Handbuch Forschungsdatenmanagement

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

BOCK + HERCHEN Verlag
Bad Honnef
2011

ISBN 978-3-88347-283-6

BOCK+HERCHEN Verlag, Bad Honnef
Printed in Germany

2.5 Forschungsdaten-Repositorien

Andreas Aschenbrenner [1], Heike Neuroth [2]
[1] Österreichische Akademie der Wissenschaften
[2] Niedersächsische Staats- und Universitätsbibliothek Göttingen

2.5.1 Einleitung

Vorangegangene Kapitel haben die zentrale Bedeutung und Rolle von Forschungsdaten in der Wissenschaft beschrieben. Die vertrauenswürdige Archivierung und Verfügbarkeit dieser Daten ist eine der Grundvoraussetzungen des wissenschaftlichen Diskurses. Repositorien spielen eine wichtige Rolle in diesem Kontext, so sind sie für die Langzeitarchivierung von Forschungsdaten verantwortlich, dienen der gemeinsamen Datenhaltung sowie ihrem Austausch und kollaborativen Nutzung innerhalb einer wissenschaftlichen *Community.*

Wissenschaftliche Daten unterlaufen in ihrem Lebenszyklus je nach wissenschaftlicher Methodik und Fach-*Community* unterschiedliche Stationen mit jeweils spezifischen Anforderungen an das Datenmanagement. Ebenso stellen die *Community* oder die Öffentlichkeit Anforderungen wie die Verifikation, Reproduzierbarkeit und Nachnutzbarkeit wissenschaftlicher Ergebnisse. Dieses Kapitel analysiert Repositorien aus technischer, organisatorischer und Nutzersicht. Angelehnt an die NESTOR Definition eines Langzeitarchivs (Dobratz & Schoger, 2010) verstehen die Autoren dieses Kapitels ein *Repository* als eine Organisation (bestehend aus Personen und technischen Systemen), die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit digitaler Objekte sowie für ihre Interpretierbarkeit zum Zwecke der Nutzung durch eine bestimmte Zielgruppe (vgl. „*designated community*“ des *Open Archival Information Systems* (OAIS) (NSSDC, o. J.) übernommen hat. Allerdings zeigt der heutige Stand, dass es sowohl weltweit, als auch national noch nicht für alle Fachdisziplinen entsprechende Repositorien gibt. Ein zumindest in den Naturwissenschaften erfolgreicher Ansatz stellt das *World Data System* (ICU WDS, 2010) dar, das aus dem *World Data Center System* (WDC) hervorgegangen ist (NGDC, o.J.). Auch hier soll eine Zertifizierung der existierenden *World Data Centers* (NGDC, 2009) für definierte organisatorische, politische, technische und inhaltliche Kriterien sorgen, damit Forschungsdaten vertrauenswürdig und nachhaltig vorgehalten werden. Auch in Deutschland gibt es eine Reihe von Forschungsdaten-Repositorien (vgl. Kap. 3.1), jedoch ist die Langzeitarchivierung von Forschungsdaten über alle wissenschaftlichen Disziplinen zurzeit noch nicht gesichert. Erste entscheidende Impulse für einen konzentrierten nationalen Ansatz kommen sicherlich von der GWK Initiative „Kommission Zukunft der Informationsinfrastruktur“ (WGL, 2011), deren im April 2011 vorgelegter Abschlussbericht als Basis für die in Vorbereitung befindlichen grundlegenden

Empfehlungen des Wissenschaftsrates zur Forschungsinfrastruktur in Deutschland dienen wird.

Es steht außer Frage, dass ohne fachspezifische Repositorien, die zum Beispiel auch komplexe Objektmodellierungen (z. B. in den Geisteswissenschaften bei kritischen Editionen oder bei Daten aus der Archäologie) oder verschiedene Versionen von Daten berücksichtigen, die Wissenschaft in den heutigen IT-gestützten Forschungsprozessen nicht optimal versorgt ist. Gerade der immer größer werdende Einsatz von Virtuellen Forschungsumgebungen für bestimmte Forschungsfragen und vernetzt arbeitende Forschergruppen zeigt, dass die Wissenschaft im Forschungsdatenmanagement unterstützt werden muss, hier spielen fachliche Repositorien eine entscheidende Rolle.

2.5.2 Definition, Funktionen und Aufgaben von Repositorien

Repositorien finden sich in den unterschiedlichsten Kontexten und mit den unterschiedlichsten Funktionsanforderungen (Aschenbrenner & Kaiser, 2005). Sie haben sich meist unabhängig voneinander entwickelt und noch heute ist der Bereich keineswegs überschaubar. Es gibt daher keine universelle Definition oder zeitlose Standards, auf die zurückgegriffen werden kann.

Heery und Anderson (2005) beschreiben Kernfunktionen von Repositorien als die technisch robuste sowie organisatorisch nachhaltige und vertrauenswürdige Verwaltung von (datei-basierten) Daten und zugehörigen Metadaten sowie die organisatorische und technische Einbettung der Schnittstellen für Ablage und Zugriff. In dieser Definition der Kernfunktionen wird das Zusammenspiel aus Technik und organisatorischen Maßnahmen deutlich.

