IANUS

Expertise – Aufbau eines Forschungsdatenzentrums Archäologie und Altertumswissenschaften

Autor:
Prof. Dr. Gerhard Schneider
Rechenzentrum der Universität Freiburg
Hermann-Herder-Str. 10
79104 Freiburg
gerhard.schneider@rz.uni-freiburg.de

Version 1.0
17. Juli 2013

Koordination

Förderung

Autor: Prof. Dr. Gerhard Schneider

Titel: Expertise – Aufbau eines Forschungsdatenzentrums Archäologie und Altertumswissenschaften

Sprache: Deutsch

DOI: 10.13149/000.ekvdia-n

Zitierhinweise:

Gerhard Schneider (2013) Expertise – Aufbau eines Forschungsdatenzentrums Archäologie und Altertumswissenschaften. [Version 1.0] Hrsg. IANUS. doi: 10.13149/000.ekvdia-n

Kontakt:

ianus-fdz@dainst.de

www.ianus-fdz.de

Version	Autoren	Datum	Beschreibung
1.0	Gerhard Schneider	17.07.2013	Erste finale Fassung

Inhaltsverzeichnis

Zusammenfassung

Das Dokument beinhaltet eine Expertise zum Aufbau und Organisation eines möglichen Forschungsdatenzentrums (FDZ) für die Altertumswissenschaften. Ausgehend von der Aufgabenstellung, einen Überblick der zu berücksichtigenden Aspekte und der nationalen sowie internationalen Anstrengungen auf dem Gebiet des Forschungsdatenmanagements zu gewinnen, werden Vorschläge für Realisierungen und Workflows herausgearbeitet. Dazu wird einerseits der Blickwinkel der Datenproduzenten und ihrer Anforderungen und Wünsche eingenommen, da die Akzeptanz dieser Gruppe ganz wesentlich die nachhaltige Existenz eines solchen FDZ beeinflusst. Für diese Aufgabe sind effektive Ingest-Workflows anzubieten, die eine möglichst breite Palette an verschiedenen Datenformaten unterstützen. Andererseits ist für zukünftige Forschungsprojekte und Kooperationen die Sicht der Datennutzer relevant, die ein Interesse daran haben, Datensätze effizient finden zu können. Ein gemeinsames FDZ hilft Standards in den Bereichen Metadaten, Persistent Identifiers, in Umgang und Präsentation von Daten zu etablieren und diese über Teildisziplinen hinweg durchzusetzen.

Die unterschiedlichen Anforderungen an das FDZ schlagen sich zudem in der Domäne des Datenmanagements als auch in der langfristig angelegten Publikationsdomäne nieder. Während die technologischen Anforderungen an eine reine Datenspeicherung und langfristige Bitstream-Preservation weitgehend geklärt sind und durch standardisierte Angebote als Vorleistung eingekauft werden können, sind die Workflows für Datenaufnahme und die spätere Datenwiedergabe ein zentraler Baustein. Das FDZ kann daher in einer 3-Tier-Architektur, die die Zugangsschicht direkt mit anbietet, oder einer 2+1–Tier-Architektur mit teilweise ausgelagerter Zugangsschicht angelegt werden. Das FDZ muss ausreichend finanziell ausgestattet sein, um die personellen Ressourcen für Beratung und vorgesehenen Dienstleistungen langfristig bereitstellen zu können. Die Kosten ergeben sich neben den reinen Aufwendungen für die Speicherung aus den Kosten für die Aufnahme der Daten, die Begleitung von Forschungsprojekten durch Schulungen und Beratungsleistungen. In der Archiv-Phase treten die Life-Cycle-Kosten der Daten in Form von Anpassungen und Bereithaltung für die Nutzung auf. Hinzu kommen je nach Ausgestaltung Kosten für Unterstützungsdienste wie IT- und allgemeine Infrastrukturen bis hin zu Raum- und Gebäudeaufwendungen.

Die Aufgabe eines FDZ in den Altertumswissenschaften ist daher mehr eine moderierende. Es sind die Leistungen der Anbieter des Bitstream-Preservation konsequent zu überwachen und Strategien einer Mehr-Anbieter-Philosophie umzusetzen. Darüber hinaus sind Strategien und Lösungsmethoden bereit zu halten bzw. bei Bedarf hinzuzukaufen, die eine Reaktivierung von Daten ermöglichen, welche in einer proprietären Umgebung gewonnen und vorgehalten wurden. Schließlich sind, zusammen mit der Fachwissenschaft, Methoden zum nachvollziehbaren Entfernen überflüssiger Daten zu entwerfen, um die Kosten für die Aufbewahrung von Daten kontrollieren zu können. Die Langlebigkeit des FDZ erlaubt darüber hinaus die zuverlässige Verwaltung der Datensätze weit über die jeweilige Forschungsförderung hinaus unter Beachtung der jeweiligen Zugriffsrechte. Zur Vermeidung von Lizenzproblemen sollte von Anfang an auf „open data“ als Grundphilosophie geachtet werden. Auf die Programmierung eigener Lösungen (wie Zugriffsportale) sollte so weit wie nur möglich verzichtet werden, um die Kostenfalle für den Pflegeaufwand solcher Ansätze zu vermeiden.

Zuverlässigkeit und solides planerisches Handeln sichern dem FDZ die Zukunft.

Vorbemerkungen

Die Thematik von Forschungsdatenzentren wird seit einiger Zeit in Deutschland an mehreren Stellen verfolgt und es gibt einige vielversprechende Ansätze, die gute Anregungen geben. Zur Erstellung dieser Expertise wurde deshalb auf eine Reihe vorausgehender Arbeiten und Literatur zum Thema zurückgegriffen. Das Handbuch „Forschungsdatenmanagement“ aus dem Jahr 2011 erstellt von den Herausgebern Stephan Büttner, Hans-Christoph Hobohm und Lars Müller verschafft einen breiten Einblick in die verschiedenen Herausforderungen und zu den möglichen Lösungsansätzen auf diesem Gebiet.

Der 2012 erschienene Sammelband „Langzeitarchivierung von Forschungsdaten - Eine Bestandsaufnahme“[1], herausgegeben von Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump und Jens Ludwig, erlaubt einen Einblick in den Stand der Langzeitarchivierung von Forschungsdaten in den verschiedenen Disziplinen. Das Kapitel 8, bearbeitet von Ortwin Dally, Friederike Fless und Reinhard Förtsch, geht dabei detaillierter auf die Altertumswissenschaften ein.

Der „Leitfaden zum Forschungsdaten-Management“ ein öffentliches Deliverable der WissGrid-Initiative[2], koordiniert, erstellt und bearbeitet von Harry Enke, Norman Fiedler, Thomas Fischer, Timo Gnadt, Erik Ketzan, Jens Ludwig, Torsten Rathmann und Gabriel Stöckle, bietet einen Überblick zum Lebenszyklus von Forschungsdaten und definiert Aufgaben des Forschungsdatenmanagements. Neben den drei zentralen Werken kamen weitere Quellen, wie beispielsweise „Digitale Wissenschaft - Stand und Entwicklung digital vernetzter Forschung in Deutschland“, herausgegeben von Silke Schomburg, Claus Leggewie, Henning Lobin und Cornelius Puschmann von 2011, zum Einsatz.

[1] Das PDF des Buches kann hier bezogen werden: http://nestor.sub.uni-goettingen.de/bestandsaufnahme

[2] vgl. dazu http://www.wissgrid.de, (zuletzt aufgerufen am 29.6.2013)

Einleitung

Die Bedeutung digitaler Daten hat in den Altertumswissenschaften in den letzten Jahren erheblich zugenommen. Zudem haben sich ganz neue Formen der Forschung und Anwendung, wie die Verwendung von 3D und Geodaten bei Grabungen, herausgebildet. Verschiedene Arten digitaler Daten speisen sich aus sehr heterogenen Quellen, wie Ausgrabungen, Fundbearbeitungen, Fotogrammetrie, unterschiedlichen Datierungsverfahren, anthropologische Untersuchungen oder Forschungen zu Klima- und Landschaftswandel. Die verschiedenen Disziplinen sind auf sorgfältige Dokumentation und Beschreibung der Primärdaten sowie auf Auswertungen, die durch Berechnungen oder Simulation gewonnen werden, angewiesen. Forschungsarbeiten werden zunehmend durch eine Verknüpfung verschiedener Datenquellen und eine zunehmende Vernetzung der Daten charakterisiert. So entstehen zunehmend digitale 2D- und 3D-Rekonstruktionen. Digitale Objekt und Raumdaten werden mit geografischen Informationssystemen verknüpft und für statistische Analysen genutzt.

Neue Forschungsvorhaben starten typischerweise mit der Literaturrecherche und der Suche nach geeigneten Datensätzen für das angestrebte Vorhaben. Eine zentrale Anlaufstelle für themen- oder disziplinspezifische Datensätze sowie eine Unterstützung beim Datenzugang erleichtern die Forschung erheblich. Ebenso vereinfacht ein zentrales Datenarchiv den Nachweis und die Zitierbarkeit von Datensätzen, was es für Datenproduzenten attraktiv macht, wiederum ihre Daten hier einzustellen, um die Sichtbarkeit der eigenen Arbeit zu erhöhen. Gleichzeitig kann ein zentrales Archiv Forschungsvorhaben durch die Bereitstellung eines klaren Nachnutzungskonzepts und die Klärung einfacher rechtlicher Fragen, beispielsweise des späteren Zugangs, unterstützen.

Die Altertumswissenschaften verfügen bereits über eine Reihe von Repositorien und Forschungsdatenbanken, die als Grundlage bzw. Standard für zukünftige Lösungen dienen können, und die durch ein Kompetenzzentrum weiterentwickelt und vereinheitlicht werden. In diese Reihe zählen *eSciDoc*[3], *Propylaeum*[4] als virtuelle Fachbibliothek oder *ArcheoInf*[5] als Frontend für GIS-Anwendungen. Das geplante Kompetenzzentrum sollte eine Forschungsdatenbank betreiben, die sowohl die Fachpublikationen als auch laufende Forschungsprojekte und insbesondere Forschungsdaten nachweisen kann. Das verhindert das bei institutionsübergreifenden Projekten und internationalen Forschungsvorhaben die gewonnenen Daten und Publikationen über viele Partner verteilt werden, da so nicht jedes Teilprojekt auf jeweils lokale Lösungen der eigenen Institution zurückgreifen muss.

Eine besondere Herausforderung stellen gemeinsame Forschungsdaten unterschiedlicher Teildisziplinen dar, da sie oft einen sehr unterschiedlichen Umfang haben und von sehr unterschiedlichem Datentyp sind. Hinzu kommt beispielsweise für Grabungsdaten das spezielle Problem der Einmaligkeit vieler Daten, da diese nicht ein weiteres Mal in genau dieser Form erhoben werden können. Der Aufbau und die Organisation eines nachhaltigen Forschungsdatenmanagements sind der Erkenntnis geschuldet, dass die Aufbereitung und Dokumentation von Forschungsdaten über das eigentliche Projektziel hinaus in vielen Disziplinen als nachrangig angesehen wird. Dabei können veröffentlichte Daten sowohl Ideen für ganz neue Forschungsfragen und Ansätze liefern als auch die laufenden Projekte sichtbarer und nachvollziehbarer machen. Mit der zunehmenden Digitalisierung der verschiedenen Disziplinen gehen neue Chancen einher, Daten in neuer Weise (nach) zu nutzen, neu zu kombinieren und neuen Hypothesen zu unterziehen. Hiervon profitiert die interdisziplinäre Forschung, die auf Daten anderer Teildisziplinen angewiesen ist, um sie neu auf neue Weise zu analysieren und miteinander in Beziehung zu setzen. Die Zentralisierung der Daten bzw. ihrer Referenzen erlaubt ganz neue Forschungsansätze, wie beispielsweise der „Data Driven Science"-Ansatz. Über ganz verschiedene Datenbestände lassen sich damit neue Zusammenhänge und Entwicklungslinien finden, was mit einzeln vorgehaltenen Forschungsdaten so nicht möglich wäre.

[3] eSciDoc - The Open Source e-Research Environment, https://www.escidoc.org (zuletzt aufgerufen am 30.6.13)
[4] Propylaeum, die Virtuelle Fachbibliothek Altertumswissenschaften, http://www.propylaeum.de (zuletzt aufgerufen am 30.6.13)
[5] Informationszentrum für die Archäologie, http://www.archeoinf.de (zuletzt aufgerufen 30.6.13)

Für eine langfristige Nachnutzung und die Referenzierung von Forschungsdaten müssen eine einheitliche Qualität sichergestellt und vereinbarte Standards eingehalten werden. Diese hängen von den einzelnen Disziplinen und der Art der Daten ab, sollten aber bei Verwaltung durch ein zentrales Forschungsdatenzentrum einem disziplinübergreifenden Qualitätsstandard genügen. Hierzu werden zunächst unabhängig von verschiedenen Fachdisziplinen Arbeitsabläufe für Einlagerung (Ingest) und den späteren Zugriff (Access) benötigt, so dass nicht nur für die ursprünglichen Datenproduzenten und -eigner, sondern auch für interessierte und berechtigte Dritte eine sinnvolle (Nach-)Nutzung möglich ist. Sinnvollerweise werden die Fachdisziplinen bei Bedarf in der Entwicklung einer DataCite-Strategie unterstützt. Die Arbeitsabläufe sollten so gestaltet sein, dass auch zukünftige, zur Laufzeit des Projektes noch nicht bestimmbare, Fragestellungen entworfen und beantwortet werden können. Gerade diese Offenheit stellt eine große Herausforderung an die Planung und Implementierung der Arbeitsabläufe dar und erfordert eine aktive, frühzeitige und dauerhafte Begleitung der jeweiligen Projekte.

Die Altertumswissenschaften orientieren sich bereits stark an Open-Access-Zielen. Zwar sind nicht alle Daten unproblematisch frei publizierbar, da sie teilweise durch urheberrechtliche oder datenschutzrechtliche Regelungen behindert werden. Dennoch sollte Ziel eines Kompetenzzentrums der Altertumswissenschaften die (internationale) Nachweisbarkeit von Daten in Absprache mit ihren Eigentümern sein, die beispielsweise durch globale DOIs referenziert werden könnten. Dadurch wird DataCite ermöglicht und es ergeben sich neben der Sekundärverwertung durch Publikationen weitere Formen der Sichtbarkeit. Hierin kann ein Kompetenzzentrum unterstützend und moderierend wirken. Insbesondere bei großen Datenmengen muss über Zeiträume hinweg betrachtet werden, welche Daten langfristig aufgehoben werden sollen.

Der Auftrag eines zentralen Datenarchivs für die Altertumswissenschaften[6] besteht in der nutzerorientierten Bereitstellung und technologisch adäquaten Ablage der Datensätze und Metadaten. Optimalerweise unterstützt das Archiv die Datenproduzenten bei der Aufbereitung und Dokumentation der Daten, um eine höchstmögliche Qualität zu erreichen.

Forschungsdaten

Der Begriff der digitalen Forschungsdaten in den Altertumswissenschaften hängt von verschiedenen Aspekten ab. Wesentliche Faktoren sind, neben den jeweiligen wissenschaftlichen Teildisziplinen, beispielsweise die verwendeten Methoden und Werkzeuge sowie Formen, Formate und Aggregationsstufen (Verarbeitungs- und Analysestadien) der Daten. Gleichzeitig sind durch moderne Arbeitsinstrumente in der Archäologie, wie 3D- und GIS-Modelle sowie neue bildgebende Verfahren, Komplexität und Umfang der Forschungsdaten angewachsen, was den sinnvollen Umgang mit und eine nutzbringende Nachnutzung von Forschungsdaten deutlich anspruchsvoller macht.

2008 befand die Allianz der deutschen Wissenschaftsorganisationen in der Schwerpunktinitiative „Digitale Information“ „einen dringenden Handlungsbedarf hinsichtlich der systematischen Sicherung, Archivierung und Bereitstellung“ von Forschungsdaten.[7] Die von der Deutschen Forschungsgemeinschaft (DFG) 1998 veröffentlichten „Vorschläge zur Sicherung guter wissenschaftlicher Praxis“[8] sehen vor, dass „Primärdaten als Grundlagen für Veröffentlichungen [...] auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden [sollten].“

Die Deutsche Forschungsgemeinschaft (DFG) hat folgende Definition für digitale Forschungsdaten und ihre Nutzung vorgelegt:

[6] Siehe hierzu: http://www.ianus-fdz.de, (zuletzt aufgerufen 12.6.2013)

[7] Allianz der deutschen Wissenschaftsorganisationen: Schwerpunktinitiative Digitale Information, http://www.dfg.de/aktuelles_presse/das_neueste/download/pm_allianz_digitale_information_ details_080612.pdf, (zuletzt aufgerufen 12.6.2013)

[8] Sicherung guter wissenschaftlicher Praxis. Denkschrift, http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf, (zuletzt aufgerufen 12.6.2013)

„Forschungsdaten sind digitale, elektronisch speicherbare Daten, die während eines wissenschaftlichen Arbeitsprozesses, z.B. durch Quellenforschungen, Experimente, Messungen, Erhebungen oder Befragungen entstehen...“

Forschungsdaten bilden einen „*…wertvollen Fundus an Informationen, die mit hohem finanziellem Aufwand erhoben wurden. Je nach Fachgebiet und Methode sind sie replizierbar oder basieren auf nicht wiederholbaren Beobachtungen oder Messungen. In jedem Fall sollten die erhobenen Daten nach Abschluss der Forschungen öffentlich zugänglich und frei verfügbar sein. Dieses ist die wesentliche Voraussetzung, dass Daten im Rahmen neuer Fragestellungen wieder genutzt werden können sowie dafür, dass im Falle von Zweifeln an der Publikation die Daten für die Überprüfung der publizierten Ergebnisse herangezogen werden können.“*

Der uneingeschränkte Zugang zu Forschungsdaten fördert Transparenz und effektive Forschung. Open Access ist für die uneingeschränkte Veröffentlichung von wissenschaftlichen Daten anzustreben, was sich für die meisten Daten in den Altertumswissenschaften aufgrund der klaren rechtlichen Lage der überwiegenden Datensätze gut ermöglichen lässt. Die Daten sind i.d.R. weder personenbezogen noch unter Datenschutzaspekten kritisch. Darüber hinaus helfen die Verwendung von offenen Standards und die Einbindung in disziplinspezifische Forschungsinfrastrukturen, wie bestehende Kataloge, Repositorien oder Nachweissysteme.

