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

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 3
Vorgehensweise

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.

Die digitale Version dieses Werkes steht unter einer Creative Commons Namensnennung-Nicht-kommerziell-Weitergabe unter gleichen Bedingungen 3.0 Unported Lizenz.

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.
© Verlag Werner Hülsbusch, Boizenburg, 2012
http://www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen.

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

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

ISBN: 978-3-86488-008-7

3 Vorgehensweise

Heike Neuroth

3.1 Untersuchungsgegenstand

Auch wenn noch viele Fragen im Zusammenhang mit der Langzeitarchivierung und Langzeitverfügbarkeit von Forschungsdaten offen sind, so wird ein erster wichtiger Schritt sein, den Status quo und die Bedürfnisse verschiedener Fachdisziplinen und ihrer Akteure zu ermitteln. So können dann die Anforderungen an Forschungsdateninfrastrukturen abgeleitet und Strategien zu deren Realisierung entwickelt werden. Bisher steht eine solche Kartierung der Wissenschaftslandschaft hinsichtlich eines nachhaltigen Forschungsdatenmanagements noch aus. Dies ist auch nur schwer umfassend zu leisten, da es zurzeit noch kein Konzept gibt, welches die Sichtbarkeit solcher Ansätze verlässlich sicherstellt – unabhängig von der Frage, welchen Erkenntnisgewinn eine flächendeckende Kartierung der Wissenslandschaft von Forschungsdaten darstellen würde. Solange ein umfassendes Bild weder vorliegt noch von einzelnen geleistet werden kann, bietet die hier vorgelegte Bestandsaufnahme einzelner Forschungsbereiche eine erste Orientierung. Die Auswahl der Fachdisziplinen unterlag dabei keinem systematischen Prozess. Stattdessen werden ausgewählte Fachdisziplinen im Sinne von Leuchttürmen aus unterschiedlichen Wissenschaftsbereichen vorgestellt, die im Umfeld von nestor[1], der eScience-Intitiative[2] oder der deutschen Grid-Initiative[3] hervorgetreten sind dadurch, dass

- der Gegenstand ihrer Forschung digital vorliegt (z.B. 3-D-Scan eines Museumsobjekts) bzw. digitale Forschungsdaten erzeugt werden,

1 Vgl. nestor (2011a).
2 Vgl. BMBF (2011e).
3 Vgl. D-Grid GmbH (2011); D-Grid (2011a).

- mit den Forschungsergebnissen auch Forschungsdaten veröffentlicht werden,
- diese Forschungsdaten langzeitarchiviert und für die Nachnutzung verfügbar gehalten werden sollen,
- sie zum Aufbau einer nachhaltigen Forschungsdateninfrastruktur aktiv beitragen und zumindest erste Überlegungen und Erfahrungen hierzu vorliegen.

Die hier vorgestellten Fachdisziplinen decken ein fachliches Spektrum von den Geistes- über die Naturwissenschaften inklusive der Medizin ab. Leider konnten nicht in allen Fachdisziplinen[4] Leuchtturmansätze in Deutschland identifiziert werden und sind in einigen Wissenschaftsbereichen (z.B. Lebenswissenschaften oder Ingenieurwissenschaften) bereits existierende Praktiken und Lösungen nicht berücksichtigt werden. Insofern erhebt die hier vorgelegte Fachübersicht keinen Anspruch auf Vollständigkeit.

