartezeit (blaue Zweige). Werden die Daten nicht wiederhergestellt (rote Zweige), kann auch kein Nutzen mehr aus den Daten gezogen werden, im Gegenteil: Durch vergebliche Suche nach den Daten werden den Nutzern Suchkosten aufgelastet.

Eine Quantifizierung des Risikos der Nicht-Archivierung ist eng mit der Frage verbunden, was die Daten wert sind. Da es für Daten der Grundlagenforschung keinen Markt gibt, auf dem ein Preis festgestellt wird, müssen andere Wege beschritten werden, um die Frage zu beantworten. Wenn diese Frage nicht ausschließlich mit einer ideellen Wertangabe, sondern auch mit einem Geldbetrag beantwortet werden soll, bieten sich zwei Wege an: Es kann versucht werden

- den Nutzen zu quantifizieren, der aus den Daten gezogen werden kann, oder
- die Kosten für eine Wiederherstellung der Daten anzugeben.

Letzteres ist selbstverständlich nur möglich, wenn sich die Daten wiederherstellen lassen.

In KRDS2 (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010) wird als Beispiel die allgemeine Haushaltsbefragung des Vereinigten Königreichs von 2001 mit Kosten von ca. 500000 £ genannt, aber auch darauf hingewiesen, dass eine Wiederholung der Befragung die Ergebnisse von 2001 nicht reproduzieren könnte, sondern Daten zum neuen Befragungszeitpunkt liefern würde. Auch viele Beobachtungen und Messdaten sind an den Beobachtungszeitpunkt gebunden, z.B. in der Medizin, Umweltforschung und Astronomie.

Aber selbst wenn das nicht der Fall sein sollte, ist eine Wiederholung einer Befragung, eines Experiments oder einer Beobachtung nicht immer möglich, z.B. weil Interviewpartner oder Experimentieranordnungen nicht mehr zur Verfügung stehen. Nicht jedes Experiment kann leicht wiederaufgebaut werden, schon aus Kostengründen nicht. Messzeiten an noch bestehenden Großgeräten und Observatorien müssen beantragt werden. Die Bewilligung ist sicherlich nicht einfacher zu erreichen, wenn es sich um eine Wiederholung handelt. Sind die Daten nicht wiederherstellbar, bleibt die Möglichkeit, den Wert der Daten als verlorengegangenen Nutzen zu quantifizieren.

7.2. Was kostet es, Daten nicht zu archivieren? Eingebüßter Nutzen

Eine Möglichkeit, die Frage in der Überschrift zu beantworten, ist die Schätzung des mit der Archivierung und Verfügbarmachung der Daten verbundenen Nutzens. Es ist generell schwieriger, den Nutzen zu quantifizieren als Kosten anzugeben. Zu Teilbereichen gibt es inzwischen trotzdem Nutzenanalysen mit Zahlenangaben.

Detaillierte Kosten-Nutzen-Analysen gibt es für Daten des öffentlichen Sektors (PSI, Public Sector Information), zu denen z.B. amtliche Wetterdaten, Landkarten und Statistiken gehören. Pira (Commercial exploitation of Europe's public sector information, 2000) schätzte das Verhältnis von Investitionen in PSI und den Nutzen für die Ökonomie in der EU auf 9,5 Mrd. € zu 68 Mrd. €, d.h. 7-facher Return of Investment. Für die USA wurde in derselben Studie sogar ein Verhältnis von 19 Mrd. € zu 750 Mrd. € angegeben. Die Zahlen wurden durch Verallgemeinerung von Fallstudien und Hochrechnung auf die Gesamtökonomie erhalten. Dekkers et al. (MEPSIR, 2006) kommen bezogen auf die EU+Norwegen nur auf ein geschätztes Marktvolumen von etwa 27 Mrd. €. Te Velde, schon an der MEPSIR-Studie beteiligt, kam sogar auf nur 3-5 Mrd. € nach Berücksichtigung weiterer Überlegungen (Public Sector Information: Why Bother?, 2009). Das ist weniger als der investierte Betrag. Te Velde hat unter anderem berücksichtigt, dass die Informationsindustrie in der EU sehr unterschiedlich entwickelt ist, und deshalb nicht einfach für jeden EU-Staat auf die Gesamtökonomie hochgerechnet.

Untersuchungen wurden auch im Zusammenhang mit der Umstellung auf kostenfreien Zugang durchgeführt. Das Australian Bureau of Statistics (ABS) hat Publikationen und

Statistiken frei verfügbar gemacht unter Creative Commons Lizensierung. Houghton (Costs and Benefits of Data Provision, 2011) hat geschätzt, dass der Zusatznutzen dieser Daten bei freiem Zugang mehr als fünfmal so hoch ist wie die Mindereinnahmen des ABS durch Verzicht auf Gebühren und Verluste durch die Creative Commons Lizensierung. Beim Nutzen hat Houghton sowohl die direkten Einsparungen der Nutzer und des ABS als auch den weiteren ökonomischen Nutzen berücksichtigt.

Anstelle der schwierigen Schätzung des Nutzens können die Nutzer befragt werden, wie hoch sie den Nutzen selbst einschätzen. In der Ölexploration und -produktion (E&P) ist der Erfolg eines Unternehmens in hohem Maße von dessen Datenbestand abhängig, wie eine Umfrage unter 20 Firmen der Branche ergab (Hawtin & Lecore, 2011). Leitende Manager gaben den Beitrag der Daten zur Wertschöpfung einer typischen E&P-Firma mit einem Viertel bis einem Drittel an. Bei einer Wertschöpfung von 100 Mill. $ pro Jahr bedeutet das, dass 25-33 Mill. $ über die petrotechnischen Daten erzielt werden. Auf die Frage, wie lange seismische Daten für brauchbar erachtet werden, gaben die Teilnehmer Antworten zwischen 4 und 20 Jahren. Der Nutzen von Daten aus Bohrungen wurde sogar über die gesamte Lebenszeit der Quelle gesehen.

Nutzer von Erdbeobachtungsdaten der Landsat-Satelliten wurden befragt, wie viel sie für ein Bild zu zahlen bereit wären, wenn die Daten kostenpflichtig wären (Miller, Sexton, Koontz, Loomis, Koontz, & Hermans, 2011). Dieses Vorgehen entspricht der Schaffung eines fiktiven Markts für die Daten. Mittelwert (760 $) und Median (218 $) der Antworten unterscheiden sich deutlich, weil der Mittelwert durch administrative Nutzer hochgezogen wird, die bereit sind, mehr zu zahlen als akademische und private Nutzer.

Nutzer der Technischen Informationsbibliothek Hannover (TIB) wurden befragt, wie hoch ihre Mehrkosten wären, wenn es die TIB nicht gäbe (Die TIB - Zukunft mit Mehrwert, 2010). In dieser Studie wurde ein Nutzen in Höhe des 3,8-fachen der Kosten für die TIB gefunden. Der weitaus größte Teil der Dienstleistungen der TIB betrifft Bücher und normale Publikationen. Die TIB übernimmt aber darüber hinaus als DataCite-Registrierungsagentur die DOI-Vergabe für den Bereich Naturwissenschaften und Technik stellvertretend für viele Institutionen, gerade bei der Publikation von Forschungsdaten.

Im allgemeinen Fall ist die Erfassung des Nutzens schwierig. Eine Klassifizierung von Komponenten des Nutzens in den drei „Dimensionen“ direkt/indirekt, lang-/kurzzeitig und privat/öffentlich hat KRDS2 (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010) vorgenommen, siehe Tabellen 4-6.

Viele der dort genannten Nutzengrößen wie z.B. neue Forschungsmöglichkeiten oder die Stimulierung neuer Netzwerke lassen sich vom Wert her kaum quantifizieren. Der Nutzen aus der Vermeidung der Doppelerzeugung von Daten lässt sich hingegen beziffern. Hier können die Forschungskosten angesetzt werden.