Es ist auch eine klare Trennung zu verwandten Systemen wie *Code-Repositories* (vgl. Apache Subversion, Git), *Registries* (vgl. oft Datenbank-basierte Kataloge wie *Service Registries*, Metadaten-*Registries*) und Anderen. Ausschlaggebend für die Unterscheidung dieser Systeme ist zumeist die Art der Daten, die sie beherbergen, und wie sie mit ihnen umgehen. Im Kontext von Repositorien für Forschungsdaten arbeitet man oft mit dem Begriff der „digitalen Objekte". Digitale Objekte sind digitale Daten, die als intellektuelle Einheiten aus (einer oder mehreren) Dateien, zugehörigen Metadaten sowie einem Netzwerk aus anderen Objekten bzw. referenzierbaren Informationen bestehen können. Ein Beispiel wäre ein digitalisierter Brief mit der zugehörigen Transkription in Volltext, die jeweils beschrieben und mit anderen Briefen zu einer Korrespondenz verknüpft sind. Objekte können alle Arten von Daten umfassen – strukturiert, semi-strukturiert (z. B. XML-basiert) oder unstrukturierte Daten wie z. B. Bilder oder Videos.

Repositorien-Systeme decken je nach Fokus und Zielgruppe unterschiedliche Funktionen[1] ab, die sich oft auch in spezifischen Bezeichnungen spiegeln (z. B. „*institutional repositories*" für Publikationsserver, „*trusted repositories*" für

Langzeitarchivierungsumgebungen, oder „*open access repositories*“ für frei zugängliche Daten):

- Verwaltung von Informationsobjekten (Speicherkonzepte, Datenarten z. B. Publikationen in PDF, Bilder über 100 MB, *stream*-bare Videos)
- Metadatenverwaltung zur Identifikation, Administration und langfristigen Erhaltung von Informationsobjekten sowie deren Einbettung in einen inhaltlichen, intellektuellen Kontext
- Vernetzung bzw. (standardisierte) Verknüpfung der Objekte untereinander mit Kontextdaten
- *Workflow*-Unterstützung zur Registrierung von Informationsobjekten (manueller *Ingest-Workflow* und automatischer Datentransfer)
- Zugang zu und Nachnutzung von Forschungsdaten durch persistente Identifikation, Suchmechanismen, Schnittstellen (z. B. *Open Archives Initiative* (OAI)[2])
- Präsentation, Einbettung in Nutzungsumgebungen, Unterstützung von kollaborativen und kooperativen Arbeitsformen
- Analyse der Nutzung (Nutzungsstatistiken) und Archivinhalte (z. B. *Text Mining*, Visualisierung)
- Berücksichtigung von rechtlichen Rahmenbedingungen (Datenschutz, Urheberrecht etc.)
- Mechanismen zur Langzeitarchivierung

Systeme können sich zum Teil erheblich darin unterscheiden, wie sie diese Kernfunktionen umsetzen und welche Zusatzfunktionalitäten sie anbieten. Gerade im Aufbau einer *Repository*-basierten Forschungsumgebung, die mitunter spezifisch auf den jeweiligen Anwendungsfall und Forschungskontext zugeschnitten sein muss, ist daher oft viel Anpassungsarbeit oder Eigenentwicklung nötig.

2.5.3 Auswahl Software

Während früher ein *Repository* eher verwendungsspezifisch und häufig ad-hoc entwickelt wurde, stellt sich die Situation heutzutage deutlich verändert dar. Eine breite *Community* teilt ähnliche Anforderungen an solche Systeme, tauscht

1. Diese kurze Auflistung kann nicht vollständig sein und listet nur einige Kern-Funktionalitäten unterschiedlicher Fokusgruppen und Ziele. Für weitere technische Funktionen siehe z. B. den ISO Standard zu einem „*Open Archival Information System*“ (OAIS) (CCSDS, 2002), das DELOS *Reference Model* (DELOS, o.J.) und andere.
2. http://www.openarchives.org/ [Zugriff am 14.08.2011].

ihre Erfahrungen hierzu aus und entwickelt gemeinschaftlich und nach dem *Open Source* Prinzip entsprechende Softwaresysteme.

Vor allem im Bereich von Publikationsservern zeichnet sich eine gewisse Konvergenz der Technologien ab. Bereits in den 90er Jahren sind erste Gesamtpakete für Repositorien aufgekommen, darunter der CERN *Document Server*[3] oder der Hochschulschriftenserver der Universität Stuttgart OPUS[4]. Andere Institutionen haben eigene Systeme entwickelt oder bestehende Systeme aufgegriffen und für ihre Bedürfnisse angepasst, wo dies sinnvoll und möglich war.