3.2 Aufbau des Sammelwerkes

Vor dem Hintergrund einer kurzen Darstellung der nationalen und internationalen Diskussionen und Entwicklungen (vgl. Kap. 2) stellen sich die jeweiligen Fachdisziplinen auf der Basis einer Liste mit Leitfragen (s.u.) für die einzelnen Fachdisziplinen vor (vgl. Kap. 4 bis Kap. 14), so dass in einer abschließenden vergleichenden Betrachtung Unterschiede und Gemeinsamkeiten bislang vorliegender bzw. erkennbarer Ansätze herausgearbeitet werden können. Daraus abgeleitet werden Schlussfolgerungen über Faktoren, die die Entstehung und den Betrieb einer generischen und fachspezifischen Forschungsdateninfrastruktur sowie die langfristige Sicherung der Forschungsdaten befördern (vgl. Kap. 15). Diese können anderen Wissenschaftsdisziplinen als Modell dienen oder ihnen Hinweise auf Entwicklungsoptionen geben. Das abschließende Kapitel 16 leitet Erkenntnisse und Thesen aus den vorangegangenen Kapiteln (insbesondere Kap. 15) ab und zeigt einige Handlungsfelder auf, die zurzeit auch national und international diskutiert werden.

4 Vgl. DFG (2011b).

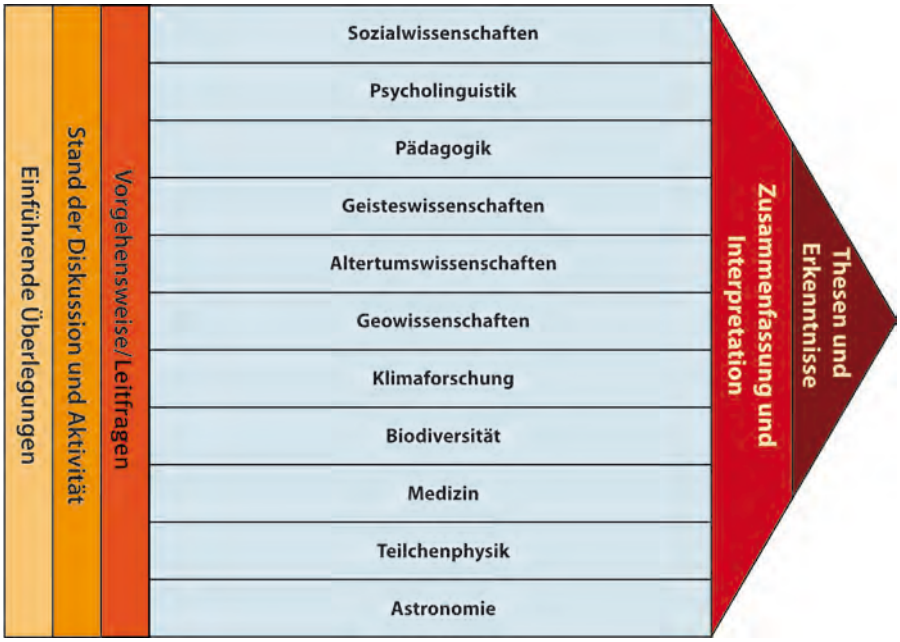

Abb. 1: Aufbau des Sammelwerkes

3.3 Leitfragen für die exemplarische Kartierung der jeweiligen Fachdisziplin

Die folgende Gliederungsstruktur samt Leitfragen wurde von den Herausgebern entwickelt und den Autorinnen und Autoren der einzelnen Fachdisziplinen zur Verfügung gestellt. Durch die einheitliche Strukturierung des jeweiligen fachspezifischen Entwicklungsstandes wird ein Vergleich des Status quo und der LZA-Praktiken möglich. Darüber hinaus sind vergleichende Analysen durchführbar, die den Aufbau von Forschungsdateninfrastrukturen in Deutschland unterstützen. Im März 2011 fand auf Einladung der D-Grid GmbH an der TU Dortmund ein Workshop mit allen Beteiligten statt, in dem erste übergreifende Lösungen und Strategien diskutiert wurden.

Im Folgenden werden die Leitfragen einzeln – zum Teil stichwortartig – vorgestellt. Erläuternde Hintergrundinformationen zu den Fragen werden stichwortartig gegeben.