Direkter Nutzen	Indirekter Nutzen (vermiedene Kosten)
Neue Forschungsmöglichkeiten	Keine Doppelerzeugung von Daten
Wissenschaftliche Kommunikation/Zugang zu Daten	Keine Einbuße von künftigen Forschungsmöglichkeiten
Nachnutzung von Daten	Geringere zukünftige Archivierungskosten
Produktivitätszunahme der Forschung	Möglichkeit der Neuausrichtung der Sammlung für neue Nutzergruppen
Stimulierung neuer Netzwerke bzw. Kollaborationen	Möglichkeit der Neuausrichtung von Methoden für neue Nutzergruppen
Wissenstransfer in die Wirtschaft	Nutzung durch neue Nutzergruppen
Erhaltung von Fähigkeiten	Schutz früherer Investitionen
Ökonomisches Wachstum	
Gute wissenschaftliche Praxis, Nachvollziehbarkeit von Forschungsergebnissen	
Erfüllung von Mandaten	

Tabelle 4: Einteilung des Nutzens in direkten/indirekten Nutzen

Kurzzeitnutzen	Langzeitnutzen
Nutzen für die derzeitigen Forscher und Studenten	Nutzen für künftige Forscher und Studenten
Kein Datenverlust beim Postdoc-Wechsel	
Kurzfristige Nachnutzung gut gepflegter Daten	
Sichere Speicherung für datenintensive Forschung	
Zugänglichmachung von Daten, die zu Fachartikeln gehören	Zusätzlicher Nutzen, wenn Archiv wächst und eine kritische Masse erreicht

Tabelle 5: Einteilung des Nutzens nach Zeitdauer

Privater Nutzen	Öffentlicher Nutzen
Nutzen für Förderer	Input für künftige Forschung
Nutzen für Forscher	Motivation neuer Forschung
Erfüllung von Förderbedingungen	Firmengründungen und Entstehung neuer Arbeitsplätze für Hochqualifizierte
Intensivere Wahrnehmung, Zitierungen	
Wirtschaftliche Verwertung von Forschungsergebnissen	

Tabelle 6: Einteilung in privaten/öffentlichen Nutzen

7.3. Was kostet es Daten nicht zu archivieren? Wiederherstellungskosten

Wiederherstellungskosten fallen bei der Wiederherstellung der Daten an, z.B. wenn

- die Daten nicht archiviert wurden, aber noch benötigt werden
- nach Datenverlust
- nach fehlender oder falscher Datenpflege und Verlust der Lesbarkeit

Im Normalfall müssen Wiederherstellungskosten für Daten mit den Forschungskosten gleichgesetzt werden, die mit der Datenproduktion verbunden sind. Kosten in zumindest im Prinzip gleicher Höhe fallen auch bei Doppelarbeit an, wenn geeignete Daten zwar vorhanden sind, dies aber nicht bekannt ist und die Daten deshalb ein zweites Mal erzeugt werden.

Die Wiederherstellung kann sehr teuer werden, wenn die Daten überhaupt nach-produziert werden können. Die Kosten dafür sind insbesondere dann hoch, wenn vor der Datenaufnahme umfangreiche Vorbereitungen erforderlich sind. Man denke nur an die Kosten für Raumsonden und Satelliten, die in die Milliarden gehen können. In einigen Fällen sind die Wiederherstellungskosten aber auch niedriger als die Kosten für eine Archivierung. Oft ist das bei Digitalisierungen und Modellrechnungen (Simulationen) so.

7.3.1. Digitalisierungen

Bei der Digitalisierung von Dokumenten werden die physischen Stücke, z.B. Bücher, Grafiken, Handschriften, eingescannt und so Bilddateien generiert. Da die Digitalisierung inzwischen weitgehend automatisch abläuft — Buchseiten können z.B. automatisch umgeblättert werden —, sind die Kosten vergleichsweise gering. Für eine Serie von etwa 200000 am UK Data Archive (UKDA) eingescannten Bildern ergab eine Kostenschätzung, dass die Wiederherstellungskosten geringer sind als die Archivierungskosten (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010). Die Bilddaten werden aber trotzdem am UKDA aufgehoben. Sollten alle vier Backups verlorengehen oder das TIFF-Format veralten und keine Tools zur Migration zur Verfügung stehen, würde das Ein-scannen wiederholt werden.

Das Herbarium am Botanischen Garten / Botanischen Museum (BGBM) in Berlin-Dahlem enthält mehr als 3,8 Millionen Papierbelege mit gepressten und getrockneten Pflanzen aus den letzten 250 Jahren. Aus dieser einzigartigen Sammlung werden jedes Jahr Tausende von Herbarbelegen an andere Institutionen verliehen. Nicht für jede Forschungsarbeit ist aber der Originalbeleg erforderlich. In 90 % der Fälle reicht ein hochaufgelöster Scan aus, bzw. setzen elektronische Auswerteverfahren ein Digitalisat sogar voraus. Am BGBM wurden deshalb bis 2007 drei Arbeitsplätze zur Digitalisierung eingerichtet, um das Herbarium in digitaler Form verfügbar zu machen. Die Kosten wurden 2008 mit 15,69 € pro Beleg angegeben[19].

Das Herbarium ist ein Beispiel dafür, dass sich nicht jedes Digitalisat kostengünstig und leicht wiederherstellen lässt, nämlich dann nicht, wenn der Herbarbeleg mit Annotationen versehen ist. Schon seit langer Zeit schreiben Forscher Zusatzinformationen mit auf die Herbarbelege oder bekleben diese mit Etiketten. Solche Zusatzinformationen sind von

[19] Kostenanalyse bei induktivem Vorgehen. Bei deduktivem Vorgehen wurden Kosten von 13,39 € ermittelt, jedoch wird die induktive Vorgehensweise für genauer gehalten. (Jaspersen, Wohlfromm, Täschner, & Wendehorst, 2008)

hohem Wert für die spätere Forschung. Handgeschriebene Label-Informationen müssen aber entziffert werden, und das ist in der Regel viel teurer als der reine Scan.

Im zurzeit laufenden Projekt Annosys wird ein System entwickelt, das eine komfortable Annotation digitalisierter Herbarbelege ermöglichen soll. Solche digitalen Annotationen sind selbstverständlich gut lesbar und können dem Digitalisat nachträglich in fast unbegrenzter Zahl hinzugefügt werden. Die gemeinsame wissenschaftliche Arbeit am selben Objekt würde dadurch erleichtert werden. Nachträgliche Annotationen erhöhen aber auch die Anforderungen an die Speicherung, denn wenn sie verloren gehen, können sie durch erneutes Einscannen nicht wiederhergestellt werden, selbst wenn das Papieroriginal des Herbarbelegs noch vorhanden ist.

7.3.2. Simulationen: Fallbeispiel Klimamodelldaten

Archivieren oder mit der noch vorhandenen Software das Modell noch einmal rechnen, wenn der Output doch noch benötigt werden sollte? Welches der kostengünstigere Weg ist, ist vom Einzelfall abhängig. Als Beispiel soll an dieser Stelle der Aufwand für CMIP5-Klimamodellrechnungen (CMIP5 - Coupled Model Intercomparison Project Phase 5 - Overview) vorgestellt werden, die am DKRZ durchgeführt wurden (Legutke, 2012).

Eine Klimamodellrechnung ist eine Simulation, ein numerisches Experiment. Globalmodelle simulieren das Klima der ganzen Erde und bestehen aus einem Atmosphärenmodell, einem Ozeanmodell und weiteren Modellkomponenten wie dem Kohlenstoffkreislauf. Die Teilmodelle sind miteinander gekoppelt und werden parallel ausgeführt. Ausgehend von einem Anfangszustand wird Zeitschritt für Zeitschritt in die Zukunft gerechnet. Danach wird der Simulation-Output in ein einheitliches Format gebracht[20].