Heute gibt es eine Vielzahl von *Repository* Systemen, wie z. B. die Auflistung von OSI (2004) oder die Überblicksarbeit von Borghoff, et al. (2005) zeigen. Die ebenso weit verbreiteten *Web-Content-Management*-Systeme (z. B. Plone[5], Drupal[6], Joomla![7]) eignen sich üblicherweise nicht als Datenrepositorien, da sie oft Workflows für Metadaten-Beschreibungen nicht unterstützen bzw. aus Langzeitarchivierungssicht nicht robust genug sind. Besonders gefragt sind zurzeit vor allem folgende drei *Repository* Systeme, die auch auf der internationalen *OpenRepositories*[8] Konferenz stark vertreten sind:

- ***EPrints***[9]. *Out-of-the-Box* Komplettsystem für Publikationen mit weitgehend vorgegebenen Strukturen und einfacher Verwaltung.
- ***DSpace***[10]. Komplettsysteme für Publikationen mit einem vorstrukturierten *Workflow*-System zur Eingabe von Metadaten, etc. beim *Ingest*.
- ***Fedora***[11]. *Middleware* zur Modellierung und Verwaltung von Daten, wobei unterschiedliche Projekte auch spezifischere Nutzerumgebungen (z. B. eSciDoc[12], Fez[13], Muradora[14]) auf Fedora aufsetzen.

Anfang 2011 weist das Verzeichnis OpenDOAR[15] z. B. über 1.800 laufende *Repository*-Installationen nach, davon nutzen ein Drittel DSpace gefolgt von

3. http://cds.cern.ch/ [Zugriff am 14.08.2011], http://www.cern.ch, [Zugriff am 14.08.2011].
4. http://elib.uni-stuttgart.de/opus/ [Zugriff am 14.08.2011].
5. http://plone.org/ [Zugriff am 14.08.2011].
6. http://www.drupal.de/ [Zugriff am 14.08.2011].
7. http://www.joomla.de/ [Zugriff am 14.08.2011].
8. http://www.openrepositories.org/ [Zugriff am 14.08.2011].
9. http://www.eprints.org/ [Zugriff am 14.08.2011].
10. http://www.dspace.org/ [Zugriff am 14.08.2011].
11. http://www.fedora-commons.org/ [Zugriff am 14.08.2011].
12. http://www.escidoc.org/ [Zugriff am 14.08.2011].
13. http://sourceforge.net/projects/fez/ [Zugriff am 14.08.2011].
14. http://www.muradora.org/ [Zugriff am 14.08.2011].
15. http://www.opendoar.org/ [Zugriff am 14.08.2011].

EPrints. *DSpace* wurde ursprünglich für das *Massachusetts Institute of Technology* (MIT)[16] entwickelt, wird inzwischen durch eine große *Community* („DSpace Federation“) weiterentwickelt und durch die Firma HP auch kommerziell vertrieben. Neben diesen drei *Open Source* Systemen hat jüngst auch z. B. Microsoft mit einem eigenen Produkt, dem Publikationsserver Zentity[17], aufhorchen lassen.

Diese Softwarepakete sind zwar als Publikationsserver weit verbreitet, aber für Forschungsdaten sind nicht alle einsetzbar. *Workflows* und Datenmodelle in *EPrints* und *DSpace* sind primär auf dokument-artige Publikationen (z. B. Dissertationen, Journale, Berichte) ausgelegt und für andere Arten von Forschungsdaten (z. B. veränderliche Objekte, bestehend aus mehreren Dateien mit komplexen Metadaten) ungeeignet.

Von den genannten Systemen ist nur Fedora so flexibel, dass es ideal für die Verwaltung und Archivierung von Forschungsdaten dienen kann. Zwei Eigenschaften seien hier speziell herausgehoben:

- (1) Die Fedora Service-Architektur[18] ist die Basis einer offenen, evolutionären Umgebung für wissenschaftliche *Workflows*, und
- (2) Fedora-Mechanismen zur Metadatenmodellierung (vgl. *Content Model Architecture* (Fedora Commons, 2007)) ermöglichen die Beschreibung unterschiedlichster Datenarten, wie es beispielsweise das Fedora-basierten eSciDoc[19] für die unterschiedlichen Disziplinen in der Max-Planck-Gesellschaft umsetzt.

Neben Fedora seien noch zwei weitere *Repository*-Pakete genannt: iRODS und Tupelo. Diese Systeme eignen sich besonders für Forschungsdaten, da sie (a) für große Datenmengen skalieren, (b) Modellierbarkeit von Daten und Metadaten unterstützen und (c) die Systeme aus Langzeitarchivierungssicht robust genug sind.

- ***iRODS***[20] – stammt von Datenzentren und ist besonders zur effizienten Verwaltung von sehr großen Datenmengen geeignet. iRODS ist ein weitgehend monolithisches System und mit zumeist proprietären Schnittstellen, wächst aber durch eine weltweite *Open Source Community.*

16. http://web.mit.edu/ [Zugriff am 14.08.2011].
17. http://research.microsoft.com/en-us/projects/zentity/ [Zugriff am 14.08.2011].
18. Vgl. z. B. das Konzept der „Disseminatoren“ im ursprünglichen Architekturkonzept (Payette & Lagoze, 1998)
19. http://www.escidoc.org/ [Zugriff am 14.08.2011].
20. http://irods.sdsc.edu/index.php/Main_Page [Zugriff am 14.08.2011].

