Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 16
Erkenntnisse und Thesen zur Langzeitarchivierung von Forschungsdaten

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Universitätsverlag Göttingen

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

nestor

Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
Förderkennzeichen: 01 DL 001 B

Universitätsverlag Göttingen
vwh Verlag Werner Hülsbusch Fachverlag für Medientechnik und -wirtschaft
D-GRID GmbH

Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme
hg. v. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig
im Rahmen des Kooperationsverbundes nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Lektorat) und Sonja Neweling (Redaktion) sowie Martina Kerzel und Lajos Herpay (Gestaltung und Montage) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte des Buches stehen auch als Onlineversion
(http://nestor.sub.uni-goettingen.de/bestandsaufnahme/)
sowie über den Göttinger Universitätskatalog
(http://www.sub.uni-goettingen.de) zur Verfügung.
Die URN lautet: http://nbn-resolving.de/urn:nbn:de:0008-2012031401.

Druck und Bindung: Kunsthaus Schwanheide
Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-86488-008-7

16 Erkenntnisse und Thesen zur Langzeitarchivierung von Forschungsdaten

Heike Neuroth, Achim Oßwald, Uwe Schwiegelshohn

Als Ergebnis der vergleichenden Bestandsaufnahme des Umgangs mit Forschungsdaten in den elf Wissenschaftsdisziplinen, die in diesem Sammelwerk exemplarisch zusammengeführt und analysiert wurden, lassen sich folgende Erkenntnisse und Thesen formulieren, die zum Teil die Ergebnisse des Kapitels 15 aufgreifen. Zum einen heben sie die Bedeutung der Langzeitarchivierung von Forschungsdaten aus wissenschaftlicher Sicht hervor und verweisen auf konzeptionelle und operative Sachverhalte, die als Zwischenergebnis festgehalten bzw. weitergehend geklärt werden sollten. Zum anderen gilt es auch eine Reihe wissenschafts- und gesellschaftspolitischer Aspekte zu berücksichtigen.

Übergreifende Sachverhalte:

1. Die Bedeutung von Forschungsdaten und deren langfristiger Archivierung bzw. Bereitstellung wird von allen hier vorgestellten Wissenschaftsdisziplinen betont.
2. Die hier skizzierten verschiedenen Ansätze bei der Langzeitarchivierung von Forschungsdaten in diesen Wissenschaftsdisziplinen sind nicht Ausdruck eines mangelnden Kooperationswillens über die Disziplingrenzen hinweg, sondern logische Konsequenz der unterschiedlichen Anforderungen und der praktizierten Methoden innerhalb der einzelnen Disziplinen.
3. Kooperative Strukturen innerhalb einer Wissenschaftsdisziplin sind bei der digitalen Langzeitarchivierung von Forschungsdaten die Regel und nicht die Ausnahme.

4. Häufig werden Infrastruktureinrichtungen wie Bibliotheken oder Rechenzentren als Kooperationspartner bei der Langzeitarchivierung von Forschungsdaten einbezogen. Deren Rolle und Funktion sind bislang aber nicht immer klar definiert.

5. In vielen Wissenschaftsdisziplinen sehen sich Forscher immer noch mit einer geringen Wertschätzung der Langzeitarchivierung und einer geringen Akzeptanz für die Weitergabe und Nachnutzung von Daten (data sharing) konfrontiert. Das Bewusstsein über den Stellenwert der Daten ist sowohl in den Wissenschaftsdisziplinen selbst als auch in der Gesellschaft sowie bei den weiteren Beteiligten (z.B. Bibliotheken, Rechenzentren etc.) eine wichtige Voraussetzung für weitere Diskussionen und Entwicklungen.

6. Der der eigentlichen Langzeitarchivierung vorgelagerte Bereich des Datenmanagements umfasst sowohl fachspezifische als auch generische Aufgaben. Eine enge Zusammenarbeit der verschiedenen Interessengruppen und Beteiligten erlaubt eine genaue Definition der Aufgaben- und Verantwortungsbereiche.

7. Über die zu archivierenden und die zur Verfügung zu stellenden Datenmengen und Stückzahlen können weder für einzelne Wissenschaftsdisziplinen noch für mehrere Disziplinen zusammen tragfähige Aussagen getroffen werden. Insgesamt ist aber quer über die Wissenschaftsdisziplinen ein rasanter Anstieg des Datenvolumens an digitalen Forschungsdaten erkennbar.