Am DKRZ wurden im Rahmen von CMIP5 391/12/79 numerische Experimente mit zusammen 8981/3228/3068 Simulationsjahren gerechnet[21]. Die Schrägstriche beziehen sich auf die verwendeten drei Modellversionen. Bei der Schätzung der Wiederherstellungskosten muss unterschieden werden, ob die Rohdaten und die ursprüngliche Hardware noch vorhanden sind.

Fall 1: Wenn beides noch da ist, muss nur die Formatanpassung wiederholt werden. Das kostet im Mittel 30 Minuten Rechenzeit (Wall-Clock-Time, WCT) pro Simulationsjahr auf einem halben Knoten und einen halben Tag Arbeit für ein Experiment[22]. Hinzu kommt noch ein kleinerer Beitrag für die Qualitätskontrolle. Wenn alle Globalmodelldaten neu erzeugt werden müssen, kostet das mindestens 120 h = ¾ Personenmonat an Arbeitszeit und 30 min · 15277 · ½ ≈ 4000 h WCT auf dann einem Knoten. Am DKRZ kostet die Stunde Wall-Clock-Time zurzeit 0,25 € pro Knoten, so dass die Kosten für die Rechenzeit mit etwa 1000 € zu Buche schlügen.

Fall 2: Die Rohdaten sind auch weg, aber die Hardware ist noch da. Dann müssen auch die Simulationsrechnungen wiederholt werden. Ein erfahrener Mitarbeiter muss dafür 9½ Personenmonate einplanen. Wenn sich der Mitarbeiter erst einarbeiten muss, verdoppelt

[20] Erzeugt wird NetCDF/CF. NetCDF (Network Common Data Form) ist ein selbstbeschreibendes, headerbasiertes Binärformat, mit dem numerische Daten als mehrdimensionale Arrays gespeichert werden können (Davis, et al.). In der CF-Konvention (Climate and Forecast) werden z.B. Koordinatensysteme und die Namen physikalischer und chemischer Größen vorgegeben (CF Metadata). Über CF hinaus gibt es im CMIP5 weitere Konventionen bezüglich der Metadaten und Variablennamen, die beachtet werden müssen.

[21] 832 TB Rohdaten und ca. 60 TB im Format NetCDF/CF mit Klimamodell MPI-ESM

[22] wenn alle Variablen (z.B. Temperatur) neu erzeugt werden müssen

sich diese Zeit etwa. Außerdem sind dafür 90/90/155 min WCT auf 4/4/10 Knoten pro Simulationsjahr erforderlich[23]. Die Schrägstriche beziehen sich wieder auf die drei verschiedenen Modellversionen. Für die Rechenzeit kommen zu den Kosten des Falles 1 also hinzu:

$$90\ \text{min} \cdot 4 \cdot 8981 + 90\ \text{min} \cdot 4 \cdot 3228 + 155\ \text{min} \cdot 10 \cdot 3068 \approx 152000\ \text{h WCT}$$

Das entspricht zurzeit 38000 €.

Fall 3: Formatangepasste Daten, Rohdaten und auch die Hardware, auf der die Simulationsprogramme liefen, sind nicht mehr da. Dann erhöht sich der Arbeitsaufwand gegenüber Fall 2 um etwa vier Personenwochen für die Neukompilierung und um eine Personenwoche für die Portierung von Skripten auf die neue Hardware, vorausgesetzt ein erfahrener Mitarbeiter wird mit diesen Aufgaben betraut. Einschließlich Wiederholung der Simulationsrechnungen und Formatanpassung muss ein erfahrener Mitarbeiter ein Jahr Arbeitszeit aufwenden, um die Daten auf neuer Hardware wiederherzustellen. Die Kosten für die Rechenzeit auf der neuen, effizienteren Hardware sollten geringer sein als jetzt.

Die Rechenergebnisse können — auf einem anderen Prozessor gerechnet — sich geringfügig von den ursprünglichen unterscheiden. Aber selbst wenn dieselbe Hardware erneut genutzt wird, können sich Unterschiede dadurch ergeben, dass z.B. die Version des Modells oder Compilers gewechselt hat. Die Qualitätskontrolle, z.B. die Ausreißerkontrolle, sollte aus diesem Grund ebenfalls wiederholt werden.

Archivierung: Die 60 TB formatangepassten Daten (482 Experimente) ins Archiv zu bringen und zehn Jahre aufzubewahren, kostet am DKRZ (Rathmann, 2013)

- 24000 € für die Bitstream-Preservation bei Berücksichtigung zweier Medienwechsel (Bandkassetten) in diesem Zeitraum
- 6-7 Personenjahre Arbeitszeit für Ingest und Kuration, aber ohne DataCite-Publikation (DOI-Vergabe)

Eine einzelne Wiederherstellung wäre damit finanziell günstiger als die Archivierung. Im Projekt CMIP5 wurde die Archivierung aber von Anfang an geplant und ist Teil des Projekts. Das hängt damit zusammen, dass der Nutzen dieser Daten erheblich ist, die Wartezeit bei einer Wiederherstellung aber untragbar lang wäre. Die Daten dienen weltweit als Referenzdaten, z.B. für neue Modelle. Auch nach den Regeln der guten wissenschaftlichen Praxis müssen sie aufbewahrt werden, da sie in den IPCC-Weltklimabericht eingehen. Dass die erneute Durchführung von Modellrechnungen finanziell günstiger ist als die Archivierung, ist auch aus der Astronomie bekannt (Enke & Wambsganß, 2012).

7.3.3. Glückliche Wiederzusammenführung verstreuter Daten

Ein Glücksfall für die Hochenergiephysik war die Wiederauffindung eines Teils der JADE-Daten aus den 80-er Jahren (Curry, 2011). Nach dem Ende des JADE-Experiments am DESY waren die Daten weltweit verstreut. Zur damaligen Zeit waren weder eine Archivierung noch eine systematische Datenpflege üblich. Dass der größte Teil der JADE-Daten gerettet werden konnte, war möglich, weil ein Physiker einen Teil der Daten ohne Auftrag alle paar Jahre liebevoll auf neuen Speichermedien gesichert hatte.

Kritische Kalibrierdaten überlebten hingegen nur als Papierausdruck und mussten in vierwöchiger Arbeit wieder eingetippt werden. Insgesamt kostete die Aufspürung und

[23] Wieder mit Klimamodell MPI-ESM

Wiederherstellung des größeren Teils der JADE-Daten mehrere Personenjahre Arbeit, die bei sorgfältiger Archivierung hätte vermieden werden können. Der kleinere Teil der Daten ist für immer verloren. Das Beispiel zeigt, dass auch die Wiederzusammenführung nicht archivierter, aber noch vorhandener Daten teuer werden kann.

7.4. Risiken der Archivierung

Daten nicht zu archivieren ist mit einem Risiko behaftet, Daten zu archivieren aber auch. Abbildung 4 ist der Versuch einer Übersicht über einige der Risiken der Archivierung ohne Anspruch auf Vollständigkeit. Datenverlust ist sicherlich ein Schreckensszenario bei der Archivierung und hat letztendlich dieselben Konsequenzen wie die Nicht-Archivierung: Werden die Daten wiederhergestellt, kostet das Geld und Zeit, werden sie nicht wiederhergestellt, fällt der Nutzen aus. Als Folge vergeblicher Suche nach geeigneten Daten werden dem Nutzer außerdem Suchkosten aufgelastet. Die Wahrscheinlichkeit des Datenverlustes ist aber bei sorgfältiger Archivierung um viele Größenordnungen kleiner als bei Verzicht auf Archivierung.

Werden die Daten nicht gepflegt, z.B. nicht auf zeitgemäße Datenträger umkopiert oder nicht in archivfähige Datenformate überführt, sind sie irgendwann nicht mehr lesbar und damit unbrauchbar. Falsche Datenpflege, beispielsweise die Migration in ein Format, für das nicht die erforderlichen Software-Werkzeuge zur Formatkonvertierung oder Weiterverarbeitung der Daten zur Verfügung stehen, kann dieselben Folgen haben.

