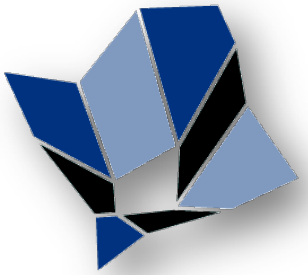

Projekt RADIESCHEN

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

Preise, Kosten und Domänen

Entspricht dem Report D4.3 „LZA-Kostenstruktur“ nach Projektantrag

Torsten Rathmann

DFG Deutsche Forschungsgemeinschaft

Inhalt

1. Einleitung

Wachsende Mengen an Forschungsdaten verbessern die Datenbasis für die Wissenschaft. So erfreulich die Datenflut ist, sie wirft auch einige Probleme auf, und die sind nicht nur technisch-organisatorischer Natur. Auch auf die Frage „Wer bezahlt wie viel für was?“ muss eine Antwort gefunden werden, und zwar in einer Weise, die Bestand hat, denn die Sicherung unseres Wissens ist eine Langzeitaufgabe und Forschungsdaten sind Teil dieses Wissens. Um die Frage beantworten zu können, müssen die Kostenstrukturen von Forschungsdatenarchiven bekannt sein, denn niemand wird bereit sein, für Kosten geradezustehen, deren Höhe völlig offen ist. Dieses Dokument, das im Rahmen des Projektes Radieschen erarbeitet wurde, will einen Beitrag dazu leisten.

In Kapitel 2 dieses Dokumentes wird untersucht, welche Kostenarten eher für große und welche eher für kleine Forschungsdatenarchive von Bedeutung sind. Grundlage dafür ist das erweiterte Domänenmodell, das auch schon in anderen Radieschen-Dokumenten verwendet wurde. Im Kapitel 3 geht es um Preise für Archivdienstleistungen. Durch Bepreisung wird eine Teilantwort auf die Frage „Wer bezahlt wie viel für was?“ gegeben, indem Auftraggeber für die Speicherung oder Nutzer an der Finanzierung des Archivbetriebs beteiligt werden. Kapitel 4 hat übergeordnete Forschungsdateninfrastrukturen zum Thema. Es werden einige Beispiele für Virtuelle Forschungsumgebungen aus dem Grid-Bereich gegeben.

Im Projekt Radieschen wurde kein neues Kostenmodell entwickelt, weil es hierzu bereits etliche Veröffentlichungen gegeben hat. Allein an Kostenmodellen in englischer Sprache sind mindestens neun bekannt (Jackson & Wheatley , 2012). In deutscher Sprache sind die KoLaWis-Studie (Dickmann, 2009) und das DP4lib-Kostenmodell (Kostenmodell für einen LZA-Dienst, 2012) erschienen. Außerdem hätte das neue Kostenmodell im Rahmen von Radieschen nicht getestet werden können.

Statt immer neuer Modelle wird viel mehr ein Referenzmodell gebraucht[1]. In den vorhandenen Modellen werden Begriffe nicht in einheitlicher Art und Weise verwendet. Für Vergleiche ist aber eine einheitliche Nomenklatur erforderlich. Ein solches Referenzmodell sollte für möglichst viele Fachrichtungen und Anwendungsfälle eine brauchbare Arbeitsgrundlage sein. Die Anforderungen an Archive sind dabei ganz unterschiedlich. Gute wissenschaftliche Praxis setzt nur Beweissicherung voraus. So beschränkte Anforderungen kann schon ein Datensarkophag (dark archive) erfüllen, auch wenn Radieschen diese Art von Archiven für Forschungsdaten nicht empfiehlt, weil in einen Datensarkophag immer nur Daten hineingehen, ein Zugriff auf die Daten aber nicht vorgesehen ist. Eine Nachnutzung der Daten durch Dritte ist damit ausgeschlossen. Ganz anders Virtuelle Forschungsumgebungen: Diese ermöglichen über den einfachen Datenzugang hinaus auf vielfältige Weise sogar eine Weiterverarbeitung von Daten innerhalb der Umgebung.

Selbst die Anforderungen an die Bitstream Preservation sind ganz unterschiedlich. Die niederländische DANS (Data Archiving and Network Services) hatte eine Web-Umfrage mit zufällig ausgewählten Teilnehmern durchgeführt (Dillo & Doorn, 2011). Dabei wurde auch gefragt, wie lange Forschungsdaten in Ihrer Fachgemeinschaft brauchbar sind. Am häufigsten wurde ein Zeitraum von 6-10 Jahren angegeben, aber 16 % der Befragten antworteten ≥50 Jahre.

[1] Dies wurde auch auf dem Knowledge Exchange Workshop „Costs and Benefits of Keeping Knowledge” am 11. Juni 2012 in Kopenhagen gefordert.

2. Kostenarten und Domänenmodell

Die bestehende Forschungsdateninfrastruktur kann in die Domänen

- Private Domäne des Wissenschaftlers (Privatarchive)
- Gruppendomäne (Archive auf Arbeitsgruppen- oder Projektebene)
- Öffentliche, dauerhafte Domäne (Archive für die gesamte Disziplin und disziplinübergreifende Archive)
- Zugangsdomäne

eingeteilt werden. Die Domänen privat bis öffentlich sind meist zugleich ihre eigene Zugangsdomäne. Einer der Gründe, weswegen das Domänenmodell um eine eigene Zugangsdomäne erweitert wurde, ist die zunehmende Zahl an Datenzugängen außerhalb der datenhaltenden Institutionen. Ein Beispiel dafür ist C3Grid, in dem sich mehrere Klimadatenarchive und Compute-Provider zusammengeschlossen haben und Daten über das C3Grid-Portal anbieten, zusätzlich zum jeweiligen hauseigenen Datenzugang. Es ist aber nicht unbedingt so, dass über solche zusätzlichen Datenzugänge Zugriff auf alle Daten besteht, die über die hauseigenen Datenzugänge verfügbar sind. C3Grid beispielsweise ist noch im Aufbau und die Auswahl entsprechend beschränkt.

Tabelle 1 enthält in der linken Spalte eine Aufstellung von Kostenarten, die bei der Archivierung und Bereitstellung von Forschungsdaten häufig auftreten. Die Kostenarten wurden überwiegend dem nestor-Handbuch (Wollschläger & Dickmann, 2010) entnommen. Die übrigen Spalten geben darüber Auskunft, ob die Kostenarten für die jeweilige Domäne Bedeutung besitzen. Die Einträge in diesen Spalten beruhen auf der Definition der Domänen, auf Ergebnissen aus Radieschen-Interviews und auf naheliegenden Vermutungen.

Die Zugangsdomäne wird als von den anderen Domänen abgetrennt betrachtet. Das soll auch dann so sein, wenn Archiv und Zugangsdomäne zu einer Institution gehören. Dementsprechend ist die Erhebung von Zugang und gewünschten Zugriffsoptionen für digitale Materialien im eigenen Haus allein Sache der Zugangsdomäne.