Das Management von Forschungsdaten bezeichnet diejenigen Maßnahmen, die sicherstellen, dass Daten (nach)nutzbar bleiben. Die notwendigen Schritte, dieses zu erreichen, variieren unter Umständen erheblich und hängen von den Einsatzzwecken ab. An dieser Stelle können verschiedene Zwecke abgegrenzt werden:

- Die Aufbewahrung dient als Dokumentation des korrekten wissenschaftlichen Arbeitens. Zudem können die Daten eine übergeordnete gesellschaftliche Relevanz besitzen, wie beispielsweise die Dokumentation von Grabungen zur Sicherung des geschichtlichen Erbes.

- Die Aufbewahrung kann weiterhin erfolgen, um rechtlichen oder anderen forschungsfremden Anforderungen, wie beispielsweise Auflagen nachzukommen oder Selbstverpflichtungen zu genügen.

- Ebenso sollten Forschungsdaten als Arbeitskopie für laufende wissenschaftliche Arbeiten nutzbar sein.

- Darüber hinaus kann es erwünscht sein, dass die Nachnutzung von Forschungsdaten für spätere Forschungsvorhaben, beispielsweise für vergleichende Studien, Dokumentation von Forschungstätigkeit, Bearbeitung neuer Fragestellungen sowie museale Zwecke, erfolgt.

Das Forschungsdatenmanagement hat zum Ziel, dass ein Datenzugriff und eine -auswertung unabhängig vom Datenproduzent möglich bleiben. Hierzu zählen einerseits die technische Speicherung und Lesbarkeit der Daten und andererseits ausreichende Informationen zu ihrer Interpretation in Metadaten. Zur Überprüfbarkeit von Forschungsergebnissen und -prozessen muss die Nachweiskette lückenlos dokumentiert sein. Dieses kann sowohl über mehrere Stufen veränderte Datensätze als Basis weiterer Forschungen beinhalten als auch die Sicherstellung der Zuverlässigkeit der Daten.

Für digitale Forschungsdaten ist eine Reihe von Bedingungen zu erfüllen - unabhängig von der vorgesehenen Nutzung. Dieses beginnt mit der Langzeitarchivierung der Daten auf Bitstream-Ebene und setzt sich über die technische und inhaltliche Nachnutzbarkeit fort. Während die Bitstream-Preservation, d.h. die Erhaltung der Bitfolge eine Grundvoraussetzung ist, genügt es noch nicht, dass die Daten auch technisch genutzt werden können. Hierzu zählt die korrekte Interpretation der Daten durch Software, die wiederum Anforderungen an Hardware und technische Infrastrukturen stellen kann. Für eine inhaltliche Nutzung wird zusätzlich Hintergrund- und Kontextwissen gebraucht, ohne dass eine Einordnung bestimmter Daten nicht sinnvoll erfolgen kann. Entsprechende Zusatzinformationen müssen bei der Erstellung der Forschungsdaten nachvollziehbar mitdokumentiert werden.

Insgesamt soll Forschungsdatenmanagement dabei helfen, die Arbeitsbedingungen und den Austausch in den Altertumswissenschaften zu verbessern, indem es neue Recherchemöglichkeiten bietet und Forschungsergebnisse einfacher auffindbar macht. Ein Forschungsdatenzentrum kann dazu dienen, eine fachdisziplinübergreifende Infrastruktur zu schaffen, die es Wissenschaftlern ermöglicht, ihre Forschungsdaten standardisiert zu dokumentieren, zu versionieren und der (Fach-)Öffentlichkeit zugänglich zu machen. Mit der Einführung einer zentralen Anlaufstelle ergeben sich etliche Vorteile: Publikationen lassen sich deutlich besser mit Forschungsdaten verknüpfen und erhalten einen erheblichen Mehrwert. Das kann einerseits eine erhöhte Aufmerksamkeit gegenüber den Ergebnissen oder eine erhöhte Zitationsrate der Publikation und Daten selbst bedeuten. Die Wissenschaftler erhalten auf diesem Wege eine Öffentlichkeitswirksamkeit und einen Qualitätsnachweis für ihre Tätigkeit.

Life-Cycle-Modell von Forschungsdaten

Für den Umgang mit Forschungsdaten wurden verschiedene Life-Cycle-Modelle entwickelt. Ein umfängliches und weitgehend von der Community akzeptiertes Modell wird von der DCC[9] vorgeschlagen.[10] Für die Zwecke dieses Gutachtens genügt das davon abgeleitete vereinfachte Modell des WissGrids:[11]

- **Planung und Erstellung:** Um das spätere Management von Forschungsdaten möglichst zu vereinfachen, ist es sinnvoll, die Daten schon geeignet zu erzeugen. Ein wichtiger Aspekt in dieser Phase ist z.B. die Wahl der richtigen Standards. Im jeweiligen Projekt wird festgelegt, welche Daten erfasst und bearbeitet werden. Im Zuge der laufenden Erhebung, Bearbeitung und Auswertung können bereits die klassischen Verfahren zum Austausch bzw. zur kurzfristigen Aufbewahrung den Projektzielen nicht genügen. Zudem besitzen Primärdaten bereits einen eigenen Wert, der zunehmend erkannt wird, was sich in den Vorgaben der Förderinstitutionen widerspiegelt. Die Daten können sich inhaltlich und von der Art ihrer Entstehung bzw. Erfassung und der involvierten Softwarewerkzeuge erheblich unterscheiden.

- **Auswahl und Bewertung:** Nicht alle Forschungsdaten müssen und können auf Dauer aufbewahrt werden. Die Gründe, Methoden und Kriterien der Selektion und die daraus resultierende Dauer der Aufbewahrung von Forschungsdaten müssen geklärt werden. Die Auswahl aufzuhebender Daten muss transparent und nachvollziehbar sowie dabei möglichst unabhängig von den Sichtweisen einer Person oder Forschergruppe sein. Optimalerweise verfügt ein Projekt selbst über ein Regelwerk zur Datenauswahl, das neben den Selektionsregeln auch die Verantwortlichkeiten für die einzelnen Datenbewertungen bestimmt. Bei der Selektion spielen Archivwürdigkeit (Relevanz der Daten) und Archivfähigkeit (Erfüllung bestimmter technischer Voraussetzungen) eine Rolle. Für die Aufbewahrung sind geeignete Kriterien für die spätere Nachnutzung zu treffen.

- **Ingest/Übernahme:** Forschungsdaten, die längerfristig aufbewahrt werden sollen, müssen in eine geeignete Umgebung wie z.B. ein Datenarchiv überführt werden. In dieser Phase werden üblicherweise zusätzliche Checks, Homogenisierungen und Anreicherungen der Daten notwendig. Die Arbeitsschritte 1. Transport der Daten zum Ingest, 2. Vorbereitung je nach Art der Daten, inklusive Vergabe eines eindeutigen Identifikators, 3. Überprüfung auf formale Korrektheit, Vollständigkeit und Richtigkeit, 4. Freiheit von Schadsoftware o.ä., 5. Validierung technischer Vollständigkeit bzw. Konsistenz, 6. Erhebung technischer Metadaten und 7. Zusammenfassung der Daten in Containerdateien können sich nach Zweck und Aufbau des Archivs unterscheiden. Am Ende steht das Einfüllen ins Archiv und die Erzeugung von Prüfsummen. Wissenschaftliche Untersuchungen, Experimente und numerische Rechnungen lassen sich nur reproduzieren, wenn alle wichtigen Schritte nachvollziehbar und alle notwendigen Softwarekomponenten vorhanden sind.

[9] Digital Curation Center in Grobritannien, siehe hierzu: http://www.dcc.ac.uk, (zuletzt aufgerufen 22.6.2013)
[10] Das DCC Curation Lifecycle Model, http://www.dcc.ac.uk/resources/curation-lifecycle-model, (zuletzt aufgerufen 22.6.2013)
[11] Checkliste zum Forschungsdatenmanagement: http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf, (zuletzt aufgerufen 22.6.2013)

- **Speicherung:** Ziel ist die langfristige Speicherung von Forschungsdaten mit Verfahren, die die Chancen von Datenverlust minimieren. Für die Speicherung sind Faktoren wie die Zahl und Größe der Datensätze und die Häufigkeit des Zugriffs auf die Datensätze wesentlich. Die Integrität der Daten wird auf der Bitstream-Ebene sichergestellt, hierzu wird üblicherweise mit Kopien gearbeitet, die in ausreichender Zahl unabhängig voneinander vorgehalten werden. Die Netzwerk-Infrastruktur und die Übertragungsprotokolle müssen zum Nutzungsszenario passen und ausreichende Kapazitäten vorhalten.

- **Erhaltungsmaßnahmen:** Es ist nicht selbstverständlich, dass digitale Forschungsdaten in anderen Umgebungen als der ursprünglichen Erstellungs- und Nutzungsumgebung nutzbar bleiben. Deshalb ist es bereits im Vorfeld sinnvoll zu bedenken und zu dokumentieren, welche Anforderungen an eine technische Umgebung zur Nutzung der Daten gestellt werden und wie mit Veränderungen der Technik umgegangen werden sollen. Erhaltungsmaßnahmen müssen dann greifen, wenn sich die Anforderungen ändern, was beispielsweise durch neue Daten- oder Dateiformate, neue Schnittstellen für neue Softwareprogramme oder Arbeitsumgebungen, neue wissenschaftliche Standards oder Arbeitsweisen oder zusätzliche Parameter, ausgelöst werden kann. Der Umgang mit solchen Änderungen sollte durch einen gründlichen Planungsprozess begleitet werden, der Anforderungen definiert, Alternativen evaluiert und Ergebnisse analysiert, die in einen Erhaltungsplan münden.

- **Zugriff und Nutzung:** Die besten Daten nutzen wenig, wenn sie nicht gefunden werden. Wie die Daten gefunden werden können, wer autorisiert auf sie zugreifen darf und mit welchen Mitteln, sind daher ebenfalls wichtige Fragen. Die gleichzeitige Suche in mehreren Archiven kann funktionieren, wenn die zur Suche erforderlichen Metadaten eine einheitliche Struktur besitzen. Ein offener Zugang zu Forschungsdaten stellt sicher, dass auch andere Forscher auf Daten zugreifen und diese nicht neu erzeugen müssen. Vielfach wird die Interoperabilität der Daten (auf verschiedenen Ebenen) erwartet, um einen einfachen Austausch zu erlauben.

Metadaten

Jedes wissenschaftliche Projekt sollte die Erstellung und Verarbeitung von Forschungsdaten umfassend so dokumentieren, dass einerseits die Entstehung der Daten selbst als auch die davon abgeleiteten Interpretationen jederzeit inhaltlich nachvollzogen werden können. Hierzu dienen typischerweise Metadaten. Sie werden für unterschiedliche Zwecke benötigt, die von der Wiederauffindbarkeit eines Datums, über die Provenienz eines Datensatzes bis hin zu technischen Daten reichen können, die die Umgebung der Entstehung von Forschungsdaten beschreiben. Bei Metadaten handelt es sich um Daten oder Informationen, die in strukturierter Form analoge oder digitale Forschungsdaten und Artefakte im weitesten Sinne beschreiben. Sie erklären, ordnen ein oder definieren Ressourcen und Informationsquellen. Sie helfen der Wissenschaft, Forschungsdaten zu erschließen, zu verstehen und damit umzugehen. Allein schon wegen der Heterogenität der Fach- und Teildisziplinen gibt es nicht „das“ Set von Metadaten oder von allen Disziplinen akzeptierte, allgemeine und übergeordnete Standards. Informationen, wie Objekte, Akteure, Quellen, Vorgänge und Ergebnisse, haben jedoch allgemeine Bedeutung. Zu den allgemeinen, deskriptiven Metadaten können technische Metadaten hinzukommen, die bereits durch Geräte bereitgestellt werden oder aus verwendeter Software resultieren. Sie sind insbesondere für einen funktionalen Archivierungsansatz relevant.

Altertumswissenschaftliche Metadaten für Forschungsdaten müssen bestimmten Standards unterliegen, insofern sie entweder in der EUROPAEANA[12] oder einem der Zuführungsprojekten wie CARARE[13] verwendet werden sollen. Eine der größeren Plattformen ist CLAROS[14], die auf dem Metdadatenschema

[12] Siehe hierzu: http://www.europeana.eu/portal/aboutus.html, (zuletzt aufgerufen 22.6.2013)
[13] Siehe hierzu: http://www.carare.eu/eng/About, (zuletzt aufgerufen 22.6.2013)
[14] Siehe hierzu: http://explore.clarosnet.org/XDB/ASP/clarosHome, (zuletzt aufgerufen 22.6.2013)

CIDOC-CRM Core[15] beruht. Zur Verlinkung und Kodierung der Forschungsdaten kommen das Ressource-Description-Format zum Einsatz oder davon abgeleitete Standards.

Zitier- und Referenzierbarkeit

Eine wichtige Voraussetzung für die Auffindbarkeit, Zitier- und Referenzierbarkeit von Forschungsdaten sind neben der Langzeitspeicherung ein geeignetes Nachweissystem. Diese Rolle übernehmen persistente Identifikatoren (PID), die sich bereits im Bereich des wissenschaftlichen Publizierens für die langfristige Kennzeichnung digitaler Objekte bewährt haben. Identifikatoren beziehen sich auf Informationsobjekte und sind damit für das Datenmanagement relevant. Für Identifikatoren bestehen keine prinzipiellen Einschränkungen, so dass sie auf einzelne Dateien, Dateicontainer, analoge und digitale Artefakte, Teile von digitalen Objekten, Funktionalitäten von Webservices etc. bezogen werden können. Hierfür muss festgelegt sein, worum es sich beim Informationsobjekt handelt, welches mit dem Identifikator bezeichnet werden soll und wie dieser Identifikator zum Objekt aufgelöst wird. Dieses hängt vom Anwendungsfall und den späteren Nutzungsszenarien ab. Hierbei spielen z.B. folgende Fragen eine Rolle: Was für Informationsobjekte existieren? Welche Informationsobjekte können sich auf andere beziehen? Welche Informationsobjekte sind wichtig? Wie werden die Identifikatoren aufgelöst? Wann werden Informationsobjekte identifiziert? Persistente Identifikatoren sind nicht automatisch persistent, sondern hängen davon ab, dass alle Veränderungen, wie beispielsweise die Lage des Objekts, im Resolver permanent nachvollzogen werden. Die Existenz eines Identifikators stellt für sich noch keine Garantie dar, dass ein Datensatz tatsächlich aufgefunden werden kann und zugreifbar ist. Insbesondere für letzteres sind die Maßnahmen der (funktionalen) Langzeitarchivierung zuständig.

Funktionale Langzeitarchivierung

Digitale Objekte bedürfen eines geeigneten Kontextes, damit auf sie zugegriffen werden kann. So lässt sich beispielsweise ein 3D-Modell oder eine Videodatei nicht ohne entsprechende Software, die üblicherweise ein bestimmtes Betriebssystem voraussetzt, abspielen. Das Betriebssystem muss über Schnittstellen zur Bildausgabe in passender Auflösung und Farbtiefe sowie im Falle des Videos über die Fähigkeit zur Audioausgabe verfügen. Das wiederum setzt geeignete Software-Hardware-Umgebungen voraus. Hierbei ergeben sich einige Abhängigkeiten, die für einen langfristigen Zugriff erhalten bleiben müssen. Die notwendigen Software- und Hardwarekomponenten zur Sichtbarmachung oder Ausführung von verschiedenen, im Langzeitarchiv enthaltenen Objekttypen sind im Moment der Archivaufnahme der digitalen Artefakte festzulegen, was durch sogenannte View-Pathes formalisiert werden kann. An dieser Stelle werden diverse Arbeitsabläufe notwendig, die es erst erlauben, dass ein späterer Archivnutzer tatsächlich mit dem gewünschten Objekt umgehen kann. Mit zunehmendem Zeitabstand zur Erstellung des Objekts sinkt die Wahrscheinlichkeit, dass dieses direkt in der digitalen Arbeitsumgebung des Nutzers ausgeführt oder betrachtet werden kann. Daraus ergibt sich eine zentrale Aufgabe von Betreibern digitaler Langzeitarchive: Ihre Zuständigkeit endet nicht mit der erfolgreichen Bewahrung und bitgetreuen Fortschreibung der Archivobjekte, sondern sie müssen zusätzlich die Nutzbarkeit zu jedem Zeitpunkt sicherstellen.

Die sogenannte funktionale Langzeitarchivierung (FLA) beschreibt hierzu eine alternative Erhaltungsstrategie, die mittels Emulation eine vollständige Softwareumgebung funktional repliziert und nutzbar macht. Im Mittelpunkt steht dabei die Emulation von Computer-Systemen. Ein Emulator ist somit ein Software-Programm, welches einerseits auf aktuellen Computer-Systemen einsetzbar ist und andererseits alle Eigenschaften eines historischen Systems nachbildet. Im Gegensatz zu Formatmigrationsstrategien kann direkt mit unmodifizierten Originalobjekten gearbeitet werden.