Nicht-Finden im Sinne von nicht gefunden werden wird vom Archiv vielleicht nicht einmal bemerkt, da der Schaden zunächst allein beim Nutzer liegt. Sind die gewünschten Daten scheinbar nicht vorhanden, wird entweder auf eine Nutzung verzichtet oder die erneute Produktion der Daten veranlasst mit der Folge unnötiger Erstellungskosten und Wartezeit. Außerdem sind beim Nutzer Suchkosten angefallen. Gründe dafür, dass Daten nicht gefunden werden, können z.B. fehlende Suchbarkeit in einem „Datensarkophag“ sein oder in den Metadaten liegen, in denen gesucht wird. Selbst wenn die Suche möglich ist, muss das Archiv bekannt sein und die Inhalte der Metadaten, in denen gesucht werden kann, müssen vom Suchenden auch verstanden werden können. Disziplinübergreifend ist das oft ein Problem.

Zunächst nicht sichtbare Mehrkosten drohen beim Outsourcing von Dienstleistungen, z.B. der Speicherung. Beim Vendor-lock-in hat sich das Archiv derart fest an den externen Dienstleister gebunden, dass es sich nicht ohne Mehrkosten aus dem Vertrag lösen kann. Beispielsweise können die Kosten für den Datentransfer bei einem Anbieterwechsel so hoch sein, dass schon aus diesem Grund der Anbieterwechsel nicht zustande kommt. Das Problem kann auch rechtlicher oder technischer Natur sein. Wenn ein einfaches Herausziehen der Gesamtheit der Daten z.B. aus der externen Cloud nicht möglich ist, ist das Archiv aus technischen Gründen an den Anbieter gebunden. Ein weiteres Risiko besteht bei vollständigem Outsourcing der Datenhaltung. Niemand gibt dann Hinweise, wie die Datenhaltung verbessert werden kann.

Das Outsourcing kann auch im Normalbetrieb zu erheblich höheren Datentransferkosten führen, da die Daten zum Dienstleister gelangen müssen. Ein damit verbundenes Problem sind unbefriedigende Geschwindigkeiten beim Transfer großer Datenmengen, die Wartezeiten und damit auch zusätzliche Kosten zur Folge haben. Der Schutz der Daten gegen Missbrauch kann beim externen Anbieter in der Praxis anders gehandhabt werden als beim Archiv selbst. Dadurch können sich die rechtlichen Risiken erhöhen. Selbst wenn der externe Anbieter ausreichende Maßnahmen zum Schutz der Daten garantiert, hat ein

Teil der Mitarbeiter des Anbieters Zugang zu den Daten, es sei denn, die Daten sind verschlüsselt.

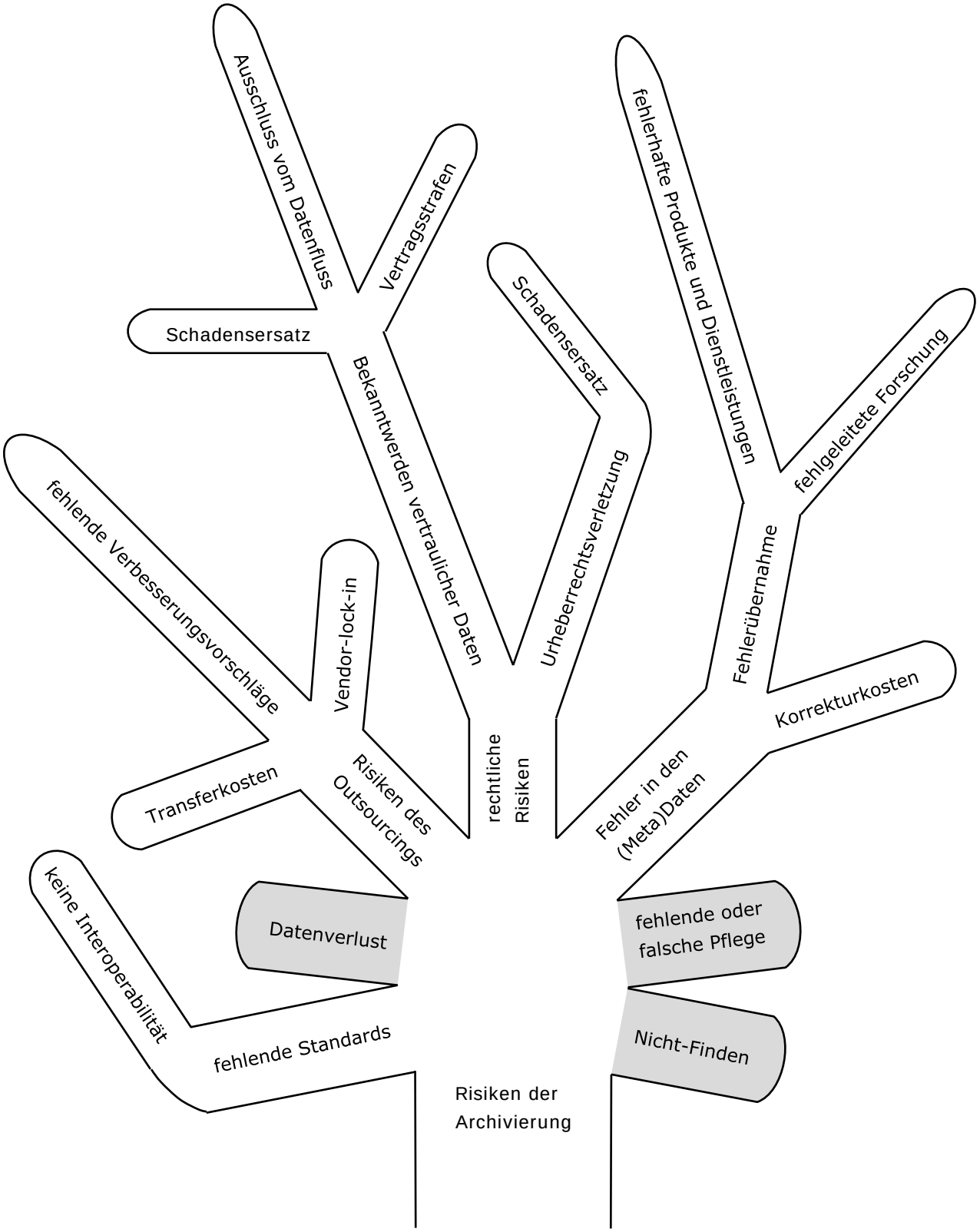

Abbildung 4: Risiken der Archivierung als Baum. An die grauen Äste könnte ein Zweig wie in Abbildung 3 angefügt werden, denn alle diese Äste, „Datenverlust“, „fehlende oder falsche Pflege“ und „Nicht-Finden“, führen dazu, dass die Daten nicht mehr oder vermeintlich nicht zur Verfügung stehen.

Fehlende Standardisierung bzw. der Verzicht auf Kontrolle des Datenmanagements ist ein weiteres, möglicherweise kostenträchtiges Risiko. Wenn jeder irgendetwas irgendwie speichert, ist es teuer und zeitaufwändig, das später zusammenzubringen.

7.4.1. Fehler in Daten und Metadaten

Daten und Metadaten könnten Fehler enthalten. Werden diese bekannt, kann überlegt werden, wie eine Korrektur durchgeführt werden könnte. Die Korrektur ist mit Kosten verbunden (Korrekturkosten). Solange die Fehler in den Daten bleiben, besteht das Risiko, dass sich die Fehler auf Analysen, die auf den Daten vorgenommen werden, unbemerkt auswirken. Die Fehler können so ungewollt in wissenschaftliche Arbeiten übernommen werden. Im schlimmsten Fall wird auf Basis falscher Forschungsergebnisse weitere Forschung auf einen falschen Pfad gelenkt oder es entstehen fehlerhafte Produkte und Dienstleistungen. Durch fehlerhafte Produkte oder Dienstleistungen in der Medizin können sogar Körperschäden entstehen. Das Risiko, dass Menschen zu Schaden kommen, ist aber auch bei Nicht-Archivierung gegeben.