3.3.1 Einführung in den Forschungsbereich

Charakterisierung des fachlichen Umfeldes, Beschreibung des Forschungsbereichs und der dort vorhandenen Strukturen. Hintergrund: Fachliche Differenzierung z.B. eher (inter)national koordinierte Verbundprojekte versus heterogene Projektlandschaft etc.; falls es sich um ein Projekt handelt, sind z.B. die Struktur, Finanzierung, Partner etc. des Projektes, Ziele, Hintergrundinformationen, Produkte des Projektes etc. relevant.

3.3.2 Kooperative Strukturen

- Ist eine ***institutionenübergreifende Zusammenarbeit*** in der Forschung die Regel oder eher eine Ausnahme? Hintergrund: Kooperative Zusammenarbeit erhöht den Bedarf für Lösungen zum Datenaustausch, dieser Druck wirkt sich positiv auf die Entwicklung gemeinsamer Datenaustauschstandards und gemeinsamer Datenmodelle aus. Im Rahmen wissenschaftlicher Zusammenarbeit entstehen auch Anreize, Daten zur Nachnutzung aufzubereiten.
- Gibt es eine Institution, die ***bereits (zentralisierte) LZA-Dienste*** für den gesamten Bereich übernimmt oder Ergebnisse des gesamten Forschungsbereichs sammelt und dokumentiert? Hintergrund: Eine bereits bestehende zentrale Einrichtung könnte deutschlandweit (zentrale) LZA-Dienste übernehmen oder koordinieren.
- Besteht in der Regel eine ***Zusammenarbeit mit einer internen oder externen Service-Infrastruktureinrichtung*** (z.B. IuD-Abteilung, Bibliothek, Rechenzentrum etc.)? Hintergrund: Beispiele für die Charakterisierung der Zusammenarbeit können sein: Sammeln und Bereitstellen von Forschungsdaten und/oder Publikationen. Eine solche Einrichtung könnte z.B. auch definierte LZA-Aufgaben übernehmen. Auch eine (zukünftige) Festlegung von Verantwortlichkeiten, Rollen/Pflichten etc. im Bereich Datenmanagement oder LZA wäre dann einfacher. Sofern solche Ansätze schon erkennbar sind, können diese hier beschrieben werden. Falls nicht, bitte dies auch benennen.

3.3.3 Daten und Metadaten

In diesem Fragenkomplex geht es nicht um die generelle LZA von Publikationen z.B. in Dokumentenservern, sondern um eine Klassifizierung der im Forschungsbereich anfallenden Datenquellen bzw. verschiedener Typen von digital vorliegenden Forschungsdaten! Die Fokussierung erfolgt hier „nur" auf Forschungsdaten, die Gegenstand einer Publikation sind bzw. sein sollen und als solche auch veröffentlicht werden.

- Welche ***Typen von Daten*** fallen im Forschungsbereich an z.B. aus Großinstrumenten (Teleskope, Beschleuniger etc.), Simulationen, Laborversuchen, Felderhebungen und -versuchen oder digitalisierten Objekten (digitale Dokumente, digitales Archivgut, digital vorliegende Fundstücke oder digitale Museumsobjekte) etc.? Hintergrund: Je nach Typ der Datenquelle müssen bestimmte LZA-Strategien definiert bzw. entwickelt werden. Falls es diese gibt, können sie hier aufgelistet werden.
- Wie werden die Forschungsdaten im Forschungsbereich ***publiziert und langfristig vorgehalten***? Gibt es z.B. bereits (etablierte) Datenzentren oder Datenarchive? Hintergrund: Wenn es bereits eine Kultur des Publizierens von Forschungsdaten gibt, ist die Etablierung von Managementstrukturen für Forschungsdaten und LZA sehr wichtig.
- Gibt es Mindestanforderungen für die ***Aufnahme von Forschungsdaten*** in ein Datenzentrum (z.B. Formatvorgabe, Qualitätsmanagement, Metadaten, Persistent Identifier etc.). Gibt es ***Managementpläne*** für Forschungsdaten? Hintergrund: Falls es sie gibt, sollten diese hier beschrieben werden, da sie vielleicht von anderen Disziplinen/Forschungsbereichen für ihre LZA-Strategie (teilweise) nachgenutzt werden können.
- Welche ***Forschungsdatenvolumina*** fallen in etwa pro Jahr an? Wie ist die Steigerungsrate? Hintergrund: Große Datenvolumina z.B. im Peta-Bereich erfordern u.U. andere LZA-Strategien als z.B. geringere Datenvolumina, die aber bezogen auf Datentyp und/oder Datenformat wesentlich heterogener sind.