In der privaten Domäne arbeitet das unmittelbare Forschungsteam mit seinen Daten und produziert Ergebnisse (Treloar & Harboe-Ree, 2008). Die private Domäne besteht personell nicht notwendigerweise nur aus einem einzelnen Wissenschaftler. Viel mehr ist die personelle Zusammensetzung spezifisch für die Institution oder sogar Arbeitsgruppe. Der privaten Domäne werden vermutlich zum Teil Zugriff auf Personal und technische Ausstattung der Institution gestattet werden.

Ergebnis aus den Radieschen-Interviews ist, dass sowohl in der Gruppendomäne als auch in der öffentlichen Domäne stets Personalkosten und, damit verbunden, Weiterbildungs- und Reisekosten angefallen sind, auch bei kleinen Forschungsdatenarchiven. Dasselbe gilt ebenso für Hardware-Kosten. Gebäudekosten und die Kosten für Strom, Kühlung und Datenleitungen werden dagegen i.d.R. von der jeweiligen Institution übernommen. Vor allem im Zuge neuer Projekte wird häufig Software entwickelt mit den entsprechenden Entwicklungskosten. Über Kosten der privaten Domäne konnten aus den Radieschen-Interviews keine Erkenntnisse gewonnen werden.

	Private Domäne	Gruppen-domäne	Öffentliche Domäne	Zugangs-domäne
Initiale Kosten				
Informationsbeschaffung über LZA-Systeme	Ja, z.B. sollte im Förderantrag be-gründet sein, warum nicht ein anderes Archiv die Aufgabe übernehmen kann			Nein
Erhebung von Bestand, Zugang und gewünschten Zugriffsoptionen für digitale Materialien im eigenen Haus	Nur Bestandserhebung, weil Zugang eigene Domäne ist			Ja
Erhebung von vorhandenen Personal- und Technikressourcen im eigenen Haus	Ja, z.B. um im Förderantrag die Eigenleistung quantifizieren zu können			
Projektplanung, ggf. Consulting, Ausschreibungen	?	Ja	Ja	Ja
Beschaffungskosten				
Hardware: Speichersysteme und sämtliche infrastrukturellen Ein-richtungskosten (Serveranbindun-gen, Datenleitungen, Mitarbeiter-rechner, ...)	Wenn größer	Ja	Ja	Ja
Lizenz(en) für Software-Systeme oder Beitrittskosten zu Konsortien	Nicht bei freier Software, ansonsten von Fall zu Fall			
Einstellung neuer Mitarbeiter	?	Ja	Ja	Ja
Entwicklungskosten	?	Meist ja		
Betriebskosten				
Personalkosten	?	Ja	Ja	Ja
Kosten für Weiterbildung, Tagungen und Reisekosten zwecks Informationsaustausch	Ja	Ja	Ja	Ja
Laufende Storage-Kosten	Wenn größer		Ja	Wenn größer
Sonstige Dauerbetriebskosten: z.B. Strom, Kühlung, Datenleitungs-kosten, Sicherheitsmaßnahmen, Backups, regelmäßige Wartungen und Tests, Software-Upgrades	Haus übernimmt oft Teil der Kosten			
Zukauf von weiteren Speichereinheiten	Wenn größer		Ja	?
Hard- und Software-Komplettersatz in Intervallen	?	Ja	Ja	Ja
laufende Lizenzkosten und/oder laufende Beitragszahlungen bei Konsortien	Nicht bei freier Software, ansonsten von Fall zu Fall			
Anteilige Gebäudekosten: Serverraum, Arbeits- und Nebenräume für Mitarbeiter, Heizung, Hausmeister	Haus übernimmt die Kosten			
Verwaltung	?	Ja	Ja	Ja

Tabelle 1: Kostenarten und deren Bedeutung für die Domänen

3. Preispolitik

Nicht jedes Forschungsdatenarchiv wird in einem Ausmaß öffentlich gefördert oder kann über so hohe Hausmittel verfügen, dass es ganz auf Leistungsentgelte verzichten kann. Häufig ist die Förderung zeitlich begrenzt und nur als Anschubfinanzierung gedacht, so dass Leistungsentgelte eine Möglichkeit sind, nach Ende der Förderung die für den Fortbestand des Archivs erforderlichen Einnahmen zu realisieren.

Natürlich sind Preise auch die Grundlage für die Abrechnung von Datendienstleistungen durch das Archiv gegenüber den Kunden. Aber Preise erfüllen noch mindestens eine weitere wichtige Funktion.

Förderinstitutionen wie z.B. die DFG gehen mehr und mehr dazu über, von Forschern Auskunft darüber zu verlangen, welche Maßnahmen zur Sicherung der Forschungsdaten ergriffen werden (DFG-Vordruck 54.01, 2012). Umgekehrt werden mit dem Projektantrag auch Gelder für die langfristige Datenhaltung genehmigt. Voraussetzung für die Genehmigung solcher Mittel ist aber deren Beantragung. Damit die Antragsteller Mittel in der später benötigten Höhe beantragen können, müssen die Preise bekannt sein, und das nicht erst wenn die Daten da sind, sondern schon zum Zeitpunkt der Antragstellung. So gesehen sind rechtzeitig bekannte Preise Voraussetzung für die Beantragung von Forschungsförderung in sinnvoller Höhe.

Wer zahlt nun was? Es können Dienstleistungen bepreist werden, die von Nutzern der Daten nachgefragt werden, oder solche, die von Auftraggebern für die Speicherung nachgefragt werden. Auftraggeber für die Speicherung sind meist die Datenproduzenten.

Dafür, die Datenproduzenten zahlen zu lassen, spricht, dass

- Datenproduzenten die Mittel hierfür vom Geldgeber bekommen, sofern diese, wie es sein soll, mit beantragt worden sind,
- Datenproduzenten vom Geldgeber zur Speicherung verpflichtet werden und sich somit schlecht wehren können,
- kostenorientierte Preise aus den Kosten des Ingest und der Speicherung hergeleitet werden können und die Preise daher auf einer soliden Grundlage stehen.

Dagegen, die Datenproduzenten zahlen zu lassen, spricht, dass

- die ihre Daten hergeben und mit den Daten schon viel Arbeit hatten und haben: die Produktion der Daten, einen Teil der Qualitätskontrolle und die Bereitstellung von Metainformationen über die Daten.

Dafür, die Datennutzer zahlen zu lassen, spricht, dass

- die den Nutzen haben. Ist der Nutzen im Falle kommerzieller Nutzung auch noch finanzieller Natur, ist eine angemessene Beteiligung sicherlich gerechtfertigt.

Dagegen, die Datennutzer zahlen zu lassen, spricht, dass

- mit Steuergeldern finanzierte Daten grundsätzlich offen zugänglich sein sollten, sofern dadurch keine Persönlichkeitsrechte (Datenschutz) verletzt werden,
- freier Zugang zumindest in einigen Bereichen zusätzliches Wirtschaftswachstum nach sich zieht und damit auch höhere Steuereinnahmen (Houghton, 2011),
- die Nutzer im Falle einer Veröffentlichung die wissenschaftliche und organisatorisch-technische Leistung der Datenbereitstellung bereits in Form eines Zitats würdigen (sollten),

- Preise schlecht aus den Kosten hergeleitet werden können, weil die Zahl der späteren Nutzer, durch die die Gesamtkosten geteilt werden sollten, möglicherweise schwer zu schätzen ist.