Welche Folgekosten fehlerhafte wissenschaftliche Daten nach sich ziehen, ist bisher kaum beziffert worden. Für das Gebiet der Proteomik, die sich mit der Gesamtheit der Proteine in einer Zelle oder einem Lebewesen befasst, hat White (The Potential Cost of High-Throughput Proteomics, 2011) diese Frage untersucht. Seiner Studie zufolge ist eine falsch-positive Proteinidentifikation insbesondere dann von Bedeutung, wenn dieses Protein für weitere Untersuchungen ausgewählt worden ist, z.B. weil es ein Kandidat für einen Biomarker eines bestimmten Krankheitsstadiums ist. Großer Schaden ist zu erwarten, wenn das vermeintlich gefundene Protein eine bisher unbekannte post-translationale Veränderung aufweist. Nach der eigentlichen Proteinsynthese am Ribosom (Translation) erfahren Proteine häufig noch Veränderungen wie Abspaltungen, den Einbau zusätzlicher funktioneller Gruppen oder die Knüpfung neuer Bindungen. Ein vermeintlich neuer Typ solcher posttranslationaler Veränderungen würde intensive Forschungstätigkeit auslösen. Diese Folgestudien können immens zeitaufwändig sein. Die Aufklärung der phänotypischen Rolle einer Phosphorylierung beispielsweise ist oft ein Manuskript für sich und kann Monate oder sogar Jahre an Forschungsaufwand bedeuten.

Auch in anderen Wissenschaftszweigen dürften fehlerhafte Daten fehlgeleitete Folge-studien zur Folge haben. Diese sehr teure Art der Richtigstellung ließe sich häufig durch Anstrengungen bei der Qualitätskontrolle vermeiden. In der Proteomik müsste diese bereits bei der Datenerzeugung im Labor ansetzen. Im ersten Schritt wird üblicherweise das Massenspektrum manuell validiert, d.h. die Zuordnung der Massen-zu-Ladungs-Quotienten *m/z* zu Fragment-Ionen wird von Hand vorgenommen (White, 2011). Der zweite Schritt wäre die chemische Synthese des Peptids. Synthetisiertes und biologisches Peptid sollten das gleiche Massenspektrum besitzen. Die volle Validierung der Protein-identifikation wäre im dritten Schritt durch ein Mischexperiment erreicht. Dabei wird das synthetisierte Protein mit der biologischen Probe gemischt und die Mischung erneut analysiert. Da Massenspektrometrie und Peptidsynthese Standardverfahren in der Proteomik sind, lassen sich die Kosten für die Qualitätskontrolle angeben. Die Validierung eines Massenspektrums (Schritt 1) erfordert Zeit (Minuten bis Stunden). Die Sachkosten für eine Peptidsynthese oder ein Mischexperiment liegen in der Größenordnung von 100 $ und erfordern mehrere Stunden Arbeit (White, 2011).

7.4.2. Rechtliche Risiken

Beim Umgang mit vertraulichen oder urheberrechtlich geschützten Daten gehen Archive und Nutzer rechtliche Risiken ein.

Gelangen vertrauliche Daten in die Öffentlichkeit und ist das Archiv dafür verantwortlich, drohen je nach Art der Daten Sanktionen. Im Fall personengebundener Daten gilt in Deutschland das Bundesdatenschutzgesetz (BDSG). In § 7 wird dort Betroffenen Schadensersatz zugesichert, bei automatisierter Datenverarbeitung durch öffentliche Stellen ist dieser jedoch nach § 8 auf 130000 € begrenzt. Wenn die Speicherung unzulässig ist, kann der Betroffene die Löschung verlangen (§ 20 und § 35). Fehlerhafte Daten sind zu korrigieren, strittige zu sperren. Auch wenn mit personengebundenen Daten korrekt gearbeitet wird, können Kosten entstehen: § 19 und § 34 sichern Betroffenen ein Auskunftsrecht zu, wenn personengebundene Daten gespeichert sind. Für Zensus- und Sozialversicherungsdaten gelten statt des Datenschutzgesetzes die Bestimmungen des Bundesstatistikgesetzes und des Sozialgesetzbuches.

Auch beim Bekanntwerden anderer, nicht personengebundener Daten wie z.B. Unternehmensdaten kann es zu Sanktionen gegenüber dem Archiv kommen, wenn dies zwischen dem Archiv und den Datenproduzenten vertraglich so festgelegt wurde. Ohne die Vereinbarung von Vertragsstrafen hätte das Archiv die Daten möglicherweise gar nicht bekommen. Dies ist z.B. häufig bei der Überlassung von Betriebs- oder Geschäftsgeheimnissen der Fall.

Auch ohne vereinbarte Vertragsstrafen und ohne Begründung können Datenproduzenten von weiteren Archivierungsaufträgen absehen, wenn es keine Lieferpflicht gibt, und ihre Daten dann beispielsweise woanders archivieren.

Verletzungen des Urheberrechts können ebenfalls Schadensersatzforderungen nach sich ziehen. Meist erreichen Forschungsdaten allein nicht die Schöpfungshöhe, die Voraussetzung dafür ist, dass das Urheberrecht greift. Ausnahmen können aber z.B. Lichtbilder sein, zu denen auch Aufnahmen im Röntgen- und anderen Wellenlängenbereichen gezählt werden. Im Zweifel sollte das Archiv gegenüber dem Schöpfer auf der Einräumung eines (nicht ausschließlichen) Nutzungsrechts bestehen. Ob das tatsächlich notwendig ist, hängt von vielen Faktoren ab (Hillegeist, 2012).

7.5. Wer ist von welchen Risiken betroffen?

Neben der Art und Klassifizierung der Risiken ist auch die Frage erörterungswürdig, wer von den beteiligten Akteuren welches Risiko zu tragen hat. Beschränken wir uns auf die Akteure Archiv und Nutzer. Werden Archive in

- die private Domäne des Wissenschaftlers (Privatarchive)
- die Gruppenebene (Archive auf Arbeitsgruppen- oder Projektebene)
- die dauerhafte Domäne (Archive für die gesamte Disziplin und disziplinübergreifende Archive)

eingeteilt, kann man sich überlegen, dass die in den Abbildungen 3 und 4 angeführten Risiken auf allen drei Domänen lasten. Die Risiken der Nicht-Archivierung, des Datenverlustes, der fehlenden oder falschen Pflege und der Fehler in den Daten treffen grundsätzlich alle; lediglich die Folgekosten sind unterschiedlich hoch. In der privaten Domäne wird der Forscher gewöhnlich Fehler in den Daten selbst korrigieren. Da er zugleich Datenproduzent und Archivmanager in einer Person ist, hat er die hierfür notwendigen

Kenntnisse und Mittel. In der dauerhaften Domäne wird anders vorgegangen. Das Archivpersonal kann die Fehler in der Regel nicht selbst korrigieren. Die Korrektur muss stattdessen delegiert werden. Fehler und erfolgte Korrektur müssen bekanntgemacht werden, zumindest in den Metadaten. Ist auf die Daten ein persistenter Identifier vergeben, ist ein neuer Identifier für die korrigierten Daten oder ein Erratum Pflicht. Alle diese Aktivitäten bedeuten Personalaufwand, der in der privaten und der dauerhaften Domäne unterschiedlich ist. An der grundsätzlichen Feststellung, dass die Risiken alle drei Domänen betreffen, ändert das jedoch nichts.

Risiken des Outsourcings lasten auf Archiven, die Aufgaben an externe Anbieter ausgelagert haben. Die Existenz des Risikos hängt nicht von der Domäne ab. Ebenso verhält es sich bei den rechtlichen Risiken. Folgekosten aufgrund von Rechtsverletzungen können alle treffen. Ob das Risiko existiert, hängt davon ab, ob mit vertraulichen oder urheberrechtlich geschützten Daten gearbeitet wird, nicht von der Domäne. Die Folgekosten können wieder unterschiedlich hoch sein, z.B. wenn Schadensersatz nach der Zahl der Nutzer bemessen wird.

