Handbuch Forschungsdatenmanagement

Herausgegeben von
Stephan Büttner, Hans-Christoph Hobohm, Lars Müller

BOCK + HERCHEN Verlag
Bad Honnef
2011

Die Inhalte dieses Buches stehen auch als Online-Version zur Verfügung:
www.forschungsdatenmanagement.de
Die Onlineversion steht unter folgender Creative-Common-Lizenz:
„Attribution-NonCommercial-ShareAlike 3.0 Unported“
http://creativecommons.org/licenses/by-nc-sa/3.0/

ISBN 978-3-88347-283-6

BOCK+HERCHEN Verlag, Bad Honnef
Printed in Germany

2.6 Langzeiterhaltung digitaler Forschungsdaten

Jens Klump
HelmholtzZentrum Potsdam Deutsches GeoForschungsZentrum GFZ – Zentrum für GeoInformationstechnologie (ZeGIT)

2.6.1 Herausforderung Langzeiterhaltung

Digitale Forschungsdaten sind sowohl die unverzichtbare Basis als auch das Resultat wissenschaftlicher Arbeit, wie sie sich seit dem 17. Jahrhundert entwickelt hat. Jede Disziplin hat ihre eigenen technischen und inhaltlichen Ansprüche, ihre Datenbestände und Konventionen, so dass Forschungsdaten alle Formate und Medientypen umfassen: von einfachen Tabellen im Textformat für z. B. physikalische Messdaten über semantisch ausgezeichnete Textsammlungen der Literaturwissenschaften bis hin zu interaktiven 3D-Modellen von Ingenieuren (Altenhöner et al., 2009).

Allen ist aber gemeinsam, dass sie eine wertvolle, jedoch schwierig zu handhabende Ressource darstellen. Forschungsdaten werden aufwendig produziert, im Falle von Beobachtungsdaten wie z. B. Wetterbeobachtungen können sie einmalig und nicht wiederherstellbar sein, sie können riesige Datenmengen umfassen. Dies gilt insbesondere für unwiederholbare Messungen, wie z. B. Umweltmessdaten (Pfeiffenberger, 2007), denn diese gehören meist zu Sammlungen, die aktiv genutzt werden und in Teilen als Referenzdaten für andere Arbeiten dienen (National Science Board, 2005). Andere Datensätze dienen lediglich der Sicherung im Sinne einer guten wissenschaftlichen Praxis und werden mit der Zeit immer weniger nachgefragt (Severiens & Hilf, 2006).

Wie alle digitalen Daten werden sie ohne langfristig wirksame Gegenmaßnahmen technisch veralten, nicht mehr nutzbar sein und damit verloren gehen. Hinzu kommt, dass keine der Infrastrukturen für eine digitale Langzeitarchivierung sich dauerhaft betreiben lässt wenn es keine Nutzer gibt, denn erst wenn eine Nachfrage der Wissenschaft nach einer digitalen Langzeitarchivierung besteht, können dauerhafte Strukturen entstehen. Für ein Forschungsdatenarchiv ist es daher eine der wichtigsten Aufgaben, seine Zielgruppe zu definieren, denn an ihr müssen sich die angebotenen Gruppen, Dienstleistungen, organisatorischen und technischen Prozesse orientieren (Parsons & Duerr, 2005).

Die nachhaltige Speicherung und auch Nachnutzung von Forschungsdaten[1] ist eine strategische, organisatorische und technische Aufgabe, die nur gemeinsam von wissenschaftlichen Anwendern und Datenzentren gelöst werden kann. (Altenhöner et al., 2009)