[15] Siehe hierzu: http://www.cidoc-crm.org/official_release_cidoc.html, (zuletzt aufgerufen 22.6.2013)

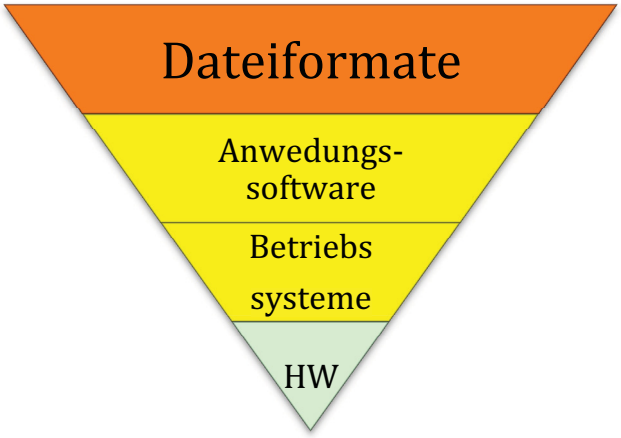

Abb. 1.: Hierarchie digitaler Objektklassen

Typischerweise sind heute eine fast nicht abzählbare Menge von digitalen Dateiformaten zu beobachten. Ebenso verhält es sich mit den entsprechenden Anwendungsprogrammen (Software). Erst auf der Betriebssystem-Ebene wird die Anzahl der vergangenen als auch aktuellen Versionen und Varianten überschaubar. Dies gilt noch stärker für die Anzahl von System-Plattformen (Hardware-Plattform) (vgl. Abb. 1). Der Einsatz von Emulationsstrategien ist gerade bei einer sehr großen Vielfalt an Formaten, die nur in kleineren Mengen vorzufinden sind, attraktiv. Für diese bietet der FLA-Ansatz eine effiziente Möglichkeit der Objektnachnutzung.

Die funktionale Langzeitarchivierung setzt daher bei der Erhaltungsplanung nicht primär am Objekt selbst an, sondern versucht die Laufzeitumgebung, d.h. den View-Path von Hardware über Betriebssystem bis hin zum Anwendungsprogramm technisch zu beschreiben und zu erhalten.[16] Im Rahmen der Erhaltungsstrategie ist es somit unerlässlich, neben dem Objekt selbst, genutzte Softwarekomponenten und Prozesswissen mitzuarchivieren und geeignet Metadaten zu hinterlegen.

Neben dem recht naheliegenden Einsatzzweck von FLA bietet ein funktionaler Ansatz aber auch die Möglichkeit, ganze wissenschaftliche Prozesse zu archivieren, d.h. beispielsweise einen konkreten Versuchsaufbau bestehend aus einem oder mehreren Computersystem(en) mit installierten und konfigurierten Softwarekomponenten so zu erhalten, dass alle Prozessschritte von dem Primärdatum bis zur publizierfähigen Auswertung funktional und langfristig nachvollziehbar bleiben.

Kostenmodelle und Finanzierung

Die langfristige Speicherung und Bereitstellung von Forschungsdaten verursacht Kosten. Die Preisentwicklung für digitale Speichersysteme verleitet hingegen dazu, davon auszugehen, dass die eigentliche Aufbewahrung von Daten eine mittelfristig vernachlässigbare Größe wird und sich der Aufwand in Grenzen hält. Die Unterschätzung dieser Kosten ist eine wesentliche Gefahr für den langfristigen Erhalt von Forschungsdaten. Das Speichern und die Pflege von Forschungsdatenbeständen ist Teil des (späteren) Nutzens. Entsprechend sollten die entstehenden Kosten von vorneherein im Projekt mitkalkuliert und nicht als optional zu leistende Zusatzkosten betrachtet werden. In der Phase der initialen Beschaffung von Speichersystemen fallen zuerst die Hardwarekosten ins Auge. Jedoch übersteigen die Aufwendungen für qualifiziertes Personal diese Kosten sehr schnell. Für die Hardwareseite kann man von sinkenden jährlichen Kosten pro Einheit aufgrund der hohen Initialkosten und der zunehmenden Effizienz der Technologien ausgehen. Auf dieser Erkenntnis beruhen die Geschäftsmodelle des „Pay once, store forever“. Hierin sind jedoch keine besonderen Dienstleistungen für den späteren Zugriff einkalkuliert.

[16] vgl. Dirk von Suchodoletz, Funktionale Langzeitarchivierung digitaler Objekte: Erfolgsbedingungen des Einsatzes von Emulationsstrategien. Nestor Edition, Cuviller, 2009.

Die allgemeinen Ausgaben für ein Forschungsdatenzentrum verteilen sich hierbei auf direkte Kosten der Speicherung und Kosten, die durch Betrieb, Beratung, Betreuung und Planung entstehen. Ein Forschungsdatenzentrum muss hierfür über ausreichend langfristig sichergestellte Ressourcen verfügen, die je nach Ausgestaltung des Betriebs sich in erster Linie in der Form von Stellen ausdrücken. Diese wird benötigt, um ein nachhaltiges Forschungsdatenmanagement zu erreichen und die Standardisierung innerhalb der Disziplin voranzutreiben. Ein Kompetenzzentrum benötigt hierfür eine Kombination aus Fach- sowie IT-Expertise, um die angeschlossenen Institutionen angemessen unterstützen und beraten zu können. Um eine langfristige Finanzierung sicherzustellen, sollten entsprechende Gelder für Datenmanagement und -haltung eingeplant werden.

Auf Basis des DCC-Curation-Lifecycle-Model haben Projekte, wie LIFE,[17] 4C[18] und Keeping Research Data Safe,[19] in einem mehrstufigen Prozess mit Modellen und Fallstudien, Kosten kalkuliert. Diese wurden für die verschiedenen Phasen bestimmt und nach Kategorien aufgegliedert:

- In der Vorarchiv-Phase sind neben den Kosten für die Generierung oder Erhebung der Daten, Kosten für Beratung, Schulung und die Planung des Datenmanagements selbst zu berücksichtigen.
- Die Archiv-Phase wird im Wesentlichen durch Kosten für alle einzelnen Lifecycle-Phasen, die von Auswahl und Bewertung bis zum späteren Zugriff und Nutzung reichen, bestimmt. Gleichzeitig können Innovationskosten für neue Werkzeugen, die Entwicklung von Standards etc. anfallen.
- Unterstützungsdienste: Aufwendungen für die Verwaltung und Steuerung der allgemeinen IT-Basisinfrastruktur und aller relevanten Aktivitäten.
- Klassische Infrastruktur: Kosten, die aus der Einrichtung und Unterhaltung der benötigten Gebäude und Räume entstehen.

Die zugrunde zu legenden Kostenmodelle unterscheiden sich nach der jeweiligen Architekturschicht. Für jede Schicht gelten spezifische Bedingungen. Im Bereich der Bitpreservation stehen verschiedene Modelle zur Auswahl, die vom eigenen Rechenzentrumsbetrieb über Modelle, wie sie vom FIZ[20] im Zuge des DFG-geförderten RADAR-Projekts[21] entwickelt werden, bis hin zu komplett externen Dienstleistern wie Amazon reichen

Überblick int. und nat. Initiativen im Forschungsdatenmanagement

Das Thema Forschungsdatenmanagement steht in etlichen Ländern bereits länger auf der Agenda. Vorreiter in diesem Gebiet sind die angelsächsischen Länder aufgrund der nationalen Vorgaben von Forschungsförderern sowie entsprechende Vorgaben durch Wissenschaftsverlage zur Archivierung der Forschungsdaten für eine mögliche Replikation bzw. Verifikation der Ergebnisse.

Auch in Deutschland gibt es in einzelnen Disziplinen bereits seit längerem etablierte Verfahren zum Management von Forschungsdaten. Zudem hat das Thema an vielen Forschungseinrichtungen an Relevanz gewonnen.

Internationale Aktivitäten

Australien stellt bereits seit einiger Zeit das Management und die Bereitstellung von Forschungsdaten sehr weit oben auf die nationale Agenda. Hierzu wurde ein zentrales Gemeinschafsprojekt zur Verbesserung und Förderung der Forschungsinfrastrukturen (National Collaborative Research Infrastructure Strategy, NCRIS)

[17] Siehe hierzu: http://www.dcc.ac.uk/projects/life, (zuletzt aufgerufen 13.7.2013)
[18] Siehe hierzu: http://www.dcc.ac.uk/projects/4c, (zuletzt aufgerufen 13.7.2013)
[19] Siehe hierzu: http://www.beagrie.com/krds.php, (zuletzt aufgerufen 23.6.2013)
[20] Siehe hierzu: http://www.fiz-karlsruhe.de, (zuletzt aufgerufen 13.7.2013)
[21] Siehe hierzu: http://www.tib-hannover.de/de/die-tib/aktuelles/aktuelles/id/409

ins Leben gerufen. Die Initiative startete 2003 durch Bedarfsevaluationen in verschiedenen Bereichen, aus denen viele Einzelprojekte hervorgingen, für die umgerechnet 420 Mio. Euro im Zeitraum von 2005-2011 vergeben wurden.[22] Der Australian National Data Service (ANDS)[23] schafft eine zentrale Anlaufstelle für den Zugang zu nationalen Forschungsdaten. Der Service wurde als Gemeinschaftsprojekt der Australia National Service Agency mit der Australia National University und Monash University zur Nachnutzung von Forschungsdaten eingerichtet. Ein zentrales Element ist hierzu der Metadatenkatalog zum Nachweis der Projekte, Daten und Ergebnisse. Das ANDS lagert die Daten nicht direkt, sondern aggregiert die Links zu den jeweiligen institutionellen Repositories.

In den USA spielen sowohl universitäre als auch disziplinspezifische Initiativen zum Forschungsdatenmanagement eine Rolle, um den Leitlinien der verschiedenen Forschungsförderern Rechnung zu tragen. Ein Großprojekt mit weltweiter Beteiligung und Aufmerksamkeit und ein frühes Beispiel für ein koordiniertes Forschungsdatenmanagement war das Human Genome Project[24], welches vom National Institutes of Health (NIH) und U.S. Department of Energy koordiniert und unter der weltweiten Beteiligung von mehreren Forschungseinrichtungen im Jahr 2003 abgeschlossen wurde. Ein wesentliches Element des 13-jährigen Projektes bildete die enge Zusammenarbeit verschiedener Wissenschaftsbereiche.

Weit fortgeschritten sind die Lösungen der Harvard University[25] und das IDEALS Repository der University of Illinois at Urbana-Champaign. Das Harvard Dataverse Network ist sowohl Repository, das Forschungsprojekte, Institute oder Aktivitäten einzelner Forscherinnen und Forscher unterstützt und gleichzeitig ein Rechtemanagement bereitstellt, das die gemeinsame Bearbeitung der Daten erlaubt. Gleichzeitig dienen Persistent Identifiers für den langfristigen Nachweis und die Zitierbarkeit der Forschungsdaten. IDEALS (Illinois Digital Environment for Access to Learning and Scholarship) ist ein institutionelles Repository, das in Kooperation zwischen der zentralen universitären IT und der Bibliothek der Universität Illinois gemeinsam betrieben wird.[26] Das Repository soll einerseits Forschungsergebnisse digital zur Verfügung zu stellen, andererseits berät es Forscher zur Veröffentlichung, zu rechtlichen Aspekten und zum Datenmanagement. IDEALS erlaubt eine formatunabhängige Aufnahme von Forschungsdaten, erwartet jedoch eine Ausweisung der Metadaten durch die Einreichenden.

In Großbritannien wird das Forschungsdatenmanagement durch die jeweiligen Policies der Forschungsförderorganisationen stark beeinflusst. Beispielsweise verlangt das Economic and Social Research Council (ESRC) seit 2011 von antragstellenden Forschern als Bedingung für die Bewilligung einen Datamanagementplan und die spätere Verfügbarkeit der Forschungsdaten. Das JISC[27] hat das Forschungsdatenmanagement wesentlich vorangebracht und finanziert, siehe hierzu beispielsweise das Digital Curation Center (DCC). Dieses gibt regelmäßig aktualisierte Handlungsanweisungen für den Umgang mit Forschungsdaten heraus.[28] Das DCC ist eine nationale Plattform, die Beratung und Unterstützung zum Forschungsdatenmanagement sowohl für Archive, Wissenschaftseinrichtungen und Hochschulen als auch für einzelne Forschergruppen anbietet. Es existieren daher inzwischen einige ernstzunehmende disziplinäre und institutionelle Ansätze auf der nationalen Ebene. Mit dem UK-Data-Archiv steht für den Bereich der Sozial- und Geisteswissenschaften eine zentrale Anlaufstelle zur Verfügung. Es arbeitet mit zentralen Einrichtungen, wie der British Library, dem National Archive und insbesondere mit dem DCC zusammen. Zusätzlich zu den JISC-Anstrengungen haben auch einzelne Universitäten, wie die University of Oxford, die University of

[22] Siehe hierzu: http://ncris.innovation.gov.au/Pages/default.aspx, (zuletzt aufgerufen 23.6.2013)

[23] Siehe hierzu: http://www.ands.org.au, (zuletzt aufgerufen 23.6.2013)

[24] HGP, http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml, (zuletzt aufgerufen 23.6.2013)

[25] Harvard DataverseNetwork: http://dvn.iq.harvard.edu/dvn, (zuletzt aufgerufen 23.6.2013)

[26] Siehe hierzu: https://services.ideals.illinois.edu/wiki/bin/view/IDEALS, (zuletzt aufgerufen 23.6.2013)

[27] Forschungsdatenmanagementstrategie des JISC (Joint Information Systems Committee): http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx, (zuletzt aufgerufen 23.6.2013)

[28] Howto-Guides des DCC zum Forschungsdatenmanagement: http://www.dcc.ac.uk/resources/how-guides, (zuletzt aufgerufen 23.6.2013)

Edinburgh und die University of Cambridge Policies zum Forschungsdatenmanagement erlassen und bieten darüber hinaus Infrastrukturen an.

In den angelsächsischen Ländern finden sich verschiedene nationale Koordinierungsorganisationen für das Management von Forschungsdaten in den Altertumswissenschaften. In den US betreibt die Organisation „Digital Antiquity“[29] ein zentrales Repository[30] „The Digital Archaeological Record“ (tDAR) und übernimmt damit die Langzeiterhaltung von Forschungsdaten. In Australien übernimmt der ANDS die disziplinübergreifende Koordinierung. Zudem entsteht ein disziplinspezifisches Datenarchiv „The Australian Historical Archaeological Database“ (AHAD), das als Bestandteil von tDAR eingerichtet wird. Der Archaeology Data Service (ADS) hat die Aufgabe des Managements von archäologischen Forschungsdaten in Großbritannien und stellt hierfür allen Archäologen als zentrales Repository offen. Zudem gibt es Empfehlungen und Standards zur guten wissenschaftlichen Praxis und bestimmte Verfahren in den Altertumswissenschaften auf der internationalen Ebene, die für Mitglieder jeweils verpflichtend sind. Dieses trägt zunehmend zu einer einheitlichen Praxis und Interoperabilität der Forschungsdaten bei.

Ebenfalls beachtenswert sind Initiativen und Portale zur Archivierung, Pflege und Nutzung von ganzen wissenschaftlichen Prozessen. Wissenschaftliche Prozesse werden zunehmend als archivierungswürdige Artefakte betrachtet[31] [32] - entweder als Teil einer wissenschaftlichen Publikation[33] oder als eigenständige Objekte. Beispielsweise bieten myExperiments[34] und CrowdLabs[35] eine entsprechende technische Plattform für die Publikation und Suche von wissenschaftlichen Workflows.

Aktivitäten in Deutschland

Das Forschungsdatenmanagement in Deutschland gelangt zunehmend in den Fokus sowohl von Forschungsförderern als auch verschiedenen Wissenschaftseinrichtungen. Die Lage unterscheidet sich in den einzelnen Wissenschaftsdisziplinen. Während beispielsweise Klimaforschung[36] und physikalische Forschungsdaten traditionell zentral erreichbar gespeichert werden stehen andere Disziplinen noch an den Anfängen. Die Entwicklung zielführender Strategien wird durch verschiedene Faktoren vorangetrieben. So gibt es die nationalen Initiativen neben den Projekten der Hochschulen und außeruniversitären Forschungseinrichtungen. Bei vielen Projekten stehen fachspezifische Herangehensweisen und Aktivitäten im Vordergrund, die jedoch durch disziplinübergreifende Projekte einiger Hochschulen und Wissenschaftseinrichtungen ergänzt werden.

Für den Bereich der Forschungsdaten hat die Schwerpunktinitiative „Digitale Information“ der Allianz der deutschen Wissenschaftsorganisationen 2010 im Rahmen der Arbeitsgruppe Forschungsdaten Grundsätze und Regeln zur Behandlung von Forschungsdaten aufgestellt. Diese wurden von allen wichtigen Forschungsinstitutionen in Deutschland, wie der Fraunhofer-Gesellschaft, der Helmholtz-Gemeinschaft, der Hochschulrektorenkonferenz (HRK), der Leibniz-Gemeinschaft, der Max-Planck-Gesellschaft und dem Wissenschaftsrat unterzeichnet. Die Grundsätze beinhalten Sicherung und Zugänglichkeit, wissenschaftliche Anerkennung, Lehre und Qualifizierung, Vereinbarung von Standards sowie den Aufbau von Infrastrukturen.

---

[29] Siehe hierzu: http://www.digitalantiquity.org, (zuletzt aufgerufen 23.6.2013)

[30] Siehe hierzu: http://www.tdar.org, (zuletzt aufgerufen 23.6.2013)

[31] Siehe hierzu: Workflow4Ever Project (EU-FP7 STREP), http://www.wf4ever-project.org/, (zuletzt aufgerufen 23.6.2013)

[32] Belhajjame K, Corcho O, Garijo D, Zhao J, Missier P, Newman DR, Palma R, Bechhofer S, Garcia-Cuesta E, Gómez-Pérez JM, Klyne G, Page K, Roos M, Ruiz JE, Soiland-Reyes S, Verdes-Montenegro L, De Roure D, Goble CA: Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse. In proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web, 2012.

[33] Sven Vlaemick, Data Management in Scholarly Journals and Possible Roles for Libraries — Some Insights from EDaWaX, Liber Quarterly, Vol. 23(1), 2013.

[34] Siehe hierzu: http://myexperiments.org, (zuletzt aufgerufen 7.7.2013)

[35] Siehe hierzu: http://www.crowdlabs.org, (zuletzt aufgerufen 7.7.2013)

[36] Beispielsweise Earth System Science Data (ESSD), http://www.earth-system-science-data.net, ist eine dem Helmholtz Open Access Projekt nahestehende Open-Access-Zeitschrift, die geowissenschaftliche Forschungsdaten publiziert.