- In welchen (standardisierten) ***Formaten*** liegen die Forschungsdaten vor? Gibt es Empfehlungen für bestimmte Formate? Hintergrund: Formate sind wesentlich für LZA-Strategien; falls sich eine Disziplin oder ein Forschungsbereich bereits auf einige (standardisierte) Formate geeinigt hat, ist dies besonders wichtig.
- Werden die Forschungsdaten in ihrer ***Nutzung beschränkt*** z.B. durch Datenschutz, rechtliche Vorgaben, Vertrauensschutz, individuelle Rechte, Urheberrechte etc. Hintergrund: Einschränkungen in der Nutzung von Forschungsdaten haben direkten Einfluss auf die LZA und müssen von Beginn an konsequent berücksichtigt werden (Policy, Technologie etc.).
- Wie wichtig ist es, ***ältere*** Forschungsartikel, Forschungsberichte etc. und Forschungsdaten nachzunutzen? Mit „älter" sind digitale Forschungsdaten gemeint, die nicht mehr direkt im öffentlichen Zugriff stehen. Diese könnten z.B. auf internen Servern gespeichert, nicht ausreichend mit Metadaten dokumentiert und daher nicht mehr richtig interpretierbar sein. Hintergrund: Wenn in einem Bereich fast nur die neuesten Artikel und Forschungsergebnisse eine Rolle spielen, dann besteht u.U. kein LZA-Bedarf für „ältere" Publikationen/Forschungsdaten. Ansonsten müssten Verfahren entwickelt und etabliert werden, um auch diese Forschungsdaten wieder in den Forschungskreislauf zu bringen.
- Werden in ***Metadaten*** (eventuell sogar standardisiert) zur deskriptiven, strukturellen und administrativen Beschreibung inklusive Metadaten zur persistenten Adressierung von Forschungsdaten verwendet? Welches Persistent Identifier-System (z.B. DOI, Handle, URN) kommt zum Einsatz? Welche (fachspezifischen) Metadaten-Schemata werden verwendet? Werden mit den Forschungsdaten auch Metadaten (eventuell sogar standardisiert) zur Beschreibung der technischen Voraussetzungen mit dokumentiert und abgelegt (z.B. um auch Hardware-, Software-Rahmenbedingungen zu dokumentieren oder gar selber zu archivieren)? Hintergrund: Wenn es keine (standardisierten) Metadaten zu den Forschungsdaten gibt und auch keine persistente Identifizierung betrieben wird, dann spricht das eher

dafür, dass die Forschungsdaten nur von einem kleinen Forscherkreis weiterverwendet werden können bzw. eventuell auch nur von diesem weiterverwendet werden sollen. Ohne deskriptive, strukturelle und administrative Metadaten sind Forschungsdaten nicht interpretierbar. Daten, die aus ihrem produzierenden System exportiert wurden und deren Struktur und Herkunft nicht dokumentiert wurden, können nur unter sehr hohem Aufwand in neuen Kontexten verwendet werden.