2.6.2 Der Daten-Lebenszyklus – Vom Labor bis ins Archiv

Im Laufe des digitalen Lebenszyklus von Forschungsdaten werden in den verschiedenen Phasen sehr unterschiedliche Anforderungen an die Persistenz der Daten und der Werkzeuge zum Umgang mit Forschungsdaten gestellt. Zwischen dem Entstehen der Daten in wissenschaftlichen Arbeitsprozessen und der sicheren, nachnutzbaren Archivierung der Daten besteht ein breites Spektrum von teilweise gegensätzlichen Anforderungen, auch *Digital Curation Continuum* genannt. Organisatorisch ist ein Kontinuum allerdings nicht handhabbar, weswegen es notwendig ist, innerhalb einer Organisation zu bestimmen, wer in welcher Phase des Lebenszyklus von Forschungsdaten für deren Pflege verantwortlich ist. Auf Grund des vorhandenen Kontextwissens reicht in den Phasen vor der Speicherung in der dauerhaften Domäne ein eingeschränktes Metadatenprofil aus, das bei der Überführung in die nächste Domäne angereichert werden muss, da in der nachfolgenden Domäne dieses Kontextwissen meist fehlt. Der Prozess der Anreicherung der Metadaten kann teilweise bis vollautomatisch erfolgen (Treloar et al., 2007; Treloar & Harboe-Ree, 2008).

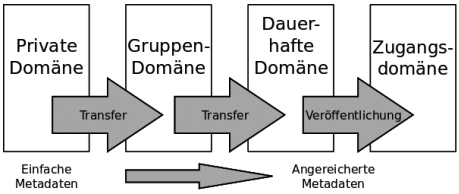

Abb. 1: Gliederung des *Data Curation Continuum* in vier Verantwortungsdomänen. Abhängig von den vorliegenden institutionellen Strukturen sind auch andere Gliederungen denkbar. Im Prozess des Datentransfers werden die vorliegenden Metadaten um wei-

1. In der Vergangenheit wurde oft von „Primärdaten“, bzw. „Forschungsprimärdaten“ (DFG 2009, 1998) gesprochen. Der Begriff „Primärdaten“ bedarf jedoch einer Klärung, denn für den einen sind „Primärdaten“ der Datenstrom aus einem Gerät, z. B. einem Satelliten, für andere sind „Primärdaten“ zur Nachnutzung aufbereitete Daten, wieder andere differenzieren nicht nach Grad der Verarbeitung sondern betrachten alle Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren, als Primärdaten. Der begrifflichen Klarheit wegen sollte daher das Präfix „Primär-“ nicht mehr verwendet und statt dessen nur noch die Begriffe „wissenschaftlichen Daten“ oder „Forschungsdaten“ verwendet werden.

tere Elemente angereichert. Dies geschieht mit der Unterstützung von Informationsfachleuten und mit informationstechnischen Werkzeugen. (Abbildung Klump, 2009)

Organisatorisch und technisch müssen sich diese Prozesse möglichst nahtlos in die wissenschaftlichen Arbeitsabläufe eingliedern. Allerdings sind wissenschaftliche Daten geprägt durch ihre Herkunft aus experimentellem Vorgehen, d.h. anders als Daten aus Arbeitsabläufen der Industrie oder Verwaltung stammen Forschungsdaten überwiegend aus informellen Arbeitsabläufen, die immer wieder ad hoc an die untersuchte Fragestellung angepasst werden (Barga & Gannon, 2007).

Der kritischste Moment im Lebenszyklus von Forschungsdaten ist, wenn das Projekt endet, denn hier endet meistens auch die Finanzierung weiterer Maßnahmen zur Datenerhaltung und das Interesse der Forscher ist bereits auf das nächste Projekt gerichtet. Um diese Klippe zu umschiffen ist es daher notwendig die zu archivierenden Daten möglichst früh in das Archiv zu überführen. Dieser Schritt ist im gesamten Betrieb eines Forschungsdatenarchivs der aufwendigste und damit auch teuerste, der ca. 45% der Gesamtkosten der Langzeitarchivierung eines digitalen Objekts verursacht (Beagrie et al., 2010). Zudem sind die Risiken unkontrollierbarer Kostensteigerungen hier am höchsten (Digital Preservation Testbed, 2005).

Da in den meisten Fällen keine Formatvorgaben bestehen werden Forschungsdaten in einer Vielfalt von Dateiformaten hergestellt. Häufig sind sie semantisch uneinheitlich strukturiert und nur lückenhaft mit Metadaten beschrieben.[2] Diese Faktoren stellen für die digitale Langzeitarchivierung von Forschungsdaten eine größere Herausforderung dar (Abrams, 2007; Klump, 2008; Lormant et al., 2005). Ein Forschungsdatenarchiv sollte daher Vorgaben machen können, welche Datenformate es für die Langzeitarchivierung annehmen kann. Auch die vom Archiv angenommenen Datenformate müssen langfristig überwacht werden um rechtzeitig Maßnahmen ergreifen zu können, sollte ein Datenformat wegen seiner Software- oder Hardwarevoraussetzungen zu veralten drohen (Curtis et al., 2007).

