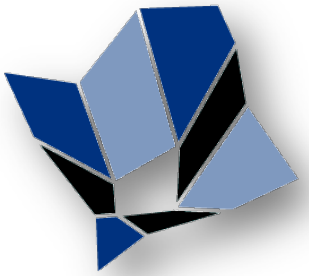

Projekt RADIESCHEN

Rahmenbedingungen einer **di**sziplinübergr**e**ifenden
For**sch**ungsdat**en**infrastruktur

Report „Technik“

Daniela Koudela, Klaus Köhler, Ralph Müller-Pfefferkorn

Inhalt

1 Einleitung

Die stetig wachsende Menge an Forschungsdaten stellt die Datenerzeuger und -nutzer vor umfassende Aufgaben: das Management der Daten muss den gesamten Prozess von der Erzeugung der Daten bis zur Langzeitarchivierung und Veröffentlichung umfassen. Für die Beschreibung dieses Prozesses stützt sich das Projekt Radieschen auf das 4-Domänen-Modell: private Domäne, Gruppendomäne, dauerhafte Domäne und Zugangsdomäne.

In allen diesen Domänen sind technische Infrastrukturen die notwendige Grundlage, um Wissenschaftler beim Datenmanagement zu unterstützen bzw. oftmals in Anbetracht der Datenmengen ein solches überhaupt erst zu ermöglich.

Im Folgenden soll deshalb dargestellt werden, wie sich der Datenworkflow im aktuellen wissenschaftlichen Arbeitsablauf verschiedener Fachrichtungen darstellt und welche funktionalen Anforderungen damit an Technik gestellt werden (Kapitel 2).

Eine Übersicht über den augenblicklichen Stand technischer Lösungen, die im wissenschaftlichen Alltag zum Einsatz kommen bietet Kapitel 3.

Als Grundlage der Analysen dienten zum einen ausführliche Interviews, die im Rahmen des Projektes mit verschiedenen Communities, Projekten und Einrichtungen zum Thema Management und Archivierung von Forschungsdaten durchgeführt wurden. Hinzu kamen weitere meist kürzere Interviews mit weiteren Projekten und Einrichtungen. Generell kann hier gesagt werden, dass der Stand der verschiedenen Communites bei diesem Thema sehr unterschiedlich ist. Während bei manchen Interviewpartnern bereits umfangreiche Hard- und Software-Lösungen im Einsatz sind und damit bereits „Best-Practice“-Erfahrungen vorliegen, befanden sich andere Communites erst in der Planungs- oder Aufbauphase.

Weitere Informationen über existierende technische Lösungen in den verschiedenen Bereichen stammen aus Recherchen in Veröffentlichungen und im Internet.

Natürlich erhebt diese Darlegung nicht den Anspruch auf Vollständigkeit. Dazu ist das Thema zu komplex und die Anzahl der existierenden Systeme zu umfangreich. Sie versucht aber die wichtigsten Aspekte der Technik und ihre Einsatzmöglichkeiten bzw. ihre tatsächliche Verwendung im wissenschaftlichen Alltag zu charakterisieren.

Kapitel 4 charakterisiert einige bereits disziplinunabhängig angebotene oder geplante Dienste im deutschen wie im europäischen Maßstab.

Das Projekt re3data.org hat eine Datenbank über im Internet verfügbare Repositories für Forschungsdaten erstellt. Eine Analyse der Angaben in Bezug auf die eingesetzten technischen Systeme findet sich in Kapitel 5.

Kapitel 6 analysiert Aussagen, Rückmeldungen und Erkenntnisse, die die Projektpartner im Rahmen zweier Workshops/Symposien mit Experten und wissenschaftlichen Anwendern verschiedener Fachgebiete zum Thema Forschungsdaten-Infrastrukturen gesammelt haben.

Schlussendlich wird im letzten Abschnitt versucht, aus den gesammelten Daten und Erkenntnissen Schlussfolgerungen zu ziehen und Empfehlungen zu geben wie die technische Seite des Forschungsdatenmanagements für wissenschaftliche Anwender zukünftig verbessert werden kann.

2 Wissenschaftliche Arbeitsabläufe

Die Interviews wurden in Bezug auf die Technik so ausgewertet, dass erst die Arbeitsschritte, die in den einzelnen Projekten und Einrichtungen beim Datenmanagement anfallen, extrahiert wurden und dann wurde geschaut, welche Technik bei den einzelnen Arbeitsschritten benutzt wird.

Es konnten die folgenden Arbeitsschritte identifiziert werden:

- Datenerfassung
- Prüfungen und Kontrollen, Qualitätssicherung
- Metadatenerzeugung/-ergänzung
- Datenspeicherung
- (Langzeit-) Archivierung
- Zugriff auf die Daten, hier gehört auch die Benutzung von Persistenten Identifikatoren dazu
- Datenanalyse
- spezielle Arbeitsschritte

Dabei ist die Reihenfolge, in der die obigen Arbeitsschritte durchgeführt werden von Projekt zu Projekt und von Einrichtung zu Einrichtung unterschiedlich. Ein Projekt stellt die gespeicherten Daten und Analysemethoden bereit, so dass die Daten direkt im Anschluss auf den Zugriff analysiert werden können. Ein anderes Projekt analysiert die Daten gleich bei der Datenerfassung, so dass auf den Rohdaten sofort weitere, von diesen abgeleitete Daten entstehen.

Fast jedes Projekt hat spezielle Arbeitsschritte, die nur bei diesem Projekt durchgeführt werden, sei es beim Ingest, bei der Analyse der Daten, oder zwischen anderen Schritten. Auch führt nicht jede Einrichtung und jedes Projekt jeden der obigen Schritte aus. Einige Projekte bieten z.B. keine Archivierung der Daten an. Ein Projekt begründet das Auslassen dieses Schrittes damit, dass es auf dem betreffenden Gebiet bereits genügend Datenarchive gibt, bei denen die ErzeugerInnen und NutzerInnen der Daten ihre Daten archivieren können.

Tabelle 1 zeigt die verschiedenen Arbeitsschritte und listet die dazu verwendete Technik auf. Da den interviewten Projekten und Einrichtungen eine vertrauliche Behandlung der Interviews zugesichert wurde, sind die Informationen, wer welche Soft-/Hardware zu welchem Schritt verwendet, anonymisiert. Zudem wurden einige Projekte und Einrichtungen hinzugefügt, die nicht interviewt wurden, da nicht alle Interviews zur Auswertung geeignet waren und die daraus folgende dürftige Datenlage auf diese Weise erweitert wurde. Bei den nicht interviewten Projekten und Einrichtungen wurden die entsprechenden Informationen im Internet recherchiert.

Bei der Auswertung der Interviews zeigte sich, dass die Arbeitsabläufe und Datenworkflows fachübergreifend in eine Reihe von Arbeitsschritten mit bestimmten Funktionalitäten eingeteilt werden können.

Tabelle 1: Übersicht über die in den Projekten angegeben Arbeitsschritte und die dazu benutzte Technik

Arbeitsschritt	Technik	Projekt
Datenerfassung		
Datenerzeugung durch Betreiber	Eigenentwicklung	Einrichtung 1
Automatische Datenerzeugung	Empfang der Daten von einer Vielzahl von Meßstationen im In- und Ausland über ISDN, DSL, Standleitungen oder Satellit	Einrichtung 5

	Mit Hilfe eines Datasubmissiontools oder eines Ticketsystems werden Datenquellen (z. B. Observatorien) direkt abgegriffen	Einrichtung 2
	Daten kommen aus experimenteller Großanlage (Beschleuniger und Detektor)	Projekt 8
Datenerzeugung durch Dritte	Daten werden von Dritten erzeugt und kommen auf Bestellung	Einrichtung 3
	Betreiber bekommt Daten von Wissenschaftlern über Datasubmissionstool oder Ticketsystem zugesendet	Einrichtung 2
Datenübertragungsprotokoll	CD1 Protokoll des Internationalen Datenzentrums (IDC) in Wien sowie SeedLink für GRSN Stationen	Einrichtung 5
Prüfungen und Kontrollen, Qualitätssicherung		
Plausibilitätskontrolle	Manuell	Projekt 5
Prüfung der Klassifikation	Manuell	Projekt 5
Überprüfung, dass Daten und Metadaten zusammenpassen	speziell dafür entwickeltes IT-Tool	Einrichtung 3
	manuell?	Einrichtung 1
Qualitätsprüfung/Qualitätskontrolle/Qualitätsmanagement	langes Skript (Eigenentwicklung)	Projekt 2
	Eigenentwicklung basierend auf dem ARB-Kurationstool	Projekt 5
	Review unter Wissenschaftlern	Einrichtung 2
	automatisch; Eigenentwicklung	Projekt 8
Metadaten		
Metadatenerzeugung/-Metadatenergänzung	manuell mit Eigenentwicklung	Einrichtung 1
	manuell mit Editorentool (Eigenentwicklung)	Einrichtung 2
	Automatische Erstellung mittels eines XQuery-Skripts, die Generierung eines neuen MODS-Objektes wird von einem Cronjob erledigt	Projekt 6
	automatisch – Klassifikation der physikalischen Ereignisse	Projekt 8
Überprüfung der Metadaten auf Vollständigkeit	Abfrage der Datenbank AMS	Einrichtung 3
Speicherung der Metadaten	relationale Datenbank	Projekt 4
	in der nativen XML-Datenbank eXist	Projekt 6
	sowohl in relationaler Datenbank als auch in Dateien	Projekt 8
Metadatenformat	auf Basis von MODS	Projekt 6

	METS/MODS	Projekt 4
	SQL-Tabellen und Dateiformat root (CERN eigene Entwicklung)	Projekt 8
Archivierungsformat für Metadaten	PDF, keine Verwendung von Metadatenstandards, Datenmaterialnr. als Identifikator	Einrichtung 3
Datenspeicherung		
Datenspeicherung	Serverbasierte relationale Datenbank	Projekt 4
	Datenbank: Postgres/PostgreSQL	Projekt 2
	relationale OpenSource Datenbank MySQL	Projekt 5
	erst auf Platte, dann auf Bänder, RAID-Systeme	Projekt 5
	Eigenentwicklung	Einrichtung 2
	OPAC- und EDOC-Server basierend auf OPUS	Projekt 6
	Im modular aufgebauten, digitalen Wissensspeicher, der aus DONATUS Language Technologie, Metadaten, Index (Lucene, Solr), Verarbeitungsmodulen (Textmining, linguistische Methoden, Technologien des Semantischen Webs), einem Übersetzungsservice und einem GUI besteht, sind die Ressourcen in einem Index registriert, aber nicht im Wissensspeicher gespeichert. Im Wissensspeicher sind nur die Metadaten der Ressourcen gespeichert, die Inhalte der Ressourcen sind über URLs erreichbar, die jeweils Teil der Metadaten sind. Ferner ist der Wissensspeicher mit externen Ressourcen, die im XML-Format in irgendwelchen Datenbanken sind, verknüpft.	Projekt 6
	Grid-basierende Speichersysteme mit SRM-Schnittstelle (dCache, Castor, DPM), Managment der Daten/Datensets durch zentrale Oracle-Datenbank; einzelner Nutzer teilweise auch im Dateisystem	Projekt 8
	Datenbank: Oracle 10g	Einrichtung 4
Datenformat	Postgres-Datenformat	Projekt 5
	WKB- oder WKT-Format für georeferenzierte Sequenzdaten	Projekt 3
	Kodierungssystem für Daten: TEI/XML	Projekt 4
	root – am CERN entwickeltes Format	Projekt 8
Anfertigung von Sicherheitskopien	iRODS	Projekt 7
	Snapshot-System: ZFS von Oracle	Projekt 5
	im Grid (WLHC) verteilte Kopien	Projekt 8
(Langzeit-) Archivierung		
Schema für den Export ins Langzeitarchiv	FOXML (internes Schema von Fedora)	Projekt 4