- ***Tupelo***[21] – ist eine kleine Initiative mit einer leichtgewichtigen Software, die sich primär auf die Daten- und Metadatenmodellierung mithilfe semantischer Technologien konzentriert.

2.5.4 Architektur, Technologien, Standards

Trotz der unterschiedlichen Systeme und der Dynamik in der *Repository-Community* mit immer neuen Entwicklungen gibt es bei allen Software-Paketen einen deutlichen technischen Trend zu Offenheit und Interoperabilität. Dieser Trend entsteht nicht nur durch eine gemeinsame Ideologie der Software-Macher, sondern begründet sich auf die Anforderungen der Organisationen, die *Repository*-Systeme betreiben, sowie der Endnutzer, die (mitunter mehrere) *Repositories* und Zusatzdienste für ihre wissenschaftliche Arbeit benötigen. Somit betreffen die im Folgenden vorgestellten Architekturkonzepte und Standards durchaus alle *Repository*-Systeme – auch kommerzielle, wie die von Microsoft oder andere Eigenentwicklungen.

Abgeleitet von den in Abschnitt „Definition, Funktionen, Aufgaben" vorgestellten Anforderungen, kann man generell drei konzeptuelle Schichten in *Repository*-Systemen unterscheiden: *Storage*, Datenmanagement und Nutzung.

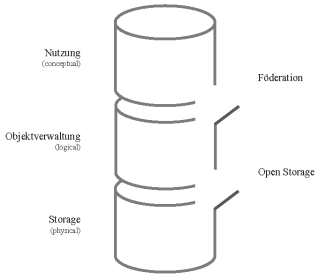

Abb. 1: Schichten-Architektur mit den drei konzeptuellen Schichten – *Storage*, Objektverwaltung, und Nutzung – angelehnt an die 3 Ebenen von Thibodeau (2002). Rechts: Bezeichnung der Interoperabilitätsebenen „Föderation" und „*Open Storage*".

2.5.4.1 Architekturschicht: Storage

Die *Storage*-Ebene beherbergt digitale Objekte – also Daten gemeinsam mit zugehörigen Metadaten. Aus Gründen der Stabilität entscheiden sich *Repository*-Systeme auf dieser Ebene zumeist für eine datei-basierte Ablage (also nicht in Datenbanken), und ermöglichen die Rekonstruktion aller Informationen aus den Dateien.

Während kleinere Repositorien mit einem lokalen Server ihre kompletten *Storage*-Anforderungen abdecken können, entscheiden sich manche Repositorien zur Auslagerung der Daten in ein Datenzentrum bzw. Rechenzentrum. Gerade für Forschungsdaten liegt ein wesentlicher Vorteil bei der Auslagerung

21. http://tupeloproject.ncsa.uiuc.edu/ [Zugriff am 14.08.2011].

des *Storage* darin, dass ggf. größere Datenmengen verwaltet werden können, mehrere Repositories auf eine gemeinsame *Storage*-Ebene zugreifen können und dass Aufgaben zur *Bit-Preservation* (z. B. Datenreplikation, *Tape-Backup*, Integritätstests) gekapselt werden können.[22]

2.5.4.2 Architekturschicht: Objektverwaltung

Das Datenmanagement in Repositorien verknüpft Daten und Metadaten zu Objekten, beschreibt Relationen zwischen Objekten, versioniert Objekte, verknüpft sie mit unterschiedlichen Darstellungs- und Zugriffsmechanismen und bettet sie in (existierende) Softwareumgebungen ein. Verbreitete Standards schließen Daten- und Metadatenbeschreibungsformate (z. B. Dublin Core[23], METS[24]) wie auch Standards für APIs (vgl. z. B. *Common Repository Interfaces Group* (CRIG)[25]) mit ein. Gerade Forschungsdaten verlangen oft eine große Flexibilität und Ausdrucksfähigkeit in der Daten- und Metadaten-Modellierung. Anforderungen an z. B. Zugriffsrechte und Veränderbarkeit der Daten können sich zwischen Forschungskontexten und Forschungsprojekten stark unterscheiden.

2.5.4.3 Architekturschicht: Nutzung

Während Publikationsrepositorien primär auf die Einfuhr und die Suche von Publikationen ausgerichtet sind, ist die Bandbreite der Nutzungsszenarien bei Forschungsdaten-Repositorien wesentlich breiter. Je nach Forschungskontext sollten Daten z. B. direkt von Messinstrumenten in das *Repository* überführt, in wissenschaftliche Workflows eingebettet oder in bestehende Forschungsapplikationen integriert werden.

Aufgrund dieser Bandbreite an Nutzungsszenarien und Forschungskontexten ist es kaum sinnvoll, generelle technische Standards auf einer Nutzungsebene zu erarbeiten. Beratungsangebote und Leitfäden wie die von WissGrid (2011) können allerdings wertvolle Erfahrungen zum Aufbau spezialisierter Forschungsumgebungen und Ratschläge zur Nachnutzung und Vernetzung von existierenden Werkzeugen geben.