Ungeachtet des in der „Berliner Erklärung“ durch die Universitäten, Wissenschafts- und Forschungsförderungsorganisationen geleisteten Bekenntnisses zum offenen Zugang zu wissenschaftlichem Wissen gibt es Gründe, aus denen manche Daten nicht offen zugänglich gemacht werden können. Anders als im kommerziellen Bereich dienen Zugriffsbeschränkungen hier nicht in erster Linie der Sicherung von Verwertungsrechten oder dem Schutz vor Betriebsspionage, sondern sie sind entweder gesetzlich vorgeschrieben (z. B. Datenschutz) oder dienen dem Schutz von Personen oder Objekten, die durch eine Veröffentlichung der Daten gefährdet würden. Auch andere Gründe, die eine Zugriffsbe-

2. Siehe Kapitel 2.4 Metadaten und Standards

schränkung rechtfertigen, sind denkbar. In der digitalen Langzeitarchivierung von Forschungsdaten sind deshalb Zugriffsbeschränkungen ein wichtiges Thema. Für geschützte Datenobjekte müssen Verfahren und Richtlinien entwickelt werden, die auch über lange Zeiträume hinweg zuverlässig die Zugriffsrechte regeln und durchsetzen können. Die Bedingungen sollten jedoch so gestaltet sein, dass dadurch weder der wissenschaftliche Erkenntnisgewinn behindert wird, noch notwendige Maßnahmen der digitalen Langzeiterhaltung der Daten blockiert werden. Die bisher üblichen „Identitätssilos" bergen auf lange Sicht die Gefahr, dass deren Inhalte veralten oder dass der Aufwand, die Inhalte zu pflegen, enorm hoch wird. Abhilfe könnten vernetzte Systeme schaffen, indem sie einen zertifikatbasierten Ansatz oder Techniken sozialer Netzwerke nutzen (Choi et al., 2006). Ebenso muss der Umgang mit „verwaisten" Datenbeständen geregelt werden. Zusätzlich sollten die Lizenzen auch maschinenlesbar hinterlegt werden, um den Umgang mit komplexen Zusammenstellungen von Daten zu erleichtern.

2.6.3 Kriterienkataloge für digitale Archive und Zertifizierung

Große Teile der wissenschaftlichen Überlieferung in Form von Daten und Texten aus wissenschaftlicher Forschung liegen heute digital vor, und in vielen Fällen ausschließlich in digitaler Form. Sie zu erhalten, sollte nicht allein Aufgabe der Wissenschaft sein, diese muss auch durch Fachleute aus den Gedächtnisorganisationen unterstützt werden, zu denen inzwischen zunehmend auch Datenzentren und andere digitale Archive gehören. Wie lässt sich jedoch die Vertrauenswürdigkeit eines digitalen Archivs feststellen?

Die Notwendigkeit systematischer Datenarchivierung wurde bereits früh erkannt. Bereits 1995 beauftragte die *International Organization for Standardization* (ISO) das *Consultative Committee for Space Data Systems* (CCSDS) damit eine Norm für die Langzeitarchivierung von Daten aus Weltraummissionen zu erarbeiten (ISO 14721:2003). Im Laufe der Arbeiten am Normentwurf wurde klar, dass ein Referenzmodell für die Entwicklung weiterer, davon abgeleiteter Normen notwendig wäre. Das Ergebnis dieser Arbeiten war das *Open Archival Information System Reference Model* (OAIS-RM), das heute den meisten als „OAIS-Modell" bekannt ist, und seit dem mehrfach überarbeitet wurde (CCSDS, 2009; Rank et al., 2010). Auch in Deutschland wurden umfangreiche Arbeiten zur Langzeiterhaltung von Forschungsdaten geleistet, insbesondere im Kontext des Kompetenznetzwerks Langzeitarchivierung (NESTOR) (z. B. Altenhöner et al., 2009; Dobratz et al., 2008; Klump, 2009; Severiens & Hilf, 2006).