Datenzugang nur gegen Bezahlung könnte die Zielsetzung von Forschung mehr hin zu Vorhaben lenken, die mit großer Wahrscheinlichkeit Ergebnisse versprechen. Forschung im Sinne von Ausprobieren ohne große Hoffnung auf Ergebnisse wird durch zusätzliche Kosten erschwert. Eine stärker auf Effizienz ausgerichtete Forschung kann von Vorteil sein, kann aber auch Nachteile haben. Manch neuer Forschungsansatz lässt sich hinsichtlich seiner Brauchbarkeit in der rauen Praxis erst nach Ausprobieren mit Daten beurteilen, wobei das Ergebnis offen ist. Dafür ist dann vielleicht kein Geld da.

Die Diskussion zeigt, dass Preise für Datendienstleistungen in jedem Falle problematisch sind und eine ausreichende Förderung der Forschungsdatenarchive besser wäre. Wenn jedoch die Förderung nicht ausreicht, können angemessene Preise helfen, die Existenz des Archivs zu sichern.

3.1. Berechnungsgrundlage für Preise einzelner Dienstleistungen

Da es hier um Datenarchive für die öffentlich geförderte Forschung geht, sollten sich alle Preise an den Kosten orientieren, wenn die Leistung schon nicht kostenlos zu haben ist. Dies ist auch die Auffassung des Wissenschaftsrats, der in seinen Empfehlungen (Wissenschaftsrat, 2012) schreibt:

Für die wissenschaftliche Nutzung von Informationsinfrastrukturen sowie für ihre nichtkommerzielle Nutzung durch angrenzende gesellschaftliche Bereiche sollten in der Regel keine oder geringe Gebühren zur Deckung des Aufwandes anfallen. Dies schließt Budgetierungsmodelle und Verfahren nicht aus, die die nachhaltige Nutzbarkeit der Ressourcen sichern.

Die Frage ist, an welchen Kosten sich der Preis orientieren soll. Infrage kommen z.B. ein Prozentanteil an den Forschungskosten oder Archivierungs- bzw. Bereitstellungskosten. Archivierungs- und Bereitstellungskosten können wiederum auf die Zahl der Datensätze oder das Datenvolumen bezogen sein. Bei Extraleistungen wie Rechenzeit im Zuge der Nachverarbeitung oder Datenveröffentlichungen können Preise z.B. nach der verbrauchten Wall-Clock-Time oder nach der Zahl der übernommenen Datenveröffentlichungen bemessen sein.

Ein Prozentanteil an den Forschungskosten ist nicht als Berechnungsgrundlage für Preise von Speicherdienstleistungen geeignet, denn sonst müssten Datenproduzenten umso mehr zahlen, je mehr ihre *eigene* Forschung gekostet hat. Geeignet sind Forschungskosten aber als Berechnungsgrundlage für die Preise von Kopien für *Nutzer*, insbesondere dann, wenn

- es sich um kommerzielle Nutzung handelt und
- die Daten oder ein Mehrwert auf den Daten an derselben Institution produziert worden sind, die auch das Archiv betreibt

Forschungskosten sind in der Regel sehr viel höher als alle Archivierungs- und Bereitstellungskosten. Im Falle kommerzieller Nutzung kann eine Beteiligung des Nutznießers an den höheren Forschungskosten aber durchaus angemessen sein.

Meist werden Archivierungs- oder Bereitstellungskosten als Berechnungsgrundlage für Preise herangezogen, weil diese die Kosten des Archivs widerspiegeln. Der Preis kann

z.B. eine Funktion der Zahl der Datensätze n, des Datenvolumens V und der Extraleistungen (z.B. Rechenzeit, DOI-Vergabe) sein.

$$P(n, V, \text{Extra})$$

An die Stelle der Zahl der Datensätze kann auch eine andere Größe treten, die die logische Gliederung der Daten quantitativ beschreiben kann. Die Preisfunktion sollte eine solche Größe enthalten, weil der Aufwand beispielsweise beim Ingest in erheblichem Maße von der Zahl der logischen Datenkomponenten abhängt. Jede logische Datenkomponente hat i.d.R. ihre eigenen Metadaten und wird einer eigenen Qualitätskontrolle unterzogen. Dass eine rein volumenabhängige Abrechnung ungeeignet ist, hat z.B. eine Schätzung der Archivierungskosten am Deutschen Klimarechenzentrum (DKRZ) ergeben. Dies wird im Fallbeispiel unten noch genauer untersucht. Die Beobachtungen am DKRZ decken sich mit denen am britischen Archaeology Data Service, wo die zwölf kleinsten Datenkollektionen 88,06 £/MB kosten, die zwölf größten aber nur 1,54 £/MB (jeweils Median) (Beagrie, Lavoie, & Woollard, Keeping Research Data Safe 2, 2010).

Hängt der Aufwand deutlich von der Art der Daten oder von unterschiedlichen Nutzerwünschen ab, sollte auch das in der Preisfunktion berücksichtigt werden. Art der Daten und Nutzerwünsche lassen sich aber gewöhnlich nur schwer in eine analytische Preisfunktion übertragen. Stattdessen sollten Leistungsstufen definiert werden, die z.B.

- die unterschiedliche Komplexität der Daten oder
- Service-Level-Agreements (SLAs), unter denen der Kunde wählen kann,

beschreiben. Existieren N Leistungsstufen, werden Datensätze und Volumina des Auftrags auf die Leistungsstufen 1, 2,…, N verteilt. Zahl der Datensätze n und Datenvolumen V sind dann Vektoren.

$$n = (n_1, \dots, n_N)$$

$$V = (V_1, \dots, V_N)$$

Damit die Preise für Kunden transparent sind, sollte sich die Preisfunktion *P* additiv aus Komponentenfunktionen zusammensetzen, die nur noch von höchstens einer Variablen abhängig sind.

$$P(n_1, \dots, n_N, V_1, \dots, V_N, \text{Extra}) = S + \sum_{i=1}^{N} P_{\text{Dataset},i}(n_i) + \sum_{i=1}^{N} P_{\text{Vol},i}(V_i) + P_{\text{Extra}}(\text{Extra})$$

Einen mengenunabhängigen Sockelpreis S anzusetzen, der pro Auftrag genau einmal abgerechnet wird, ist ratsam, weil dann keine mengenunabhängigen Kosten auf mengenabhängige Preiskomponenten verteilt werden müssen. Über den mengenunabhängigen Sockel *S* können alle Kosten eingepreist werden, die unabhängig von der Datenmenge sind, aber häufig in Zusammenhang mit einem Auftrag anfallen, z.B. allgemeine Beratung. Ebenso können dort Bereitschaftskosten wie z.B. Gebäudekosten eingehen.