Das Risiko fehlender Standards wird in der privaten Domäne möglicherweise nicht im gleichen Maße gesehen wie in der dauerhaften Domäne. Die Existenz des Risikos hängt aber nicht davon ab, ob es bemerkt wird. Spätestens wenn der Wissenschaftler versucht, Daten seines Privatarchivs in ein Gruppen- oder übergeordnetes Archiv zu integrieren, werden sich Insellösungen rächen.

Beim Risiko des Nicht-Findens ist es etwas komplizierter. Selbstverständlich sollte der Wissenschaftler in der Lage sein, seine eigenen Daten in seinem Privatarchiv wiederzufinden. Lässt man als Akteure aber auch externe Nutzer zu, so ändert sich das Bild. Externe Nutzer sind in allen Domänen vom Risiko des Nicht-Findens betroffen. Externe Nutzer sollten in die Betrachtung aber mit einbezogen werden, wenn gemeinsame fachbezogene oder gar fachübergreifende Nachnutzung von Daten angestrebt wird.

Von welchen Risiken ist nun der (interne oder externe) Nutzer betroffen? Den Nutzer treffen Nicht-Archivierung, Datenverlust, fehlende oder falsche Datenpflege und Nicht-Finden direkt, da beim Eintreten eines dieser Risiken die Daten, die er braucht, nicht oder vermeintlich nicht zur Verfügung stehen. Fehler in den Daten treffen ihn ebenfalls direkt, denn er ist es, der die Daten einsetzt.

Rechtliche Risiken müssen sowohl Nutzer als auch Archiv berücksichtigen. Beide können mit Schadensersatzforderungen oder Vertragsstrafen konfrontiert und auch vom Datenfluss ausgeschlossen werden. Vertragsstrafen kann die Datenübernahmevereinbarung zwischen Datenproduzent und Archiv aber auch der Vertrag zwischen Archiv und Nutzer vorsehen. Datenproduzenten können das Archiv vom Datenfluss ausschließen, sofern keine Pflicht besteht, dort zu archivieren. Sie können sich einfach für ein anderes Archiv entscheiden und dort ihre neuen Forschungsdaten archivieren. Nutzern kann die Nutzungsmöglichkeit genommen werden, wenn der Nutzungsvertrag zwischen Nutzer und Archiv eine solche Sanktion vorsieht oder indem dieser Vertrag beendet wird, z.B. durch Kündigung.

Die Risiken des Outsourcings trägt zunächst das Archiv. Der Nutzer ist aber indirekt mit betroffen, denn tritt eines der Risiken tatsächlich ein und entstehen unvorhergesehene Mehrkosten, so wird sich das Archiv nicht in dem Umfang weiterentwickeln können, wie das ohne diese Mehrkosten möglich gewesen wäre. Das wird dann später möglicherweise auch der Nutzer zu spüren bekommen.

Fehlende Standards treffen den Nutzer ebenfalls indirekt, weil der Komfort gemeinsamer Datendienste mehrerer Archive bei fehlender Interoperabilität nicht realisiert werden kann. Insgesamt ist der Nutzer von allen Risiken der Nicht-Archivierung und Archivierung direkt oder indirekt betroffen.

In Radieschen-Dokumenten wird zusätzlich zu den schon genannten drei Domänen eine Zugangsdomäne betrachtet. Dies erleichtert die Untersuchung von Fällen, in denen nur ein Teil der Daten frei zugänglich ist oder es mehrere Institutionen gibt, die den Zugang zu denselben Archivdaten anbieten. Die Zugangsdomäne ermöglicht den Zugang zu den frei verfügbaren Daten in technischer Hinsicht. Sie sammelt selbst keine Daten und speichert diese — wenn überhaupt — nur kurz zwischen im Rahmen der Bereitstellung. Die Zugangsdomäne kann aber sehr wohl Nachprozessierung anbieten und so auf Basis der Daten einen Mehrwert schaffen.

Die Abtrennung der Zugangsdomäne hat bezüglich der Risiken nur dann Konsequenzen, wenn die Zugangsdomäne rechtlich eigenständig ist, denn andernfalls trägt die Institution, der Archiv und Zugangsdomäne gemeinsam angehören, die Risiken. Als Zwischenstation zwischen speicherndem Archiv und Nutzer ist die Zugangsdomäne in der gleichen Lage wie der Nutzer, denn sie handelt wie ein Nutzer. Die Zugangsdomäne bezieht Daten vom Archiv und tut damit etwas (gibt die Daten an den Nutzer weiter, gegebenenfalls nach Nachprozessierung). Insofern ist die Zugangsdomäne von allen Risiken betroffen, die auch den Nutzer treffen. Beispielsweise ist sie von einem Datenverlust beim Archiv betroffen, wenn sich die Daten nicht zufällig noch im Zwischenspeicher befinden und aus diesem gerettet werden können.

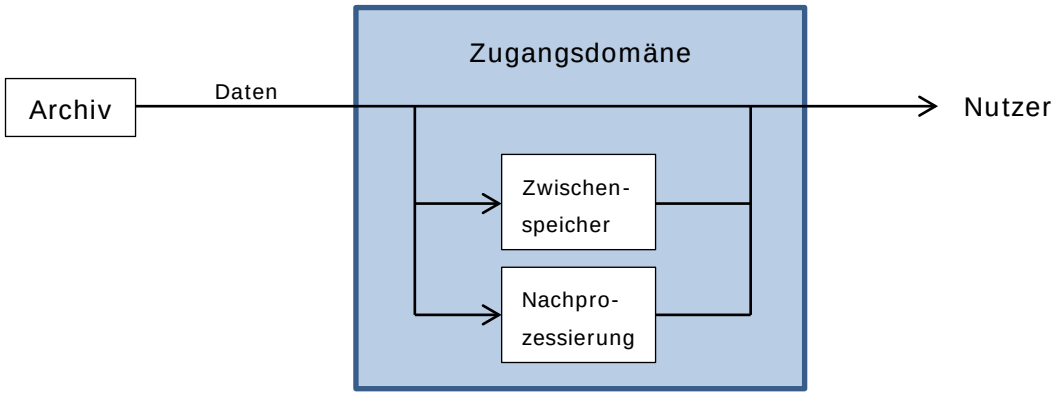

Abbildung 5: Datenfluss über eine Zugangsdomäne

In einigen Punkten trifft die Zugangsdomäne aber eine besondere Verantwortung, die sich aus ihrer Tätigkeit und Zielsetzung ergibt. Das Risiko von Fehlern in den Daten erhöht sich

- durch zusätzlichen Transfer und Zwischenspeicherung
- gegebenenfalls durch Nachprozessierung

Aus fehlerfreien Daten können fehlerbehaftete Daten erzeugt werden, z.B. durch zufällige Schreib-Lese-Fehler oder durch Fehler in der Nachprozessierungs-Software. Andererseits kann die Zugangsdomäne durch eine zusätzliche Qualitätskontrolle das Fehlerrisiko auch verringern.

Das Risiko des Nicht-Findens klein zu halten sollte der Zugangsdomäne besonders am Herzen liegen, denn für den Zugang ist sie da und der setzt normalerweise eine Suchmöglichkeit voraus. Ausnahme sollte nur der Zugang über eine Schnittstelle im Rahmen

der Maschine-Maschine-Kommunikation sein, wenn davon ausgegangen werden kann, dass die automatisch arbeitenden Programme, die diesen Zugang nutzen, eigene Such- und Auswahlmechanismen haben.

Auch in Sachen Standards darf eine besondere Aufmerksamkeit seitens der Zugangsdomäne erwartet werden, denn Interoperabilität ist zum großen Teil eng mit dem Zugang verbunden.

8. Schlussfolgerungen

An dieser Stelle sind die wichtigsten Ergebnisse zusammengefasst und mit Schlussfolgerungen, Hinweisen und Empfehlungen versehen.