Forschungsdatenzentren:

1. In jenen Wissenschaftsdisziplinen, in denen sich (zentrale) Strukturen für das Datenmanagement gebildet haben, sind die Prozesse zur Langzeitarchivierung von Forschungsdaten bereits besser etabliert als in den anderen Wissenschaftsdisziplinen.

2. Datenzentren werden von vielen Disziplinen als die ideale Lösung angesehen, die Verfügbarkeit und die effiziente Nachnutzung von Forschungsdaten zu verbessern und langfristig zu sichern. Sie können zentral oder in einem dezentralen Verbund organisiert sein. Außerdem können sie innerhalb der jeweiligen Wissenschaftsdisziplin bei der Entwicklung von Standards und in der Beratung eine wichtige Rolle übernehmen.
3. Es besteht Klärungsbedarf hinsichtlich der Vertrauenswürdigkeit von Datenzentren und welche Kriterien dafür erfüllt sein müssen. Die Frage, wie diese Vertrauenswürdigkeit überprüft werden kann (z.B. Zertifizierung von Datenzentren) und wer dafür zuständig ist, ist offen.

Metadaten und Formate:

1. Nahezu jede Wissenschaftsdisziplin verwendet eigene Metadatenformate, von denen die meisten auf XML basieren. Viele Wissenschaftsdisziplinen haben in den letzten Jahren fachspezifische Metadatenformate entwickelt.

2. Forschungsdaten sind in einer fast unüberschaubaren Vielzahl und Vielfalt von Datenformaten verfügbar. Fast allen Fachdisziplinen ist dabei gemeinsam, dass über die allgemein bekannten Formate hinaus auch zahlreiche fachspezifische und proprietäre Formate genutzt werden.

3. Die einzelnen Wissenschaftsdisziplinen gehen mit der Vielfalt und Heterogenität der Formate sehr unterschiedlich um: Die verschiedenen Formate sind entweder durch eine Richtlinie vorgegeben bzw. sonst faktisch eingeschränkt oder die Formatwahl ist offen bzw. kann aus disziplinbezogenen Gründen nicht eingegrenzt werden.

4. Insgesamt setzen die Wissenschaftsdisziplinen – wo möglich – offene Formate ein. Allerdings wird dies durch vorgegebene Software oder Geräte zum Teil stark eingeschränkt. Bei einer Standardisierung ist es

hilfreich, etablierte industrielle und kommerzielle Verfahren zu berücksichtigen.

Technische Datensicherung:

1. Die technische Datensicherung stellt einen ersten Schritt der Langzeitarchivierung von Forschungsdaten dar: Durch die rein technische Speicherung von Forschungsdaten kann die Integrität der Daten unabhängig von Datei- und Metadatenformaten gesichert werden. Allerdings ist eine inhaltliche Nachnutzung der Forschungsdaten damit nicht gewährleistet.

2. Eine Begrenzung der Vielfalt von Datei- und Metadatenformaten reduziert die Anzahl der zur Wiedergabe notwendigen technischen Umgebungen (Hardware und Software) und erleichtert die Nachnutzung der Daten.

3. Die andauernde Technologiebeobachtung (technology watch) und die Beobachtung von Bedarf und technischen Gegebenheiten im Bereich der Langzeitarchivierung bei den verschiedenen Zielgruppen (community watch) dienen zur Sicherung der technischen und inhaltlichen Nachnutzbarkeit von Forschungsdaten.

Nachnutzung von Forschungsdaten:

1. Forschungsdaten werden aus unterschiedlichen Gründen für die Nachnutzung bereitgestellt, z.B. für die Kooperation innerhalb eines Forschungsprojekts, für externe Wissenschaftler oder bei der Publikation auch für die breite (Fach)Öffentlichkeit.

2. Die Wissenschaften, ihre Förderer und die (politische) Öffentlichkeit verfolgen die Diskussionen im Bereich der Nachnutzung von Forschungsdaten und die Regelungen dazu. Dies resultiert u.a. in nachdrücklichen Forderungen, Forschungsdaten zugänglich zu machen und ihre Nachnutzung langfristig sicherzustellen.