Für die nur noch von der Zahl der Datensätze n_i abhängigen Funktionen $P_{\text{Dataset},i}$ werden am besten lineare Funktionen angesetzt, sofern keine genauere Abhängigkeit bekannt ist.

$$P_{\text{Dataset},i}(n_i) = P_{\text{D},i} \cdot n_i$$

Solche linearen Funktionen sind auch für Kunden am einfachsten nachvollziehbar. Die Koeffizienten $P_{\text{D},i}$ sind Konstanten, Preise pro Datensatz auf der Leistungsstufe i.

Für die nur noch volumenabhängigen Komponenten $P_{\text{Vol},i}$ wird am besten ebenfalls eine lineare Funktion angesetzt, sofern keine genauere Abhängigkeit bekannt ist.

$$P_{\mathrm{Vol},i}(V_i) = P_{\mathrm{V},i} \cdot V_i$$

Die Faktoren $P_{\mathrm{V},i}$ sind Konstanten, die Preise pro Volumeneinheit auf der Leistungsstufe i. In die $P_{\mathrm{V},i}$ sollten die Kosten für die Bitstream Preservation eingehen. Werden in einem Archiv stets nur kleine Datenvolumina beispielsweise im Megabyte-Bereich angeliefert, kann auf die volumenabhängigen Komponentenfunktionen $P_{\mathrm{Vol},i}$ auch ganz verzichtet bzw. $P_{\mathrm{V},i} = 0$ gesetzt werden.

Zusammengefasst sei die folgende Preisfunktion empfohlen, sofern keine genauere Abhängigkeit bekannt ist:

(1)
$$P(n_1, \ldots, n_N, V_1, \ldots, V_N, \mathrm{Extra}) = S + \sum_{i=1}^{N} P_{\mathrm{D},i} \cdot n_i + \sum_{i=1}^{N} P_{\mathrm{V},i} \cdot V_i + P_{\mathrm{Extra}}(\mathrm{Extra})$$

Die Preisfunktion P_{Extra} für Extraleistungen ist hier nicht näher spezifiziert, kann aber für viele Arten von Extraleistungen auf ähnliche Weise ausgestaltet werden.

3.2. Preisfindung und -gestaltung

Von Forschungsdatenarchiven der Gruppen- und öffentlichen Domäne wird gewöhnlich erwartet, dass Preise kostendeckend, aber nicht höher sind. Von daher sollten die wesentlichen Kosten bekannt sein. Es müssen aber nicht alle Kosten berücksichtigt werden. Forschungsdatenarchive sind nicht dazu verpflichtet, ihre Preise auf Basis einer Vollkostenrechnung festzulegen. Eine Teilkostenrechnung genügt und ist zweckmäßig, wenn darin nicht enthaltene Kosten anderweitig abgedeckt sind, z.B. durch den allgemeinen Haushalt der Institution. Schließlich kostet die Kostenrechnung auch etwas, und die Erfassung aller Kosten in einer Vollkostenrechnung ist teurer als eine Teilkostenrechnung. Selbstverständlich ist es von Vorteil, alle Kosten zu kennen, aber nur zum Zwecke der Bepreisung ist eine Vollkostenrechnung nicht zwingend erforderlich.

Die Kostenrechnung kann deduktiv oder induktiv erfolgen. Bei der deduktiven Methode wird ausgehend von den Gesamtkosten versucht, durch Abgrenzung zu Einzelkosten zu kommen. Der induktive Ansatz geht von den einzelnen Arbeitsschritten aus. Die Einzelschritte werden zu Teilprozessen und schließlich zu Hauptprozessen verdichtet. Die Kosten für die Einzelschritte müssen bekannt sein oder geschätzt werden, damit die Kosten für die Teil- und Hauptprozesse berechnet werden können. Umgekehrt müssen beim deduktiven Ansatz die prozentualen Anteile bekannt sein, um die Abgrenzung durchführen zu können.[2]

Insgesamt sollte die Preisgestaltung möglichst einfach und übersichtlich sein, damit sie von Kundenseite gut nachvollzogen werden kann. Außerdem sollten die Risiken beachtet werden, die mit der Bepreisung von Archivdienstleistungen verbunden sind.

Neben dem Risiko eines zu hohen oder zu niedrigen Preises, das mit jeder Preiskalkulation verbunden ist, ist dies vor allem das Risiko sich schnell ändernder Kosten. Deutliche Preissprünge nach oben sind schädlich für die Forschung, da Mittel in ausreichender Höhe für die Archivierung beantragt werden müssen, lange bevor die Daten vorhanden sind. Im Prinzip gehen von den beiden größten Kostenblöcken, den Personal- und den Hardware-Kosten keine besonderen Risiken aus, die nicht auch in anderen Teilen

[2] Beide Ansätze wurden z.B. in einer Kostenrechnung zur Digitalisierung von Herbarien am Botanischen Garten / Botanischen Museum Berlin-Dahlem angewandt (Jaspersen, Wohlfromm, Täschner, & Wendehorst, 2008).

der Forschungsinfrastruktur eingegangen werden müssen. Hardware-Kosten nehmen in der Regel stark ab. Personalkosten wachsen langsam. Risiken lauern jedoch dort, wo sich Anforderungen plötzlich erhöhen und das auf die Preise durchschlägt. Es ist daher wichtig, frühzeitig sicherzustellen, dass alle kostenintensiven Arbeitsschritte den Anforderungen der nächsten Jahre gerecht werden können. Das gilt insbesondere für die Qualitätssicherung und die Kuration.

Regeln für die Festlegung von Preisen auf der Basis der Prozesskostenrechnung können bei DP4lib gefunden werden (Kostenmodell für einen LZA-Dienst, 2012).

3.3. Fallbeispiel: World Data Center for Climate

Am DKRZ wird das ICSU World Data Center for Climate (WDCC) betrieben. Dort sind hauptsächlich numerische Ergebnisse von Klimasimulationen gespeichert. Ein weiterer Schwerpunkt sind Messdaten, und zwar solche, die für die Kalibrierung von Klimamodellen benötigt werden.

Die Kosten für die Archivierung müssen in der Regel von denjenigen getragen werden, die den Auftrag dafür gegeben haben. Das Herunterladen von Daten über das Web ist hingegen kostenfrei, sofern die Daten nur für wissenschaftliche Zwecke verwendet werden. Bisher wurden hauptsächlich Klimadaten von Institutionen archiviert, die zugleich DKRZ-Nutzer sind. In diesem Fall werden die Kosten mit den bewilligten Kontingenten verrechnet, und die Archivierung ist de facto kostenfrei. In letzter Zeit sind die Archivierungsdienste des WDCC auch für externe Kunden interessant geworden, z.B. für Forschungsinstitute, die sich den Aufbau eines eigenen Langzeitarchivs ersparen wollen. Deshalb wurde am WDCC ein Preismodell aufgestellt.

Die Preise (Luthardt, 2010) basieren auf einer Teilkostenrechnung. Berücksichtigt werden die Arbeitszeiten der Mitarbeiter für die Einzelschritte, die zur Abarbeitung eines Auftrags erforderlich sind, sowie Medien- und Dauerbetriebskosten. Alle anderen Kosten, z.B. Entwicklungs-, Fortbildungs- und Gebäudekosten, bleiben unberücksichtigt. Für die Arbeitszeiten wurde die induktive Methode gewählt und die Arbeitszeit gemeinsam mit den Mitarbeitern geschätzt, die mit dem jeweiligen Arbeitsschritt befasst sind. Die Einzelschritte (Luthardt, 2010) wurden zu den Teilprozessen der Tabelle 2 verdichtet.