---

22. Für *Cross-Repository* Interoperabilität reicht eine *Storage*-Ebene zur Dateiablage nicht aus. Die *Repository-Storage*-Ebene bezieht auch standardisierte Mechanismen zur Ablage von Metadaten, Datenversionierung, *Locking*, etc mit ein. Vgl. z. B. *Fedora High Level Storage* (Fedora Repository Development, 2007).
23. http://dublincore.org/ [Zugriff am 14.08.2011].
24. http://www.loc.gov/standards/mets/ [Zugriff am 14.08.2011].
25. http://www.ukoln.ac.uk/repositories/digirep/index/CRIG [Zugriff am 14.08.2011].

2.5.4.4 Offene Repository-Umgebungen

Technisch gesehen eröffnet der Trend zu Offenheit und Interoperabilität ganz neue Möglichkeiten, die vor allem im Umfeld von Forschungsdaten noch weiter erforscht werden müssen. Dieser Trend wird allein schon dadurch gefördert, dass manche Institutionen mehrere Installationen von unterschiedlichen Systemen bei sich führen, um unterschiedlichen Anforderungen in ihrer Organisation gerecht zu werden. Aber auch die Sichtbarkeit der *Open Access* Bewegung (Berliner Erklärung, 2003) und aufkommende *e-Science* Mechanismen zur Vernetzung unterschiedlichster Daten und Dienste untereinander[26] fördern die Offenheit und Interoperabilität von *Repository* Systemen.

Für die Interoperabilitäts-Ebene „*Open Storage*“ (vgl. Abb. 1: Schichten-Architektur mit den drei konzeptuellen Schichten – *Storage*, Objektverwaltung, und Nutzung – angelehnt an die 3 Ebenen von Thibodeau (2002). Rechts: Bezeichnung der Interoperabilitätsebenen „Föderation“ und „*Open Storage*“.) gibt es derzeit noch keine eindeutigen Standards. Derzeit arbeitet z. B. das *Duraspace*-Projekt (Minton Morris, 2008) an einer generellen *Cloud*-basierten *Storage*-Ebene für Fedora und DSpace, die für den Produktivbetrieb geeignet ist und auch Anforderungen der Langzeitarchivierung (bzw. zumindest *Bit-Preservation*) abdecken wird.

Förderationsstandards wie OAI-PMH (Open Archives, o.J.), OAI-ORE (Pepe et al., 2009) und Zing[27] verschränken das Datenmanagement unabhängiger Repositorien zu einem übergreifenden, virtuellen Repositorium. Nutzer von Föderationen wie DRIVER (*Digital Repository Infrastructure Vision for European Research*)[28] oder Europeana[29] haben dadurch unmittelbaren Zugriff zu einer Vielzahl von institutionellen und thematischen Repositorien. Auch im Bereich von Forschungsdaten werden diese Standards bereits vereinzelt eingesetzt (WissGrid, 2010). Allerdings werden erst die Entwicklungen der nächsten Jahre zeigen, wie diese Standards für neue Anwendungen im Kontext von Forschungsdaten eingesetzt werden können – z. B. Analyse und Visualisierung von Forschungsdaten sowie Rechtemanagement und Aufgabensteuerung für Forschergruppen – und wie Repositorien-basierte Infrastrukturen den Aufbau und die Vernetzung von virtuellen Forschungsumgebungen verändern (Aschenbrenner et al., 2010).

26. Zum Beispiel die Verknüpfung von Publikationen mit den zugrunde liegenden wissenschaftlichen Rohdaten und Diensten zur Analyse der Daten. Vgl. DRIVER (2009).
27. Im Rahmen der ZING-Initiative (Z39.50 International Next Generation) entstand der technische Standard SRU Search/ Retrieval via URL (Library of Congress, 2011).
28. http://www.driver-repository.eu/ [Zugriff am 14.08.2011].
29. http://www.europeana.eu/ [Zugriff am 14.08.2011].

2.5.5 Weitere Aspekte

Neben technologischen Aspekten gibt es eine Reihe weiterer Überlegungen, die frühzeitig berücksichtigt werden müssen und Einfluß nehmen auf den Aufbau und die (Weiter-) Entwicklung von Forschungsdaten-Repositorien.

Dies beinhaltet zum Beispiel Vorüberlegungen[30] zu Strategie und Management und umfaßt Definition (*mission statement*), Zielgruppe(n), notwendige Kooperationen (z. B. Rechenzentrum, Bibliothek) und Regelungen für den potentiellen Nachfolgebetrieb im Notfall. Sogenannte *Service-Level-Agreements* (SLA) müssen ausgearbeitet werden und die verschiedenen Stufen des Angebotes (von *bitstream preservation* bis hin zu „echter *data curation*“) verständlich und transparent dokumentiert sein. Ein Betriebsplan, der auch Qualitätskontrolle und Überwachung im Sinne von Monitoring umfaßt, ist ebenfalls integraler Bestandteil eines Repositoriums. Ein stabiler Finanzierungsplan und mittel- bis langfristige Überlegungen zu Personalplanungen inklusive Aufbau notwendiger Qualifikationen und Kompetenzen gehören ebenfalls dazu.