Datenarchivierung	Archivierung der Daten in flachen Formaten (EBCDIC, ASCII oder CSV), Datensatzbeschreibungen in XML, Daten und dazugehörende Metadaten werden zusammen in sogenannten Datenpaketen archiviert, Datenbank: Oracle, Archivierungsmanagementsystem	Einrichtung 3
	Format im Langzeitarchiv: Überführung der Daten, die innerhalb der relationalen Datenbank liegen, in ein TEI-konformes Schema und Übernahme des TEI-konformen Schemas als Datei	Projekt 4
	Format der Daten für die Archivierung: MiniSeed und GSE	Einrichtung 5
	im Grid verteilte Kopien auf Bandsystemen	Projekt 8
	Archivierung auf Festplattensystemen und Langzeitspeichern	Einrichtung 5
Repository-Software	Fedora mit Framework Hydra	Projekt 4
Zugriff		
Datenzugriff	SRM-Schnittstelle zum Datei-Management (z.B. kopieren) im Grid; root-Server, NFS oder lokaler Zugriff zum Lesen der Dateien	Projekt 8
	Zugriff auf Archivdaten: Suchmaschine - Apache Solr, Suchoberfläche - Blacklight	Projekt 4
	Zugriff auf das Archiv über eine Metadatenbank	Einrichtung 5
	Zugriff auch per E-Mail über den AutoDRM-Service (E-Mail-basierte Schnittstelle für die Anforderung von seismischen Wellenform- und Parameterdaten) möglich; in diesem Fall werden die Daten bis zu einem bestimmten Umfang als E-Mail zurückgesendet oder bei größeren Datenmengen über Anonymous-FTP bereitgestellt	Einrichtung 5
Webzugriff auf die Daten	Web-Interface/Webserver: Apache Tomcat	Projekt 2
	Webfrontend: Apache mit PHP, Content-Management-System: Typo3	Projekt 5
	mittels Wissensbrowser (Eigenentwicklung); dieser ermöglicht einen komplexen Suchprozeß mit visualisierten Ergebnissen und ist als typische Client-Server-Architektur gebaut	Projekt 6
	Interface zum Webserver wurde mit Python programmiert	Projekt 1
	GUI ist mit einem auf Python basierenden Django-Framework implementiert, Apache Tomcat Server für Nutzereingaben, Antworten des Servers werden in JSON kodiert, Webseiten für Nutzer sind mit Django-Templates gemacht und sind standardisiertes HTML5 mit CSS und JavaScript; Webschnittstelle zur Datenabfrage: Apache Lucene mit Lucene Abfrage Sprache	Projekt 6

	Java-basiertes Mehrsprachen-Content-Management-System	Projekt 3
	Zugriff auf Datenmanagement-System über Webbrowser	Projekt 8
	Nutzer kann über ein WEB-Formular suchen und auf (Geo)Webdienste wie Web Map Services (WMS) und ein Catalogue Service Web (CSW) zugreifen	Einrichtung 5
Datenformat für die Nachnutzung	Umformatieren der Daten aus dem relationalen System in das, das der Nutzer möchte	Einrichtung 2
	Downloadformat der Daten: ARB, FASTA	Projekt 5
	GSE2.0-Format bei Abfrage über AutoDRM; GSE-Daten als ASCII-Dateien	Einrichtung 5
Verschicken von Festplatten an Nachnutzer der Daten		Projekt 3
Persistent Identifier	DOI	Einrichtung 1, Einrichtung 2
	EPIC/Handle-System	Projekt 7
Analyse		
Datenanalyse	angebotene Analysemethoden basieren auf der Programmiersprache R	Projekt 2
	Die Sequenzdaten werden mit Spezialtools auf clusterbasierten Rechenumgebungen analysiert	Projekt 3
	Computing Grid (WLCG) basierend auf der Middleware gLite; Methoden zum Zugriff auf die Daten sind implementiert; Nutzer entwickelt seine Analyse-Methoden selber; Grundlegende Methoden sind bereits für alle Nutzer implementiert	Projekt 8
	Automatische Bearbeitung der eingehenden Daten durch eine Vielzahl interaktiver Prozesse; ausgewählte Einrichtungen werden über signifikante Ereignisse per SMS und E-Mail informiert	Einrichtung 5
Aufbereitung des plausibilisierten Materials, Auswertung der Statistiken und ggf. Erstellung von Analysen	teils Eigenentwicklung, teils Auswertungstool SAS	Einrichtung 3
Sonstiges		
Datenintegration (Extract, Perform, Load)	Linuxbasierte Bash-Skripte (Eigenentwicklung?)	Projekt 3
Abspeichern der Abstracts aus Pubmed und Speichern der	relationale Datenbank Postgres	Projekt 1

Schlüsselwörter		
Schlüsselwort- und sequenzbasierende Suchen um rRNA-Daten zu extrahieren	Eigenentwicklung	Projekt 5
Verschlagwortung	manuell?	Einrichtung 1
Indexierung	Lucene, Solr, der Import der Informationen für den Index ist mittels eines XML-Parsers, der auf Saxon basiert, realisiert	Projekt 6

3 Technischen Systeme für Forschungsdateninfrastrukturen

In diesem Kapitel werden existierende Software-Technologien zum Management von Forschungsdaten vorgestellt. Diese gliedern sich in fünf Bereiche: Speichertechnologien (Kap. 3.1), Speicherlokalität (Kap. 3.2), Systeme zur Speicherung von Metadaten (Kap. 3.3), Repository-Systeme (Kap. 3.4) und Systeme zur dauerhaften Identifizierung von Forschungsdaten (Kap. 3.5). Die beiden letzteren spielen insbesondere in der Langzeitarchivierung von Daten eine Rolle.

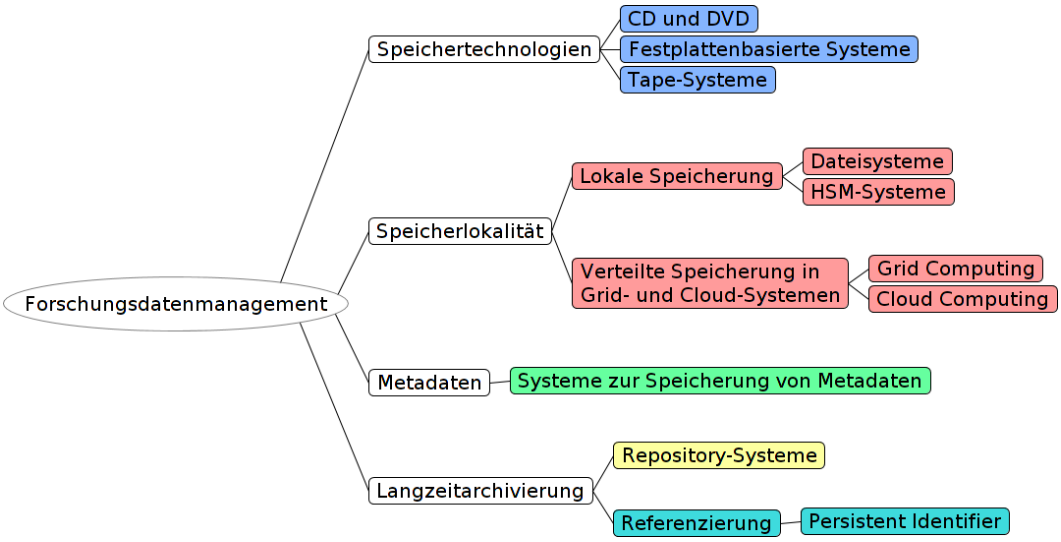

Abbildung 1: Einordnung der Software zum Forschungsdatenmanagement

Kapitel 3.1 beschreibt die verschiedenen Hardware-Systeme zur Speicherung von Daten. Kapitel 3.2, in dem es darum geht, wo die Daten gespeichert werden, untergliedert sich weiter, da man Daten sowohl lokal (Kap.3.2.1), als auch auf verteilten Systemen (Kap. 3.2.2) speichern kann.

Zu allen Technologien werden ein oder mehrere aktuelle Systeme beispielhaft vorgestellt. Dabei ist die Auswahl subjektiv und nicht vollständig. Abbildung 1 verdeutlicht den Zusammenhang der einzelnen Unterkapitel.

3.1 Speichertechnologien

Im gesamten Lebenszyklus digitaler Forschungsdaten muss es jeweils ein Medium geben, auf dem diese gespeichert sind. In diesem Kapitel geht es um die verschiedenen Möglichkeiten, Daten zu speichern. Die folgenden Kapitel (Kap. 3.2 bis Kap. 3.5) behandeln dann die Software, die die einzelnen Bits Dateien und Verzeichnissen zuordnet und somit Zugriff auf die Daten ermöglicht.

3.1.1 CD und DVD

CDs [2][3] und DVDs [3][4] gehören zu den optischen Speichermedien mit Kapazitäten bis zu 900 MB bzw. 8,5 GB. Sie haben den Vorteil, dass die Daten beim Lesen berührungsfrei abgetastet werden und somit keinerlei Abnutzungserscheinungen entstehen. Auch sind sie sehr robust gegen äußere Einflüsse, wohingegen jedoch Temperatureinflüsse starke Auswirkungen auf solche Medien haben können. Des Weiteren sind Zugriffszeiten auf CDs und DVDs relativ groß (50 bis 65 ms).

Anwender:

Im Privatbereich haben sich CDs und DVDs aufgrund ihrer geringen Kosten und einfachen Handhabbarkeit zum Speichern z. B. von digitalen Fotografien etabliert. Auch im wissenschaftlichen Bereich werden deshalb oft optische Medien von Einzelanwendern zur Archivierung eingesetzt (meist als einfaches Kopieren der Daten von Festplatte auf DVD). Dies ist kritisch zu betrachten, da

zum einen die Lebensdauer von CDs und DVDs stark von den Lagerungsbedingungen abhängt. Des Weiteren werden dabei oft die anderen Bedingungen zur Langzeitarchivierung wie Bitstream Preservation oder das Anlegen von Metadaten nicht beachtet. Auch eine Suche auf diesen Medien ist umständlich.

Als Transportmedium für Daten wurden CD/DVDs mittlerweile von USB-Speichersticks verdrängt.

3.1.2 Festplattenbasierte Systeme

Die Speicherung der Forschungsdaten geschieht auf sehr unterschiedlichen Level. Das geht von einer einfachen Speicherung der Daten in einem Dateisystem bis zu einem Archivierungssystem mit Metadatenverwaltung.