Ein Teil der Arbeitsschritte fällt nur einmal pro Auftrag an, z.B. Beratung oder die Erstellung eines Konzeptes. Andere Arbeitsschritte, wie z.B. die Qualitätskontrolle der Daten und Metadaten, sind einmal für jedes Experiment vorgesehen. Als Experiment wird in der Klimaforschung eine Simulationsrechnung bzw. deren Output bezeichnet. Jedes Experiment besteht aus Datensätzen. Am WDCC ist ein Datensatz nahezu immer eine Zeitreihe für eine Variable auf einem Höhenlevel, z.B. die Temperatur auf dem 500 mbar-Luftdruck-Level. Zu jedem Zeitschritt der Simulation enthält die Zeitreihe ein zweidimensionales Array von Werten der Variable auf einem Netz von Gitterpunkten. Das Gitter wird von den Ortskoordinaten aufgespannt. Ortskoordinaten sind üblicherweise Längen- und Breitengrade.

In einem zweiten Schritt wurden die Teilprozesse weiter verdichtet zu den Hauptprozessen „Daten- und Metadatenarchivierung“ und „10 Jahre Speicherung inklusive Pflege“ (grau unterlegt in Tabelle 2), die dann bepreist wurden.

Die Kostenschätzung bezieht sich auf Experimente mit 500 Datensätzen, einer für das WDCC typischen Größenordnung. Die Zahl der Datensätze geht jedoch nicht in den Preis ein. Entscheidend ist stattdessen die Zahl der Experimente. In Tabelle 2 werden dabei

zwei Leistungsstufen unterschieden. Die Spalte „Pro Experiment bei gleichen Datenstrukturen“ (Leistungsstufe 2) wird mit herangezogen, wenn ein Auftrag zur Archivierung mehrerer Experimente gegeben wurde, die mit dem gleichen Klimamodell, mit identischem Gitter und den gleichen Klimavariablen (Temperatur, relative Luftfeuchte,…) gerechnet wurden. Dann ergeben sich Synergieeffekte vor allem bei den Metadaten. Gleiche Datenstruktur haben z.B. Experimente mit verschiedenen Szenarien der anthropogenen CO_2-Emission, aber ansonsten gleichen Vorgaben.

	Pro Auftrag	**Pro Experiment**	**Pro Experiment bei gleichen Datenstrukturen**
Daten- und Metadatenarchivierung (Ingest)			
Information und Beratung	4		
Projekt-Spezifikation (Festlegung Datenumfang, Formate, Datenorganisation, Speicherstrategie, Weg der Daten zum WDCC, Data-Policy, Zugriffsbedingungen)	2		
Erstellen eines Konzeptes (Metadatenumfang, Preprocessing, Zeitplan) und Kostenabschätzung	4		
Erfassen, Einfüllen und Qualitätskontrolle der Metadaten	10	5	3
Aufsetzen Datentransfer und Einfüllen der Daten	7	1	1
Qualitätskontrolle der Daten einschl. Prüfung der Konsistenz von Metadaten und Daten		10	4
Freischaltung und Abschluss-Report	6		
insgesamt	33	16	8
10 Jahre Speicherung inklusive Pflege			
Aktualisierung der Metadaten	10	10	8
Pflege der Datensätze innerhalb der Datenbank		10	5
Anpassung der Zugriffsberechtigungen	8	2	2
Laufende Anpassung an die DKRZ-Infrastruktur	10	5	3
insgesamt	28	27	18

Tabelle 2: Personalaufwand am WDCC für Teilprozesse in Personenstunden **(Luthardt, 2010)**

Sollen z.B. fünf Experimente gleicher Datenstruktur archiviert werden, gehen in den Preis

1x Personalaufwand nach Spalte „Pro Auftrag“ ein,
1x Personalaufwand nach Spalte „Pro Experiment“ und
4x Personalaufwand nach Spalte „Pro Experiment bei gleichen Datenstrukturen“

ein. Haben nur vier der fünf Experimente die gleiche Datenstruktur, wird

1x nach Spalte „Pro Auftrag“,

2x nach Spalte „Pro Experiment“ und
3x nach Spalte „Pro Experiment bei gleichen Datenstrukturen“

bepreist. Für die Umrechnung der Arbeitszeit in Euro wird ein Faktor von zurzeit 31,25 €/h verwendet. Abschließend kommen noch die Medien- und Dauerbetriebskosten für das Archiv hinzu, die bisher mit 400 €/TB für 10 Jahre Speicherung angesetzt werden. Dabei sind zwei Medienwechsel berücksichtigt.

Alle diese Größen eingesetzt in (1) ergeben ungefähre Preisformeln für die angebotenen Hauptprozesse, für den Ingest

$$P = 1031{,}25\ € + 500{,}00\ € \cdot n_1 + 250{,}00\ € \cdot n_2$$

und für 10 Jahre Speicherung inklusive Pflege

$$P = 875{,}00\ € + 843{,}75\ € \cdot n_1 + 562{,}5\ € \cdot n_2 + 400\ €/\mathrm{TB} \cdot V$$

Dabei sind P der Preis, n_1 die Zahl der Experimente auf Leistungsstufe 1 (verschiedene Datenstrukturen), n_2 die Zahl der Experimente auf Leistungsstufe 2 (gleiche Datenstrukturen) und V das Datenvolumen. In die genauen Preise, die auf der Rechnung ausgewiesen sind, gehen noch Rundungsschritte ein. Für den Ingest lautet die Preisfunktion

$$P = \mathrm{runde}\left(5000\ €/\mathrm{Monat} \cdot \mathrm{runde}\left(\frac{33\ \mathrm{h} + 16\ \mathrm{h} \cdot n_1 + 8\ \mathrm{h} \cdot n_2}{160\ \mathrm{h/Monat}}\right)\right)$$

und für 10 Jahre Speicherung inklusive Pflege

$$P = \mathrm{runde}\left(400\ €/\mathrm{TB} \cdot V + 5000\ €/\mathrm{Monat} \cdot \mathrm{runde}\left(\frac{28\ \mathrm{h} + 27\ \mathrm{h} \cdot n_1 + 18\ \mathrm{h} \cdot n_2}{160\ \mathrm{h/Monat}}\right)\right)$$

Alle Preise sind netto, d.h. ohne Mehrwertsteuer.

Am DKRZ wird seit 2013 auch die reine Bandspeicherung ohne Ingest und Datenpflege gegen Bezahlung angeboten. Aus diesem Anlass wurden die Medien- und Dauerbetriebskosten neu berechnet. Die Neuberechnung ergab 165 € pro Terabyte und Kopie für die zehnjährige Speicherung. Da im WDCC immer zwei Kopien vorhanden sind, betragen die Kosten dort 330 €/TB für die reine Bitstream-Preservation. Alter und neuer Kostenpreis liegen damit recht nahe beieinander. Auch die tabellierten Arbeitszeiten werden noch überprüft und aktualisiert werden. Auf der einen Seite haben Weiterentwicklungen an der Software zu Einsparungen durch weitergehende Automatisierung geführt. Auf der anderen Seite ist der Aufwand für die Qualitätssicherung gestiegen.