Die DFG ist ein treibender Akteur des Forschungsdatenmanagements. Sie unterstützt verschiedene Aspekte des Umgangs mit Forschungsdaten. So verbessert beispielsweise das DFG geförderten Projekt PubFlow[37] den Umgang mit Forschungsdaten während des wissenschaftlichen Arbeitsprozesses am Beispiel geo- und meereswissenschaftlicher Arbeiten, das in weiteren Schritten auf andere Disziplinen ausgeweitet werden soll. Unterstützende Dienste, wie beispielsweise PANGEA[38], werden in den Workflow der Wissenschaftler integriert. PANGEA bietet eine zentrale Anlaufstelle für digitale Forschungsdaten aus dem Bereich der Erdsystemforschung und sorgt für ihre langfristige Verfügbarkeit durch die Vergabe von DOIs. Diese Data Object Identifier[39] erlauben den Nachweis von Forschungsdatensätzen und werden beispielsweise vom Registrierungsservice der Technischen Informationsbibliothek Hannover (TIB) angeboten. Die TIB stellt gemeinsam mit anderen DataCite-Partnern mehrere für das Forschungsdatenmanagement relevante Aspekte wie den Zugang zu und die Nachnutzung von digitalen Forschungsdaten mit einer langfristigen Perspektive bereit.

Beispielsweise in den Wirtschafts- und Sozialwissenschaften wurden hierfür Forschungsdatenzentren etabliert, die eine enge Verbindung von Datenproduzenten und -nutzern anstreben. Mit dem Leibniz-Institut für Sozialwissenschaften[40] (GESIS) wurde eine Infrastruktureinrichtung ins Leben gerufen, die grundlegende, überregional und international bedeutsame forschungsbasierte Dienstleistungen erbringt. Es erarbeitet Expertisen für den gesamten Forschungsdatenzyklus innerhalb von Projekten. Im DFG geförderten Projekt da|ra[41] bietet GESIS in Kooperation mit der ZBW[42] und dem DataCite-Konsortium einen DOI-Registrierungsservice für den sozialwissenschaftlichen Bereich an. Diese Zentren erlauben es, Daten sehr heterogener Quellen, die von primären Forschungsinstituten bis hin zu Kommunalen-, Landes- und Bundeseinrichtungen reichen, systematisch und unter Berücksichtigung aller rechtlichen Bestimmungen und fachlicher Standards, bereitzustellen.

Neben diesen grundsätzlichen Betrachtungen haben das DFG-geförderte Projekt Radieschen[43] und die Überlegungen zum Curation Continuum[44] verschiedene Domänen der Datenverwendung identifiziert, die von der privaten Domäne, in die Gruppendomäne, anschließend in die dauerhafte Domäne und zuletzt in die Domäne des Zugangs und der Nachnutzung reichen. Nicht alle Fragen dieser Lebenszyklen sind für ein Forschungsdatenzentrum von gleicher Relevanz.

Mit der Max Planck Digital Library[45] (MPDL) hat die MPG 2007 begonnen, Forschungsdatenmanagement im weiteren Sinne für ihre einzelnen Institute zu organisieren. Die Aufgaben reichen vom bestmöglichen Zugang zu digitalen Ressourcen bis hin zur wissenschaftlichen Informationsversorgung, eScience-Services, Dissemination und die Unterstützung und die Umsetzung von Open Access für die MPG. Die MPDL nutzt als technologische Basis Fedora und eSciDoc.

---

[37] Siehe hierzu: http://www.pubflow.uni-kiel.de, (zuletzt aufgerufen 25.6.2013)

[38] Data Publisher for Earth & Environmental Science, http://www.pangaea.de, (zuletzt aufgerufen 25.6.2013)

[39] Persistenter Identifikator für Forschungsdaten, der speicherort-unabhängig ist und das Wiederfinden und Zitieren von Datensätzen erlaubt, siehe auch: http://www.tib-hannover.de/de/dienstleistungen/doi-service, (zuletzt aufgerufen 25.6.2013)

[40] Siehe hierzu: http://www.gesis.org/home, Mitglied in der Leibnitz-Gemeinschaft, (zuletzt aufgerufen 25.6.2013)

[41] Siehe hierzu: http://www.da-ra.de, (zuletzt aufgerufen 25.6.2013) bzw. den Aufsatz von Brigitte Hausstein und Wolfgang Zenk-Möltgen „Ein Service der GESIS für die Zitation sozialwissenschaftlicher Daten“ in „Digitale Wissenschaft Stand und Entwicklung digital vernetzter Forschung in Deutschland“.

[42] Leibnitz-Informationszentrum Wirtschaft, http://www.zbw.eu(zuletzt aufgerufen 8.7.2013)

[43] Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur: http://www.forschungsdaten.org/uber-radieschen, (zuletzt aufgerufen 22.6.2013)

[44] Siehe hierzu den Artikel von A. Treloar , D. Groenewegen und C. Harboe-Ree: http://www.dlib.org/dlib/september07/treloar/09treloar.html, (zuletzt aufgerufen 22.6.2013)

[45] Siehe hierzu: http://www.mpdl.mpg.de, (zuletzt aufgerufen 22.6.2013)

Die Deutsche Nationalbibliothek (DNB) entwickelt einen Langzeitarchivierungsdienst[46], der gesammelte Erfahrungen der DNB mit denen der Nutzer zusammenbringen soll. Der Service-Katalog des Basisdienstes umfasst dabei die Bereitstellung von Massenan- und -auslieferungsschnittstellen, eine Integritätsprüfung aller digitalen Objekte, die Generierung technischer Metadaten, eine (technische) Qualitätsprüfung aller digitaler Objekte sowie eine ortsunabhängige Mehrfachspeicherung. Hinzu kommt die Bereitstellung von Zugriffs- und Suchfunktionen, eine Bereitstellung von Kontrollmöglichkeiten und die Möglichkeit von Preservation Planning inklusive Risikomanagement. Als Anlieferungsschnittstellen werden Hotfolder und das OAI-PMH-Protokoll genutzt, für die Auslieferung sind SOAP- und REST-Schnittstellen sowie als Benutzerschnittstelle eine WEB-GUI im Angebot. Dieser Dienst wendet sich an Institutionen mit Langzeitarchivierungsbedarf ohne eigene oder mit eingeschränkter Infrastruktur. Er implementiert eine Reihe von Qualitätssicherungsmaßnahmen bei der Datenaufnahme, um daraus Garantien für den späteren Zugriff ableiten zu können. Weiterhin ist die DNB im EU-geförderten 4C Project, das dabei helfen soll, Langzeitarchivierungsprojekte zuverlässiger zu planen und durchzukalkulieren, [47] involviert. Hierzu sollen verlässliche Kostendaten ermittelt und geeignete Kostenmodelle entwickelt werden. Weiterhin sollen relevante Initiativen auf dem Gebiet „Kosten in der LZA“ identifiziert und existierende Werkzeuge zur Kostenberechnung für die Anwender in öffentlichen Institutionen und der Privatwirtschaft gebrauchstauglicher werden.

Das Projekt DA-NRW[48] (Digitales Archiv Nordrhein-Westfalen) ist ein Landesprojekt, das vom Ministerium für Familie, Kinder, Jugend, Kultur und Sport gefördert und durch die Historisch-Kulturwissenschaftliche Informationsverarbeitung (HKI) der Universität zu Köln koordiniert und technisch entwickelt wird. Hierzu arbeitet das HKI mit anderen Gedächtnis- und Forschungseinrichtungen an einer prototypischen Infrastruktur zur digitalen Langzeitarchivierung. Im DA-NRW Prototyp sind eine Reihe von Systemschichten realisiert, die eine langfristig ausgerichtete, mehrfach redundante OAIS-konforme Speicher-Architektur mit Schnittstellen zu anderen nationalen und internationalen Archivierungsinitiativen sowie deren Metadatenformaten umsetzen. Das DA-NRW soll damit eine zentrale Infrastruktur in Nordrheinwestfalen für Langzeitarchivierungsaufgaben werden. Das Digitale Archiv NRW implementiert eine Reihe von Tools und Anwendungen, nutzt Fedora als Präsentations-Repository und stellt ein OAI-PMH-Interface zur Verfügung.[49] Als Backend werden ContentBroker zusammen mit iRODS betrieben.

Parallel zu den Entwicklungen in den großen Wissenschaftseinrichtungen entwickeln einige Universitäten disziplinübergreifende Strategien, die fachspezifische Einzelinitiativen unter hochschulweit geregelten Voraussetzungen zusammenfasst. Erweitert und unterstützt werden diese Ansätze durch fachübergreifende Projekte und andere Aktivitäten.

Aktivitäten zur Funktionalen Archivierung

Im Laufe der Zeit entstehen in den verschiedenen Fachdisziplinen zunehmend komplexe Objekte, die zusätzlich unterschiedliche Medienarten zusammenführen. Hierzu zählen - neben den verschiedenen in der Wissenschaft genutzten Verfahren - digitale Kunst, wissenschaftliche Simulationen, elektronische Lernumgebungen oder Laborbücher. Sie spielen daher eine zunehmende Rolle in der aktuellen wissenschaftlichen und kulturellen Literatur- und Informationsversorgung für Forschungsprojekte. Heutige FDZ müssen in der Lage sein, solche Objekte im Rahmen ihrer Aufgabe, langfristig auf effektive Weise bereitstellen zu können.[50]

---

[46] Siehe hierzu die Präsentation: http://files.dnb.de/nestor/veranstaltungen/Praktikertag2013/2013-06-dnb-schmitt.pdf, (zuletzt aufgerufen 30.6.2013)

[47] Siehe hierzu: http://4cproject.net, (zuletzt aufgerufen 12.7.2013)

[48] Siehe hierzu: http://www.danrw.de, (zuletzt aufgerufen 22.6.2013)

[49] Siehe hierzu das Benutzerhandbuch für den technischen Betrieb der Archivierungsknoten: http://www.danrw.de/wp-content/uploads/TechnischeDokumentationDANRW.docx.pdf, (zuletzt aufgerufen 30.6.2013)

[50] Vergleiche hierzu “Empfehlungen zu wissenschaftlichen Sammlungen als Forschungsinfrastrukturen”, Wissenschaftsrat, Drucksache 10 46 4 -11, Januar 2011, Berlin.

Die digitale Langzeitarchivierung im Allgemeinen und die Migration im Besonderen ist insbesondere für Publikationen und digitaler (Ab-)Bilder aller Art ein eingeführter Bestandteil der Agenda von Forschungsdatenzentren.[51] Die praktische Beschäftigung mit der Emulation hingegen zu Bewahrungs- und Zugriffszwecken hat erst in den letzten Jahren begonnen. Zwar beschrieb bereits 1995 Jeff Rothenberg die Bedeutung der Emulation für diese Zwecke[52], doch waren eine Reihe von Anläufen wie das englisch-amerikanische CAMiLEON-Projekt[53] (1999 – 2003) oder die EU-geförderten Großprojekte PLANETS (2006-10) und KEEP (2009-12) erforderlich, um theoretische Konzepte formulieren und erste praktisch Ansätze realisieren zu können.[54]

Unter den Forschungsprojekten hat PLANETS wesentliche konzeptionelle Überlegungen zu Charakterisierung, View Paths[55] und Workflows für die Objektbehandlung beigesteuert sowie die Implementierung eines Referenz-Frameworks für Networked Services[56] erreicht. KEEP, das einen starken Fokus auf die Emulation als Strategie für den langfristigen Zugriff auf digitale Objekte richtete, beschäftigte sich mit Frameworks zur Einbindung konkreter Emulatoren, um diese über abstrakte Schnittstellen ansprechen zu können. Hierbei standen insbesondere Fragen nach notwendigen technischen Metadaten und geeigneter Tool-Registries im Vordergrund. Sie ermöglichen die (automatische) Rekonstruktion von derzeit wenigen Originalumgebungen, die alle notwendigen Komponenten enthalten, um ein bestimmtes digitales Objekt sichtbar zu machen oder ablaufen zu lassen. Mit einem prototypischen Emulationsframework und auf der Basis von TOTEM als Technical Registry wurde eine erste Operationalisierung des View-Path-Konzepts von PLANETS vorgeschlagen, das eine mögliche konzeptionelle Grundlage für die funktionale Bereitstellung von Forschungsdaten in ihrer Originalumgebung bilden könnte. Neben den konzeptionellen Projekten existieren ganz konkrete Prototypen für die Wiederherstellung von Originalumgebungen. So beschäftigt sich die Arbeitsgruppe um G. Brown an der University of Illinois[57] mit einem zentralen Aspekt, der (automatisierten) Wiederherstellung bestimmter Umgebungen für Objektklassen in Windows-basierten Umgebungen. Mit der Integration von in KEEP entwickelten Emulationstechniken in Tessellas Safety Deposit Box[58] sind erste (kommerzielle) Endkundenprodukte und Dienstleistungen für eine funktionale Langzeitarchivierungsstrategie verfügbar.

TIMBUS (Digital Preservation for Timeless Business Processes and Services)[59] ist ein 2011 gestartetes und aktuell von der EU gefördertes Forschungsprojekt, das einen interdisziplinären Ansatz für die Langzeitarchivierung dynamischer Objekte und Prozesse verfolgt. Laufende Unternehmensprozesse und Forschungsdaten stellen gegenüber klassischen statischen Daten neue Herausforderungen dar, da vernetzte Strukturen und die Einbindung von externen Dienstleistern eine zunehmende Rolle spielen. Im Zuge von TIMBUS soll ein Toolkit entstehen, mit dem Zugang, Abruf und Absicherung von Geschäftsprozessen und Forschungsdaten realisiert werden kann. Mit dem Legalities Lifecycle Management soll zudem eine umfassende rechtliche Lösung für die langfristige digitale Erhaltung des Zugriffs auf Daten und die dazu notwendige (proprietäre) Software entwickelt werden.

---

[51] Siehe hierzu: beispielsweise nestor-Handbuch "Kleine Enzyklopädie der digitalen Langzeitarchivierung", (zuletzt aufgerufen 9.7.2013)

[52] Rothenberg, Ensuring the Longevity of Digital Information, Scientific American, 272, 1995; eine neuere Version unter: http://www.clir.org/pubs/archives/ensuring.pdf, (zuletzt aufgerufen 9.7.2013)

[53] Siehe hierzu: http://www2.si.umich.edu/CAMILEON, (zuletzt aufgerufen 9.7.2013)

[54] Research on Digital Preservation within projects co-funded by the European Union in the ICT programme, siehe hierzu: http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf, (zuletzt aufgerufen 9.7.2013)

[55] View Paths beschreiben einen virtuellen Zeiger vom Objekt in eine originale oder kompatible Ausführungs- oder Präsentationsumgebung.

[56] Das Framework diskutiert Möglichkeiten verteilter Dienste zur Langzeitarchivierung, die zu Workflows zusammengefasst werden können.

[57] Siehe http://www.ijdc.net/index.php/ijdc/article/view/153/216

[58] Siehe http://www.digital-preservation.com/solution/safety-deposit-box, (zuletzt aufgerufen 7.7.2013)

[59] Siehe hierzu: http://timbusproject.net/about, (zuletzt aufgerufen 29.6.2013)

Einen anderen Weg bestreitet das baden-württembergische Landesprojekt bwFLA[60]. Ziel des Projekts ist es, Technologie und ein Betriebsmodel für Emulation als Dienstleitung (Emulation-as-a-Service) zu entwickeln. Auf Emulationstechnologie spezialisierte Dienstleister übernehmen die technische Vorbereitung sowie die langfristige Erhaltungsplanung der Emulatoren. Gedächtnisorganisationen nutzen die Dienstleistungen bzw. die Technologie als Werkzeug in den institutionseigenen Workflows, z.B. für Ingest und Access, verwalten aber Objekte und Metadaten weiterhin selbst.

Zusammenfassung & Schlussfolgerungen

Aus der Analyse der Entwicklungen in verschiedenen Ländern und Disziplinen kann zunächst beobachtet werden, dass die Geldgeber zunehmend Wert auf ein sinnvolles, auf Nachnutzung ausgelegtes Forschungsdatenmanagement legen. Insbesondere die Forschung aus öffentlichen Mitteln soll einerseits möglichst kosteneffizient durch Nachnutzung von Ergebnissen sein und andererseits deren Nutzen langfristig dokumentieren. Die Nachnutzung von Wissen und Ressourcen ist letztendlich die Ausweitung der klassischen Bibliotheksidee auf Forschungsdaten.

Generell geht der Trend zu einer Zentralisierung der Anlaufstellen für Forschungsdaten - entweder mit einer Ausrichtung nach den Fachdisziplinen, wie PANGEA oder GESIS, oder als zentrale fächerübergreifende Serviceeinrichtung an Universitäten. In allen betrachteten Projekten und Initiativen treten spezialisierte Einrichtungen oder Dienstleister auf. Dabei kann das Datenzentrum bei relativ überschaubaren Größenordnungen die Daten selbst vorhalten, wie das GESIS. Es muss jedoch nicht zwingend die Primärdaten selbst speichern wie PANGEA zeigt. Der Umfang der Daten scheint eines der wichtigen Entscheidungskriterien für eine zentrale Speicherung zu sein. Jedoch spielen die zentrale Referenzierung und die Bestückung mit geeigneten Metadaten zur Erleichterung von Suchen eine wichtige Rolle. Die dauerhafte Referenzier- und Auffindbarkeit ist eine zwingende Voraussetzung für die Zitierbarkeit von Forschungsdaten. Mit DA-NRW und dem LA-Service der DNB haben zwei Dienste den Betrieb aufgenommen, die wichtige (Teil-)Aufgaben eines FDZ abdecken. Als Repository wurde auf Fedora gesetzt und als Schnittstellen nach außen OAI-PMH angeboten.

Eine Zentralisierung und Spezialisierung scheint vor allem langfristige Effizienzvorteile und Skaleneffekte zu versprechen, da ein Teil des Risiko- und Qualitätsmanagements der Daten durch eine breite Nachnutzung an die Wissenschaftscommunity ausgelagert wird. Eine gemeinsame Institution hilft zudem sich über Metadaten und Qualitätsstandards zu einigen und erleichtert die Kooperation in verteilten Forschungsgruppen oder disziplinübergreifenden Projekten. Für ein zentrales Datenzentrum selbst sinkt jedoch nicht notwendigerweise der Managementaufwand, da mit einer breiten Nutzung wahrscheinlich auch recht breit gefächerte Anforderungen an technische und organisatorische Dienstleistungen gestellt werden. Dies umfasst sowohl die stetige Entwicklung und Anpassung der zur Verfügung gestellten Schnittstellen und Zugriffswerkzeuge, aber auch der Koordinationsaufwand steigt mit der Menge externer Repositorien, die eingebunden und referenziert werden müssen.