Die Speicherung der Daten erfolgt bei großen Datenaufkommen in der Regel in Systemen mit „Tiered Storage“. Dabei werden die Daten je nach Prioritäten (Policies) auf Speichermedien mit unterschiedlichen Zugriffszeiten abgelegt. Prinzip des „Tiered Storage“ ist eine Datenspeicherung auf wenigstens zwei verschieden Klassen von Speichermedien. Man kann heute von den folgenden vier Hauptklassen von Speicher ausgehen

- SSD / Solid State Disk
- SAS Disk (bzw. FC-Disk)
- NLSAS Disk (Nearline SAS, bzw. SATA-Disk)
- Tape

Die Verwaltung der unterschiedlichen Speicherklassen erfolgt in der Regel über allgemeine und nutzerdefinierte Policies.

Eingesetzt werden diese Systeme in vielen Anwendungen. Die meisten freien Archivierungsysteme wie dSpace, Fedora Commons usw. setzen in der jeweiligen Installation auf derartigen Dateisystemen auf.

Ein Nachteil dieser Speichersysteme ist aus Sicht der Archivierung die ungenügende Unterstützung der „Bitstream Preservation“. Dieser Support der „Bitstream Preservation“ kann in der Regel nur durch die darüberliegende Anwendung erfolgen.

3.1.3 Tape-Systeme

Archivierungs-Systeme mit direkter Integration von Tape/Tape-Libraries sind in vielen Fällen auf spezielle Anwendungen bzw. Archivierungsumgebungen zugeschnitten. Diese meist durch den Hersteller und die Zielgruppe geprägten Systeme sind meist nicht universell einsetzbar.

Beispiele

Name	**Hersteller**
Atempo digital archiv	Atempo Deutschland GmbH
Livearc	ARCITECTA PTY LTD

3.2 Speicherlokalität

3.2.1 Lokale Speicherung

Von lokaler Speicherung spricht man, wenn man die Daten auf einem System, z.B. auf einem Rechner, den man in seinem Büro stehen hat oder auf dem institutseigenen Rechencluster speichert. Die folgenden Abschnitte stellen anhand von Dateisystemen, HSM-Systemen, sowie CD und DVD beispielhaft einige Möglichkeiten vor.

3.2.1.1 Dateisysteme

Das Dateisystem organisiert den Zugriff auf die Daten, die auf einer Hardware (meist Festplattensysteme) gespeichert sind. Je nach Hardware und darauf laufendem Betriebssystem gibt es unterschiedliche Dateisysteme. Das Dateisystem erlaubt den Zugriff auf gespeicherte Daten mittels Dateinamen und verwaltet die Information, in welchem Speicherbereich die Daten zu einer bestimmten Datei gespeichert sind. Insofern braucht jeder, der mit digital gespeicherten Daten arbeitet, ein Dateisystem. Ref. [1] bietet eine nach Hardware und Betriebssystem sortierte Liste an Dateisystemen.

3.2.1.2 *HSM-Systeme*

Ein hierarchischen Speichermanagement (kurz HSM) ist ein System, welches Dateien, auf welche über längere Zeit nicht zugegriffen wurde, auf ein Speichermedium mit einer niedrigeren Speicherhierarchiestufe (z. B. Medium mit einer größeren Zugriffszeit) auslagert. Eine solche Hierarchie kann beispielsweise aus Festplatten und Magnetbändern bestehen. Da die niedrigere Hierarchiestufe meist kostengünstiger ist, erlauben HSM den Aufbau großer Speichersysteme, allerdings mit dem Nachteil einer geringeren mittleren Zugriffszeit.

Für die Archivierung sind HSM-Systeme meist nur bedingt geeignet. Dafür sollte die Anzahl der Kopien der Dateien auf Tapes aus Redundanzgründen wenigstens zwei sein oder durch Parameter konfigurierbar sein.

Hauptnachteil der HSM-Systeme aus dem Blickwinkel der Archivierung ist die fehlende Metadatenverwaltung/Repositoryfunktionalität und das Fehlen der „Bitstream Preservation“. Das liegt darin begründet, dass die HSM-Systeme in der Regel von außen (Nutzersicht) als Dateisysteme repräsentiert werden.

Beim Speichern der Daten (Files) werden keine nutzerdefinierten Metadaten erzeugt, bzw. in Datenbanken gespeichert, somit gibt es keinerlei Möglichkeiten einer Recherche in diesen Systemen.

Bei den meisten HSM-Systemen fehlt eine Konsistenzprüfung der gespeicherten Daten. Damit ist die Integrität und Authentizität der Daten nur bedingt gegeben. Die HSM-Systeme speichern in der Regel die Daten in mehreren Kopien auf Tapes. Bei den meisten HSM-Systemen fehlt eine Konsistenzprüfung dieser auf Tape verteilt gespeicherten Daten. Damit ist die Integrität und Authentizität der auf Tape ausgelagerten Daten nur bedingt gegeben. Weitere wichtige Eigenschaften der „Bitstream Preservation“(Tools zur Migration, Diversität der Speichertechnik usw.) sind nur teilweise oder gar nicht realisiert.

Beispiele zu HSM-Systemen: Im universitären Umfeld eingesetzte Systeme sind u. a.

Name	**Hersteller**
DMF	SGI
GPFS/TSM	IBM
SAM/QFS	Oracle/SUN
StorNext	Quantum
HPSS	IBM

3.2.2 Verteilte Speicherung in Grid- und Cloud-Systemen

3.2.2.1 *Grid Computing*

Seit Mitte der 1990er Jahre gerieten verteilte Systeme in den Fokus der Forschung und Entwicklung für die Speicherung großer Datenmengen. Ursachen dafür waren zum einen wissenschaftlicher Natur

– nämlich der rapide Anstieg der Datenmengen in wissenschaftlichen Experimenten (z. B. der Hochenergiephysik oder der Genetik) und Simulationen. Zum anderen wurde es auch für kleinere Einrichtungen möglich sich mit preiswerten, von-der-Stange-Systemen eine Hardware-Basis aufzubauen. Dies führte zur Entwicklung des Grid Computing Paradigmas, bei dem Ressourcen-Anbieter ihre Systeme mit Nutzern und miteinander teilen, um die Systeme besser auszunutzen bzw. größere Datenmengen zu speichern als auf ihrer lokalen Hardware. Dabei entstand eine Reihe von Speicherinfrastrukturen, auf die im Folgenden eingegangen werden soll.

3.2.2.1.1 Storage Resource Management (SRM)

Die Diversität an Speicherlösungen mit unterschiedlichen API im Grid führte zu Standardisierungs-Anstrengungen – zunächst auf Initiative einiger Gruppen, später im Rahmen des Open Grid Forum (früher Global Grid Forum). Es entstand das Storage Resource Management Protokoll [5], vorangetrieben vor allem von der Hochenergiephysik-Community. Im Folgenden implementierten einige Grid-Speicher-Systeme (z. B. dCache, DPM, StoRM oder Castor aber auch in kommerziellen Produkten) dieses Protokoll zum Zugriff und der Verwaltung von dateibasierten Speichern. Die Schnittstelle sind dabei Web Services. Die aktuelle Version ist 2.2.

SRM bietet unter anderem folgende Fähigkeiten:

- voneinander unabhängige, dateibasierte Speichersysteme
- dynamische Speicherallokation und -reservierung
- Verwaltung von Ordnern und Zugriffslisten
- Abstraktion des Dateibegriffs (SURL = Site URL)
- Dienste zum Transfer von Daten im Netz
- Third-Party-Transfer

Systeme, die SRM anbieten, werden z. B. im Grid der European Grid Initiative (EGI) eingesetzt und von verschiedenen Communities genutzt. Hauptnutzer ist die Hochenergiephysik.

SRM-Systeme bieten zwar eine Abstraktion des Dateibegriffs, aber keine globalen Dateinamen. Deshalb werden sie meist mit Datei-Katalogen wie z. B. dem LHC File Catalogue (LFC) [6] gekoppelt. Diese übersetzen globale Dateistrukturen (Verzeichnisse, Dateinamen) in Lokalität (Ort der Speicherung) und lokale Namen.

3.2.2.1.2 dCache

dCache [7][8][9] ist ein Storage-Element für das Grid. Es ist ein skalierbares virtuelles Dateisystem, das über verschiedene Grid-typische Protokolle (SRM, GridFTP, xrootd) und weitere Standards (NFSv3, NFSv4.1, WebDAV, HTTP(s)) sowie ein dCache-eigenes Protokoll ((gsi-)dcap) ansprechbar ist. Es ist jedoch kein Archivsystem. dCache wird genutzt, um unter anderem die Datenmengen, die an den Beschleunigeranlagen der Hochenergie-Physik, wie z. B. dem Large Hadron Collider (LHC), anfallen, zu speichern. Die Geschichte von dCache begann im Jahr 1999 als sich das Fermi National Accelerator Laboratory (FERMI) und das Deutsche Elektronen Synchrotron (DESY) auf ein gemeinsames Projekt einigten, um den Zugriff auf Datenspeicher im Grid zu verbessern, zu vereinheitlichen und zu vereinfachen mit dem Ziel, riesige Datenmengen auf eine große Anzahl heterogener Server zu verteilen.

dCache ist java-basiert und aus Modulen aufgebaut. Die Nutzer-Authentifikation erfolgt über X.509 Grid-Zertifikate (GSI) oder Kerberos. dCache organisiert die Verteilung der Daten im Hintergrund über beliebig viele Speicherknoten. Um Datenverlust vorzubeugen, werden von jeder Datei automatisch mehrere Kopien an verschiedenen Orten gespeichert. Dem Nutzer gegenüber erscheinen die verteilt gespeicherten Daten als wären sie in einem einzigen Dateisystem gespeichert.

Ein solches Dateisystem nennt man virtuelles Dateisystem. Das Mapping zwischen Pfad/Dateinamen im virtuellen Dateisystem und dem physischen Speicherplatz wird für den Nutzer transparent über eine Datenbank gemacht, in welcher auch die Systemmetadaten wie z. B. das Datum, an dem die Datei angelegt wurde, oder dCache-spezifische Metadaten (z. B in welchem Stadium sich die Datei befindet oder welchen Status sie hat) gespeichert sind.

Es steht eine Vielzahl der Standard-Zugriffsprotokollen zur Verfügung (z.B. SRM, NFS4, eigene Klienten). Anhand seiner Konfiguration entscheidet dCache autonom, welche Daten an welche Server geschickt werden abhängig von CPU-Auslastung, verfügbarem Speicherplatz usw. (Load Balancing). Optional können Daten auf ein angeschlossenes Bandsystem geschrieben werden, so dass sie, wenn sie schon lange nicht mehr verwendet wurden, von den Festplatten gelöscht werden um Platz für neue Daten zu schaffen.

Anwender:

dCache ist das am stärksten genutzte Storage-Element im Worldwide LHC Computing Grid (WLCG). Fast 50 % der Daten des WLCG sind in dCache-Instanzen gespeichert. dCache wird an 7 Tier-I und 40 Tier-II-Zentren im WLCG eingesetzt (Stand 2012), mit derzeit insgesamt ca. 94 PB gespeicherter Daten. Weiterhin wird dCache auch außerhalb des WLCG eingesetzt. Abbildung 2 zeigt die Anzahl der einzelnen dCache-Instanzen (linkes Bild) und der einzelnen Sites, die dCache installiert haben (rechtes Bild) im Vergleich zu anderen Systemen, die ähnlich zu dCache sind. Ein paar der in der Abbildung genannten Systeme sind im Folgenden kurz aufgeführt.