3. Der Bereitstellung und damit Nachnutzung von Forschungsdaten stehen meistens folgende Gründe im Wege: drohender Kontrollverlust über die Daten, ungeklärte Rechteverhältnisse an den Daten und datenschutzrechtliche Restriktionen. Auch der bei der Erhebung der Daten geleistete (finanzielle) Aufwand beeinflusst potentielle Szenarien einer Nachnutzung.

4. Die langfristige Zitierbarkeit und Referenzierbarkeit von Forschungsdaten ist einer von mehreren Beweggründen für die Langzeitarchivierung von Forschungsdaten. Dabei spielen persistente Identifier eine große Rolle.

Kosten, Finanzierung, Effizienz und Institutionalisierung:

1. Da Wissenschaft insgesamt eine gesellschaftliche Aufgabe ist, sind ihre Kosten von der Gesellschaft zu tragen. Die Gesellschaft kann im Gegenzug eine effiziente Verwendung der zur Verfügung gestellten Mittel erwarten. In Bezug auf Forschungsdaten und ihre Nachnutzung gibt es zwei Ansätze:

    - Die Archivierung von Forschungsdaten nach ihrer Erhebung für eine spätere Verwendung oder
    - die Wiederholung der Forschungsdatenerhebung. Hierbei ist zu beachten, dass sich manche Prozesse nicht wiederholen lassen (z.B. Erhebung von Klimadaten).

2. Bei Gleichwertigkeit in Bezug auf die Qualität der Forschungsdaten ist dem jeweils kostengünstigeren Ansatz der Vorzug zu geben. Für eine diesbezügliche Entscheidung müssen fundierte Kostenabschätzungen vorliegen.

3. Belastbare Aussagen über die Kosten und Kostenfaktoren der Langzeitarchivierung von Forschungsdaten gibt es bisher nur ansatzweise. Insofern können bislang auch noch keine konkreten Aussagen über Kostenstrukturen getroffen werden. Bisherige Untersuchungen deu-

ten darauf hin, dass die Personalkosten den überwiegenden Teil der Gesamtkosten ausmachen, wobei dieses Personal bisher fast überwiegend aus Projektmitteln finanziert wird.

4. Nur für einen Teil der Wissenschaftsdisziplinen konnte für die Langzeitarchivierung von Forschungsdaten bisher eine (anteilige) finanzielle Absicherung in Form einer institutionellen Grundfinanzierung etabliert werden. Die meisten Wissenschaftsdisziplinen finanzieren diese Aktivitäten (noch) aus Projektmitteln, wobei die Projekte zum Teil recht lange Laufzeiten haben.

5. Es besteht ein dringender Handlungsbedarf, die Kosten und Kostenfaktoren für die einzelnen Dienstleistungen der Langzeitarchivierung von Forschungsdaten zu klären. Nur so können nachhaltige Organisations- und Geschäftsmodelle (inklusive Finanzierungsmodelle) in den einzelnen Wissenschaftsdisziplinen entwickelt und umgesetzt werden.

6. Die Sicherung und Pflege von Forschungsdaten ist Teil des wissenschaftlichen Arbeitens. Bei einer Kostenabschätzung eines Forschungsprojektes ist der dafür notwendige Aufwand einzuplanen.

7. Bezüglich der Effizienz der Langzeitarchivierung von Forschungsdaten können Skalierungseffekte über die Bildung von Datenzentren genutzt werden. Dies kann zu neuen Organisationsstrukturen führen, die über Institutionsgrenzen hinweg angelegt sind.

Qualifizierung:

1. Es besteht dringender Qualifizierungsbedarf im Bereich der Langzeitarchivierung von Forschungsdaten, insbesondere auch im theoretisch-konzeptionellen Bereich. Bisher gibt es außer den nestor Aktivitäten wenig bis keine systematischen Qualifizierungsangebote, weder fachwissenschaftsbezogen noch für Informationsspezialisten.

2. Ein Transfer von langzeitarchivierungsrelevanten Forschungsergebnissen oder best practice-Beispielen erfolgt u.a. auch mangels systematischer Transferangebote nur eingeschränkt. Fallbasiertes Agieren

und eine z.T. sehr auf die eigene Wissenschaftsdisziplin und deren vermeintliche Singularität fokussierte Sicht verhindern bislang die praktische Etablierung disziplinenübergreifender Qualitätskriterien und Qualifizierungsmaßnahmen.