Angaben über die zu archivierenden Sammlungen und Objekte müssen dokumentiert sein inklusive notwendiger Standards (z. B. Metadatenstandards) und rechtlicher Rahmenbedingungen. Die Anforderungen zum Beispiel in Bezug auf Authentizität, Integrität, Nachnutzbarkeit, Sicherheit und Verfügbarkeit sind klar zu definieren. Ein stetiger Abgleich der Anforderungen mit dem bestehendem Dienstleistungsangebot ist zu leisten. Vereinbarungen und Verträgen über Rechte, Verpflichtungen, Haftungen und Umsetzungen zwischen den unterschiedlichen Akteuren sind zu treffen und zu dokumentieren. Die einzelnen Arbeitsabläufe sind mit klarer Rollenverteilung und Festlegung von Verantwortlichkeiten zu regeln. Die Erfordernisse bei der Umsetzung durch eine IT-Infrastruktur und Technologie inklusiver langfristiger Technologiestrategie sind festzulegen.

Die hier beschriebenen Aspekte geben nur einen kleinen Einblick in die nötigen (Vor-)Überlegungen wieder und zeigen auf, dass ein wesentlicher Bereich im Vorfeld, abhängig von den unterschiedlichen Beteiligten und den organisatorischen sowie strukturellen Rahmenbedingungen, zu klären ist. Die demnächst veröffentlichten DIN[31] Richtlinien und ISO Standard[32] im Bereich der vertrauenswürdigen Zertifizierung von Repositorien geben einen umfassenden Einblick. Beispiele für Forschungsdaten-Archive in Deutschland wie das Deutsche

---

30. Nach Ludwig, J. & Strathmann, S.: „Zehn-Punkte-Plan zum Aufbau eines Angebots zur Langzeitarchivierung und zum Forschungsdatenmanagement“, Veröffentlichung in Vorbereitung.
31. DIN 31644, vgl. auch NESTOR (2010).
32. ISO 16363 für vertrauenswürdige Langzeitarchive.

Fernerkundungszentrum (DFD[33]), Pangaea[34] für die Geo- und Umweltwissenschaften oder die *World Data Center* (WDC MARE[35], WDC Climate[36], WDC RSAT[37]) zeigen, dass die intensive Zusammenarbeit mit den jeweiligen Fachdisziplinen unerläßlich für die Akzeptanz solcher Repositorien ist. Einerseits müssen die Fachwissenschaftler eng bei der Formulierung der Anforderungen eingebunden werden, andererseits müssen sie klar den Nutzen und den Mehrwert solcher Langfrist-Archive erkennen, um ihre Daten dort abzulegen. Die Aufgabe der Langzeitarchivierung von Forschungsdaten muss als *Community*-Aufgabe verstanden werden. Nicht umsonst finden sich in bereits gut organisierten, zum Teil international vernetzten Fachdisziplinen mit einem in der Regel überdurchschnittlich hohen Aufkommen von Forschungsdaten bereits erste stabile Ansätze von Forschungsdaten-Repositorien.

2.5.6 Aktuelle Entwicklungen, Diskussionen und Ausblick

In den letzten Jahren hat es eine Reihe von Aktivitäten, Entwicklungen und Diskussionen im Bereich von Forschungsdaten gegeben. So hat zum Beispiel die Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen im Juni 2010 im Rahmen der Arbeitsgruppe Forschungsdaten (Allianz, o.J.) Grundsätze (Allianz, 2010) zum Umgang mit Forschungsdaten veröffentlicht, die unter anderem von den Organisationen Deutsche Forschungsgemeinschaft (DFG), Fraunhofer-Gesellschaft, Helmholtz-Gemeinschaft, Hochschulrektorenkonferenz (HRK), Leibniz-Gemeinschaft, Max-Planck-Gesellschaft und Wissenschaftsrat unterschrieben wurden. Diese Grundsätze beginnen mit einer Präambel, in der festgehalten wird, dass „Qualitätsgesicherte Forschungsdaten ... einen Grundpfeiler wissenschaftlicher Erkenntnis [bilden] und ... unabhängig von ihrem ursprünglichen Erhebungszweck vielfach Grundlage weiterer Forschung sein [können]“. Weiter heißt es „Die nachhaltige Sicherung und Bereitstellung ... bildet eine strategische Aufgabe, zu der Wissenschaft, Politik und andere Teile der Gesellschaft gemeinsam beitragen müssen“. Die Eckpunkte der Grundsätze beinhalten Sicherung und Zugänglichkeit, Unterschiede der wissenschaftlichen Disziplinen, Wissenschaftliche Anerkennung, Lehre und Qualifizierung, Verwendung von Standards sowie Entwicklung von Infrastrukturen.