Die oben diskutierten Projekte zeigen diesbezüglich noch kein klares Bild, da das Thema Forschungsdatenmanagement größtenteils noch in den Kinderschuhen steckt, insbesondere was längerfristige praktische Erfahrungen betrifft. Derzeit sind viele Vorhaben und Projekte noch in der Aufbauphase - deren Fokus liegt somit stark auf der Aufnahme und Erfassung von Datenbeständen. Der nachhaltige Erfolg eines Forschungsdatenzentrums zeigt sich jedoch erst mit Beginn der Datennutzungsphase. Erst dann werden Strategien, Standards und Qualitätssicherung an den Bedürfnissen und der dann vorherrschenden wissenschaftlichen Praxis gemessen werden können.

Die beiden unterschiedlich angelegten Ansätze zur funktionalen Langzeitarchivierung von KEEP und bwFLA zeigen die mögliche Bandbreite an organisatorischer und technischer Integration einer funktionalen Archivierungsstrategie sowie auch die unterschiedlichen Risiken in der Erhaltungsplanung. KEEP setzt auf

[60] K. Rechert, I. Valizada, D. von Suchodoletz, J. Latocha, bwFLA – A Functional Approach to Digital Preservation, Praxis der Informationsverarbeitung und Kommunikation (PIK), Vol 35(4), 2012.

die Integration der Emulatoren direkt in die (System-)Umgebung des Nutzers. Dies hat zum Vorteil, dass die emulierte Umgebung unter direkter Kontrolle des Nutzers ist, aber auch gleichzeitig den Nachteil, dass ein solches Verfahren auch Wartung und Unterstützung auf Seiten des Nutzers benötigt. Bei einem Dienstleistungsansatz werden die Emulatoren direkt über technische Schnittstellen angesprochen und genutzt. Die Integration in lokale Workflows ist leichtgewichtig. Die Wartung und Erhaltungsplanung sind ausgelagert, was einerseits die Kosten kalkulierbar macht, aber das interne Risikomanagement verkompliziert.

Anforderungsanalyse

Ausgangspunkt der weiteren Untersuchungen dieser Expertise ist zunächst die Konkretisierung der Problemstellung und Ziele des zu entwickelnden Forschungsdatenzentrums (FDZ) sowie die Charakterisierung dessen Nutzer. Ein Ziel der 2011 angestoßenen Untersuchungen für die zukünftige Organisation des Forschungsdatenmanagementsbesteht in der Einrichtung eines Kompetenzzentrums, das als feste Größe in den Altertumswissenschaften etabliert und als Anker für deren Forschungsdaten dienen soll. Das Zentrum sollte sowohl die Begleitung von laufenden Projekten realisieren als auch die Nachhaltigkeit der Forschungsdaten nach Projektabschluss sicherstellen.

Problemstellung

Aus der schnell fortschreitenden technischen Entwicklung folgt auch die Einführung neuer Mittel und Methoden in der Erfassung und Verarbeitung von Forschungsdaten. Aus dieser Entwicklung folgen einerseits schnell wachsende Datenmengen und andererseits eine ebenso schnell wachsende Vielfalt von technischen Formaten und Datenstrukturen. Diese Vielfalt an Daten, sowohl quantitativ als auch qualitativ mittel- und langfristig zu managen, stellt die besondere Herausforderung eines modernen FDZ dar. Da ein Ende der technischen Weiterentwicklung nicht absehbar ist, können technische Lösungen eher nur punktuell wirken, und müssen daher in strukturelle und organisatorische Konzepte, weit über technische Lebenszyklen hinaus, integriert werden.

Ein FDZ mit dem Anspruch, eine zentrale Anlaufstelle einer Disziplin zu werden, muss der anvisierten Wissenschaftscommunity zielgerichtete Angebote und Strukturen liefern, um die entsprechende Akzeptanz zu erlangen und damit die langfristige Finanzierung sicherstellen zu können. Zudem stellt sich die Frage, wie Forschungsdaten möglichst effizient aufbewahrt werden können, um einerseits die Kosten gering und andererseits den Nutzen hoch zu halten. Es kann zudem nicht im Interesse des FDZ liegen, zu einem Datengrab nachrangiger Daten zu werden, die lediglich aus formalen Gründen hinterlegt werden, an deren Pflege und Nachnutzung von Seiten des ursprünglichen Eigentümers nur wenig Interesse besteht oder dieser voraussichtlich nicht über die entsprechenden Ressourcen verfügen wird, die Daten weiter zu pflegen. Daher sind Fragen zu stellen, wie ein End-of-Life von Datensätzen definiert und/oder festgestellt werden kann (d.h. über klassische, statische Haltefristen hinaus) oder wie potenziell ähnliche Datensätze für eine effizientere Nutzung im Laufe der Zeit kuratiert und gegebenenfalls konsolidiert werden können. Damit folgt die Notwendigkeit von organisatorischen und technischen Neu- und Weiterentwicklungen zur Erreichung dieser Ziele, insbesondere aber die disziplinspezifische Anpassung an die speziellen Anforderungen der Fachcommunity.

Anforderungen

Um den Widerspruch zwischen einem hohen Digitalisierungsgrad sowie einem hohen Grad an Kooperation und Data-Sharing einerseits und der hohen Ausdifferenzierung der Fachdisziplinen und bisher zurückgebliebenen Standardisierung und Vereinheitlichung andererseits zu überwinden, sollte ein gemeinsames, übergreifendes Kompetenzzentrum mit den folgenden fachlichen/technischen Anforderungen geschaffen werden:

1. Zentralisierte Datenablage von aktuellen, dynamischen Projektdaten (mittlere Priorität)
2. Online Präsentationsmöglichkeit für Projekte (niedrige Priorität)
3. (Langzeit-)Archivierung von statischen Forschungsdaten (hohe Priorität)
4. Unterstützung von Rechnerintensiven Analyseverfahren (niedrige Priorität)
5. Online-Bereitstellung von Forschungsdaten via Portalen (hohe Priorität)
6. Online-Bereitstellung von Forschungsdaten via (Web-)Services (hohe Priorität)
7. Spiegelung von Datenbeständen externer Institutionen (mittlere Priorität)

Zur weiteren Analyse der geforderten Punkte und Ziele werden diese im Folgenden bezüglich der anvisierten Nutzer-und Zielgruppen, der möglichen Betriebs- und Architekturkonzepte sowie bezüglich der technischen und organisatorischen Anforderungen der funktionalen Langzeitarchivierung diskutiert. Aus dieser Diskussion folgt eine abschließende Bewertung und Zusammenfassung der Empfehlungen für den technischen und organisatorischen Betrieb des geplanten FDZ.

Nutzer und Zielgruppen

Die Altertumswissenschaften sind stark durch kooperative Strukturen sowohl auf nationaler als auch auf internationaler Ebene geprägt. Neben der traditionellen universitären Forschung und Ausbildung in Deutschland findet ein nicht unerheblicher Teil durch außeruniversitäre Einrichtungen, wie das DAI oder regional verankert durch die Landesdenkmalämter, statt. Zudem läuft an verschiedenen Akademien eine Reihe von langfristigen Projekten oder Forschungsvorhaben, die durch DFG-Forschungsverbünde organisiert sind.[61] Während auf institutioneller Ebene langjährige Strukturen bestehen steht das Forschungsdatenmanagement noch weitgehend am Anfang. Für eine zielgerichtete Bewertung der oben gestellten Anforderungen sind die einzelnen Nutzergruppen zunächst anhand ihrer Interaktion mit dem FDZ zu unterscheiden. Daher werden zunächst zwei Gruppen gebildet: *Datenproduzenten* und *Datenkonsumenten*. Beide Gruppen stellen sehr unterschiedliche Erwartungen an ein FDZ und haben verschiedene Sichten auf Forschungsdaten. Die genaue Untersuchung der möglichen Zielgruppen (Kunden) des FDZ mit dem Ziel, die Bedürfnisse und Anforderungen dieser Gruppen zu verstehen, ist notwendig, da die erfolgreiche Einbindung von Nutzergruppen und insbesondere die Berücksichtigung von deren Interessen und Anforderungen der wohl entscheidendste Erfolgsfaktor in der Planung eines neuen, zentralen FDZ ist.

1. **Datenproduzenten:** In der ersten Phase eines entstehenden FDZ sind die Datenproduzenten zunächst die entscheidende Nutzergruppe. Daher ist es von entscheidender Bedeutung, führende Akteure frühzeitig zu identifizieren und in die Planungen und Entwicklung einzubinden, bis eine kritische Masse erreicht wird.

   Primäre Nutzer- und Zielgruppen des geplanten FDZ (Datenproduzenten):

   I. Universitäten und universitäre Forschungsprojekte

   II. Außeruniversitäre Forschungseinrichtungen

   III. Internationale Forschungsverbünde und Initiativen (Konsolidierung/einheitliches Management)

   Datenproduzenten haben in erster Linie ein Interesse daran, ihre Forschungsdaten effizient abliefern zu können, um mindestens den Anforderungen ihrer Geldgeber und Förderer zu entsprechen. Sie arbeiten in der Regel mit aktuellen technischen Werkzeugen und nach derzeitigem Stand der Wissenschaft. Sie sind daher ggf. darauf angewiesen, dass es bei der Erfassung möglichst wenige Einschränkungen bei den erlaubten Datenformaten gibt. Insbesondere verteilt arbeitende (internationale) Projekte benötigen ein effektives Datenmanagement und einen einfachen Austausch ihrer Forschungsdaten.

   Datenkonsumenten: Zwar erfolgt die Interaktion mit Datenkonsumenten erst in einer späten Phase des FDZ (Nachnutzungsphase), aber die Bedürfnisse und Anforderungen möglicher Datenkonsumenten sind bereits in der Planungsphase notwendig, da diese maßgeblich das Qualitäts- und Risikomanagement des FDZ steuern.

   Primäre Nutzer- und Zielgruppen des geplanten FDZ (Datenkonsumenten):

   I. Universitäten und universitäre Forschungsprojekte

[61] Vergleiche hierzu beispielsweise Kapitel 8 in „Langzeitarchivierung von Forschungsdaten - Eine Bestandsaufnahme“.

a. Forschung/Verifikation etc.

b. Einsatz von FD in der Lehre

c. Interdisziplinärer Einsatz

II. Außeruniversitäre Forschungseinrichtungen

III. Institutionelle Nutzer (Städte, Ämter, etc)

IV. Museen, Archive und Bibliotheken

V. Internationale Forschungsverbünde und Initiativen

Datennutzer wollen möglichst gut und schnell passende Daten zu ihrem Forschungsvorhaben finden oder sich schnell einen Überblick über ein bestimmtes Gebiet verschaffen, ohne dabei ganz konkrete Datensätze im Blick zu haben.

Die Aufgabe des FDZ liegt darin, mit technischen und organisatorischen Mitteln die Interessen beider Gruppen auszugleichen. Für das FDZ ist eine Definition von formalen Verantwortlichkeiten, organisatorischen Konventionen und technischen Regeln notwendig. Diese regelt den Umgang mit den Datenproduzenten und -nutzern, organisieren, sichern und kontrollieren die produzierten Daten und Metainformationen.

Datenorganisation und Datenmanagement

Ein Forschungsdatenzentrum muss die eigentlichen Datensätze nicht zwingend selbst speichern, sollte aber die komplette Kontrolle über den Ingest und den Zugriff behalten und hierfür die notwendigen Workflows bereitstellen. Das FDZ sollte hierzu in die Planungen der Projekte für die jeweiligen Maßnahmen zur Erhebung, Sicherung, Verarbeitung und Dokumentation von Projektdaten einbezogen werden. Auf beiden Seiten sind die grundsätzlichen Verantwortlichkeiten im Datenmanagement festzulegen und entsprechend im Datenmanagementplan zu beschreiben. Die Zuständigen seitens des FDZ ist das Planen, die Organisation und Speicherung von Daten samt Metadaten. Die Projektseite legt fest, wie mit den Daten in der täglichen Arbeit umgegangen wird, und welche Dateien mit welchem Status (z. B. Entwurf, Zwischenversion, Endversion, Originaldatei oder abgeleitete Kopie) wie lange gespeichert werden sollen. Darüber hinaus sind alle getroffenen Maßnahmen zur Qualitätssicherung zu dokumentieren, welche die Nutzbarkeit und Qualität der gespeicherten Daten gewährleisten und die gesicherte Speicherung der digitalen Informationen garantiert.

Ein solch institutionalisiertes Forschungsdatenmanagement bindet die jeweils laufenden Forschungsprojekte möglichst frühzeitig ein, überwacht beratend deren Datenmanagement, beispielsweise im Sinne einschlägiger Richtlinien, und unterstützt diese bei der laufenden Erfassung von Metadaten. Die wesentlichen Dienstleistungen reichen von der Bereitstellung zitierbarer Primärdaten bis hin zur Aufnahme aller entstandenen Publikationen und Forschungsarbeiten. Zusätzlich unterstützt das FDZ die strukturierte Speicherung wissenschaftlicher Daten, Referenzen und Kontextinformationen.

Die Erfüllung, insbesondere der beratenden Funktionen, setzt voraus, dass frühzeitig Kontakt zu Projekten und Forschergruppen besteht. Allgemeine Datendienstleistungen für diese könnten hier eine Brückenfunktion übernehmen. Die oben genannten Aufgaben lassen sich mit der Bereitstellung einer zentralisierten Datenablage (vgl. Anforderung 1) verbinden und damit wesentlich effizienter umsetzen. Im Sinne eines konsistenten Datenmanagements und folgender Langzeitarchivierung scheint eine (vor-)strukturierte Datenablage in Verbindung mit Beratung sehr erfolgsversprechend. Somit sollte bei der Wahl des einzusetzenden Systems geprüft werden, ob dieses erweiterbar ist (Plug-Ins etc.), um beispielsweise die Erstellung von Meta-Daten zu unterstützen oder die Daten im Sinne der Langzeitarchivierung (LZA) vorzustrukturieren.

Zudem sind in laufenden Forschungsprojekten die Daten nicht fix, sondern ändern sich durch Aggregation, Umformatierung, Normalisierung oder durch die Beseitigung von Fehlern. Dieses gilt ebenfalls für einen Teil

der mit ihnen verknüpften Metadaten. Insbesondere bei größeren Forschungsprojekten mit verteilt arbeitenden Wissenschaftlern ist die Versionierung eine wichtige Hilfe zur Nachvollziehbarkeit von Änderungen bezogen auf Zeitpunkte und involvierte Akteure. Diese Informationen können die Grundlage für Provenance-Metadaten bilden und damit die Bewertung der Daten durch Dritte hinsichtlich Vertrauenswürdigkeit und Korrektheit erlauben. Diese Aufgabe könnte das FDZ im Zuge einer Unterstützung von dynamischen Daten als Datendienstleister übernehmen (Anforderung 1).

Datenhaltung und Publikation

Eine weitere grundlegende Vorbedingung für die Datenpublikation bildet das Datenmanagement, eine der zentralen Forderungen aus der Aufgabenstellung der Expertise. Für die eigentliche Erfassung, Bearbeitung, Anreicherung mit Metadaten und die spätere Datenbereitstellung bieten Datenmanagementsysteme (DMS) die notwendige technische Basis. Die vom DMS zu erbringenden Dienste lassen sich dabei in Teilaufgaben, wie die Speicherung der Daten, Suche und Verwaltung und Durchsetzung von Zugriffsrechten untergliedern. Nach dem Modell des „Data Curation Continuum“ von A. Treloar, D. Groenewegen und C. Harboe-Ree[62] übernimmt ein DMS die *Domäne des Datenmanagements* für Kooperation und eigentliche Forschung. Für die Publikation und verbundene Dienste kommen spezialisierte Systeme in Betracht, da publizierte Daten dauerhaft auffindbar bleiben müssen, damit sie zitierfähig (zugänglich für DataCite) und in anderen Publikationen referenzierbar werden. Publizierte Daten dürfen deshalb nicht mehr verändert werden können, was sie von aktuell in der Forschung verwendeten Daten unterscheidet. Zu publizierende Daten werden deshalb in einem eigens vorgenommenen Kuratierungsschritt, der beispielsweise nur Teile der vorliegenden Daten auswählt, in die *Publikationsdomäne* übertragen (Anforderung 3). In DMS werden hingegen alle Daten vorgehalten, unabhängig davon, ob sie jemals publiziert werden. Hierzu zählen Datensätze mit einer begrenzten Haltefrist, die beispielsweise aufgrund der Empfehlungen der DFG zur guten wissenschaftlichen Praxis zehn Jahre aufbewahrt werden. DMS halten Daten für Kooperation, innerhalb einer Einrichtung, Instituts oder Forschergruppe vor, um auf diesen direkt arbeiten zu können (Anforderung 1). Datensätze können sich beispielsweise durch Aggregation oder Weiterverarbeitung ändern und/oder in mehreren Versionen vorliegen.

Eine redundante Datenhaltung ist notwendig. Diese Problematik wurde jedoch bereits im Kontext verschiedener anderen LZA-Fragen weitestgehend konzeptionell als auch praktisch gelöst. Im Sinne einer resilienten FDZ-Architektur sollte diese Aufgabe an mehrere externe Dienstleister ausgelagert werden und als Vorleistung betrachtet werden. Die Auslagerung von archäologischen Forschungsdaten wird nicht durch den notwendigen Schutz personenbezogener Daten oder spezielle Datenschutzanforderungen verkompliziert. Standardisierbare externe Modelle, wie sie derzeit von diversen Initiativen entwickelt werden, sind Eigenentwicklungen zu bevorzugen. Doppelentwicklungen in diesem Bereich sind wenig empfehlenswert und führen zu keinem langfristigen Mehrwert für das Zentrum und seine Nutzer. Das geplante FDZ sollte und müsste daher seine Ressourcen auf die Entwicklung und Bereitstellung höherwertiger Dienstleistungen konzentrieren. Storage-Dienstleister und Datenzentren sind in der Regel generisch ausgelegt, so dass das FDZ selbst die Workflows gestalten und langfristig überwachen muss. Dies ist auch notwendig um die Hoheit über die Abläufe, z.B. bei Anbieterwechsel und Anbieter-Diversifikation, zu behalten. Zum pro-aktiven Risikomanagement gehören auch eine laufend aktualisierte Expertise in der Bewertung von FLAs sowie die Erarbeitung von Ausstiegs- und Migrationsszenarien.