3. Perspektivisch ist anzustreben, dass das Thema „Langzeitarchivierung digitaler Forschungsdaten" in Studiengängen oder Studienschwerpunkten (z.B. data librarian, data curator[1]) und Forschungskontexten in die methodische Basisqualifizierung einfließt. Ergänzende Qualifizierungsangebote, z.B. Studienschwerpunkte oder disziplinenübergreifende Masterstudiengänge sind eine wichtige Infrastrukturunterstützung.

Gesellschaftliche Bedeutung:

1. Wissenschaftliche Ergebnisse beeinflussen in steigendem Maße gesellschaftspolitische Entscheidungen (z.B. Kernenergie, Präimplantationsdiagnostik, Pandemien, Gesundheitsrisiken). Eine spätere Überprüfung dieser Entscheidungen benötigt zur Gewährleistung der Transparenz jene Forschungsdaten, die die Grundlage für die Entscheidungen gebildet haben.

2. Die Bewahrung des kulturellen Erbes ist eine anerkannte gesellschaftliche Aufgabe. Forschungsdaten sind Teil des kulturellen Erbes.

3. Die Aufklärung von Verstößen gegen die „gute wissenschaftliche Praxis" oder auch methodischer Fehler setzt die Verfügbarkeit von Forschungsdaten voraus, die der jeweiligen Publikation oder Forschungsarbeit zugrunde liegen.

Insgesamt zeigen die in diesem Buch dokumentierte Bestandsaufnahme und die hier formulierten Erkenntnisse und Thesen, welche große (zukünftige) Bedeutung der Langzeitarchivierung von Forschungsdaten zukommt. Entsprechend hat die EU-Expertengruppe

1 Vgl. z.B. JISC Study (2011).

„High Level Expert Group on Scientific Data“[2] dazu sinngemäß formuliert: Daten sind die Infrastruktur und ein Garant für innovative Forschung.

Handlungsempfehlungen, die diese Zielsetzung aufgreifen, müssen in der (wissenschafts)politischen Arena erarbeitet und in politische und fördertechnische Programme auf nationaler und internationaler Ebene umgesetzt werden. Einige Handlungsfelder in diesem Bereich sind bereits von Forschung und Politik identifiziert worden:

So formuliert die oben genannte „High Level Expert Group on Scientific Data“ sechs Handlungsempfehlungen[3], die die Etablierung eines internationalen Rahmens für die Entwicklung von kooperativen Dateninfrastrukturen, die Bereitstellung von Fördermitteln für die Entwicklung von Dateninfrastrukturen und die Entwicklung von neuen Ansätzen und Methoden umfassen, um den Wert, die Bedeutung und die Qualität der Nutzung von Daten zu messen und zu bewerten. Des Weiteren wird auf die Bedeutung der Qualifizierung einer neuen Generation von sog. „data scientists“ hingewiesen sowie die Verankerung von Qualifizierungsangeboten in (neuen) Studiengängen. Die Schaffung von Anreizsystemen im Bereich der „green technologies“, um dem gesteigerten Bedarf an Ressourcen (z.B. Energie) aus umwelttechnischen Gesichtspunkten gerecht zu werden, spielt ebenfalls eine Rolle. Zuletzt wird auch die Etablierung eines internationalen Expertengremiums vorgeschlagen, das die Entwicklungen von Dateninfrastrukturen vorantreiben und steuern soll.

Der KII-Bericht[4] betont aus einer nationalen Perspektive die Notwendigkeit von Datenmanagementplänen und Datenrichtlinien (policies) als Voraussetzung für den Austausch und die Nachnutzung von Forschungsdaten. Dabei sollten auch die Verantwortlichkeiten, Funktionen und Rollen aller Beteiligten klar definiert sein. Es werden darüber hinaus gezielte Förderprogramme für die verschiedenen Aspekte im Bereich der Langzeitarchivierung von Forschungsdaten gefordert, wobei zwischen Entwicklungskosten für den Aufbau bzw. Ausbau von Dateninfrastrukturen und Betriebskosten für den dauerhaften Betrieb inklusive Datenpflege unterschieden wird.