---

33. http://www.dlr.de/ [Zugriff am 14.08.2011].
34. http://www.pangaea.de/ [Zugriff am 14.08.2011].
35. http://www.wdc-mare.org/ [Zugriff am 14.08.2011].
36. http://www.mad.zmaw.de/wdc-for-climate/ [Zugriff am 14.08.2011].
37. http://wdc.dlr.de/ [Zugriff am 14.08.2011].

Im Jahr 2010 wurde die *„Kommission Zukunft der Informationsinfrastruktur"* (WLG, 2011) gebildet mit dem Auftrag, ein nationales Gesamtkonzept für die Informationsinfrastruktur in Deutschland zu erarbeiten und 2011 vorzulegen. Zu den insgesamt acht eingesetzten thematischen Arbeitsgruppen findet sich auch eine AG Forschungsdaten, die im Oktober 2010 dem Steuerungsgremium der KII einen Bericht vorgelegt hat, der Aspekte wie Status Quo in Deutschland, internationaler Kontext, Nutzererwartungen, Handlungsbedarf, Visionen, Querschnittsthemen, Ressourcenabschätzung und Aufgaben und Rahmenbedingungen abdeckt. Letztendlich sollen daraus auch für den Themenbereich Forschungsdaten Handlungsempfehlungen für den Gesamtbericht[38] der KII abgeleitet werden, die darüber Auskunft geben, wie in Deutschland das Thema Forschungsdaten und Forschungsdaten-Repositorien gesamtheitlich angegangen und umgesetzt werden kann. Bei diesen Diskussionen hat sich klar herauskristallisiert, dass jede datenintensive Disziplin einen Datenmanagementplan entwickeln sollte und dass eine Initial- und Grundfinanzierung für den Aufbau und den Betrieb von Dateninfrastrukturen nötig ist. Die daraus abgeleiteten Handlungsempfehlungen umfassen technische (z. B. Diensten für die Zitierbarkeit von Forschungsdaten), organisatorische (z. B. Festlegung von klaren Verantwortlichkeiten und organisatorischen Strukturen), finanzielle (z. B. Grundfinanzierung) rechtliche (z. B. transparente rechtliche Regelungen) und sonstige Aspekte (z. B. Etablierung von Anreizsystemen für die Wissenschaftler). Dabei ist die Anerkennung der Forschungsdaten als nationales Kulturgut eine wesentliche Grundbedingung.

Insgesamt kann festgehalten werden, dass sich bei dem Thema Forschungsdaten-Repositorien in Deutschland in den letzten Jahren viel bewegt hat, auf fachwissenschaftlicher, technologischer und politischer Ebene. Dabei hat sich auch gezeigt, dass die Technologie nur eine Seite der Herausforderungen darstellt. Die andere Seite besteht darin, sowohl die politischen als auch strukturellen Rahmenbedingungen für den Aufbau und den dauerhaften Betrieb von fachwissenschaftlichen Forschungsdaten-Repositorien zu schaffen, als auch die Fachwissenschaftler sowie die weiteren Akteure (Infrastruktureinrichtungen wie Rechenzentren und Bibliotheken) in einem organisatorischen Gesamtkonzept sinnvoll einzubeziehen. Es bleibt abzuwarten, wie die Öffentlichkeit und die Politik auf den Gesamtbericht der KII reagieren und welche konkreten Maßnahmen in Deutschland ergriffen und umgesetzt werden.

38. Der Bericht der Arbeitsgruppe „Forschungsdaten" ist im „Gesamtkonzept" publiziert, vgl. WGL, 2011.

Literaturhinweise

Allianz der deutschen Wissenschaftsorganisationen, 2010. *Grundsätze zum Umgang mit Forschungsdaten.* Online. http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/grundsaetze/ [Zugriff am 14.08.2011].

Allianz der deutschen Wissenschaftsorganisationen, o.J. *Forschungsprimärdaten*. Online: http://www.allianzinitiative.de/de/handlungsfelder/forschungsdaten/ [Zugriff am 14.08.2011].

Aschenbrenner, A. & Kaiser, M., 2005. *White Paper on Digital Repositorys. reUSE! Deliverable.* Online: http://www2.uibk.ac.at/reuse/docs/reuse-d11_whitepaper_10.pdf [Zugriff am 14.08.2011].

Aschenbrenner, A. Blanke, T. Küster, M. W. & Pempe, W., 2010. Towards an Open Repository Environment. *Journal of Digital Information* (JoDI), 11(1).

Berliner Erklärung, 2003. *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.* (Stand: 22.10.2003) Online: http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/ [Zugriff am 09.08.2011].

Borghoff, U. M. et al., 2005. *Vergleich bestehender Archivierungssysteme*. (NESTOR-Materialien, 3) Online: http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20050117016 [Zugriff am 14.08.2011].

CCSDS (Consultative Committee for Space Data Systems), 2002. *Reference Model for an Open Archival Information System (OAIS)*. (CSSDS 650.0-B-1) (Jan. 2002) Online: http://public.ccsds.org/publications/archive/650x0b1.pdf [Zugriff am 14.08.2011].