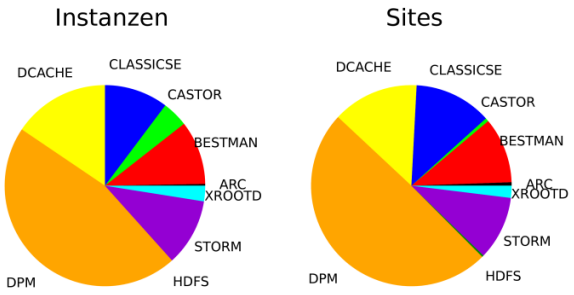

Abbildung 2 Nutzung der verschiedenen Grid-Speichersysteme im WLCG

3.2.2.1.3 StoRM

Eine Alternative zu dCache ist StoRM (STOrage Resource Manager)[11]. Dieses Programm ist jedoch für kleinere Installationen gedacht und unterstützt weniger Funktionen als dCache.

Anwender:

StoRM wird vor allem in Italien benutzt, aber auch in anderen europäischen Ländern gibt es StoRM-Installationen (siehe [12]).

3.2.2.1.4 DPM

Auch der Disk Pool Manager (DPM) ist eine Speicherlösung für das Grid [13]. Genauso wie StoRM ist diese leichtgewichtige Speicherlösung für kleinere Installationen gedacht und unterstützt weniger Funktionen als dCache.

3.2.2.1.5 CASTOR

Der CERN Advanced STORage manager (CASTOR) [14] ist ein Datenspeicher des CERN (Europäisches Labor für Teilchenphysik, ehemals Conseil Européen pour la Recherche Nucléaire) und erledigt als hierarchisches Speichermanagementsystem vergleichbare Aufgaben wie dCache. Daten werden (üblicherweise nach 24 bzw. 8 Stunden) auf Bänder kopiert.

3.2.2.1.6 iRODS – integrated Rule-Oriented Data System

iRODS [15][16] ist eine Open Source Grid-Middleware, um verteilte Forschungsdaten von Kollaborationen zu verwalten, sowie Zugriff auf die Daten und deren Langzeitarchivierung zu ermöglichen. Man kann es auch als Grid-Datenmanagementsystem bezeichnen. Es wird unter einer BSD-Lizenz veröffentlicht und seine Entwicklung erfolgt hauptsächlich vom DICE Center.

Man kann sich iRODS wie ein „Daten-Grid“vorstellen, bei dem man auf Daten zugreifen kann, welche sowohl auf lokalen als auch auf entfernten Rechnern gespeichert sein können. iRODS zeigt diese Daten als vereinheitlichte „virtuelle Sammlung“ und ist somit ein verteiltes Dateisystem mit einheitlichem Namensraum. Der Zugriff kann auf Attributen oder physikalischen Speicherplätzen anstatt auf dem Namen basieren.

iRODS basiert auf einer Client-Server-Architektur und bietet verschiedene Benutzerschnittstellen (grafische und Kommandozeilenklienten, Browser-Interface, Zugriffsprotokolle wie SRM, NFS4, GridFTP, WebDAV), um auf Daten und Metadaten zugreifen und diese managen zu können. Komponenten sind iRODS-Server zur Datenspeicherung, eine Regelmaschine und eine Datenbank, die einen Metadatenkatalog enthält. Es kann Sicherheitskopien erstellen, Daten synchronisieren und archivieren, und so als digitales Datenarchiv verwendet werden.

Der Zugriff auf Dateien erfolgt entweder direkt oder durch eine Suche über Metadaten. Für dSpace und Fedora (siehe Kapitel 3.4) wurde Software zur Integration von iRODS entwickelt.

Der iRODS-Server, auf dem die Daten gespeichert sind, kann vor dem Lesen oder Schreiben der Daten auf diese sogenannte Mikroservices anwenden, die über Regeln gesteuert werden. So können definierte Arbeitsabläufe ausgeführt werden. Ferner stellt iRODS auch eine Prüfungsumgebung für Beurteilungskriterien (audit trails) sowie ein Konsens-Erstellungssystem um Kollaborationen aufzubauen, zur Verfügung. Dabei bezieht sich der Konsens unter anderem auf Policies, Datenformate und verteilte Sammlungen.

In iRODS gibt es zwei verschiedene Sorten Metadaten. Das sind zum einen Systemmetadaten, zum anderen benutzerdefinierte Metadaten. Letztere sind vor allem Schlüssel-Wert-Einheit-Tripel und Erläuterungen, die als relationale oder XML-Metadaten in domänen-spezifischen Schemen (Dublin Core, Darwin Core, FITS, DICOM gespeichert sind.

Anwender:

iRODS wird weltweit eingesetzt, wobei die Anzahl der Nutzer wächst. Unter anderem wird iRODS von mehreren Datengrids in Petabytegröße verwendet, die damit ihre verteilten Sammlungen verwalten. Beispiele hierfür sind das National Optical Astronomy Observatories Data Grid (NOAO) und das Cyber Square Kilometer Array (CyberSKA), beide aus der Astronomie, aus der Hochenergiephysik das BaBar high energy physics data grid und KEK, im Bereich der Erdsysteme der NASA Center for Climate Simulation, im Bereich der Genomforschung das UNC-CHIRENCI und als nationales Datengrid der Australian Research Collaboration Service. Ferner gibt es eine Reihe von institutionellen Repositories und Bibliotheken, die iRODS verwenden, zum Beispiel das Carolina Digital Repository, die French National library, die Texas Digital libraries und der Seismology-Southern California Earthquake Center. Ein Beispiel für ein Archiv, das iRODS benutzt, ist die Ocean Observatories Initiative[18]. In Deutschland wird iRODS z. B. von einer Kollaboration zwischen dem ZIH und dem Max-Planck-Institut für Molekulare Zellbiologie und Genetik angewandt, die sich mit Hochdurchsatzmikroskopie und Hochdurchsatzbildanalyse beschäftigt. Des Weiteren hat sich das EU-Projekt EUDAT auch auf die Benutzung von iRODS geeinigt [19].

3.2.2.1.7 UNICORE

Die Grid-Middleware UNICORE [21] bietet neben Computing-Diensten auch die Möglichkeit Daten im Grid zu speichern. Ein API - das Storage Management System (SMS) - bietet einen einheitlichen Zugriff auf verschiedene Speichersysteme (Dateisystem, iRODS, SRM, Hadoop …). Dadurch sind diese Systeme nahtlos ins Gesamtsystem integriert. Für den Datentransfer stehen verschiedene Protokolle zur Verfügung wie z.B. parallele Transfer-Streams mit uFTP oder http. Metadaten werden als JSON-Dateien bei den eigentlichen Daten mit abgelegt. Mit Apache Tika [22] werden Metadaten automatisch extrahiert. Die OpenSource Volltext-Suchmaschine Apache Lucene [23] wird verwendet, um die Metadaten zu indizieren und danach zu suchen.

Anwender

UNICORE wird weltweit eingesetzt. Ein Beispiel aus Deutschland ist MoSGrid [24], eine Community aus der Chemie. Über ein Portal (Science Gateway) wird Nutzern ein einfacher Zugriff auf mit UNICORE verbundene Ressourcen geschaffen. Daten werden in einem verteilten Dateisystem (XtreemFS), das an UNICORE angebunden ist, gespeichert und dazu Metadaten über chemische Workflows abgelegt und indexiert.

3.2.2.2 Cloud Computing

Das Cloud Computing kann erst auf eine kürzere Historie zurückblicken. Ab etwa 2006 begannen große Firmen ihre riesigen Hardware-Ressourcen, die sie zum Abfangen von Spitzenlasten aufgebaut hatten, an andere Konsumenten zu „vermieten“. Dazu schufen sie eine meist auf Virtualisierung basierte Software-Infrastruktur, die einen einfachen und skalierbaren Zugriff auf die Ressourcen ermöglicht. Gekoppelt mit Abrechnungsmechanismen hat sich die Speicherung von Daten in der Cloud aufgrund des einfachen und oft preiswerten Modells sowie der zunehmenden Vernetzung und Mobilität der Konsumenten weit verbreitet.

3.2.2.2.1 Amazon Web Services

Amazon bietet Nutzern verschiedene Speicherdienste in ihrer Cloud an [20]. Auf diese Speicher kann man entweder weltweit zugreifen und/oder sie können innerhalb der anderen Amazon Web Services genutzt werden. Der bekannteste Dienst ist der Simple Storage Service S3 mit dem Dateien gespeichert werden können. Weiterhin kann man mit der SimpleDB eine einfache verteilte relationale Datenbank aufbauen, der Amazon Relational Database Service RDS bietet eine auf einer virtuellen Amazon-Computing-Instanz laufende relationale Datenbank an. Alle Dienste werden auf einer Pay-per-Use Basis abgerechnet. Zum Management der Dienste können Web-Schnittstelle und APIs genutzt werden.

In S3 können die Nutzer Dateien mit einer Größe von 1 Byte bis 5 Terabyte schreiben und lesen. Diese Daten können in unterschiedlichen Zonen gespeichert werden (z. B. Europa oder USA). Dabei sind dem Speicherplatz keine Grenzen gesetzt. Das Speichern geschieht redundant, selbst wenn man nur eine Zone zum Speichern auswählt. Der Nutzer kann seine Daten als privat oder öffentlich kennzeichnen, Access Control Lists erlauben es die Zugriffsrechte feingranularer zu definieren (z. B. für Gruppen). Eine Datei wird als ein sogenannter Bucket abgelegt, auf den über eine global eindeutige ID mittels http, SOAP, REST oder BitTorrent-Schnittstellen zugegriffen werden kann. Abgerechnet wird die Speichermenge, die Anzahl der Zugriffe und die Menge der gelesenen Daten.

Anwender:

Konkrete Informationen über Anwender der Amazon Web Services aus dem wissenschaftlichen Bereich liegen nicht vor. Aufgrund der einfachen Zugänglichkeit und Abrechenbarkeit der Dienste kann jedoch davon ausgegangen werden, dass sie auch genutzt werden.

3.2.2.2.2 Dropbox

Der kommerzielle Dienst Dropbox [21] ist ein Datenaustausch-Dienst zwischen Benutzern und Rechnern. Daten von Anwendern werden über Online-Speicher zwischen verschiedenen Rechnern eines Nutzers synchronisiert oder können mit anderen Anwendern ausgetauscht werden. Dropbox nutzt dazu die Amazon-Datendienste. Klienten existieren für verschiedene Betriebssysteme. Auch per Webbrowser ist ein Zugriff möglich. Nutzern steht 2 GByte Speicher kostenfrei zur Verfügung, weiterer Speicherplatz kann gekauft werden.

Anwender:

Dropbox wird sowohl im privaten Bereich als auch in der Arbeitswelt angewendet. Die einfache Handhabung und Integration in die verschiedenen Betriebssysteme haben den Dienst sehr populär gemacht.