Für die allgemeine Verfügbarmachung von Daten sind *Persistent Identifier* zu vergeben. Dieses erfolgt typischerweise bei der Überführung in den dauerhaften Speicherbereich, in dem Daten nicht mehr verändert und für die Publikation zur Verfügung gestellt werden. Das geschieht damit relativ spät im Lebenszyklus der Objekte und umfasst eher ausgewählte, aufbereitete Datensätze als die Gesamtheit der Daten eines Projekts. Unter Umständen macht es bereits Sinn, Daten bereits früher in der Datenmanagementdomäne zu referenzieren, dieses liegt im Ermessen der entsprechenden Projekte und Teildisziplinen. Für die Publikation von Daten existieren derzeit mehrere Systeme von sogenannten Identifiern nebeneinander, zu denen Digital

[62] Siehe hierzu den Artikel in der Dlib: http://www.dlib.org/dlib/september07/treloar/09treloar.html

Object Identifier[63] (DOI), persistent URLs[64] (PURL), Uniform Resource Names[65] (URN) oder Archival Resource Key[66] (ARK) zählen. Um die Hürden für Datenproduzenten und -nutzer gering zu halten und gleichzeitig die Auffindbarkeit der referenzierten Objekte sicherzustellen, ändert sich ein zugeteilter Identifikator in Bezug auf das Objekt nicht. Die eigentliche Objekt-URL muss hingegen nicht persistent sein. Sie wird durch das jeweilige interne Resolving-System des Identifiers immer zur aktuellen URL aufgelöst. Dieses sicherzustellen ist Aufgabe des FDZ.

Unter den verschiedenen Systemen scheint sich der DOI verstärkt durchzusetzen, da es Vorteile gegenüber anderen Systemen aufweist. So hängen beispielsweise die in Deutschland und den Niederlanden oft eingesetzten Uniform Resource Names for National Bibliography Numbers[67] (URN:NBN) von der expliziten Angabe eines nationalen Resolvers ab, wohingegen DOI einen globalen Handle nutz, der über den zentralen Proxy http://dx.doi.org ermittelt wird. Seit 2012 wurde DOI durch die International Organization for Standardization (ISO) zum ISO-Standard 26324:2012 für die Dokumentation digitaler Objekte erhoben. So nutzen beispielsweise PANGEA, das GESIS, DataCite[68] oder die International Association of Scientific, Technical & Medical Publishers (STM) dieses System. Hinzu kommt die überwiegend auf DOI-Vergabe für wissenschaftliche Artikel ausgerichtete CrossRef-Organisation, womit sich das DOI-System schrittweise zum de facto-Standard des elektronischen Publizierens etabliert. In Entwicklung befindet sich zudem ein „DOI Content Negotiation"-Service, der es künftig erlauben soll, die Metadaten von durch CrossRef registrierten Publikationen und von durch DataCite registrierten Forschungsdaten automatisiert auszutauschen, um so „Enhanced Publications" quasi automatisch entstehen zu lassen.

Erhaltungsplanung und Risikomanagement

Für den langfristigen Zugriff auf Forschungsdaten können aufgrund des technologischen Wandels oder die Änderung der Anforderungen seitens der Datenkonsumenten Maßnahmen notwendig werden, die der technischen und intellektuellen Nachnutzbarkeit dienen. So können beispielsweise Datenformate obsolet werden und nicht mehr in den aktuellen Arbeitsumgebungen nutzbar sein. Ebenso könnten bestimmte Workflows und Verfahren nicht mehr funktionieren, weil keine geeignete Ablaufumgebung mehr zur Verfügung steht. Um feststellen zu können, dass es relevante Änderungen in der Zielgruppe oder der Technologie[69] gab, sind die notwendigen Metadaten und Informationen bei der Einlagerung der Daten zu erheben. Den Umgang mit solchen Änderungen regeln ein geeigneter Planungsprozess und dazugehöriges Risikomanagement.[70] Während technische Erhaltungspläne bereits gut etabliert sind und in der Regel unproblematisch umsetzbar sind, liegen die Anforderungen für eine effiziente inhaltliche Nachnutzung der Daten wesentlich höher.

Gegebenenfalls verteilte Infrastruktur und die dezentrale Veröffentlichung von Forschungsdaten kann bedeuten, dass sich Datenerzeuger, Manager und Nutzer nicht zwingend kennen und zwischen Dateneinlagerung und Nachfrage erhebliche Zeitdifferenzen liegen können. Daher kann es sinnvoll sein, die Entscheidung über die weitere Aufbewahrung von Daten nicht allein von fachinternen Kriterien abhängig zu

[63] Siehe hierzu: http://www.doi.org, (zuletzt aufgerufen 12.6.2013)
[64] Siehe hierzu: http://www.purlz.org, (zuletzt aufgerufen 12.6.2013)
[65] Siehe hierzu: http://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml, (zuletzt aufgerufen 12.6.2013)
[66] Siehe hierzu: https://wiki.ucop.edu/display/Curation/ARK, (zuletzt aufgerufen 12.6.2013)
[67] Siehe hierzu: http://www.persistent-identifier.nl für die Niederlande und http://nbn-resolving.org/ für Deutschland, Österreich und Schweiz (jeweils zuletzt aufgerufen 12.6.2013)
[68] Vergleiche hierzu: http://www.datacite.org, (zuletzt aufgerufen 12.6.2013)
[69] In der Fachliteratur als „Technology and Community Watch" bezeichnet.
[70] Christoph Becker and Andreas Rauber , Decision criteria in digital preservation: What to measure and how. Journal of the American Society for Information Science and Technology (JASIST), 2011.

machen. Das FDZ kann über fachbezogene Kriterien hinaus, Entscheidungshilfen definieren, welche eine kostengünstige und effiziente Datenhaltung ermöglichen. Neben gesetzlichen oder fachspezifischen Vorgaben wäre eine mögliche Metrik das Interesse seitens der Datenkonsumenten, deren Zugriffe oder die Zahl der Zitate eines Datensatzes. Eine geeignete Berücksichtigung dieser Problematik bei der Umsetzung von Anforderung 2 (Online-Präsentation) und Anforderung 5 (Bereitstellung über Portale) können somit wesentlich dazu beitragen, Datenbestände effizient zu pflegen. Durch die regelmäßige Nutzung und Überprüfung durch die Fachcommunity lassen sich die Qualität und Nutzbarkeit der Daten nachhaltig verbessern und ggf. Datenfriedhöfe vermeiden.

Der Erhaltungsplan kann zudem um eine funktionale Komponente ergänzt werden. In diesem Fall werden die Softwareumgebungen, die für die Erstellung der jeweiligen Objekte genutzt wurden, bereits bei der Datenaufnahme berücksichtigt.

Metadaten

Forschungsprimärdaten müssen mit Metadaten versehen werden, da diese überhaupt erst eine sinnvolle Nachnutzung ermöglichen. Hierzu zählen neben technischen Metadaten, wie Größe, Format, Versionierung, die Art der Daten (Rohdaten oder weiterverarbeitet), die Entstehungs- und Bearbeitungsgeschichte, sowie bei technisch erzeugten Datensätzen Konfigurations- und Kalibrierungsdaten. Eine große Herausforderung besteht in der qualitativ hochwertigen Bereitstellung von Metadaten, da dieses oft mit einem hohen Aufwand verbunden ist, den viele Wissenschaftler, insbesondere für Datensätze, die nicht primär für die Veröffentlichung vorgesehen sind, scheuen. Ohne Beschreibung verlieren Datensätze schnell an Wert, da nur der Datenproduzent eine sinnvolle Interpretation vornehmen kann. Derzeit werden allgemeine Verfahren untersucht und entwickelt, die versuchen möglichst früh, idealerweise schon im Augenblick der Entstehung, Metadaten automatisch zu erzeugen. Insbesondere für technische Metadaten gibt es erste Ansätze, wie beispielsweise FITS.[71] Wird das DMS von vorneherein involviert (Anforderung 1), können Provenance-Metadaten bereits durch dieses erhoben und mit den jeweiligen Datensätzen abgelegt werden.

Schnittstellen

Metadaten dienen in den Altertumswissenschaften auch dazu, Informationen effizient auszutauschen, und Nutzern zur Verfügung stellen zu können, ohne dass dafür aufwändige Abstimmungen zwischen einzelnen Systemen notwendig sind (Anforderung 5). Stattdessen sollen standardisierte Datenschnittstellen wie OAI/PMH[72] eingesetzt werden (Anforderung 6), in die dann unterschiedliche semantische Ontologien implementiert und auf die Ausgangssysteme abgebildet werden können (etwa CIDOC-CRM, Dublin Core Metadatenschema etc.).

So sind zwar einerseits die Altertumswissenschaften in der Digitalisierung und ganz allgemein im Einsatz von technischen Hilfsmitteln weit vorangeschritten, jedoch existieren andererseits keine zentralen Strukturen für ein fachspezifisches Forschungsdatenmanagement oder für allgemein übergreifend genutzte Standards für Datenformate und Metadaten. Die technischen Folgen der Wahl bzw. Festlegung auf eine Art von Metadaten zeigen sich hauptsächlich in der Bereitstellung von Forschungsdaten. Während mit Datenproduzenten, insbesondere durch frühe Beratung, ein gemeinsamer Standard gefunden werden kann, sind die Anforderungen der Datenkonsumenten weniger vorhersehbar. Angesichts der Anforderungen 5 (Bereitstellung durch Portale) und 6 (Bereitstellung durch Webservices) sollte bei der Wahl der vom FDZ unterstützten Metadaten-Schemata die Interoperabilität und die Konvertierungsstrategien eine wesentliche Rolle spielen. Eine breite Unterstützung von Formaten für den Datenzugriff machen die Daten des FDZ für Portale etc. einfach integrierbar und damit attraktiv (Anforderung 5).

[71] File Information Tool Set, (Stern & McEwen, 2009)

[72] Schnittstellendefinition der Open Archives Initiative: http://www.openarchives.org/pmh, (zuletzt aufgerufen 22.6.2013)

Alternative Erhaltungsstrategien - Funktionale Archivierung

Ein funktionaler Ansatz für Archivierung und Zugriff macht die spätere Nutzung höchst unterschiedlicher Objekttypen insofern von der technologischen Weiterentwicklung unabhängig als dass die originalen Software-Hardware-Umgebungen, in denen die nachgefragten Objekte erstellt wurden, aufgehoben werden. Die funktionale Archivierung wird i.d.R. durch sekundäre Workflows charakterisiert. Die archivarische Erfassung und Beschreibung der Objekte, insbesondere hinsichtlich Biterhaltung, sind daher zuerst durchzuführen. Wenn ein Objekt in das Langzeitarchiv übertragen wird, erfolgt zunächst seine Klassifizierung. Hierzu kann es notwendig sein, dass der Objekttyp umfassend (technisch) bestimmt wird und falls notwendig fehlende Informationen hinzugefügt werden. Falls das Objekt aus mehreren Teilobjekten besteht, muss diese Bestimmung für jedes Teilobjekt ausgeführt werden. Auf jeden Fall ist der Dateityp zu ermitteln, damit eine geeignete Einordnung in eine oder mehrere Archivierungsstrategien erfolgen kann.

Für die Ermittlung des Dateityps und weitergehender Angaben bieten sich verschiedene, bereits vorhandene Formate Registries oder -Validatoren an:

- PRONOM[73] ist eine öffentlich nutzbare Registratur des Britischen Nationalarchivs, welches ursprünglich für interne Zwecke eingerichtet, später aber allgemein öffentlich zugänglich gemacht wurde. Für bekannte Dateitypen liefert PRONOM Metainformationen, wie den Ersteller des Dateiformats, Informationen zur Erstellungsapplikation, Versionsnummer und technische Details über den Aufbau des Formats. Die Registratur verwendet sogenannte PUIDs (PRONOM Persistent Unique Identifier), die ein ähnliches Konzept wie Unix Magic Numbers1 umsetzen. Bisher steht eine Webschnittstelle zur Nutzung offen, die derzeit um einen entfernt einbindbaren Web-Service erweitert wird. Zusätzlich wird mit „DROID" eine separate Java Applikation zur Nutzung angeboten, die diese Registratur benutzt.

- JHOVE (JSTOR/Harvard Object Validation Environment) ist ein Werkzeug zur Formaterkennung und Validierung. Das Programm wurde in Java erstellt, das die Version 1.4 auf dem jeweiligen System voraussetzt. Es lässt sich als Bibliothek in eigene Projekte einbinden. Die von diesem Programm erzeugten technischen Metadaten können sowohl in XML mit definiertem XML-Schema als auch in Standard-Textform ausgegeben werden.

Wenn sich Objekte nicht für Migration eignen oder alternative (redundante) Strategien notwendig sind, kann die funktionale Langzeitarchivierung zum Einsatz kommen. Dieses impliziert eine Reihe von Workflows, die von der Objektaufnahme, über die Haltezeit des Objekts im Archiv bis hin zur Bedienung der Datenkonsumenten reichen. Um eine funktionale Archivierungsstrategie zu implementieren, sind insbesondere drei Workflows durch das FDZ zu implementieren: 1. Erfassung der Laufzeitumgebung eines Objekts, 2. Erstellung einer Laufzeitumgebung und 3. Zugriff auf das Objekt in der gewählten Laufzeitumgebung.

Erfassung der Laufzeitumgebung eines Objekts (FLA-Ingest)

Ein typischer FLA-Ingest Lauf beginnt mit dem Import eines Artefakts aus einem (möglicherweise externen) Repository. Abhängig von dessen konkreter Manifestation und Metadaten müssen ggf. noch vorbereitende Maßnahmen für die spätere Einführung in die emulierte Umgebung getroffen werden. Anschließend kann nach passenden Laufzeitumgebungen gesucht werden. Dies kann (halb-)automatisch anhand der bereits vorhandenen Metadaten (z.B. Dateiformat, Erstellungsdatum, Provenienzinformationen, etc.) geschehen oder durch manuelle Wahl eines Archivars (Abb. 1: Workflow-Schritte 0-2). Hierbei kann sich herausstellen, dass für das gegebene Objekt noch keine Umgebung existiert. An dieser Stelle ist zu überlegen, ob das Objekt zurückgewiesen wird oder die notwendigen Komponenten zur Unterstützung in das Archiv eingestellt werden. Dieses kann spezielle Software sein, mit der das Objekt erstellt oder bearbeitet wurde. Das wiederum kann bedeuten, dass ein rekursiver Softwarearchivierungs-Workflow angestoßen wird, der schrittweise alle

[73] Vergleiche hierzu: http://www.nationalarchives.gov.uk/PRONOM, (zuletzt zugegriffen 8.7.2013)

notwendigen Softwarekomponenten zur Wiederherstellung der Objektumgebung inklusive Metadaten erhebt und mitarchiviert. Ist die passende Umgebung gefunden bzw. die technische Beschreibung dieser, können die technischen Metadaten erzeugt, dem Objekt zugeordnet und gespeichert werden. Optimalerweise wird das Rendering des Objekts in seiner zukünftigen Zugriffsumgebung verifiziert. Dieses geschieht mithilfe des Datenerstellers bzw. -eigentümers, der auf diese Weise die durch ihn erwartete korrekte Wiedergabe sicherstellt und ggf. noch objektspezifische Anpassungen nehmen kann (Abb. 1, Workflowschritt 1).

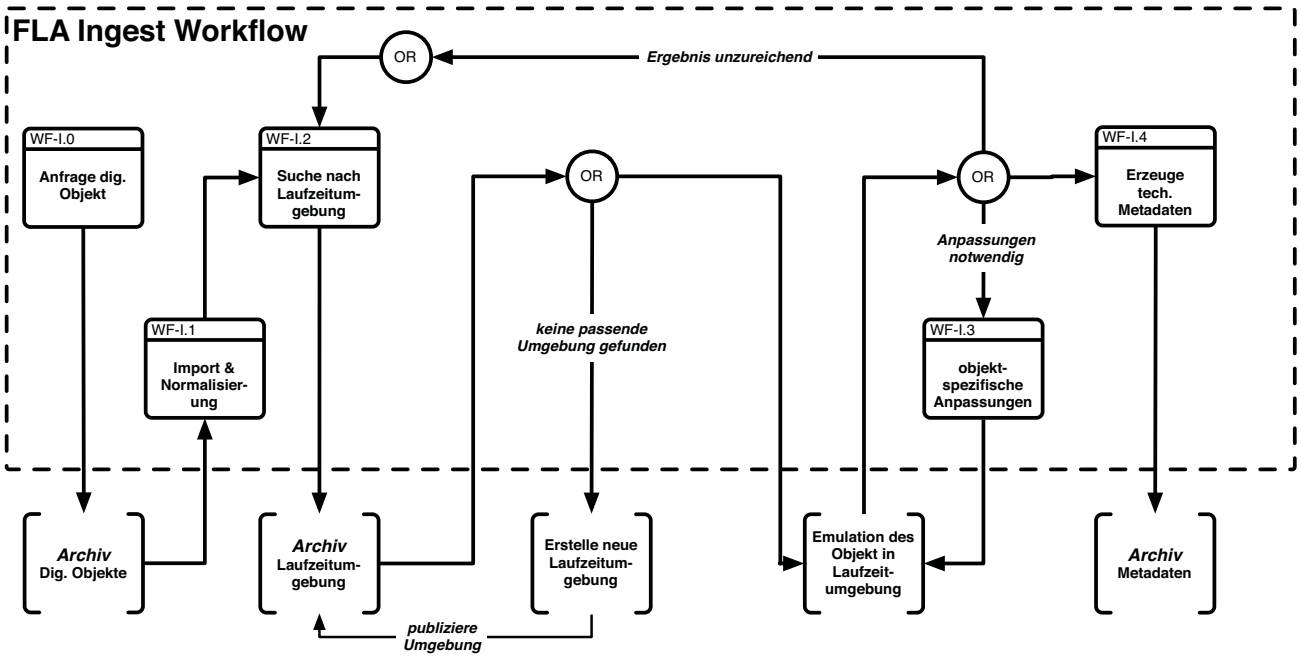

Abbildung 1: Workflow zur Erfassung einer Laufzeitumgebung eines Objekts

Erstellung und Pflege von Laufzeitumgebungen (Image-Archive)

Der zweite wesentliche und konzeptionell wichtige Workflow beschreibt die Erstellung einer Laufzeitumgebung (vgl. Abb. 2). Mit Hilfe einer emulierten Umgebung werden so lange Software-Komponenten installiert und getestet bis die Umgebung vollständig und frei von Konflikten ist. [74] Dabei können, möglichst automatisch, technische Metadaten erzeugt werden, die eine Replikation dieses Vorgangs ermöglichen. Für die benötigten Softwarekomponenten kann es sinnvoll sein, ein eigenes Archiv für die Wiederherstellung von Originalumgebungen bereitzuhalten oder einen solchen Dienst von Dritten zu beziehen. Letzteres empfiehlt sich für Standardsoftware, insbesondere für disziplinspezifische Software ist eine unabhängige Archivierung, insbesondere auf Grund der möglicherweise geringen Verbreitung der Komponenten, empfehlenswert.