DELOS, o.J. *A Reference Model for Digital Library Management Systems*. Online: http://www.delos.info/index.php?option=com_content&task=view&id=345&Itemid= [Zugriff am 14.08.2011].

Dobratz, S. & Schoger, A., 2010. Kapitel 8.3 Evaluierung der Vertrauenswürdigkeit digitaler Archive. In: Heike Neuroht et al., Hrsg. 2010. *NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* (Version 2.3.) Online: http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_78.pdf [Zugriff am 14.08.2011].

DRIVER (Digital Repository Infrastructure Vision for European Research), 2009. *Enhanced Publications*. Online: http://www.driver-repository.eu/Enhanced-Publications.html [Zugriff am 14.08.2011].

Fedora Commons, 2007. *The Fedora Content Model Architecture (CMA)*. (Version 3.0 Beta 1) Online: http://www.fedora-commons.org/

documentation/3.0b1/userdocs/digitalobjects/cmda.html [Zugriff am 14.08.2011].

Fedora Repository Development, 2010. *High Level Storage.* (Stand: 07.12.2010) Online: https://wiki.duraspace.org/display/FCREPO/High+Level+Storage [Zugriff am 14.08.2011].

Heery, R. & Anderson, S., 2005. *Digital Repositorys Review.* Online: http://www.jisc.ac.uk/uploaded_documents/digital-Repositorys-review-2005.pdf [Zugriff am 16.08.2011].

ICU WDS (International Council for Science World Data System), 2010. *ICSU World Data System (Home)*.Online: http://icsu-wds.org/ [Zugriff am 14.08.2011].

Library of Congress, 2011. *SRU Search / Retrieval via URL*. (Stand: 04.08.2011) Online: http://www.loc.gov/standards/sru/ [Zugriff am 14.08.2011).

Minton Morris, C., 2008. *DSpace Foundation and Fedora Commons Receive Grant from the Mellon Foundation for DuraSpace*. (Stand: 11.11.2008, 9:21 am) Online: http://expertvoices.nsdl.org/hatcheck/2008/11/11/dspace-foundation-and-fedora-commons-receive-grant-from-the-mellon-foundation-for-duraspace/ [Zugriff am 14.08.2011].

NESTOR, 2010. *AG Vertrauenswürdige Archive – Zertifizierung (aufgegangen in DIN NABD 15)*. (Stand: 14.12.2010) Online: http://www.langzeitarchivierung.de/arbeitsgruppen/agkritkat.htm [Zugriff am 14.08.2011].

NGDC (National Geophysical Data Center), o.J. *World Data System*. Online: http://www.ngdc.noaa.gov/wdc/wdcmain.html [Zugriff am 14.08.2011].

NGDC (National Geophysical Data Center), 2009. *List of current WDCs*. (Last Revised: 30.06.2006) Online: http://www.ngdc.noaa.gov/wdc/list.shtml [Zugriff am 14.08.2011].

NSSDC (National Space Science Data Center), o. J. *ISO Archiving Standards*. Online: http://nssdc.gsfc.nasa.gov/nost/isoas/ [Zugriff am 16.8.2011].

Open Archives, o.J. *Open Archives Initiative – Protocol for Metadata Harvesting*. Online: http://www.openarchives.org/pmh/ [Zugriff am 14.08.2011].

OSI (Open Society Institute), 2004. *Guide to Institutional Repository Software*. 3. ed. Online: http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf [Zugriff am 14.08.2011].

Payette, S. & Lagoze, C., 1998. Flexible and Extensible Digital Object and Repository Architecture (FEDORA). In: Nikolaou, C., ed. 1998. Research and advanced technology for digital libraries, *Second European Conference on Research and Advanced Technology for Digital Libraries*. (LNCS 1513) Heraklion, Kreta, Griechenland 21.-23. Sept. 1998. Berlin: Springer, S. 41–59. Online: http://www.cs.cornell.edu/payette/papers/ECDL98/FEDORA.html [Zugriff am 14.08.2011].

Pepe, A., Mayernik, M., Borgman, C. L. & Van de Sompel, H., 2009. From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web. *JASIST Journal of the American Society for Information Science and Technology,* 61(3). Online: http://arxiv.org/ftp/arxiv/papers/0906/0906.2549.pdf [Zugriff am 14.08.2011].

Thibodeau, K., 2002. *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years.* Online: http://www.clir.org/pubs/reports/pub107/thibodeau.html [Zugriff am 14.08.2011].

WissGrid, 2010. *WissGrid-Spezifikation: Grid-Repository.* Online: http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.5.2-grid-repository-spezifikation.pdf [Zugriff am 14.08.2011].

WissGrid, 2011. *Grid für die Wissenschaft*. (Stand: 18.04.2011) Online: http://www.wissgrid.de [Zugriff am 14.08.2011].

WGL (Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz e.V./ Leibniz Gemeinschaft), 2011. *Informationsstruktur*. Online: http://www.wgl.de/?nid=infrastr&nidap=&print=0 [Zugriff am 14.08.2011].