1. Die Personalkosten sind gewöhnlich der größte Kostenblock, mit dem es Forschungsdatenarchive zu tun haben. Der Personalaufwand ist aber nicht gleichmäßig auf die einzelnen Schritte im Datenlebenszyklus verteilt. Primärarchive, d.h. solche Forschungsdatenarchive, die Daten direkt vom Forscher nehmen, haben deutlich mehr Arbeit mit dem Ingest als mit der Kuration (Datenpflege) und der eigentlichen Speicherung, der Bitstream-Preservation. Das liegt zum einen am hohen manuellen Aufwand für die Kommunikation mit den Datenproduzenten, zum anderen auch an anderen arbeitsintensiven Schritten wie der Qualitätskontrolle. Insgesamt besteht der Ingest bei den meisten befragten Datenprojekten aus einer Vielzahl von Schritten, die zum Teil maschinell, zum Teil manuell abgearbeitet werden. Durch eine weitergehende Automatisierung könnten Kosten reduziert und Arbeitszeit für andere Aufgaben freigemacht werden. In KRDS2 wurde vorgeschlagen, entsprechende Software-Werkzeuge zu entwickeln (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010). Im DFG-Projekt PubFlow (Brauer, 2012) wird nun an einer disziplinübergreifenden Software gearbeitet, die eine möglichst weitgehende Automatisierung des Ingest zum Ziel hat. Darüber hinaus plant PubFlow die Verzahnung von maschinellen und manuellen Schritten durch die Verbindung mit einem Ticketing-System. Manuelle Schritte sollen durch Öffnen eines Tickets angefordert werden können.
2. Weitergehende Sparmaßnahmen beim Ingest sind problematisch, weil hier auch die Qualitätskontrolle angesiedelt ist. Nachlassende Anstrengungen bei der Qualitätskontrolle würden früher oder später zu einer höheren Zahl von Fehlern in den Daten und Metadaten führen. Dies kann zu hohen gesellschaftlichen Folgekosten z.B. durch fehlgeleitete Forschung führen.
3. Das gilt auch für Sekundärarchive, d.h. solche Forschungsdatenarchive, die Daten nicht direkt vom Forscher, sondern von anderen Archiven nehmen. Diese haben einen geringeren Kommunikationsaufwand beim Ingest, weil sie weniger mit den einzelnen Datenproduzenten kommunizieren müssen. Unter den von Radieschen befragten Sekundärarchiven ist der Ingest sogar mit dem geringsten Personalaufwand verbunden, verglichen mit der Bereitstellung (Access) und der zusammengefasst betrachteten Kuration und Speicherung. Die Bemerkung eines Interviewpartners, dass sein Ingest unterfinanziert sei, sollte aber zu denken geben. Auch Sekundärarchive haben Aufgaben beim Ingest, die sich nicht beliebig zusammenstreichen lassen.
4. Wer auf eine Archivierung seiner Forschungsdaten verzichtet, lässt sich auf ein hohes Datenverlustrisiko ein. Selbst wenn die Daten noch irgendwo auf Daten-

trägern liegen, werden die Suche nach den Daten, die Migration und die Neuzusammenstellung teuer. Sind die Daten erst einmal verloren, können die Folgekosten noch wesentlich höher liegen. Werden die Daten wiederhergestellt, fallen dafür Kosten an, häufig in Höhe der Forschungskosten. Außerdem muss mit erheblichen Wartezeiten gerechnet werden. Werden die Daten nicht wiederhergestellt, fällt deren Nutzen aus, und den Nutzern werden Kosten für die vergebliche Suche nach den Daten aufgelastet.

5. Nicht immer ist die Wiederherstellung der Daten teurer als die Archivierung. Digitalisierungen können häufig kostengünstig wiederholt werden, wenn die Originale noch vorhanden sind. Es gibt aber Ausnahmen: Annotationen lassen sich weniger leicht oder gar nicht wiederherstellen. Teuer werden Wiederherstellungen auch, wenn Handschriften erneut entziffert werden müssen. Selbst wenn die Wiederherstellung finanziell günstiger sein sollte, kann eine Archivierung aus anderen Gründen trotzdem angezeigt sein, z.B. zur Schonung des Originals.
6. Auch die Neuberechnung von Simulationen ist häufig billiger als die Archivierung. Aber selbst wenn die Wiederherstellung finanziell günstiger sein sollte, kann eine Archivierung aus anderen Gründen trotzdem nötig sein, z.B. weil die Simulationsdaten als Referenzdaten dienen.

Literaturverzeichnis

(kein Datum). Von International Nucleotide Sequence Database Collaboration: www.insdc.org abgerufen

(kein Datum). Von DNA Data Base of Japan: http://www.ddbj.nig.ac.jp/ abgerufen

Costs of Digital Preservation. (Mai 2005). Von nationaalarchief: www.nationaalarchief.nl/sites/default/files/docs/kennisbank/codpv1.pdf abgerufen

da|ra. (2012). Abgerufen am 4. Dezember 2012 von Registrierungsagentur für Sozial- und Wirtschaftsdaten da|ra: http://www.da-ra.de/

DNA Sequencing Costs. (2013). Abgerufen am 26. April 2013 von National Human Genome Research Institute: http://www.genome.gov/sequencingcosts/

Ashley, K. (1999). *Digital Archive Costs: Facts and Fallacies.* Von DLM Forum'99: http://ec.europa.eu/archives/ISPO/dlm/fulltext/full_ashl_en.htm abgerufen

Beagrie, N. (Januar 2007). *E-Infrastructure strategy for research: final report from the OSI Preservation and Curation Working Group.* Abgerufen am 28. November 2012 von National e-Science Centre (NeSC): http://www.nesc.ac.uk/documents/OSI/preservation.pdf

Beagrie, N., Chruszcz, J., & Lavoie, B. (2008). *Keeping Research Data Safe, A Cost Model and Guidance for UK Universities.* Abgerufen am 28. November 2012 von JISC: http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

Beagrie, N., Lavoie, B., & Woollard, M. (April 2010). *Keeping Research Data Safe 2.* Abgerufen am 8. August 2012 von JISC: http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf

Björk, B.-C. (2007). *Evaluation of the Costing Activities and Economic Models for Digital Curation Using the LIFE Methodology.* Von UCL Discovery: http://eprints.ucl.ac.uk/7684/1/7684.pdf abgerufen

Brauer, P. (9. Februar 2012). *PubFlow*. Abgerufen am 13. März 2013 von Christian-Albrechts-Universität zu Kiel: http://www.pubflow.uni-kiel.de/

CF Metadata. (kein Datum). Abgerufen am 15. November 2012 von CF Metadata: http://cf-pcmdi.llnl.gov/

CMIP5 - Coupled Model Intercomparison Project Phase 5 - Overview. (kein Datum). Abgerufen am 8. November 2012 von http://cmip-pcmdi.llnl.gov/cmip5/

Curry, A. (11. Februar 2011). *Rescue of Old Data Offers Lesson for Particle Physicists.* Abgerufen am 20. August 2012 von Science: http://www.sciencemag.org/content/331/6018/694.full

Davis, G., Rew, R., Hartnett, E., Caron, J., Heimbigner, D., Emmerson, S., et al. (kein Datum). *NetCDF (Network Common Data Form)*. Abgerufen am 15. November 2012 von unidata: http://www.unidata.ucar.edu/software/netcdf/

Dekkers, M., Polman, P., te Velde, R., & de Vries, M. (24. April 2006). *MEPSIR (Measuring European Public Sector Resources).* Abgerufen am 12. Dezember 2012 von European Public Sector Information Platform: http://ec.europa.eu/information_society/policy/psi/mepsir/index_en.htm

Dickmann, F. (21. April 2009). *AP5 - Kosten der elektronischen Langzeitarchivierung.* Abgerufen am 31. August 2012 von KoLaWiss ("Kooperative Langzeitarchivierung für Wissenschaftsstandorte"): http://kolawiss.uni-goettingen.de/projektergebnisse/AP5_Report.pdf

Dokumentation Bedrohter Sprachen. (kein Datum). Abgerufen am 6. Dezember 2012 von Dokumentation Bedrohter Sprachen: http://www.mpi.nl/DOBES/

Enke, H., & Wambsganß, J. (2012). Astronomie und Astrophysik. In H. Neuroth, S. Strathmann, A. Oßwald, R. Scheffel, J. Klump, & J. Ludwig (Hrsg.), *Langzeitarchivierung von Forschungsdaten* (S. 289). Boizenburg: Verlag Werner Hülsbusch.