2 Vgl. High Level Expert Group on Scientific Data (2010).
3 Vgl. ebenda.
4 Vgl. Leibniz-Gemeinschaft (2011).

Der EU-GRDI2020-Report[5] geht davon aus, dass in den nächsten zehn Jahren globale Forschungsdaten-Infrastrukturen aufgebaut werden müssen um über sprachliche, politische und soziale Grenzen hinweg zu operieren. Sie sollen Forschungsdaten zur Verfügung stellen und ihre Nutzung (discovery, access, use) unterstützen. Dabei wird das Modell „Digital Science Ecosystem" eingeführt, an dem die folgenden (neuen) Akteure beteiligt sind: Digital Data Libraries, Digital Data Archives, Digital Research Libraries und Communities of Research. Dieses Modell impliziert eine zum Teil komplett neue Rollen- und Aufgabenverteilung der bisherigen Beteiligten und fordert die Schaffung neu definierter Aufgabenbereiche. Im Mittelpunkt steht immer die Sicherung und Nachnutzung von Forschungsdaten, wobei die Nachnutzung auch über disziplinäre Grenzen hinweg ermöglicht werden soll. Dazu werden insgesamt elf Empfehlungen und Handlungsoptionen formuliert, die u.a. auch vorsehen, dass neue Berufszweige und Qualifizierungswege etabliert werden müssen. Darüber hinaus wird empfohlen, neue Werkzeuge (z.B. in den Bereichen Datenanalyse oder Datenvisualisierung) und Dienste (z.B. zur Datenintegration, zum Datenretrieval oder Ontologie-Dienste) für den Umgang mit und die Nutzung von Daten zu entwickeln sowie Aspekte der „open science"- und „open data"-Konzepte zu berücksichtigen.

Die vorliegende Bestandsaufnahme der elf Wissenschaftsdisziplinen unterstreicht die oben erwähnten Aussagen. Aus dem Gesamtbild lässt sich ein dringender Handlungsbedarf erkennen, der vor allem folgende Themenfelder umfasst:

- Nationale und internationale Programme müssen initiiert werden, um den neuen und großen Herausforderungen im Bereich der Forschungsdaten gewachsen zu sein.
- Eine Neudefinition von Rollen, Aufgabenverteilungen und Verantwortungsbereichen ist nötig, um die unterschiedlichen Handlungsfelder bei der Zugänglichkeit und Nachnutzung sowie Langzeitarchivierung von Forschungsdaten abdecken zu können.
- Neue Berufsfelder und Qualifizierungsmaßnahmen müssen entwickelt werden und das Qualifikationsprofil zum Management von

5 Vgl. GRDI2020 Roadmap Report (2011).

Forschungsdaten muss in (neuen) Studiengängen[6] und -schwerpunkten berücksichtigt werden, um professionelle Handhabung von Forschungsdaten zu gewährleisten.

- Die Veröffentlichung von Forschungsdaten muss als unverzichtbarer Teil des Forschungsprozesses gewertet werden, um die Verifikation und Weiterentwicklung der Resultate zu unterstützen.

Abschließend bleibt festzuhalten, dass Forschungsdaten Resultat und ebenso unverzichtbare Basis wissenschaftlicher Arbeit sind. Sie müssen auch als Ressource begriffen werden, die sowohl für zukünftige Forschergenerationen als auch disziplinübergreifend immer mehr Bedeutung gewinnen. In diesem Sinne sind sie Teil des (inter)nationalen Kulturguts. Dies verlangt ihre Pflege (data curation) über ihren gesamten Lebenszyklus hinweg.

Auch wenn schon international vielversprechende Ansätze existieren und national einige Entwicklungen und Diskussionen initiiert wurden, bedarf es einer großen, national koordinierten Anstrengung, bevor die Vision „Daten als Infrastruktur" Wirklichkeit werden kann. Dieser Prozess wird sowohl disziplinspezifische als auch disziplinübergreifende Aspekte enthalten und ist in internationale Bemühungen einzubetten. Dabei sollten rechtliche, finanzielle und organisatorische Aspekte nicht behindern, sondern unterstützen. Hier ist besonders auch die Politik gefordert.

6 Vgl. z.B. Büttner; Rümpel; Hobohm (2011), S. 203f.