3.2.2.2.3 Weitere Cloud-Dienste

Durch den Erfolg der Amazon-Dienste oder Dropbox folgten viele weitere Anbieter im kommerziellen Bereich diesem Modell der Speicherung von Daten in einem Datencenter und dem ortsunabhängigen Zugang für die Nutzer. Mittlerweile existiert eine große Menge an ähnlichen Angeboten weltweit, wie z.B. in Deutschland von TeamDrive, PowerFolder oder T-Systems.

Im OpenSource- Bereich wurden einige Software-Pakete entwickelt, die ähnliche Funktionalität liefern aber nicht von einer Firma betrieben werden, sondern die von Anwendern oder Institutionen installiert und betrieben werden können. Der bekannteste Vertreter ist dabei OwnCloud [22]. Es bietet einen dropboxartigen Dienst für ortsunabhängigen Speicher auch zum Austausch mit anderen Nutzern und mit Datenverschlüsselung. Eine Integration in die verschiedenen (auch mobilen) Betriebssysteme ist vorhanden.

3.3 Systeme zur Speicherung von Metadaten

Will man Daten archivieren, so muss man diese mit Metadaten versehen, damit man zum einen auch noch in mehreren Jahren weiß, was diese Daten bedeuten, und zum anderen, damit auch andere Personen als der Wissenschaftler, der diese Daten erzeugt hat, wissen, wie die Daten entstanden sind und was sie bedeuten. Man unterscheidet technische-administrative und inhaltliche Metadaten. Für beide Aspekte gibt es eine Reihe von Standards. Während erstere meist disziplinunabhängig sind, reflektieren letztere den fachlichen Inhalt und sind somit fachspezifisch.

Im Folgenden werden einige gebräuchliche Systeme vorgestellt, die explizit zur Metadaten-speicherung dienen.

3.3.1 Dateibasierte Speicherung

Viele Anwender legen ihre Metadaten direkt mit den Daten in Dateien ab. Entweder werden sie direkt in die Daten-Dateien eingebettet, als Beschreibungsdatei dazugelegt oder z.B. in Verzeichnisstrukturen kodiert. Typische Formate sind dabei ASCII, XML oder RDF. Auch einige Systeme wie z.B. UNICORE (siehe Abschnitt 3.2.2.2) speichern sie in Dateien.

3.3.2 Datenbanken

In vielen Wissenschaftsbereichen werden Metadaten in (meist relationalen) Datenbanken abgelegt inklusive eines Verweises auf die eigentlichen Daten. Dadurch wird eine schnelle Suche auf den Metadaten ermöglicht. Leistungsfähige relationale Datenbanken sind kostenfrei erhältlich.

Gelegentlich werden fach- oder gruppenspezifische Schemata verwendet. Meist sind jedoch eigene Definitionen im Einsatz.

3.3.3 ICAT

Der information catalogue ICAT [21] unterstützt das Management verschiedenster Metadaten von der Erzeugung der Daten bis zur Publikation mit einem Schwerpunkt auf Datenkollektionen von Einrichtungen. Er basiert auf einer relationalen Datenbank, einem API und Metadaten-Schematas. Daten-Kollektionen können registriert, gesucht oder mit anderen Daten assoziiert werden.

3.3.4 Stellaris

Stellaris ist ein Metadaten-Management-Service, der innerhalb des AstroGrid-D-Projektes entwickelt wurde [22]. Es wurde speziell für die Benutzung im Grid entwickelt und ermöglicht ein flexibles Metadaten-Management für das Grid.

Es bietet eine flexible Lösung um Metadaten, die für e-Science und Grid-Computing relevant sind, zu speichern und abzufragen. Die Daten werden im RDF-Format gespeichert, die Abfragesprache ist SPARQL. Es bietet eine Authentifizierung mittels X.509-Zertifikaten, wobei das Autorisierungssystem gruppenbasiert ist.

3.3.5 Repository-Systeme

Zwar eignen sich Repository-Systeme (siehe Kap. 3.4) auch zur Verwaltung von Metadaten, doch Repository-Systeme kommen erst bei der Archivierung der Daten ins Spiel. Werden Daten nicht gleich bei ihrer Entstehung oder Erzeugung mit Metadaten versehen, so wirkt sich dies nachteilig auf die Archivierbarkeit und insbesondere auf die Nachnutzbarkeit der Daten aus, da wichtige Informationen zur Erzeugung und Bedeutung der Daten bis dahin bereits in Vergessenheit geraten. Aus diesem Grund ist es wichtig, bereits bei der täglichen Forschungsarbeit ein System für die Verwaltung der Metadaten zur Verfügung zu haben.

3.4 Repository-Systeme

Im internationalen Rahmen werden Standards für die Archivierung durch die OAI (Open Archives Initiative [23]) vorangetrieben. Die beiden derzeit laufenden Projekte sind „Protocol for Metadata Harvesting“ (OAI-PMH) und „Object Reuse and Exchange“ (OAI-ORE). In dieser Organisation sind die unten vorgestellten Systeme durch das Systemdesign einer der Hauptträger der Standardisierung.

Ein Repository-System oder auch kurz Repository genannt, ist ein System zur Speicherung und Verwaltung von Metadaten. Gleichzeitig beinhaltet es in der Regel die Funktionalitäten zur Recherche und zum Einbringen der digitalen Objekte. Die Speicherung der digitalen Objekte, auf die sich die Metadaten beziehen, ist ebenso ein Bestandteil dieser Systeme.

Eine ausführliche Beschreibung ist u. a. im „Nestor Handbuch, Kapitel 11.2 Repository Systeme – Archivsoftware zum Herunterladen“ [24] zu finden.

Die Repository-Systeme realisieren in der Regel den Ingest-Prozess, die Speicherung der Daten und die Verwaltung der Metadaten sowie den Zugriff der Nutzer (Recherchefunktionalität). Ein Hauptproblem der Langzeitarchivierung, die Bitstream-Preservation, wird meist nicht durch diese Systeme realisiert.

Auch Bibliotheken benutzen Repository-Systeme um ihre Bestände zu archivieren. In den Bibliotheken werden unterschiedliche Systeme, oft basierend auf Eigenentwicklungen, eingesetzt. Daraus folgt eine Vielzahl von Systemen, die derzeit in den Bibliotheken eingesetzt oder erprobt werden. Das Ziel der Archivierung der Bibliotheksbestände ist durch die Aufgaben der Bibliotheken in der Regel anders gelagert als bei wissenschaftlichen Daten.

Im Folgenden werden einige der derzeit aktuellen Repository-Systeme näher erläutert.

3.4.1 DSpace

DSpace [25] ist konzipiert als Software zum Betrieb eines Dokumentenservers. Sie wird in vielen Fällen zum Betrieb eines „Institutional Repository“ eingesetzt.

Entwicklung DSpace

- Entwickler: Massachusetts Institut of Technology (MIT) / Hewlett-Packard (HP)
- Projektstart 2002 mit Release 1.0
- November 2009: weltweit mehr als 700 Projekte
- Mai 2012: weltweit mehr als 1300 Projekte
- weltweit gibt es viele Service-Provider für DSpace (registriert bei DSPACE)
- Aktuelles Release 1.8.2 (Stable Release/seit Februar 2012)

3.4.2 Fedora Commons

Fedora Commons [26] ist ein Open-Source-Repository, mit dessen Hilfe man digitale Objekte in elektronischen Archiven verwalten und zugänglich machen kann. Dieses System ist kompatibel zu den Anforderungen von OAI.

Dieses System ist eine Java-Implementierung, damit ist es flexibel einsetzbar. Neben der Recherchefunktionalität ist auch eine Zugriffsverwaltung enthalten.

Entwicklung Fedora Commons

- Entwickler: Cornell University / University of Virginia
- Projektstart 2003 mit Release 1.0
- November 2009: weltweit mehr als 160 Projekte
- Mai 2012: weltweit mehr als 300 Projekte
- Mai 2012 Release 3.5 (Stable Release/seit August 2011)

3.4.3 EPrints

Eprints [27] ist eine Open-Source-Software zum Aufbau und zur Verwaltung von Open-Access-Repositories. Diese Repositories sind kompatibel mit dem „Open Archive Information System“ (OAIS).

Einer der Hauptanwender in Deutschland ist die Fraunhofer Gesellschaft, die mit dem „Fraunhofer-ePrints“ das offizielle institutionelle Repositorium der Fraunhofer-Gesellschaft betreibt [28]. Es verzeichnet frei zugängliche Veröffentlichungen der Fraunhofer-Gesellschaft, ihrer Institute sowie deren Mitarbeiterinnen und Mitarbeiter.

Entwicklung EPrints

- Entwickler: Universität Southhampton
- Projektstart 2003 mit Release 1.0
- November 2009: weltweit mehr als 260 Projekte
- Mai 2012: weltweit mehr als 320 Projekte
- Mai 2012 Release 3.3.10 (Stable Release 3.3/seit September 2011)

3.4.4 LOCKSS

LOCKSS (Lots of Copies Keep Stuff Safe) ist als Langzeitarchivierungssystem an der Stanford University, Kalifornien entwickelt worden [30][31]. Das System basiert auf dem OAIS-Modell. Der Grundgedanke ist die verteilte Speicherung von Daten auf verteilten Servern (7 Kopien und eine Masterkopie), wobei das LOCKSS-Netzwerk auf der aktiven Beteiligung der Partner an der Speicherung beruht.

Mit diesem Speicherkonzept stellt LOCKSS ein alternatives Konzept zu der sonst geschlossenen Speicherarchitektur der Repository-Systeme vor.

Um die Anzahl der beteiligten Betreiber der Datenspeicher einzuschränken, wurde das System CLOCKSS (Closed LOCKSS) [32] sowie die „PLNS“ (Private LOCKSS Networks) geschaffen.

3.4.5 UrMEL

Die Universal Multimedia Electronic Library (UrMEL) ist die zentrale Zugangsplattform für multimediale Angebote der Thüringer Universitäts- und Landesbibliothek Jena (ThULB) und weiterer Partner [33]. Basis dieses Systems ist „MyCoRe“ [34], ein Content Repositoriy, auf dessen Grundlage Dokumenten- und Publikationsserver aufgebaut werden können. MyCoRe verwendet Java- und XML/XSL-Technologien und unterstützt für die Verwaltung und Suche in den Inhalten verschiedene Backend-Systeme. Neben Open Source Backends wie MySQL und Apache Lucene werden auch kommerzielle Backend-Systeme wie IBM DB2 unterstützt. Die in den Anwendungen verwalteten Objekte können aus einer oder aus mehreren Dateien oder auch aus ganzen Dateibäumen bestehen und vielfältige Datenformate besitzen. Metadaten werden gemäß Dublin Core verwaltet.

3.4.6 OPUS

OPUS [38] ist eine Open-Source-Software unter der GNU General Public License für den Betrieb von institutionellen Dokumentenservern bzw. Repositorien und dient als solche der Veröffentlichung, Erschließung, Administration, Recherche sowie Verbreitung elektronischer Publikationen mit und ohne Volltext. Die Metadaten zu einem Dokument können während des Veröffentlichungsprozesses über ein Online-Formular eingegeben und von entsprechend berechtigten Personen administriert werden.