Das OAIS-RM war auch die Grundlage für eine Reihe von Kriterienkatalogen, mit denen die Vertrauenswürdigkeit eines digitalen Langzeitarchivs ermit-

telt werden soll. Trotz scheinbar paralleler Arbeiten sind die wichtigsten Kataloge (TRAC[3], NESTOR[4] und *Data Seal of Approval*[5]) auf einander abgestimmt. Zwischen den CCSCS[6], DIN NABD 15[7], AG Vertrauenswürdige Archive und DANS / *Data Seal of Approval* wurde im Sommer 2010 eine weitere Zusammenarbeit und Abstimmung der Aktivitäten vereinbart (Klump, 2011).

2.6.4 Zusammenfassung

Umfangreiche Vorarbeiten der vergangenen Jahre haben dazu geführt, dass uns eine ganze Reihe von Werkzeugen für die Langzeiterhaltung von Forschungsdaten zur Verfügung steht. Diese Werkzeuge sind jedoch noch unzureichend in die wissenschaftlichen Arbeitsabläufe integriert, die daraus entstehenden Medienbrüche verursachen immer noch einen Mehraufwand, der selten geleistet wird. In einzelnen Fachgebieten gibt es jedoch bereits vorbildliche Lösungen, die sich auch auf andere Fachgebiete übertragen lassen.

Um das Ziel einer nachhaltigen digitalen Langzeitarchivierung von Forschungsdaten zu erreichen, muss eine Strategie verfolgt werden, die Langzeitarchivierung von Daten zu einem anerkannten Beitrag zur wissenschaftlichen Kultur macht und in den institutionellen Arbeitsabläufen verankert. Diese organisatorische Strategie muss gleichzeitig von einer technischen Strategie unterstützt werden, die den Akteuren für die digitale Langzeitarchivierung von wissenschaftlichen Forschungsdaten geeignete Werkzeuge in die Hand gibt. Wissenschaft und Wissenschaftsorganisationen haben dies erkannt, so dass in den nächsten Jahren mit erheblichen Fortschritten gerechnet werden darf.[8]

---

3. Trustworthy Repositories Audit & Certification: Criteria and Checklist (Ambacher et al., 2007).
4. (Dobratz et al., 2008).
5. (Sesink et al., 2008).
6. Consultative Committee for Space Data Systems, s.o..
7. DIN Normenausschuss Bibliotheks- und Dokumentationswesen 15, Schriftgutverwaltung und Langzeitverfügbarkeit digitaler Informationsobjekte
8. Verweis auf Kapitel 3.4 Archivierung von Forschungsdaten

Literaturhinweise

Abrams, S., 2007. File Formats. In: S. Ross & M. Day, S., Hrsg. 2007. *DCC Digital Curation Manual*., Glasgow: Digital Curation Centre, 53 S. Online: http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/file-formats/file-formats.pdf [Zugriff am 18.08.2011].

Altenhöner, R. et al., 2009. *Digitale Forschungsdaten bewahren und nutzen – für die Wissenschaft und für die Zukunft*. Göttingen: Niedersächsische Staats und Universitätsbibliothek. urn:nbn:de:0008-2009071031

Ambacher, B. et al., 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC).* Chicago, IL.: CRL Center for Research Libraries. Online: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf [Zugriff am 18.08.2011].

Barga, R. & Gannon, D. B., 2007. Scientific versus business workflows. In: I. J. Taylor, E. Deelman, D. B. Gannon, & M. Shields, Hrsg. 2007. *Workflows for e-Science*. London: Springer, S. 9–16. doi:10.1007/978-1-84628-757-2_2

Beagrie, N., Lavoie, B. F. & Woollard, M., 2010. *Keeping research data safe 2*. Bristol: Joint Information Systems Committee (JISC). Online: http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx [Zugriff am 18.08.2011].