[74] vgl. Rechert, K., Valizada, I., von Suchodoletz, D.: Future-proof preservation of complex software environments. In: Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES2012), University of Toronto Faculty of Information (2012) 179–183

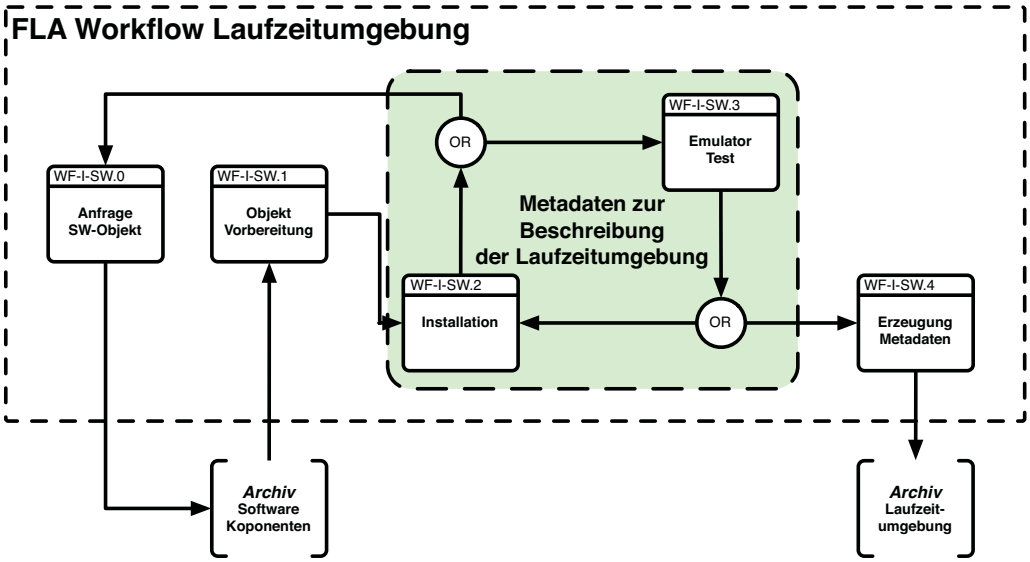

Abbildung 2: Workflow zur Erstellung einer Laufzeitumgebung

Funktionaler Zugriff auf digitale Objekte

Zu einem gegebenen Zeitpunkt wird ein Primärobjekt von einem Archivbenutzer nachgefragt werden. In jedem Fall benötigt das entnommene Objekt einen Kontext, der dem Nachfrager eine Interpretation erlaubt. Mit zunehmendem Zeitabstand zur Erstellung des Objekts sinkt die Wahrscheinlichkeit, dass dieser Kontext in der digitalen Arbeitsumgebung des Datenkonsumenten bereits vorliegt. Hierzu wird eine geeignete Umgebung erwartet, in der ein Archivnutzer das Objekt betrachten oder ausführen kann. Die verschiedenen Kunden des FDZ werden sich im Grad ihrer Kenntnisse unterscheiden. Da bei einem durchschnittlichen Nutzer nicht zwingend von einem erfahrenen Computeranwender auszugehen ist, sind Überlegungen anzustellen, wie dieser geeignet an die notwendige Software und ihre Schnittstellen herangeführt werden kann, um mit ihr umgehen zu können.

Für die Wiederherstellung von Originalumgebungen, welche benötigt werden, damit Datenkonsumenten auf archivierte Objekte zugreifen können, wird eine Reihe von weiteren Workflows benötigt. Diese reichen von der Beschaffung aus dem Archiv, der Interpretation der technischen Metadaten, über den Transport in die Originalumgebung bis zur Konfiguration und Start eines geeigneten Emulators oder virtueller Maschine. Mit der fortschreitenden Technik sind die jeweils aktuellen Arbeitsumgebungen von einem permanenten Wandel betroffen, der den Zugriff gefährdet. Über den Zeitraum der Notwendigkeit der Verfügbarkeit der jeweiligen Umgebung sind die Überwachung des technologischen Wandels und der Anforderungen seitens der Nutzercommunity notwendig. Die Aufgaben bzgl. der technischen Erhaltungsplanung sind idealerweise an externe Dienstleister zu delegieren.

Mögliche Umsetzungsstrategien

Aus der obigen Aufgabendefinition ergeben sich für die Umsetzung bzw. Integration von funktionalen Strategien drei mögliche Umsetzungskonzepte:

1. ***Minimal-Ansatz***: Das FDZ fordert bei der Aufnahme von Forschungsdaten die Listen von verwendeten Software-Komponenten. Zusammen mit dem Datenproduzenten wird entschieden, welche Software zusätzlich für die Einlagerung in Frage kommen. Die Auswahl sollte sich dabei auf disziplinspezifische Spezialsoftware beschränken, da populäre und weit verbreitete Softwarekomponenten wie zum Beispiel Excel o.ä. auch in Zukunft sicherlich ohne größere Probleme beschafft werden können (beispielsweise durch Softwarearchive von Staats- und Nationalbibliotheken mit entsprechendem (ggf. gesetzlichen) Auftrag).
   Dieser Ansatz ist sehr kostengünstig und bedarf wenig zusätzlichen Aufwands. Nachteilig jedoch ist ein Verschieben der notwendigen Arbeiten (und damit auch Kosten) in die Zukunft. Bei Bedarf

müsste die vom Produzenten abstrakt skizzierte Umgebung manuell rekonstruiert werden. Ein Rückgriff auf das Fachwissen des Produzenten oder auf derzeit allgemeines Fachwissen zur Konfiguration und Nutzung der entsprechenden Softwarekomponenten ist nicht garantiert. Somit bestehen erhebliche Kostenrisiken und eine ungewisse, aber nicht unwahrscheinliche Erfolgsaussicht, die Daten funktional nutzbar zu machen. Technische Kompetenz kann an spezialisierte Dienstleister ausgelagert werden, die im Falle des Zugriffs mit der technischen Umsetzung vom möglichen Datenkonsumenten beauftragt werden. Die Bereitstellung der technischen Infrastruktur, u.a. die Bereitstellung eines passenden Emulators, ist ebenfalls über einen technischen Dienstleister sicherzustellen.

2. ***Strukturierte funktionale Erfassung:*** Dieser Ansatz erfordert den geführten bzw. strukturierten Nachbau der genutzten Forschungsumgebung. Bei diesem Prozess, durchgeführt durch den Datenproduzenten, werden alle funktionalen Abhängigkeiten der Daten automatisch erfasst und mittels technischer Meta-Daten beschrieben. Jeder Installations- und Konfigurationsschritt wird gleich in einer emulierten Umgebung getestet und die Funktionalität der Umgebung bestätigt. Zuletzt werden die Daten in die Umgebung eingebracht, so dass der Nutzer deren Darstellung, Verarbeitung etc. bewerten und dokumentieren kann.
   Dieser Ansatz ist bei der Datenerfassung aufwändiger als der oben beschriebene Minimal-Ansatz. Der Datenproduzent muss eine Reihe von Tests durchführen sowie alle Prozesse und Konfigurationen seiner Forschungsumgebung reproduzieren. Zudem liefert dieser Ansatz Garantien bzgl. der Vollständigkeit und Funktionalität der Softwareumgebung im Zusammenwirken mit dem entsprechenden Datensatz. Damit beschränken sich die zukünftigen Kostenrisiken bei diesem Ansatz auf die Bereitstellung der Emulationsumgebung und ggf. der notwendigen Softwarelizenzen.

3. ***Integrierter funktionaler Ansatz:*** Um die Kosten des Datenproduzenten, insbesondere in Großprojekten zu reduzieren, empfiehlt es sich die technische Beschreibung von Forschungsumgebungen und deren Prozessen frühzeitig, möglichst „nebenbei“, zu erzeugen. Dafür empfehlen sich insbesondere zentral „gemanagte“ Computersysteme, so dass bei der Vorbereitung der Systeme die notwendigen Prozesse zur Erzeugung der technischen Meta-Daten integriert werden können, so dass nur projektspezifische Einstellungen noch zusätzlich dokumentiert werden müssen.

Diskussion

Ein Vorteil eines funktionalen Ansatzes ist der Verzicht auf eine Einschränkung der erlaubten Datenformate. Sofern die Daten einer Laufzeitumgebung technisch zugeordnet werden können, lassen sich diese wieder nutzen. Diese Nutzung jedoch ist zunächst auf eine Verwendung und Bearbeitung innerhalb der Laufzeitumgebung beschränkt. Eine direkte Nutzung mit aktuellen technischen Mittel ist zunächst nicht direkt möglich. Somit liegt der primäre Einsatzzweck zunächst in der Sichtung von Daten, der Replikation oder Verifikation und der interaktiven Nutzung in der Originalumgebung. In einem zweiten Schritt kann die Originalumgebung dazu genutzt werden, Formatmigrationen durchzuführen und einzelne Datensätze oder verarbeitete Daten zu „exportieren“, d.h. in der aktuellen Arbeitsumgebung nutzbar zu machen.

Ein ggf. erhebliches Risiko stellen kurz und mittelfristig Urheberrechte auf Software dar. Bereits beim Ingest ist eine Prüfung der notwendigen Lizenzen notwendig. Ebenso problematisch sind Softwarekomponenten, die mittels Lizenzserver o.ä. für die Nutzung freigeschaltet werden müssen. Auf Grund der breiteren Aufmerksamkeit dieser Problematik[75] ist die Einrichtung von dedizierten Softwarearchiven mit entsprechendem Geschäftsmodell wahrscheinlich. Unabhängig davon kann die Erfassung der Laufzeitumgebung von Forschungsdaten, insbesondere für die kurz- und mittelfristige Nachnutzung, aber auch für die langfristige Dokumentation von wissenschaftlichen Prozessen, sehr sinnvoll sein. Gerade kurz- und mittelfristig sind die Risiken der Lizenzproblematik, wahrscheinlich durch die bestehenden Lizenzen und

[75] Vgl. u.a. TIMBUS Projekt, http://www.timbusproject.net:
Elisabeth Weigl, Johannes Binder, Stephan Strodl, Barbara Kolany, Daniel Draws and Andreas Rauber, A Framework for Automated Verification in Software Escrow, iPres 2013.

Rahmenverträge, beherrschbar. Zudem wird hinsichtlich der recht kurzen Lebenszyklen der kommerzielle Wert von archivierten Software-Komponenten in den für FDZ entscheidenden Zeiträumen schnell irrelevant, so dass längerfristig neue Lösungen wahrscheinlich sind.

Ein weiterer Einsatzzweck einer funktionalen Strategie in einem FDZ ist die Archivierung von wissenschaftlichen Prozessen. Durch die Archivierung einer ganzen Umgebung wird meist auch die Replikation von wissenschaftlichen Prozessen möglich. Zusammen mit den Forschungsprimärdaten können Zwischenergebnisse sowie alle Prozessschritte beobachtet und validiert werden. Weiterhin können archivierte Umgebungen in der universitären Lehre einen realistischen und ggf. funktionalen Eindruck in die Arbeiten eines Forschungsprojekts geben.

Heutige Forschungsumgebungen sind u.U. auch auf externe Dienste angewiesen. Diese können Datendienste, aber auch Dienste für Berechnungen oder weitere Arten der Datenverarbeitung, sein. Eine funktionale Archivierungsstrategie findet an solchen externen Abhängigkeiten zunächst ihre technischen Grenzen. Auch in diesen Fällen kann eine Erfassung für die funktionale Archivierung sinnvoll sein, da solche Abhängigkeiten explizit werden und damit erst ein entsprechendes Risikomanagement und entsprechende Anpassungen in der Erhaltungsplanung erlauben.

Insgesamt kann eine breite Einführung und konsequente Nutzung einer funktionalen Archivierungsstrategie für zusätzliche Redundanz auch auf der Erhaltungs- und Zugriffsebene sorgen und nicht nur für Fälle, die mit Migration zu ineffizient oder nicht möglich sind.

Technischer Aufbau und Architektur

Die gestellten Anforderungen können mit einem zwei- bzw. dreiteiligen Schichtenmodell erreicht werden. Für das geplante FDZ kommen die folgenden drei Schichten in Frage: (1) Speicherschicht, (2) Managementschicht und (3) Zugriffsschicht:

1. **Speicherschicht:** Auf der untersten Schicht, der Speicherschicht, werden die digitalen Manifestationen der Objekte, d.h. deren Daten abgelegt. Je nach Anlage der Architektur kann das gemeinsam mit den zugehörigen Metadaten erfolgen. Viele Repository-Systeme nutzen auf dieser Ebene eine einfache Datei-basierte Ablage, die eine Rekonstruktion aller Informationen aus den Dateien erlaubt und nicht auf weitere Komponenten wie ein Datenbankmanagementsystem angewiesen sind. Wegen Umfang und Größe der Dateisammlungen wird man die Daten zur Erfüllung der Storage-Anforderungen an (externe) Dienstleister in ein Datenzentrum bzw. Rechenzentrum auslagern. Das vereinfacht die Aufgaben zur Bitstream-Preservation, wie Integritätstests, Datensicherheit, Datenreplikation, Tape-Backup. Service-Level-Agreements sichern solche Vorleistungen ab. Aufgabe des FDZ bleiben alle vorbereitenden Maßnahmen, die Überwachung der Dienstleister und deren Risikomanagement.

2. **Managementschicht:** Die Managementschicht stellt die wesentliche Schicht eines Forschungsdatenzentrums dar. Fachwissen, angepasste Workflows, Mittel und Methoden zur Verwaltung der Daten etc. werden auf dieser Schicht organisiert. Das Datenmanagement im Allgemeinen und im speziellen seitens Repositorien stellt die Verlinkung zu Objekten in der Storage-Schicht mit den Metadaten, die u.U. direkt im Forschungsdatenzentrum verwaltet werden, her. Zudem liefert es die Relationen zwischen Objekten, erlaubt eine Versionierung von Datensätzen. Je nach Systemarchitektur verknüpft es die jeweiligen Objekte mit unterschiedlichen Darstellungs- und Zugriffsmechanismen (definiert und betrieben in der Zugriffsschicht) oder bettet diese in (existierende) Softwareumgebungen für den funktionalen Zugriff ein. Abhängig von Forschungsprojekt und -kontext können sich die Anforderungen an diese Schicht unterscheiden und erfordern zusätzliche Funktionalität, wie die Abbildung von Zugriffsrechten (Identity Management) und Schutz der Daten.

3. **Zugriffsschicht:** Auf Grund der heterogener Daten und Benutzergruppen lassen sich Zugriffsstrategien und -angebote sowie deren Schnittstellen schlecht komplett vorausplanen. Hier

sollte ein iterativer Ansatz genutzt werden, der auf Best-Practice-Erfahrungen beruht, die beispielsweise durch die engere Begleitung von verschiedenen (laufenden) Forschungsprojekten erhoben werden. Optimalerweise unterscheiden sich diese Projekte sowohl in der Disziplin als auch im Umfang und den verschiedenen Formaten der Daten. Mit der Datenerhebung, -bearbeitung und -auswertung sind typischerweise Softwareprodukte verbunden, die oft proprietär sind, jedoch als Standard in der Teildisziplin angesehen werden. Hiermit können Risiken für den späteren Zugriff verbunden sein, da einerseits die Software selbst und entsprechende Lizenzen als auch eine geeignete Ablaufumgebung vorhanden sein müssen.

Umsetzung als 3-Tier Architektur

Das geplante FDZ kann als 3-Schicht-Architektur geplant werden, bestehend aus getrennten Zugangs-, Management- und Bitstream-Preservation-Schicht. Alle drei Schichten kommunizieren mit der darüber- und ggf. mit der darunterliegenden Schicht über definierte Schnittstellen, so dass alle drei Schichten als organisatorisch und technisch eigenständige Einheiten betrieben werden können. Dadurch werden auch die Voraussetzungen für ein Out-Sourcing an spezialisierte Dienstleister, beispielsweise der Bit-Stream-Preservation, geschaffen. Insbesondere die Zugangsschicht wird als eigenständiger Geschäftsbereich angesehen und geht über die Bereitstellung einfacher Zugangsarten und Beratung hinaus. Dies setzt eine permanente Anpassung an die Anforderungen potentieller Datenkonsumenten voraus. Aus Anforderung 5 (Portale) und 6 (Web-Services) folgt bei Betreiben und Pflege einer Zugangsschicht ein nicht unerheblicher Arbeitsaufwand, der über Definition und Bereitstellung von Schnittstellen hinausgeht. Im Gegensatz dazu ist die Ausgliederung der Speicherebene organisatorisch recht unproblematisch. Weder Datenkonsumenten noch Datenproduzenten interagieren typischerweise direkt mit dieser Schicht, so dass Anbieterwechsel, Anpassungen der Schnittstellen, veränderte technische Anforderungen lediglich in einem kleinen Kreis verhandelt werden müssen.