3.3.4 Interne Organisation

- Wird die LZA von Forschungsdaten bereits nach ***festen Prozessen und Regeln*** durchgeführt? Gibt es etablierte LZA-Strategien, Policies, Verfahren, Umsetzungen etc.? Wird hier konkret mit anderen (nationalen, internationalen) Partnern zusammengearbeitet? Hintergrund: Ohne einen solchen strukturellen Rahmen lässt sich ein Datenarchiv nicht nachhaltig betreiben.
- Wie wird das Datenarchiv ***finanziert***? Werden im Haushalt feste Mittel für die Langzeiterhaltung von Forschungsdaten eingeplant? Ist eine Bund/Land-Finanzierung gesichert oder angestrebt? Wird eine Datenpauschale (z.B. pro definierter Datenmenge) bei (externen) Projekten erhoben? Hintergrund: Erst wenn eine dauerhafte Finanzierung gesichert ist, kann von einer geregelten, institutionalisierten LZA gesprochen werden.
- Wie hoch sind schätzungsweise die ***Kosten*** für die Entwicklung (Initialaufwand) und den Betrieb Ihres Datenarchivs? Hintergrund: Für einige Datenarchive in Deutschland liegen bereits grobe Schätzungen vor. Erst wenn bei Politik, Förderern, Forschungsbereich, Service-Infrastruktureinrichtungen etc. Klarheit über eine Kostenabschätzung herrscht, kann es zu einer „nationalen Strategie“ und damit zur Nachhaltigkeit im Bereich LZA von Forschungsdaten kommen.
- Gibt es ***speziell geschultes Personal*** und speziell damit beauftragte Mitarbeiter (z.B. Wissenschaftler, Datenmanager, Informationsspezialisten, IT-Experten, LZA-Experten etc.), die sich schwerpunktmäßig

mit der LZA von Forschungsdaten befassen? Hintergrund: Ansonsten müssten sich die Wissenschaftler in viele und wesentliche Aspekte selbst einarbeiten, was insgesamt einem homogenen Ansatz in diesem Bereich nicht förderlich ist. Darüber hinaus sind die Wissenschaftler dann selbst für die LZA von Forschungsdaten verantwortlich, was weniger nachhaltig sein kann.

- Sind die Forscher und/oder die für das Datenmanagement Zuständigen eher ***fest bei einer Institution angestellt*** oder wird der Forschungsbedarf überwiegend durch Projekte mit zeitlich befristetem Personal abgedeckt? Hintergrund: Bei starker Personalfluktuation wächst der Bedarf an standardisierter Datenarchivierung und standardisierter Dokumentation der Daten. Gleichzeitig könnte jedoch die Motivation des Personals sinken, sich um eine qualitativ hochwertige Datenarchivierung zu kümmern, wenn den Wissenschaftlern klar ist, dass sie nur für eine kurze Zeit an einem speziellen Problem arbeiten.
- Werden für die Langzeiterhaltung der Forschungsdaten ***externe Dienste Dritter*** in Anspruch genommen? Hintergrund: Für kleinere Institutionen oder Fachbereiche kann Outsourcing von LZA-Aufgaben an größere Einheiten eine mögliche Lösung für die langfristige Erhaltung von Forschungsergebnissen sein.

3.3.5 Perspektiven und Visionen

- Gibt es ***spezifische Probleme*** und Herausforderungen, die bisher noch nicht benannt wurden und für die LZA relevant sind?
- Welche ***Möglichkeiten*** gibt es, die gemeinsame und langfristige Nutzung von Forschungsdaten (*data sharing*, *data reuse*, *data publication*) zu ***initiieren und zu befördern***? Z.B. durch die Unterstützung von verschiedenen Personengruppen (Wissenschaftler, geschulte IT/LZA-Experten etc.), von bestimmten Infrastrukturbereichen (z.B. persistente Identifizierung, Authentifizierung, technische Bereitstellung von Daten-Repositories) und Sonstiges wie z.B. Forschungsdateninfrastrukturen, Förderorganisationen, EU-Richtlinien, Anreizsystemen, Ausbildungsprogrammen etc.

- Welche ***Wunschvorstellungen/Visionen*** für die LZA von Forschungsdaten bestehen und wer kann bei der Umsetzung helfen? Was fehlt z.B. dafür, wo und wie kann auf externe Unterstützung (z.B. auf nationaler Ebene) zurückgegriffen werden?