OPUS steht für Online Publikationsverbund Universität Stuttgart und wurde dort Ende der 90er Jahre vom Rechenzentrum und der Universitätsbibliothek entwickelt. Mittlerweile erfolgt die Weiterentwicklung unter der Ägide des kooperativen Bibliotheksverbund Berlin-Brandenburg (KOBV) und des Zuse-Instituts Berlin.

3.5 Persistent-Identifier-Systeme für Forschungsdaten

Inhalte, auf die über das Internet zugegriffen werden kann, sind üblicherweise durch eine URL adressiert. Das Problem hierbei ist, dass sich URLs im Laufe der Zeit ändern, wenn Inhalte verschoben werden bzw. URLs sind allgemein nicht mehr gültig, wenn der Inhalt von der ursprünglichen Stelle entfernt wird. Dieses Problem wird durch Persistent-Identifier-Systeme behoben. Für Anwender ähneln Persistent-Identifier einer URL, die in den Browser eingegeben wird, doch intern wird der Persistent-Identifier von einem Auflöse-Service aufgelöst und der Nutzer wird zum aktuellen Speicherplatz der Ressource weitergeleitet. Dabei ist es wichtig zu bemerken, dass sich die Persistenz nicht aus einer technischen Eigenschaft ableitet, sondern eine organisatorische Eigenschaft ist.

Beispiele für Persistent-Identifier sind:

- Digital Object Identifier (DOI) [35]
- Uniform Resource Name (URN) – der Persistent-Identifier der Deutschen Nationalbibliothek [36]
- Persistent Uniform Resource Locator (PURL) [37]

Eine ausführlichere Liste zu verschiedenen Persistent-Identifier-Systemen findet sich in Referenz [38].

Als Beispiel für ein Persistent-Identifier-System soll hier der Digital Object Identifier (DOI) näher besprochen werden.

3.5.1 Digital Object Identifier (DOI)

Der DOI ist ein internationaler ISO Standard. Das System wird von der internationalen DOI-Stiftung, welche ein Konsortium aus kommerziellen und nicht-kommerziellen Mitgliedern ist, geleitet. Die DOIs werden von einer DOI-System-Registrierungsagentur vergeben. Der DOI besteht aus folgenden Komponenten [39]:

- einer speziellen Nummerierungssyntax
- einem Auflöse-Service basierend auf dem CNRI Handle-System
- einem Datenmodell
- Richtlinien und Prozeduren für die Implementierung von DOI-Namen durch ein Bündnis von Registrierungsagenturen

DOIs werden vor allem bei wissenschaftlichen Artikeln benutzt.

4 Disziplinübergreifende Dienste

In Europa und Deutschland werden bereits einige Dienste für Forschungsdaten disziplinübergreifend angeboten. Sie werden hier beispielhaft aufgeführt.

Zum ersten seien hier Universitäts- und Hochschulrechenzentren zu nennen, die oft Datendienste für ihre breite Nutzerschaft aber meist nur einrichtungsintern anbieten. Dies umfasst u. a. Festplattenspeicher, Backup oder zentrale Datenbankserver.

Einige Einrichtungen bieten diese Dienste auch für externe Nutzer an. Zu nennen seien hier beispielsweise das Archivierungsangebot für Forschungsdaten der GWDG Göttingen oder die Datenspeicherdienste der Höchst und Hochleistungsrechenzentren der Gauß-Allianz (meist jedoch nur in Zusammenhang mit HPC-Computing-Diensten). Die Large Scale Data Facility LSDF des KIT und der Universität Heidelberg koppeln ihre Daten-Dienste campusübergreifend.

Das EU-Projekt EUDAT plant Datendienste disziplinübergreifend und europaweit anzubieten. Die Planungen beinhalten u. a. Datenspeicher, Datenreplizierung, schneller Datentransfer, einen zentralen Metadatenkatalog und Authentifizierung und Autorisierung.

Globus-Online [40] ist ein auf einer Grid-Middleware basierender Dienst, der den schnellen und verlässlichen Transfer von Daten über das Internet ermöglicht. Er wird weltweit angeboten.

5 Auswertung der Ergebnisse des Projektes re3data

Das interviewte Projekt re3data konnte nicht für die Auswertung in Tabelle 1 verwendet werden, da es kein Forschungsdatenmanagement in dem Sinne betreibt, dass Forschungsdaten in einer Infrastruktur erfasst, analysiert und/oder archiviert werden. Trotzdem ist dieses Projekt für das Arbeitspaket Technik von Radieschen interessant, denn re3data erstellt eine Registry für Forschungsdatenarchive. Einzelheiten zu den registrierten Archiven kann man unter www.re3data.org recherchieren. Zum Zeitpunkt der Auswertung (1.2.2013) waren 116 Archive bereits von re3data analysiert und mit Einzelheiten online gestellt.

Das Arbeitspaket Technik extrahierte aus [41] den Namen der Repository-Software, ob ein System für Persistente Identifier benutzt wird und wenn ja, welches, sowie die Art des Inhalts des Repositorys (siehe Anhang A). Als Ergebnis ergibt sich folgende Liste verwendeter Software: 1x dLibra, 5x DSpace, 1x EPrints, 34x „other“ und 75x „unknown“. Dabei bezeichnet „other“ eine Repository-Software, die nicht in der Liste CKAN, Digital Commons, DSpace, EPrints, eSciDoc, OPUS, dLibra enthalten ist und „unknown“ eine Software, die unbekannt ist (siehe [42], Seite 21, Tabelle 25.1).

Dieses Ergebnis bestätigt die Ergebnisse der Analyse der Interviews, nämlich dass für die Speicherung und das Management von Forschungsdaten derzeit überwiegend Eigenentwicklungen verwendet werden.

Analysiert man die Daten in Bezug auf Systeme für Persistente Identifier, so fällt auf, dass alle Archive, die DSpace verwenden, Daten über CNRI-handle identifizierbar machen. Allerdings bietet der überwiegende Teil der Archive (90 von 116) keine Persistenten Identifier an.

In Bezug auf den Inhalt der Archive haben alle Repositories, die mit DSpace arbeiten, nur Standard-Office Dokumente, d. h. Text Dokumente, Spreadsheets und Präsentationen (siehe [42], Seite 18, Tabelle 14). Bei den Archiven, die als Repository-Software Eigenentwicklungen benutzen, ist die Variationsbreite der Art des Inhalts am größten. Hier finden sich Konfigurationsdaten, Datenbanken, Bilddateien, Netzwerkbasierte Daten wie E-Mails, Webseiten oder die Historie von Chats, verschieden encodierte *.txt-Dateien, gerätespezifischer Output, wissenschaftliche und statistische

Datenformate, Standard-Office Dokumente, strukturierte Graphiken, sowie strukturierte Textdateien (z. B. XML).

6 Auswertung der Radieschen-Workshops

Das Projekt Radieschen veranstaltete zwei Workshops, die beide am GeoForschungsZentrum Potsdam (GFZ) stattfanden. Der erste Workshop trug den Titel „DFG-Projekt Radieschen – Experten-Workshop: Elemente einer übergreifenden Forschungsdaten-Infrastruktur: Eine für Alle? “, der zweite Workshop nannte sich „Symposium Forschungsdaten-Infrastruktur (FDI 2013)“und wurde von den DFG-Projekten Radieschen, re3data, KomFor, EWIG und BoKeLa gemeinsam organisiert.

Der erste Workshop behandelte Policies und Anreize, die Einbindung der Datenzyklen in den Forschungsprozess, generische versus disziplin-spezifische Dienste, sowie die Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten. Aus den zahlreichen Diskussionen ergaben sich die folgenden Erkenntnisse für das Arbeitspaket Technik (entnommen aus [43]):

- In vielen Bereichen fehlt die für ein geregeltes Forschungsdatenmanagement notwendige Infrastruktur. Infrastrukturen sind bis jetzt nur in den Communities etabliert, die aus unterschiedlichen Gründen großen Druck hatten, Datenstrukturen zu schaffen.
- Infrastrukturen müssen aus der Community heraus entstehen, damit sie genutzt werden.
- Vielen Wissenschaftlern und Wissenschaftlerinnen fehlen die notwendigen Kenntnisse und Fertigkeiten des Datenmanagements.
- Die Erfassung der Metadaten von Forschungsdaten ist noch zu arbeitsaufwendig und die Werkzeuge dazu sind noch zu wenig benutzerfreundlich. Das Ziel muss eine weitgehende Automatisierung der Annotation von Forschungsdaten mit Metadaten sein, die verbleibenden Arbeiten müssen sich nahtlos in den Arbeitsalltag der Forscher einfügen.
- Für Persistent Identifier wird ein Service zur Speicherung und Verwaltung der Metadaten, der Suche in den Metadaten, sowie zur Auflösung der gefundenen Referenz benötigt.
- Der Datentransport ist eine Herausforderung, da die Datenmenge schneller wächst als die Bandbreiten im Internet. → Abwägung zwischen lokaler und zentraler Speicherung.
- Bitstream Preservation ist disziplinübergreifend.
- Daten „inhaltlich lesbar machen“ ist disziplinübergreifend.
- Kuration, Indexierung und Auffindbarkeit der Daten sind disziplinspezifisch.
- Software, die einfach zu bedienen und praktisch ist, wird auch benutzt. Beispiele hierfür sind Skype und Dropbox, die jedoch aus Datenschutzgründen im wissenschaftlichen Bereich meist nicht gestattet sind. Deswegen sollte die Wissenschaft eigene Dienste analog zu Skype und Dropbox entwickeln, um nicht in den Konflikt der Datenschutzproblematik der kommerziellen Dienste zu kommen.
- Stabilität und Standardisierung sind wichtig, dies ist aber noch nicht genug kommuniziert.
- Der Zugang zu Expertise ist wichtig. Deswegen sind lokale Rechenzentren mit Expertise wichtig, da erfahrungsgemäß der lokale Ansatz von Nutzern bevorzugt wird. Das bedeutet jedoch nicht, dass auch die Daten dort liegen müssen.
- Software bzw. Anwendungen müssen verständlich und leicht zu bedienen sein.
- Defizite bestehen im Datenbereich bei Smart Tools und an Knowhow.
- Für manche Probleme fehlen die Dienste, obwohl die Rechenleistung vorhanden ist.
- Infrastrukturen müssen aufgebaut werden.
- Zurzeit existiert eine Vielfalt von Formaten sowohl für große als auch kleine, heterogene als auch homogene Datensätze, welche schwer zu verwalten ist.

Folgende Fragen blieben in Bezug auf die Technik offen (entnommen aus [43]):

- Welche Dienste können auf sinnvolle Weise zentral von disziplinübergreifenden Infrastrukturen angeboten werden und welche Dienste werden besser lokal aufgebaut?
- Wie kann Nachhaltigkeit in der privaten Domäne erreicht werden?
- Was ist der adäquate Nachfolger des Laborbuches, das früher alle Daten fasste?