European Nucleotide Archive. (kein Datum). Von European Bioinformatics Institute: http://www.ebi.ac.uk/ena/ abgerufen

GenBank Overview. (kein Datum). Von National Center of Biotechnology Information (NCBI): http://www.ncbi.nlm.nih.gov/genbank/ abgerufen

Hawtin, S., & Lecore, D. (2011). *The business value case for data management - a study, Common Data Access Ltd., London, United Kingdom.* Abgerufen am 7. November 2012 von Oil & Gas UK: http://www.oilandgasuk.co.uk/datamanagementvaluestudy/

Hillegeist, T. (2012). *Rechtliche Probleme der elektronischen Langzeitarchivierung wissenschaftlicher Primärdaten* (Göttinger Schriften zur Internetforschung Ausg., Bd. 8). (S. Hagenhoff, D. Hogrefe, E. Mittler, M. Schumann, G. Spindler, & V. Wittke, Hrsg.) Universitätsverlag Göttingen.

Houghton, J. (September 2011). *Costs and Benefits of Data Provision.* Abgerufen am 12. Dezember 2012 von Australian National Data Service: http://ands.org.au/resource/houghton-cost-benefit-study.pdf

ICSU World Data System. (kein Datum). Abgerufen am 5. Dezember 2012 von ICSU World Data System: http://www.icsu-wds.org/

Jaspersen, T., Wohlfromm, B., Täschner, M., & Wendehorst, S. (2008). *Kostenanalyse zur Digitalisierung von Herbarbelegen im Botanischen Garten / Botanischen Museum in Berlin-Dahlem.* Abgerufen am 7. November 2012 von Herbar-Digital: http://www.yasni.de/ext.php?url=http%3A%2F%2Fopus.bsz-bw.de%2Ffhhv%2Fvolltexte%2F2009%2F257%2Fpdf%2F080627_Zwischenbericht_Kostenanalyse_Herbar_Digital_A1a.pdf&name=Marc+T%C3%A4schner&cat=filter&showads=1

Jensen, U. (Juli 2012). *Leitlinien zum Management von Forschungsdaten, Sozialwissenschaftliche Umfragedaten.* Abgerufen am 19. Dezember 2012 von gesis: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-07.pdf

Kahn, S. D. (11. Februar 2011). *On the future of genomic data.* Abgerufen am 24. August 2012 von Science: http://www.sciencemag.org/content/331/6018/728.full

Klump, J. (2012). Geowissenschaften. In H. Neuroth, S. Strathmann, A. Oßwald, R. Scheffel, J. Klump, & J. Ludwig (Hrsg.), *Langzeitarchivierung von Forschungsdaten* (S. 184). Boizenburg: Verlag Werner Hülsbusch.

Kunert, P. (25. November 2011). *Disk drive prices swell 5% every DAY in floods aftermath.* Abgerufen am 28. März 2013 von The Channel: http://www.channelregister.co.uk/2011/11/25/disk_drive_pricing/

Legutke, S. (2012). Privatmitteilung.

Miller, H., Sexton, N., Koontz, L., Loomis, J., Koontz, S. R., & Hermans, C. (2011). *The Users, Uses, and Value of Landsat and Other Moderate-Resolution Satellite Imagery in the United States—Executive Report.* Abgerufen am 13. Dezember 2012 von USGS: http://pubs.usgs.gov/of/2011/1031/pdf/OF11-1031.pdf

Pira. (30. Oktober 2000). *Commercial exploitation of Europe's public sector information.* (D. G. European Commission, Hrsg.) Abgerufen am 12. Dezember 2012 von European Public Sector Information (PSI) Platform:

http://epsiplatform.eu/content/commercial-exploitation-europe-s-public-sector-information-pira-study

Rathmann, T. (23. April 2013). *Preise, Kosten und Domänen.* http://dx.doi.org/10.2312/RADIESCHEN_006

Rosenthal, D. (22. Januar 2013). *Talk at IDCC2013.* Abgerufen am 28. März 2013 von DSHR's Blog: http://blog.dshr.org/2013/01/talk-at-idcc2013.html

te Velde, R. (2009). *Public Sector Information: Why Bother?* Abgerufen am 13. Dezember 2012 von The National Academic Press: http://www.nap.edu/openbook.php?record_id=12687&page=25

TIB und TNS Infratest Business Intelligence. (2010). *Die TIB - Zukunft mit Mehrwert.* Abgerufen am 13. Dezember 2012 von Technische Informationsbibliothek: http://www.tib-hannover.de/de/die-tib/aktuelles/aktuelles/id/181/

White, F. M. (15. Februar 2011). *The Potential Cost of High-Throughput Proteomics.* Abgerufen am 21. August 2012 von Science: http://stke.sciencemag.org/cgi/content/full/sigtrans;4/160/pe8

Wittenburg, P. (Januar 2010). *WG2-9 Cost Estimations.* Von Steven Krauwer, Universiteit Utrecht: http://www-sk.let.uu.nl/u/D2R-9a.pdf abgerufen

Wittenburg, P. (2012). Privatmitteilung.

Wittenburg, P., Drude, S., & Broeder, D. (2012). Psycholinguistik. In H. Neuroth, S. Strathmann, A. Oßwald, R. Scheffel, J. Klump, & J. Ludwig (Hrsg.), *Langzeitarchivierung von Forschungsdaten* (S. 93). Boizenburg: Verlag Werner Hülsbusch.

Wittenburg, P., Váradi, T., & Tadić, M. (2008). Budapest Meeting of the Alliance for Permanent Access. Budapest.

Anhang: Interviewfragen zum Thema Kosten

Eingangsfragen

- Welche Stationen im Lebenszyklus von Daten — Datenerzeugung, Auswahl, Ingest, Speicherung, Bereitstellung — werden von Ihrer Institution angeboten und welche davon führt Ihre Institution selbst durch?
- Wie viele Personen sind daran beteiligt?

Kernfragen

- Was für Daten werden bei Ihnen wie erzeugt? Wie viele Personen sind daran beteiligt? Wie vielen Stellen entspricht das allein für die Datenerzeugung?
- Wie selektieren Sie Daten vor dem Ingest? Wie viele Personen sind daran beteiligt? Wie vielen Stellen entspricht das allein für die Datenauswahl?
- Welche Datenmengen werden pro Jahr zum Ingest ausgewählt? Wie vielen Ingest-Vorgängen entspricht das?
- Aus welchen Schritten besteht Ihr Ingest? Wie viele Personen sind daran beteiligt? Wie vielen Stellen entspricht das allein für den Ingest?
- Welche Dienstleistungen bieten Sie im Bereich Speicherung an (z.B. Datendokumentation, Langzeitarchivierung, Datenpublikation, Datenpflege)? Wie viele Personen sind daran beteiligt? Wie vielen Stellen entspricht das allein im Bereich Speicherung?
- Welche Möglichkeiten des Zugriffs und Postprocessing werden angeboten bzw. wie werden Daten bereitgestellt? Wie viele Personen sind daran beteiligt? Wie vielen Stellen entspricht das?
- Haben Sie Lizenzgebühren zu tragen? In welchen Bereichen?
- Welche Anschaffungskosten hatten Sie für Ihre aktuelle Hardware-Ausstattung? Wie häufig wird Ihre Hardware erneuert? Wie häufig Ihre Speichermedien?
- Welche anderen Kosten haben Sie zu tragen (z.B. Fortbildung, Gebäude)?

Weitere Fragen je nach Verlauf des Gesprächs