3.4. Sammelpreise für zusammengezogene Leistungen

Datendienstleistungen müssen nicht einzeln abgerechnet werden, sondern können auch zusammengezogen verkauft werden. Eine bemerkenswerte Spezialität ist *Pay once, store forever* (Goldstein & Ratliff, 2010). Hier wird die Speicherung für einen unendlich langen Zeitraum angeboten. Bei der Preisfindung wird ausgenutzt, dass

1. die Preise für Speichermedien exponentiell fallen
2. alle anderen Kosten, wie z.B. die Personalkosten, bezogen auf das Datenvolumen ebenfalls exponentiell fallen, wenn das Archivvolumen exponentiell wächst
3. die Potenzreihe der Kosten unter diesen Umständen konvergiert

Dass die Preise für Speichermedien exponentiell fallen, ist empirisch beobachtet worden und hängt mit dem Kryderschen Gesetz zusammen, einer Faustregel, nach der die

Speicherdichte von Festplatten über Jahrzehnte hinweg exponentiell gewachsen ist. Da sich der Preis für eine Festplatte nicht wesentlich verändert hat, sind die Preise pro Datenvolumen exponentiell gefallen.

Dass die Bedingungen 1 und 2 ewig lange erfüllt sind, ist aber äußerst fraglich. Das Krydersche Gesetz kann voraussichtlich nicht bis in alle Ewigkeit Gültigkeit besitzen, da irgendwann atomare Dimensionen für die Speicherung eines Bits erreicht sind. Schon jetzt sind Abweichungen vom gleichbleibend schnellen, exponentiellen Wachstum und der Übergang zu einem weniger rasanten Wachstum der Festplatten-Speicherdichte absehbar (Rosenthal, 2013). Auch Bedingung 2 ist von der Gültigkeit des Kryderschen Gesetzes abhängig, weil ohne dieses ewiges exponentielles Volumenwachstum nicht realisierbar wäre.

Pay once, store forever funktioniert aber, wenn die Kunden ihre Daten nicht unendlich sondern nur *beliebig* lange im Archiv lassen und die Dauer der Speicherung einer statistischen Verteilungsdichte folgt, die exponentiell oder schneller als exponentiell gegen null geht. Das wird aber im Falle von Forschungsdaten sicher nicht erfüllt sein, da wichtige, nicht wiederherstellbare Daten unbeschränkte Zeit aufbewahrt werden müssen. Deshalb ist *Pay once, store forever* für die Langzeitarchivierung von Forschungsdaten nicht geeignet.

Ein ganz anderer Ansatz der Zusammenziehung von Leistungen ist die *Flatrate*. Diese erlaubt den Zugang zu allen Daten einer bestimmten Kollektion oder eines oder mehrerer ganzer Archive für einen bestimmten Zeitraum. Die Flatrate hat für den gebuchten Zeitraum einen festen Preis. In den Vertragsbedingungen sollte geregelt sein, welche Leistungen in der Flatrate enthalten sind. Das Abonnement ist nicht notwendigerweise auf den Zugang zu Daten beschränkt. Andere Leistungen wie Ingest und Speicherung könnten auch in dieser Form angeboten werden. Wenn die Daten leicht und in großer Menge produziert werden können, muss eventuell durch Vertrag sichergestellt werden, dass das Archiv nicht überfordert wird, z.B. durch eine Mengenbegrenzung.

Die unbefristete Variante der Flatrate ist das *Abonnement*. Wenn es nicht von einer Vertragspartei bis zu einem bestimmten Termin gekündigt wird, verlängert es sich automatisch um einen zuvor festgelegten Zeitraum. Das dem Abonnement zugrundeliegende Preismodell wird in der KoLaWiss-Studie *Versicherungsmodell* genannt und der Abopreis *Prämie* (Dickmann, 2009).

Die *Mitgliedschaft* entspricht in etwa dem Abonnement, setzt aber die Aufnahme des Kunden in eine Organisation, z.B. einen Verein, voraus. Die Mitglieder können in dieser Organisation ihre satzungsgemäßen Rechte ausüben und zahlen einen Mitgliedsbeitrag. Beispielsweise tragen Mitgliedsbeiträge von über 700 Institutionen zur Finanzierung des Forschungsdatenarchivs des Inter-university Consortium for Political and Social Research (ICPSR) an der University of Michigan bei (Lyle, Alter, & Vardigan, 2013).

Die wohl stärkste Bindung an das Archiv bringt das *Miteigentum* mit sich. Die Rechte und Pflichten von Miteigentümern sind je nach Gesellschaftsform unterschiedlich geregelt. Mit dem Erwerb der Eigentumsanteile können z.B. Pflichten wie Haftung und Nachschusspflicht verbunden sein.

Sowohl vom Open Access als auch bei bestehender Flatrate oder Mitgliedschaft können bestimmte Leistungen ausgeschlossen sein und extra kosten. Beispielsweise ist am WDCC und bei GESIS der Datenzugang über das Web kostenfrei, für die Lieferung der Daten auf CD ist aber ein Preis festgelegt.

4. Schaffung einer übergeordneten Forschungsdateninfrastruktur

Eine Vergrößerung der Zahl der Datenkollektionen von 10 auf 60 am britischen National Digital Archive of Datasets hat die Kosten nicht versechsfacht, sondern nur um 325 % steigen lassen (Beagrie, Chruszcz, & Lavoie, Keeping Research Data Safe, A Cost Model and Guidance for UK Universities, 2008). Über die Etablierung eines gemeinsamen Datenzentrums oder einer gemeinsamen Dateninfrastruktur ließe sich dieser Skalierungseffekt vermutlich nutzen. Als einige weitere Beispiele aus der Fülle der Kooperationsmöglichkeiten seien hier die folgenden genannt:

- Suche und Zugriff über ein gemeinsames Portal
- Gemeinsame Nutzung von Entwicklungsergebnissen, z.B. von Postprozessing-Workflows
- Gemeinsame Arbeit an Dokumenten
- Gegenseitiges Backup
- Gemeinsamer Betrieb von Speicher-, Compute- und Netzwerkressourcen
- Kooperative Verwaltung von freier Software

Es gibt also Anzeichen dafür, dass eine kollaborative Dateninfrastruktur Datendienste kostengünstiger anbieten kann. Eine abschließende Antwort auf die Frage, ob und unter welchen Bedingungen über den Zusammenschluss mehrerer Institutionen zu einer virtuellen Forschungsumgebung, z.B. einem Daten-Grid, tatsächlich Kosteneinsparungen realisiert werden können, kann in dieser Studie aber nicht gegeben werden.