Die Vorteile einer vollen 3-Schicht-Architektur liegen insbesondere in der Organisation. So sind alle Bereiche unter Kontrolle. Dienstleistungen werden aus einer Hand angeboten. Dies ermöglicht einheitliche Qualitätssicherung und zentrales Rechtemanagement (IdM). Eine solche Lösung birgt aber auch die Gefahr einer langfristigen Überforderung. Technischer Fortschritt und heterogene Benutzergruppen halten die Zugangsschicht im stetigen Wandel. Für ein effektives Risikomanagement sind alle drei Schichten zu betrachten. Zudem stellt der Verbund nochmals eine zu überwachende Einheit dar.

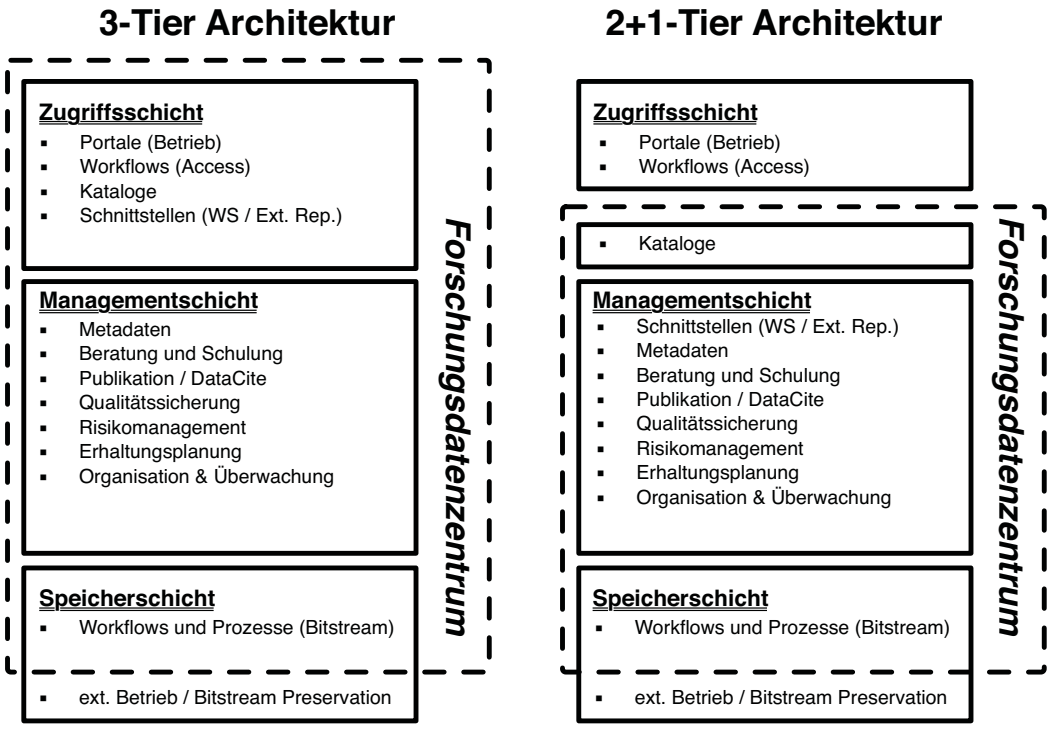

Abbildung 3: Architekturkonzepte FDZ

Umsetzung als 2 + 1-Tier Architektur

Als alternative zu einer klassischen 3-Schicht-Architektur kann eine 2+1-Schicht-Architektur in Erwägung gezogen werden. Dabei wird die Zugangsschicht auf die notwendigen Basiszugänge beschränkt, zum Beispiel ein einfacher Katalogzugriff. Die Definition und Bereitstellung von technischen Schnittstellen wird Teil des organisatorischen Mittelbaus. Durch die Beschränkung auf Schnittstellen und Beratung bzw. Hilfestellung bei der Entwicklung von Zugängen können angesichts des schnellen technischen Wandels und angesichts der heterogenen Anforderungen der möglichen Datenkonsumenten erheblich Ressourcen eingespart werden. Der Aufbau und die Entwicklung von Portalen kann durch Projekte und nach Bedarf finanziert werden. Der Betrieb und die Pflege kann durch die Fachcommunity erfolgen. Dadurch ergeben sich erhebliche organisatorische Vorteile, u.a. offene Nachnutzungsszenarien (auch im Sinne von OpenData Initiativen) und Konzentration auf die wesentliche Aufgaben (Management und Koordination von Daten und Datenströmen sowie deren Nutzern), aber auch Nachteile durch ggf. höhere Hürden für die Nachnutzung sowie ein komplexeres Rechtemanagement.

Organisationsstruktur

Die Architektur des Forschungsdatenzentrums sollte über die rein technologischen Aspekte hinaus weitere Überlegungen, beispielsweise zu Strategie und Management, möglichen Geschäfts- und Finanzierungsmodellen, anstellen, insbesondere zu Aufbau und Entwicklung eines Forschungsdatenzentrums. Die verschiedenen Datenproduzenten und -nutzer der Altertumswissenschaften spielen auch eine wichtige Rolle. Für externe Dienstleister sind Service-Level-Agreements (SLA) vorzusehen, die das Angebot beschreiben, die die Qualitätskontrollen und die Überwachung im Sinne von Monitoring beinhalten und die die Vereinbarungen für den potentiellen Nachfolgebetrieb im Notfall treffen. Die Anforderungen definieren je nach Einbeziehung von externen Dienstleistern Kriterien in Bezug auf Sicherheit, Verfügbarkeit, Authentizität, Integrität und Nachnutzbarkeit. Alle Vereinbarungen und Verträge zwischen den unterschiedlichen Akteuren über Rechte, Pflichten und Haftung sind geeignet niederzulegen und zu dokumentieren. Hierbei sollte berücksichtigt werden, dass ein stetiger Abgleich der Anforderungen mit dem bestehendem Dienstleistungsangebot erfolgt. Die verschiedenen Workflows müssen auf einer klaren Rollenverteilung mit Festlegung von Verantwortlichkeiten basieren. Zu guter Letzt sind die Erfordernisse bei der Umsetzung durch eine IT-Infrastruktur inklusive langfristiger Technologiestrategie zu bestimmen. Ebenso wichtig erscheint ein wissenschaftlicher Beirat, der ein fachgerechtes Dienstleistungsangebot sichert.

Softwarelösungen

Optimalerweise können Schnittstellen für die Datenabgabe so gebaut werden, dass sie sowohl kein Spezialpersonal im Forschungsprojekt benötigen, das die Daten liefert, als auch keine spezialisierten Personen auf Seiten des FDZ erfordert, die die Daten sinnvoll nutzbar einlagern. Hier wäre eine Standardisierung von Übergabeprozessen hilfreich. An dieser Stelle ist zu überlegen, ob es sich aus Beratersicht des FDZ lohnt in bestimmte Standardverfahren zu investieren.

Eine geringe Einschränkung von erlaubten Formaten geht mit einer höheren Flexibilität und Akzeptanz einher, kann aber die spätere Behandlung und den Zugang komplexer machen. Alle Softwarelösungen (unabhängig auf welcher Ebene sie angesiedelt sind) sollten keine externen Abhängigkeiten (z.B. Client-seitig, bei Treibern oder anderen OS-Anforderungen, etc) erzeugen. Offene OS-neutrale Lösungen sind notwendig. Hier sollen nur exemplarisch die heute empfehlenswerten Ansätze aufgezeigt werden:

Fedora

Die Universitäten University of Virginia und die Cornell University entwarfen die Software Fedora (Flexible Extensible Digital Object and Repository Architecture)[76] zunächst als Repository-System. Es handelt sich dabei eher um ein erweiterbares Basis-Framework als eine direkt einsetzbare Komplettsoftware. Mit Fedora können alle notwendigen Dienste bereitgestellt werden, mit denen sich ein an die institutionellen Bedürfnisse angepasstes Repository entwickeln lässt. Das in Java umgesetzte Framework ist als Serviceorientierte

[76] Siehe hierzu: http://www.fedora-commons.org, (zuletzt aufgerufen 29.6.2013)

Architektur (SOA) angelegt, die sowohl Flexibilität als auch Modularisierbarkeit erlaubt. Fedora bildet einzelne digitale Objekte im Repository mittels RDF (Resource Description Framework) ab und nutzt für Import und Export das Metadatenframework METS und ein eigenes entwickeltes Format, Fedora Object XML (FOXML). Mit einigen zentralen Webservices, die mittels SOAP erreicht werden können, stellt Fedora lediglich Basisdienste für den Objektzugriff und Operationen auf Objekten bereit. Weitere Komponenten müssen auf dieser Basis selbst implementiert werden.

DSpace

Bei DSpace[77] handelt es sich um ein ursprünglich am MIT (Massachusetts Institute of Technology) in Kooperation mit Hewlett-Packard entwickeltes Open-Source-System, das inzwischen durch die DSpace-Foundation aktuell fortgeführt wird. In der Foundation sind große Forschungseinrichtungen zusammengeschlossen, die die Weiterentwicklung koordinieren. DSpace ist ein Java-Framework mit Nutzermanagement, welches rollenbasierte Publikations-Workflows unterstützt. Das System erlaubt für Objekte, die Festlegung von Autorisierung, Urheber- und Verwertungsrechten durch nutzergruppenspezifische Anpassungen.

Ausbildung, Beratung und Öffentlichkeitsarbeit

Neben der aktiven Hilfestellung bei laufenden Projekten sollen aber auch Beratungsleistungen und Dienstleistungen für Forscher und Forschergruppen entwickelt werden, die den vollständigen Zeitraum eines Forschungsvorhabens umfassen und sich je nach Art des Projekts auch auf die Antragsphase ausdehnen kann. Beratungsleistungen umfassen beispielsweise die Beantwortung von Fragen nach einem Data-Management-Plan und helfen, einen solchen Plan konform zu den Regeln der jeweiligen Förderinstitution aufzustellen und ihn in den Antragstext einzubringen. Hierzu zählt ebenfalls die Information der Antragsteller über bereits verfügbare Datenquellen, die das beantragte Vorhaben überhaupt ermöglichen bzw. ergänzen können. Weitere Schwerpunkte wären die Unterstützung bei Fragen für die Datenübergabe:

- Wie soll generell vorgegangen werden, worauf ist dabei zu achten?
- Wie sieht ein geeignetes Rechtemanagement aus und wie wird sowohl Open Access als auch den Dateneigentümern Rechnung getragen?
- Wie und wo werden die Daten mittel- und langfristig gesichert? Wer bekommt Zugriff darauf? Wie kann im Nachgang entschieden werden, wer Zugriff bekommt?
- Festlegen der Aufbewahrungsdauer und erneute Entscheidung nach Ablauf dieser.

Ein zukünftiges Kompetenzzentrum wäre mit der Erfassung und Koordinierung von fachspezifischen Richtlinien und Best-Practices befasst und würde Expertisen für den Umgang mit Daten und Definitionen von Standards entwickeln. Gleichzeitig sollte es sich um die Erprobung und Bereitstellung von fachspezifischen und nicht fachspezifischen Tools kümmern.

Neue Forschungsmethoden und Techniken erfordern neue Fähigkeiten sowohl bei den Forschern selbst als auch für die mit dem Forschungsdatenmanagement befassten Personen. Bisher differenziert die Ausbildung zu bibliothekarischen, dokumentarischen und archivarischen Tätigkeiten stark in voneinander abgegrenzte Bereiche, obwohl sich zunehmende Überschneidungen bei Fragen um Erschließung, Bewertung, Erhaltung, Nutzbarmachung und Vermittlung von Wissen ergeben. Zunehmend kommen Objekte in den Altertumswissenschaften sowohl in der analogen als auch digitalen Domäne vor. Bisher hat in Deutschland eine integrative Ausbildung (Bibliothekare, Archivare, Dokumentare) immer noch Modellcharakter (Modell der Fachhochschule Potsdam). Zudem finden Themen wie digitale und funktionale Langzeitarchivierung erst langsam Einzug in die Kerncurricula einschlägiger Studiengänge.

[77] Siehe hierzu: http://www.dspace.org, (zuletzt aufgerufen 29.6.2013)

Die Rolle von Data L**ibrariens** unterscheidet sich zunehmend von den Aufgaben der klassischen Bibliothekare, die erst mit den fertigen Publikationen in Berührung kamen. Die Aufgaben verschieben sich zunehmend nach „vorne“, da sie Dienstleistungen im Forschungsprozess selbst erbringen und nicht erst später Forschungsergebnisse verwalten. Ein Paradigmenwechsel ist notwendig, der die Sichtweise eines Dienstleisters der Publikationsphase auf eine Dienstleistung und Partnerschaft im gesamten Forschungsprozess verschiebt. Aus Nutzerperspektive wird weiterhin die bibliothekarische Expertise zur geeigneten Aufbereitung und Erschließung von Publikationsdaten benötigt, die dafür sorgt, dass die Inhalte über Suchmaschinen und Bibliothekskataloge optimal auffindbar sind. Ähnliche Entwicklungen lassen sich im Bereich archivarischer Tätigkeiten beobachten, die zunehmend mit digitalen Daten und ihren Verarbeitungssystemen in Berührung kommen. Informationstätigkeit setzen nun nicht mehr erst nach Abschluss des Verwaltungshandelns ein, sondern beginnen bereits bei der Entstehung von archivierbaren Strukturen. Mehr auf der technischen Seite arbeitet der Data Manager, der sich um Probleme wie die Aufbewahrung und den Zugriff auf Daten kümmert. Dieses Feld kann sowohl von IT-Spezialisten als auch von IT-affinen Informationswissenschaftlern bearbeitet werden. In jedem Fall ist eine enge Kooperation mit den Fachwissenschaftlern notwendig, wenn es um die Auswahl und Zurverfügungstellung von IT-Komponenten für Datenhaltung, Zugriff oder virtuelle Forschungsumgebungen geht.

Schlussfolgerungen und Empfehlungen

1. Der wesentlichste Erfolgsfaktor eines neuen, insbesondere zentralen FDZ sind gute Kenntnisse der Anforderungen der möglichen Nutzer.

2. Die größten Herausforderungen liegen in der organisatorischen und strukturellen Entwicklung und nicht in der technischen Problemstellung, was sich aus Ziffer 1 ergibt. Dies sollte sich in den Arbeitsplänen und auch in der mittelfristigen Budgetplanung widerspiegeln. Hierzu gehören insbesondere die frühzeitige Nutzereinbindung sowie die Beratung, die Qualitätssicherung und das Risikomanagement.

3. Nicht nur die Erhaltung der Daten, sondern auch Kontext- und Prozesswissen sind wertvoll (Data-Driven-Science).

4. Nur eine frühzeitige Einbindung von Wissenschaftlern (Daten-Produzenten) bereits im Stadium der Datenerstellung sichert den Zugang zu Kontextwissen und erweitert ggf. damit die Nachnutzungsmöglichkeiten.

5. Ein Schichtenmodell (2/3-Tier-Architektur) führt zu effizienteren Betriebsmodellen, führt aber auch zu einem komplexeren Risikomanagement, da nicht alle Akteure direkt kontrolliert und gesteuert werden können. Alle Ebenen müssen getrennt betrachtet werden, zudem ist der Verbund als Ganzes noch zusätzlich zu betrachten.

6. Eine sinnvolle Trennung der Forschungsdaten und deren archivarischen Meta-Daten ist notwendig, was sich bereits aus Ziffer 3 ergibt. Die Verwendung einer Data-Cite-Strategie reduziert organisatorischen Aufwand und Komplexität der einzusetzenden Software-Lösungen.

7. Eine Auslagerung der Bit-Preservation-Ebene an spezialisierte Dienstleister ist zu empfehlen, da ein Aufbau an lokaler Expertise in diesem Bereich ggf. das FDZ überfordert und für die gegebene Aufgabenstellung zu keinem zusätzlichen Mehrwert führt.

8. Bei der Auslagerung von Aufgaben an externe Dienstleister sind Redundanzen entscheidend. Eine Abhängigkeit von einzelnen Anbietern ist durch vorzeitig definierte Ausstiegs- und Notfallszenarien zu begegnen.

9. Redundanz ist auf allen Ebenen notwendig. Nicht nur auf der Speicher/(Bit-Preservation)-Ebene, sondern auch hinsichtlich Erhaltungs- und Zugriffstrategien sollten redundante/alternative Strategien entwickelt und verfolgt werden.

10. Die funktionale Langzeitarchivierung kann eine alternative und effiziente Strategie für seltene und komplexe Objektarten darstellen und bietet Möglichkeiten, Prozesse und vernetzte Umgebungen zu bewahren.

11. Für das geplante FDZ ist es wahrscheinlich effizienter, technische Dienstleistungen wie beispielsweise die Erhaltungsplanung und den Betrieb von Emulatoren an externe Dienstleister zu delegieren. Funktionale Strategien, unabhängig von der konkreten Ausgestaltung, bedürfen allerdings disziplinspezifisch angepasster Konzepte, die durch das FDZ in Zusammenarbeit mit Datenproduzenten als auch Datenkonsumenten ausgearbeitet werden müssen.

12. Externe, spezialisierte und international organisierte Repositorien sind möglichst zu unterstützen. Sie bieten zusätzliche Redundanz in der Speicherung und ein aktives Risikomanagement bzgl. der Zugriffs- und Erhaltungsstrategien der dort hinterlegten Daten. Eine pro-aktive Unterstützung, beispielsweise durch Bereitstellung von Schnittstellung, Datenspiegelung etc., ist sinnvoll.

13. Eine Open-Data-Strategie (sofern rechtlich möglich) ist ebenso im Sinne des Risikomanagements und der Zugriffssicherung empfehlenswert.

14. Es ist genau zu prüfen, inwieweit die Entwicklung und Bereitstellung einer Zugangsschicht – über einfache Basis-Zugänge hinaus – Aufgabe des zu entwickelnden FDZ sein soll. Selbst spezialisierte FDZ stehen einer heterogenen Nutzergruppe sowie schnell fortschreitendem technischen Wandel und daraus folgend diversen Anforderungen an einen effizienten Zugang zu Forschungsdaten gegenüber.