CCSDS (Consultative Committee for Space Data Systems), 2009. *Audit and certification of trustworthy digital repositories.* Draft Recommended Practice, Red Book. Grenbelt, MD: CCSDS. Online: http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206520R1/Attachments/652x0r1.pdf [Zugriff am 18.08.2011].

Choi, H. et al., 2006. Trust Models for Community Aware Identity Management. *WWW2006.* Edinburgh. Online: http://hdl.handle.net/10379/488 [Zugriff am 19.09.2011].

Curtis, J. et al., 2007. AONS – An obsolescence detection and notification service for Web archives and digital repositories. *New Review of Hypermedia and Multimedia*, 13(1), S. 39–53. doi:10.1080/13614560701423711

DFG (Deutsche Forschungsgemeinschaft), 1998. *Sicherung guter wissenschaftlicher Praxis*. Online: http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf [Zugriff am 18.08.2011].

DFG (Deutsche Forschungsgemeinschaft), 2009. *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. Online: http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/veroeffentlichungen/dokumentationen/download/ua_inf_empfehlungen_200901.pdf [Zugriff am 18.08.2011].

Digital Preservation Testbed, 2005. *Costs of Digital Preservation*. From digital volatility to digital permanence. Den Haag: Nationaal Archief. Online: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf [Zugriff am 18.08.2011].

Dobratz, S. et al., 2008. *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive*. NESTOR-Materialien. 2. Aufl. Frankfurt am Main: Deutsche Nationalbibliothek. urn:nbn:de:0008-2008021802.

Klump, J., 2008. *Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten*. NESTOR-Materialien. Expertise. Frankfurt am Main: Kompetenznetzwerk Langzeitarchivierung (NESTOR). urn:nbn:de:0008-2008040103.

Klump, J., 2009. Digitale Forschungsdaten, In: H. Neuroth et al., Hrsg. 2009. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. S. 104–115. urn:nbn:de:0008-20100305375.

Klump, J., 2011., Criteria for the Trustworthiness of Data Centres. *D-Lib Magazine*, 17(1/2). doi:10.1045/january2011-klump.

Lormant, N. Huc, C. Boucon, D. & C. Miquel, 2005. How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information? *Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005)*. Edinburgh, 11 S. Online: http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/003.pdf [Zugriff am 18.08.2011].

National Science Board, 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Washington, DC.: National Science Foundation. Online: http://www.nsf.gov/pubs/2005/nsb0540/ [Zugriff am 18.08.2011].

Parsons, M. A., & R. Duerr, 2005. Designating user communities for scientific data: challenges and solutions. *Data Science Journal*, (4), S. 31–38. doi:10.2481/dsj.4.31.

Pfeiffenberger, H., 2007. Offener Zugang zu wissenschaftlichen Primärdaten. *Zeitschrift für Bibliothekswesen und Bibliographie*, 54(4-5), S. 207–210. hdl:10013/epic.28454.d001

Rank, R. H., Cremidis, C. & K. R. McDonald, 2010. Archive Standards: How Their Adoption Benefit Archive Systems, In: L. Di & H. K. Ramapriyan, Hrsg. 2010. *Standard-Based Data and Information Systems for Earth Observation.* Berlin: Springer, S. 127–142. doi:10.1007/978-3-540-88264-0_8.

Sesink, L. Horik, R. van & Harmsen, H., 2008. *Data Seal of Approval*. Den Haag: Data Archiving and Networked Services (DANS). Online: http://www.datasealofapproval.org/

Severiens, T. & Hilf, E. R., 2006. *Langzeitarchivierung von Rohdaten.* (NESTOR-Materialien 6). Frankfurt am Man: NESTOR – Kompetenznetzwerk Langzeitarchivierung. urn:nbn:de:0008-20051114018

Treloar, A. Groenewegen, D. & C. Harboe-Ree, 2007. The Data Curation Continuum – Managing Data Objects in Institutional Repositories. *D-Lib Magazine*, 13(9/10), S. 13. doi:10.1045/september2007-treloar.

Treloar, A. & Harboe-Ree, C., 2008. Data management and the curation continuum: how the Monash experience is informing repository relationships. *VALA2008 14th Biennial Conference*. Melbourne, Australien 5.-7. Feb. 2008. Online: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf [Zugriff am 18.08.2011].