Sicher ist jedoch, dass kollaborative Strukturen das Angebot an Datendiensten verbessern und vereinheitlichen können und damit der Forschung dienen können. Fächerübergreifende Strukturen können multidisziplinäre Forschung erleichtern. An dieser Stelle sollen einige Beispiele für übergeordnete Forschungsdateninfrastrukturen gegeben werden, die schon genutzt werden oder kurz vor der allgemeinen Nutzung stehen.

Die Beispiele zeigen, dass schon für verwandte und sogar einzelne Fächer der Aufbau einer leistungsfähigen, übergeordneten Forschungsdateninfrastruktur eine Millioneninvestition ist. Das gilt auch für den Betrieb.

4.1. TextGrid

In der ersten Ausbaustufe war TextGrid eine Arbeitsumgebung für die Geisteswissenschaften, in der Textressourcen gemeinsam editiert, annotiert, analysiert und veröffentlicht werden können. Mit der Komponente TextGrid Repository (TextGrid-Rep) wird eine Infrastruktur für die Datenhaltung zur Verfügung gestellt. Über die Komponente TextGridLab wird eine Vielzahl von Werkzeugen für den gesamten Forschungsprozess, insbesondere für die Erstellung digitaler Editionen, angeboten.

Die erste stabile Version 1.0 wurde im Sommer 2011 zum kostenfreien Download angeboten. Die Implementierung wurde vor allem von Editionsphilologen und Linguisten genutzt. Im Zuge der Erweiterungen, die mit der Version 2.0 vom Mai 2012 verfügbar wurden, nutzen heute auch Kunstgeschichte, klassische Philologie und Musikwissenschaft TextGrid als Virtuelle Forschungsumgebung. Schwerpunkt der gegenwärtigen Förderphase (TextGrid III, 2012-2015) ist die Etablierung eines nachhaltigen Dauerbetriebs (TextGrid, Das Projekt, 2012).

	TextGrid I	TextGrid II	TextGrid III	Summe
insgesamt	2015954	3022979	2866049	7904982
davon Personalkosten	1889804	2586279	1997401	6473484

Tabelle 3: Kosten in Euro für die drei Projektphasen von TextGrid. Förderungszeitraum 9 Jahre

Eine Kostenaufstellung zeigt Tabelle 3. Zusätzlich und nur für TextGrid III wird eine Projektpauschale von 429580 € gezahlt. Dadurch erhöhen sich die Gesamtkosten auf 8334562 €. Alle drei Projektphasen mit einbezogen wird TextGrid neun Jahre lang gefördert. Die Zahlen in Tabelle 3 enthalten nicht nur Aufbau- und Entwicklungskosten, sondern auch Produktionskosten. Im Mittel wurden bzw. werden 926062 €/Jahr für TextGrid ausgegeben.

4.2. CLARIN (Common Language Resources and Technology Infrastructure)

CLARIN soll Geisteswissenschaftlern auf europäischer Ebene einen verbesserten Zugang zu Sprachressourcen bieten. Es handelt sich um eine elektronische Forschungsinfrastruktur, die den Zugriff auf Datenrepositorien und zeitgemäße Software-Werkzeuge mit einschließt. Die folgende Schätzung (Wittenburg, 2010) führt Kostenkomponenten auf, die pro CLARIN-Teilnehmerstaat (bei Ressourcen- und Service-Center-Kosten pro Zentrum) anfallen:

	Kosten in Euro/Jahr	Kostenart/-entwicklung	Personal (Stellen)
Ressourcen und Service-Center	500000	Betriebskosten pro Zentrum	7
Ressourcen- und Tool-Integration	600000 → 200000	Kosten werden fallen bis auf ein Basisniveau	10 – 2
Infrastruktur	800000 → 400000	Anfangs höhere Kosten	12 – 7
Support	0 → 200000	Kosten werden steigen mit zunehmender Nutzerzahl	0 – 2
Training und Fortbildung	200000	Etwa konstant	2
Koordination und Management	100000	Konstant	1½
Externe Dienste und Lizenzen	100000	Konstant	—
insgesamt	2300000 → 1700000		

Tabelle 4: CLARIN-Kostenprognose für teilnehmende Länder einschließlich. zeitlichem Verlauf, Kostenprognose für Ressourcen und Service-Center pro Zentrum

4.3. C3Grid (Collaborative Climate Community Data and Processing Grid)

Im C3Grid kooperieren mehrere Archive und Compute-Provider, um eine Plattform für den Zugang zu Klimadaten aufzubauen. Über ein einheitliches Web-Portal können Daten gesucht, zugeschnitten (eine geographische und zeitliche Auswahl getroffen werden) und heruntergeladen werden. Außerdem sind im C3Grid etliche Postprocessing-Workflows verfügbar, mit denen Klimadaten weiterverarbeitet werden können.

In der ersten Projektphase (September 2005 bis August 2008) wurde ein Prototyp aufgebaut. Im Nachfolgeprojekt C3-INAD[3] (Oktober 2010 bis September 2013) werden Funktionalität und verfügbare Datenmenge deutlich erweitert. Ziel ist die Aufnahme der Produktion. Beide Projektphasen kosten zusammen etwa 6 Mill. €.

5. Schlussfolgerungen

1. Auf die Frage „Wer bezahlt wie viel für welche Forschungsdaten-Dienstleistung?“ muss eine nachhaltige Antwort gefunden werden. Das betrifft sowohl die Förderpolitik als auch die Archive selbst, die darüber entscheiden müssen, ob und in welcher Höhe Preise vom Nutzer oder vom Dateneinsteller verlangt werden. Solche Preise für Datendienstleistungen sind in jedem Falle problematisch. Wenn sie denn verlangt werden, müssen diese wenigstens frühzeitig bekannt sein, damit Forscher Mittel in der erforderlichen Höhe im Rahmen der Forschungsförderung beantragen können.
2. Falls einzelne Forschungsdaten-Dienstleistungen mit einem Preis versehen werden müssen, sei hier eine affine Preisfunktion empfohlen. Diese besteht aus einem konstanten Sockelbetrag, in dem mengenunabhängige Kosten berücksichtigt werden, und einer Summe linearer Teilfunktionen. Der Preis sollte linear mit der Zahl der Datensätze ansteigen. An die Stelle der Zahl der Datensätze kann auch eine andere Größe treten, die die Zahl der logischen Dateneinheiten wiedergibt. Wenn die Datenvolumina größer sind, sollte als zweite Variable das Datenvolumen in die Preisfunktion aufgenommen werden, ebenfalls in Form einer linearen Teilfunktion. Eine affine Preisfunktion lässt sich leicht an erfasste oder geschätzte Kosten anpassen und besitzt darüber hinaus den Vorteil, transparent und nachvollziehbar für die Kunden zu sein.
3. Anstelle von Einzelpreisen können Flatrate, Abo, Mitgliedschaft oder Miteigentum angeboten werden.
4. Statt immer neue Kostenmodelle aufzustellen sollte der Fokus auf der Schaffung eines Referenzmodells liegen. In den vorhandenen Modellen werden Begriffe nicht in einheitlicher Art und Weise verwendet. Für Vergleiche ist aber eine einheitliche Nomenklatur erforderlich.
5. Über die Kosten der privaten Domäne ist wenig bekannt. Auch die ganz großen Strukturen sollten weiter untersucht werden, damit die Dateninfrastruktur der Zukunft, die es mit Datenvolumina im Maßstab Petabyte pro Tag zu tun haben wird, von der Kostenseite her geplant werden kann.