Auf dem Symposium, das ein knappes Jahr nach dem Experten-Workshop stattfand, wurden Aspekte der Finanzierung, Organisation und Technologie der zu schaffenden Forschungsdaten-Infrastrukturen, sowie deren rechtliche und politische Rahmenbedingungen diskutiert. In Bezug auf die Technik wurden dort folgende Aussagen gemacht (entnommen aus [44][45][46]):

- Werkzeugentwicklung ist wichtig, um die Arbeit von Wissenschaftlerinnen und Wissenschaftlern zu vereinfachen.
- IT-Abteilungen können kleine Dinge tun, um Wissenschaftlerinnen und Wissenschaftlern das Datenmanagement zu erleichtern.
- Derzeit ist in der privaten Domäne das Auffinden von früher datierten Datensets, die Verwaltung der Daten, das Vermeiden von Überschreibungen, sowie die Nicht-Reproduzierbarkeit der Daten aufgrund fehlender oder veralteter Software eine Herausforderung.
- Derzeit ist der überwiegende Teil der Software in Forschungsdateninfrastrukturen selbst entwickelt. Der Grund dafür ist, dass kommerzielle Lösungen häufig so komplex sind, dass man nur 20% davon braucht. Zudem ist die Anpassung der kommerziellen Software so aufwändig, dass schlanke Eigenentwicklungen, die gerade das bieten, was man braucht, bevorzugt werden.
- Eine Forschungsdaten-Infrastruktur ist eigentlich nie fertig.
- Eine systematische und ganzheitliche Erfassung, welche Lösungen es in welchen Bereichen gibt, wäre hilfreich als Handreichung für Einrichtungen die auf der Suche nach geeigneten Lösungen sind. Bis jetzt gibt es hierfür nur Vorarbeiten von einzelnen Bereichen wie z. B. Nestor für Langzeitarchivierung oder CARPET für ePublishing.
- Bis jetzt gibt es keine Software, die als Gesamtlösung alle Stufen des Forschungsprozesses unterstützt.
- Softwarelösungen sollten möglichst alle Phasen des Forschungsprozesses abbilden.
- Bis jetzt gibt es keine Standardlösung für einzelne Bereiche (z. B. Langzeitarchivierung, Repository, VRE, …)
- Forschungsdatenmanagement muss sich an den Arbeitsweisen der Forscherinnen und Forscher orientieren, nicht anders herum. Dafür sollte ein intensiver (evtl. disziplinspezifischer) Dialog zwischen WissenschaftlerInnen und Infrastrukturbetreibern etabliert werden.
- Bei der Konzeption einer eigenen Software-Lösung sollte beachtet werden, dass der wissenschaftliche Workflow vom Beginn (z. B. Datenerhebung) und nicht vom Endprodukt (z. B. Text- oder Daten-Publikation) her gesehen werden sollte. →Lücken bestehen auch noch bei den Datenmanagement- Werkzeugen und bei deren Integration in die Arbeitsabläufe der Wissenschaftler.

7 Schlussfolgerungen

Im Folgenden sollen Erkenntnisse und Schlussfolgerungen dargestellt werden, die sich aus den umfangreichen erhobenen Daten ergeben.

1. Aus der Analyse der Daten-Workflows der verschiedenen Fachgebiete (Kapitel 2) lassen sich eine Reihe von Arbeitsschritten identifizieren, die disziplinunabhängig sind. Daraus können eine Reihe von Funktionalitäten abgeleitet werden, die disziplinübergreifende Werkzeuge zur Verfügung stellen sollten. Allerdings treten nicht alle Arbeitsschritte bei allen Disziplinen auf.
2. Eine der wesentlichen Erkenntnisse aus den Interviews und Recherchen ist die Tatsache, dass die meisten Fachdisziplinen Eigenentwicklungen gegenüber existierenden Systemen bevorzugen. Dies kann jedoch nicht allein aus den individuellen Anforderungen abgeleitet werden, da viele Arbeitsschritte allgemeiner Natur sind und entsprechende Werkzeuge dafür existieren, die an die individuellen Bedürfnisse lediglich angepasst werden müssten. Für die Eigenentwicklungen werden einige generische Werkzeuge als Grundlage benutzt, die allerdings eher auf der unteren technischen Ebene einzuordnen sind, z.B. Datenbanken zum Speichern von Metadaten/Daten oder Webserver und damit das http-Protokoll für die Zugangsdomäne.
3. Überwiegend disziplinspezifisch sind die folgenden Arbeitsschritte. Hier erscheint eine vollständige Unterstützung durch allgemeine Werkzeuge schwierig.
   - Datenerfassung
   - Qualitätskontrolle der Daten
   - Disziplinspezifische Metadaten
4. Für disziplinübergreifende Werkzeuge erscheinen folgende Funktionalitäten geeignet:
   - Datenformate
   - Allgemeine Metadatenerzeugung/-ergänzung (z.B. jhove für Standarddateiformate)
   - Daten- und Metadatenspeicherung
   - Datentransfer
   - Datenreplizierung
   - (Langzeit-)Archivierung
   - Webbasierter Zugriff auf die Daten
5. In den Diskussionen in den von Radieschen veranstalteten Workshops wurden oft aktuelle Dienste kommerzieller Anbieter, die von den Wissenschaftlern im privaten Bereich verwendet werden, als Vorbilder und Anforderungen für die umfangreichen Aufgaben des Forschungsdatenmanagements genannt. Beispiele waren vor allem kollaborative Cloud-Dienste wie Dropbox zur Datenreplizierung und Synchronisation oder zum Datenaustausch. Neben der Funktionalität und einfachen Handhabbarkeit wurde hier die Integration in die (private) Arbeitsumgebung hervorgehoben. Allerdings ist den meisten Wissenschaftlern bewusst, dass letzteres für die verschiedenen wissenschaftlichen Arbeitsumgebungen schwieriger zu erreichen ist und dass die Nutzung kommerzieller Cloud-Dienste Probleme der Datensicherheit, des geistigen Eigentums und Kosten nach sich ziehen. Werkzeuge zum Forschungsdatenmanagement sollten deshalb die Integration in die wissenschaftliche Arbeitsumgebung und eine einfache Handhabbarkeit als eines der Hauptziele definieren.
6. Während Wissenschaftler in der jüngeren Vergangenheit Wert darauf legten ihre Dienste selbst zu betreiben, so hat sich diese Meinung in letzter Zeit geändert. Es besteht die Bereitschaft Dienste anderer in Anspruch zu nehmen. Dieser Wandel erfolgte u.a. durch Erfahrungen mit Diensten aus dem privaten Bereich, aber auch durch die zunehmende Komplexität der Dienste und den damit erhöhten Aufwand des Betriebes.

7. Ein wichtiges Thema im Bereich Technik sind Standards und Schnittstellen. Technologien sind einer stetigen Weiterentwicklung unterworfen und ändern sich. Standards und standardisierte Schnittstellen ermöglichen es einzelne Implementierungen auszutauschen ohne das ganze System verwerfen zu müssen. Dies ist eine wichtige Voraussetzung für komplexe Systeme zum Forschungsdatenmanagement, die eine Vielzahl von Einzelaufgaben erfüllen müssen.
8. Die Referenzierung von Daten mit persistenten Identifikatoren hat sich noch nicht allgemein durchgesetzt. Es existieren hier jedoch technische Systeme mit dem zugehörigen organisatorischen Hintergrund, deren Dienste durch die Anwender einfach nutzbar sind.
9. Ein vielfach geäußerter Punkt zur Nicht-Nutzung von generischen Software-Komponenten war deren Handhabbarkeit (Software-Pakete, Installation, Dokumentation) und Nachhaltigkeit. Die Software muss eine Perspektive bezüglich Support und Pflege bieten.
10. Das Wissen über disziplinübergreifende technische Lösungen zum Forschungsdatenmanagement muss verbreitet werden. In den Symposien und Workshops zeigte sich, dass der fachgebietsübergreifende Austausch zu diesem Thema noch nicht ausreichend ist. Existierende Lösungen waren oft nicht bekannt. Thematische Kompetenzzentren könnten (ähnlich dem Nestor-Netzwerk für Langzeitarchivierung) Wissen konzentrieren und Ansprechpartner für die Nutzer sein.
11. Auf der untersten technischen Ebene werden Daten im Wesentlichen als Dateien oder in Datenbanken gespeichert. Festplattensysteme und Bänder spielen dabei die Hauptrolle. Optische Medien sind kaum im Einsatz.

8 Literatur

[1] Wikipedia: Dateisysteme, URL: http://de.wikipedia.org/wiki/Liste_von_Dateisystemen

[2] Wikipedia: Compact Disc, URL: http://de.wikipedia.org/wiki/Compact_Disc

[3] Nestor-Handbuch Kapitel 10, URL: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010062493

[4] Wikipedia: DVD, URL: http://de.wikipedia.org/wiki/DVD

[5] Storage Resource Management Working Group: SRM v2.2 Specification, September 2009, URL: https://sdm.lbl.gov/srm-wg/ (zugegriffen im Juli 2012)

[6] Baud, J.-P.; Casey, J.; Lemaitre, S.; Nicholson, C.; Performance analysis of a file catalog for the LHC computing grid, High Performance Distributed Computing, 2005. HPDC-14. Proceedings. 14th IEEE International Symposium on , vol., no., pp. 91- 99, 24-27 July 2005, doi: 10.1109/HPDC.2005.1520941

[7] Deutsches Elektronen Synchrotron Hamburg: dCache, Webseite, URL: http://www.dcache.org/

[8] Deutsches Elektronen Synchrotron Hamburg: dCache summary, URL: http://www-dcache.desy.de/summaryIndex.html

[9] Deutsches Elektronen Synchrotron Hamburg: dCache White Paper, URL: http://www.dcache.org/manuals/dcache-whitepaper-light.pdf

[10] WLCG REsource, Balance & USage, URL: http://gstat-wlcg.cern.ch/apps/gt/dm/ (zugegriffen im Sept. 2012)

[11] INFN – CNAF, EGRID – ICTP: Storage Resource Manager, URL: http://storm.forge.cnaf.infn.it/home (zugegriffen Sept. 2012)

[12] INFN – CNAF, EGRID – ICTP: Storm, URL: http://storm.forge.cnaf.infn.it/storm_maps.html (zugegriffen Sept. 2012)