[3] INAD steht für „towards an INfrastructure for general Access to climate Data”.

6. Ausblick

Wenn Forscher, Archive und Förderer mit gutem Willen an den Ausbau der Forschungsdateninfrastruktur gehen, könnte der Blick ins Jahr 2020 etwa so aussehen.

Für Forscher ist es selbstverständlich geworden, die eigenen Forschungsdaten bei einem Archiv zu speichern und so anderen die Nachnutzung zu ermöglichen. Forscher des Jahres 2020 erzeugen die zu den Daten gehörigen Metainformationen gleich mit und helfen so den Archiven, die Kosten im Griff zu behalten. Dabei werden die Forscher von Archiven, kollaborativen Strukturen und Geräteherstellern durch geeignete Software und Beratung unterstützt.

Die jahrelange schwierige Suche der Fachdisziplinen nach einheitlichen Metadatenstandards und Vokabularen trägt allmählich Früchte. Viele Fachdisziplinen haben sich auf entsprechende Normen verständigt, andere ringen noch um eine Lösung. Einheitliche Datenstrukturen vereinfachen Software-Entwicklung, Qualitätskontrolle und nicht zuletzt die Suche in Metadaten und Daten und helfen, die Kosten zu senken.

Forschungsdatenarchive verstehen sich wie Bibliotheken und Verwaltung als Dienstleister für die Wissenschaft. Sie achten auf Kosteneffizienz und entwickeln ihre Dienste den Anforderungen entsprechend weiter. Vergleichbare Kostenmodelle, ein Controlling und disziplinübergreifende Archivierungssoftware unterstützen sie dabei.

Die Förderer haben einen Teil der Forschungsförderung umgewidmet und stellen diesen Teil jetzt dauerhaft zur Finanzierung der Forschungsdateninfrastruktur zur Verfügung. Dadurch lassen sich Langzeitarchivierung und Datenpflege nun endlich langfristig planen.

Der gute Wille und die gedeihliche Zusammenarbeit aller Beteiligten werden dringend gebraucht, denn auch im Jahr 2020 sind nicht alle Probleme gelöst. Die Menge an Forschungsdaten wächst ständig weiter und hat ein noch nie dagewesenes Maß erreicht. Die an sich erfreuliche Datenflut stellt die vorhandene Infrastruktur vieler Fachdisziplinen auf eine harte Probe.

Automatische Sensorsysteme liefern unablässig Daten: Umweltdaten, astronomische Daten, medizinische Daten, Wirtschafts- und Verbrauchsdaten, z.B. die des intelligenten Stromnetzes, des „*smart grid*“. Aber auch Messgeräte, die vom Menschen ausgelöst werden müssen, erzeugen riesige Datenmengen, vom Großgerät bis zum Mini-Sequenzierer im Datenstickformat. Soziale Netzwerke quellen über vor Daten, die nicht nur für die Sozialwissenschaften interessant sind. Durch Weiterverarbeitung von Daten entstehen wieder neue Daten. Die zunehmende Verknüpfung von Daten lässt die Datenmengen ebenfalls anschwellen. Durch die Verknüpfung wird die Speicherung auch solcher Daten notwendig, die an sich wiederherstellbar wären, aber natürlich möchte niemand auf die Wiederherstellung warten.

Die Datenflut wird kommen. Sie wird uns nicht wegspülen. Es besteht aber die Gefahr, dass aus Kostengründen nicht alle Forschungsdaten, die aufbewahrt werden müssten, auch aufbewahrt werden können. Dieses Problem kann nur gelöst werden, wenn alle Beteiligten vertrauensvoll zusammenarbeiten. Innerhalb der Fachdisziplinen sollten sich die Wissenschaftler auf effiziente Datenstrukturen, Metadaten und ein gemeinsames Vokabular einigen. Datenarchive müssen ihre Workflows weiter automatisieren und Software gemeinsam nutzen. Auch die Forschungsförderung ist gefragt, denn ganz ohne eine bessere Förderung der Forschungsdateninfrastruktur wird es nicht gehen. Angesichts der Datenmengen wird es teurer werden. Durch Effizienzsteigerung bei den Archiven allein werden die Datenmengen nicht zu bewältigen sein.

Literaturverzeichnis

DFG-Vordruck 54.01. (Oktober 2012). Abgerufen am 6. Februar 2013 von Deutsche Forschungsgemeinschaft: http://www.dfg.de/formulare/54_01/54_01_de.pdf

Kostenmodell für einen LZA-Dienst. (Mai 2012). Abgerufen am 7. Februar 2013 von DP4lib: http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-Kostenmodell_eines_LZA-Dienstes_v1.0.pdf

TextGrid, Das Projekt. (2012). Abgerufen am 31. Januar 2013 von TextGrid: http://www.textgrid.de/ueber-textgrid/projekt/

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

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

Dillo, I., & Doorn, P. (2011). *The Dutch data landscape in 32 interviews*. Abgerufen am 6. Februar 2013 von DANS: http://www.dans.knaw.nl/sites/default/files/file/publicaties/The_Dutch_Datalandscape_DEF.pdf

Goldstein, S. J., & Ratliff, M. (27. August 2010). *DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data.* Abgerufen am 6. Februar 2013 von Princeton University: http://dspace.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf

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

Jackson, A., & Wheatley , P. (19. Dezember 2012). *Digital Preservation and Data Curation Costing and Cost Modelling*. Abgerufen am 1. Februar 2013 von OPF Knowledge Base Wiki: http://wiki.opf-labs.org/display/CDP/

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

Luthardt, H. (22. November 2010). *Datenspeicherung und Verteilung von Projektdaten am DKRZ (≥ 10 Jahre).* Abgerufen am 16. Januar 2013 von DKRZ: https://www.dkrz.de/daten-en/long_term_archiving/LZA_Kostenabschaetzung_generell_v05a.pdf/at_download/file

Lyle, J., Alter, G., & Vardigan, M. (17. Januar 2013). *“The Price of Keeping Knowledge” Workshop: ICPSR Position Paper.* Abgerufen am 7. Februar 2013 von Knowlegde Exchange: http://www.knowledge-exchange.info/Default.aspx?ID=571

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

Treloar, A., & Harboe-Ree, C. (2008). *Data management and the curation continuum: how the Monash experience is informing repository relationships.* Abgerufen am 14. März 2013 von vala: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

Wissenschaftsrat. (13. Juli 2012). *Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020.* Abgerufen am 6. Februar 2013 von http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf

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

Wollschläger, T., & Dickmann, F. (2010). Kosten. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, & K. Huth (Hrsg.), *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* Göttingen, Niedersachsen: Neben der Online Version 2.3 ist eine Printversion 2.0 beim Verlag Werner Hülsbusch, Boizenburg erschienen.