[13] CERN: Disk Pool Manager DPM, URL: https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm (zugegriffen im Sept. 2012)
[14] CERN: CASTOR, URL: http://castor.web.cern.ch/ (zugegriffen im Sept. 2012)
[15] iRODS, Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems, URL: https://irods.org (zugegriffen Sept. 2012)
[16] iRODS overview, URL: https://www.irods.org/pubs/iRODS_Overview_0903.pdf (zugegriffen im Sept. 2012)
[17] US National Library of Medicine, National Institutes of Health: PubMed, URL: http://www.ncbi.nlm.nih.gov/pubmed/
[18] iRODS User Group Meeting 2012, URL: https://www.irods.org/index.php/iRODS_User_Group_Meeting_2012
[19] The project EUDAT, URL: http://www.eudat.eu/safe-replication
[20] Amazon Web Services LLC: Amazon Web Services, URL: http://aws.amazon.com/ (zugegriffen Juli 2012)
[21] Dropbox, URL: https://www.dropbox.com/ (zugegriffen im Januar 2013)
[22] OwnCloud, URL: http://owncloud.org/ (zugegriffen im Januar 2013)
[23] UNICORE, URL: http://www.unicore.eu/ (zugegriffen Juli 2012)
[24] Apache Tika, URL: http://tika.apache.org/ (zugegriffen im Januar 2013)
[25] Apache Lucene, URL: http://lucene.apache.org/ (zugegriffen im Januar 2013)
[26] Molecular Simulation Grid, URL: https://mosgrid.de/portal (zugegriffen im Januar 2013)
[27] ICAT, URL: http://www.icatproject.org (zugegriffen im Januar 2013)
[28] Nestor-Handbuch Kapitel 6, URL: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010062454
[29] Konrad-Zuse-Zentrum für Informationstechnik Berlin: Stellaris, URL: http://www.zib.de/de/projekte/aktuelle-projekte/projekte-detail/article/stellaris.html
[30] OpenArchive, URL: http://www.openarchives.org/
[31] Nestor-Handbuch Kapitel 11.2, URL: http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-20100305219
[32] dSpace, URL: http://www.dspace.org/
[33] Fedora-Commons, URL: http://fedora-commons.org/
[34] University of Southampton: ePrints, URL: http://www.eprints.org/
[35] Fraunhofer-Gesellschaft: Open Access Server der Fraunhofer-Gesellschaft, URL: http://eprints.fraunhofer.de/
[36] Registry of Open Access Repositories, URL: http://roar.eprints.org/?action=home&type=institutional
[37] Stanford University: LOCKS, URL: http://www.lockss.org/
[38] Stanford University:LOCKSS, URL: http://lockss.stanford.edu/
[39] CLOCKSS, URL: http://www.clockss.org/clockss/Home
[40] Thüringer Universitäts- und Landesbibliothek Jena: URMEL, http://www.urmel-dl.de/ (zugegriffen Sept. 2012)
[41] KOBV: Opus4, URL: http://www.kobv.de/opus4 (zugegriffen Sept. 2012)
[42] Regionales Rechenzentrum der Universität Hamburg: MyCoRe, URL: http://www.mycore.de/index.html
[43] International DOI Foundation: Digital Object Identifier, URL: http://www.doi.org/
[44] Deutsche Nationalbibliothek: Persistent Identifier, URL: http://www.persistent-identifier.de/
[45] OCLC: Persistent Uniform Resource Locators, URL: http://purl.oclc.org/

[46] Nestor-Handbuch Kapitel 9, http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2010062482
[47] International DOI Foundation: DOI Factsheet, URL: http://www.doi.org/factsheets/DOIHandle.html
[48] University of Chicago, Argonne National Laboratory: Globus Online, URL: www.globusonline.org
[49] Re3sata.org: Liste der Archive http://service.re3data.org/search/results?term=
[50] Paul Vierkant et al., Vocabulary for the Registration and Description of Research Data Repositories, Version 2.0 (December 2012), doi:10.2312/re3.002,
[51] Projekt Radieschen: Report des Expertenworkshops 2012 „Elemente einer übergreifenden Forschungsdaten-Infrastruktur: Eine für Alle?“, 17. April 2012, GFZ Potsdam, Deutschland, http://www.forschungsdaten.org/uber-radieschen/publikationen/
[52] Projekt Radieschen: Workshop-Materialien FDI 2013, DOI: 10.2312/RADIESCHEN_001, URL: http://www.forschungsdaten.org/uber-radieschen/publikationen/
[53] Projekt Radieschen: „Symposium Forschungsdaten-Infrastrukturen“, Protokolle der Workshops (intern)
[54] Projekt Radieschen: „Symposium Forschungsdaten-Infrastrukturen“, Februar 2013,Vortragsfolien , URL: http://www.forschungsdaten.org/uber-radieschen/projektveranstaltungen/symposium-forschungsdaten-infrastrukturen/

A. Anhang

Liste der von re3data analysierten Repositories

Die folgende Tabelle zeigt eine Auflistung der von re3data reviewten Repositories inklusive einiger in Bezug auf die verwendete Technik interessanter Angaben.

In der 1. Spalte ist der Name des Repositorys gelistet. Die 2. Spalte gibt an, welche Repository-Software verwendet wird, in der 3. Spalte steht, ob die Daten mit einem Persistenten Identifier versehen werden und wenn ja, mit welchem. In der letzten Spalte der Tabelle ist angegeben, welches Format die Daten im Repository haben. Die Bedeutungen der einzelnen Tabelleneinträge in der 2. bis 4. Spalte sind [42] zu entnehmen.

Name des Archivs	Name der Repository-Software	Persistent Identifier-System	Art des Inhalts
Access to Archival Databases	unknown	none	Structured text
Aktuelle Wetterwerte deutscher Stationen	unknown	none	Networkbased data
Archaeology Data Service	unknown	DOI	Standard office documents
Atmospheric Science Data Center	other	none	Standard office documents
Australian Social Science Data Archive	other	other	Standard office documents
Barbara A. Mikulski Archive for Space	other	none	Standard office

Telescopes			documents
bii	other	none	Raw data
Blue Obelisk Data Repository	unknown	none	Configuration data
British Atmospheric Data Centre	unknown	none	Standard office documents
Carbon Dioxide Information Analysis Center	other	DOI	Raw data
Cell Centered Database	unknown	none	Images
Chemical Database Service	unknown	none	Scientific and statistical data formats
Chesapeake Bay Environmental Observatory	unknown	none	Standard office documents
CISL Research Data archive	other	none	Raw data
CiteSeerX	unknown	none	Plain text
Climate and Environmental Retrieval and Archive	other	DOI	Plain text
Coastal Data Information Program of the Scripps Institution of Oceanography	unknown	none	Raw data
Communication Portal for Accessing Social Statistics	other	none	Scientific and statistical data formats
Comprehensive Epidemiological Data Resource	unknown	none	keine Angabe
CoRIS	unknown	none	Standard office documents
CR-EST	unknown	none	Images
Crystallography Open Database	unknown	none	keine Angabe
CUAHSI Hydrologic Information System	other	none	Standard office documents
CyberCell Database	unknown	none	Standard office documents
Das Sozio-oekonomische Panel	unknown	none	Raw data

Data.gov	other	none	Raw data
Datenbank Gesprochenes Deutsch	other	none	Audiovisual data
Der Karlsruher Wolkenatlas	unknown	none	Images
DOE Joint Genome Institute Genome Web Portal	other	none	Standard office documents
Earth Resources Observation and Science Center	unknown	none	Images
Ecological Society of America esa Ecological Archives	unknown	URN	Standard office documents
eCrystals	EPrints	DOI	Structured graphics
Edinburgh DataShare	DSpace	handle	Standard office documents
Encyclopedia of Astronomy and Astrophysics	unknown	DOI	Standard office documents
EnvBase	unknown	none	Standard office documents
Environmental data explorer	unknown	none	Structured graphics
ESO Science Archive Facility	unknown	none	Raw data
European Climate Assessment & Dataset project	other	none	Standard office documents
Federal Reserve Economic Data	unknown	none	Standard office documents
figshare	other	DOI	Standard office documents
Forschungsdatenzentrum der Rentenversicherung	unknown	none	Scientific and statistical data formats
geocommons	other	none	Standard office documents
GeoConnections - Discovery Portal	unknown	none	Images
GeoGratis	unknown	none	Standard office documents
Global Change Master	other	DOI	Standard office

Directory			documents
Global Initiative on Sharing Avian Influenza Data	unknown	none	Raw data
High Energy Astrophysics Science Archive Research Center	other	none	Raw data
Historical hydrographic data from BSH	unknown	none	Plain text
HubbleSite	unknown	none	Images
Inorganic Crystal Structure Database	unknown	none	Structured graphics
Integrated Climate Data Center	unknown	none	Standard office documents
IPB MassBank	unknown	none	Structured graphics
IQSS Dataverse network	other	handle	Standard office documents
IUBio-Archive	unknown	none	Standard office documents
JASPAR	unknown	other	Structured graphics
JEDI	DSpace	handle	Standard office documents
Journal of applied econometrics Data Archive	unknown	none	Plain text
KNMI Climate Explorer	unknown	none	Standard office documents
Kujawsko-Pomorska Digital Library	dLibra	none	Standard office documents
LOGKOW	unknown	none	Scientific and statistical data formats
Mansfeld's World Database of Agriculture and Horticultural Crops	unknown	none	Plain text
MetaCrop	unknown	other	Standard office documents
MIRAGE	other	none	Images
MolTable	unknown	none	Scientific and statistical data formats
MorphoBank	other	none	Images
Multimission Archive at	unknown	none	Raw data

STScI			
Multimodal Learning Corpus Exchange	unknown	none	Standard office documents
NASA Distributed Active Archive Center at National Snow & Ice Data Center	unknown	PURL	Standard office documents
NASA/IPAC Infrared Science Archive	unknown	none	Images
National Center for Ecological Analysis and Synthesis Data Repository	other	none	Standard office documents
National Forestry Database	unknown	none	Scientific and statistical data formats
National Snow and Ice Data Center	unknown	none	Images
National Space Science Data Center	other	none	Images
NEEShub	other	none	Databases
NeuroMorpho	other	other	Standard office documents
Neuroscience Information Framework	unknown	other	Standard office documents
Ocean Biogeographic Information System	other	none	Standard office documents
Odum Institute Dataverse Network	unknown	none	Standard office documents
Open Context	unknown	none	Standard office documents
OpenEI	other	none	Standard office documents
Open Research Data	unknown	DOI	Images
Paleobiology Database	unknown	none	Structured text
PANGAEA	other	DOI	Standard office documents
PDS	other	none	Images
Polar Data Center	unknown	none	Standard office documents
Proteome Commons	other	none	Raw data
PubChem	unknown	other	Databases

RNA Abundance Database	unknown	none	Scientific and statistical data formats
SAFER-Data	other	none	Standard office documents
SEDAC	unknown	none	Standard office documents
SeSam	other	none	Standard office documents
ShareGeo open	DSpace	handle	Standard office documents
SIMBAD Astronomical Database	other	none	Standard office documents
SkyView	unknown	none	Images
SMOKA Science Archive	unknown	none	Standard office documents
Southeast Asian Climate Assessment & Dataset	unknown	none	Standard office documents
Spec Patterns	unknown	none	Structured graphics
St. Lawrence Global Observatory Data	unknown	none	Standard office documents
Swedish National Data Service	unknown	none	Standard office documents
The Biological and Chemical Oceanography Data Management Office	unknown	handle	Structured text
The Cell	unknown	none	Images
The Durham HepData Project	other	ARK	Scientific and statistical data formats
The Knowledge Network for Biocomplexity	other	none	Standard office documents
The University of Oxford Text Archive	unknown	none	Standard office documents
The World Atlas of Language Structure	unknown	none	Standard office documents
The World Data Center for Remote Sensing of the Atmosphere	unknown	DOI	Databases
UPSpace	DSpace	handle	Standard office documents
USU Institutional repository	DSpace	handle	Standard office documents

VegBank	unknown	none	Scientific and statistical data formats
Victoria Experimental Network Under the Sea	unknown	none	Images
Wetter, Wolken, Klima	unknown	none	Networkbased data
World Data Center for Human Interactions in the Environments	unknown	none	Standard office documents
World Data Centre for Greenhouse Gases	unknown	none	Scientific and statistical data formats
World Data Centre for Precipitation Chemistry	unknown	none	Standard office documents
XCOM	unknown	none	Standard office documents
X-Ray Database	unknown	none	Plain text