Anforderungen eines nachhaltigen, disziplinübergreifenden Forschungsdaten-Repositoriums

Jan Potthoff[1], Jos van Wezel[1], Matthias Razum[2], Marius Walk[1]

[1]Karlsruher Institut für Technologie (KIT)
[2]FIZ Karlsruhe - Leibniz-Institut für Informationsinfrastruktur
jan.potthoff@kit.edu, jos.vanwezel@kit.edu,
matthias.razum@fiz-karlsruhe.de, marius.walk@kit.edu

Abstract: Wissenschaftliche Erkenntnisse basieren zunehmend auf digitalen Daten, deren Publikation, Verfügbarkeit und Nachnutzung im Rahmen der guten wissenschaftlichen Praxis gewährleistet werden muss. Da auf die Daten in weiteren Arbeiten Bezug genommen werden kann und neue Erkenntnisse daraus entstehen, muss der Datenquelle ein besonderes Vertrauen entgegengebracht werden können. Das Projekt Research Data Repositorium (RADAR) geht diese Herausforderung durch die Etablierung einer generischen Infrastruktur für die Archivierung und Publikation von Forschungsdaten an. Der Schwerpunkt wird auf Forschungsdaten aus Bereichen mit kleineren Datenmengen (small sciences) gelegt, da in diesen Forschungsdateninfrastrukturen meist noch fehlen. Die Nachhaltigkeit und Vertrauenswürdigkeit wird durch Bitstream Preservation, einer angestrebten Zertifizierung und einem sich selbst tragenden Betrieb mit einer Kombination aus Einmalzahlungen und institutionellen Angeboten gewährleistet.

1 Einleitung

Unter digitaler Langzeitarchivierung wird das Bewahren von digitalen Informationen über einen technologischen Wandel hinaus verstanden [Ne12]. Mit digitaler Information ist dabei jegliche Art von Information gemeint, die durch eine Sequenz von Bits (Bitstream) darstellgestellt werden kann. Die Information lässt sich im digitalen Bereich in drei Ebenen unterteilen. Die physische Ebene stellt die kodierte Information auf einem Speichermedium dar. Art und Weise hängt dabei vom verwendeten Speichermedium ab. Auf der logischen Ebene liegen die Informationen in einer Form vor, welche durch ein Informationssystem, anhand einer Formatspezifikation, die Semantik und Syntax beschreibt, interpretiert werden. Auf der konzeptuellen Ebene ist die Information schlussendlich für den Menschen interpretierbar. Eine Aufgabe der digitalen Langzeitarchivierung ist daher der Erhalt der drei Informationsebenen [Th02].

Neben der rein technischen Ausrichtung der digitalen Langzeitarchivierung sind des Weiteren organisatorische Prozesse zu berücksichtigen. Mit dem Open Archival Information System (OAIS) wird beispielsweise ein Referenzmodell beschrieben, dass nicht nur die technischen Komponenten eines solchen Systems berücksichtigt, sondern

auch Personen, die an den Prozessen, wie Einlagerung (Ingest), Lesen (Retrieval) und der Administration, beteiligt sind [CC12]. Es werden also sowohl Aufgabenbereiche und Verantwortlichkeiten definiert als auch Modelle der Funktionen und Informationen in einem digitalen Archiv. Durch die allgemeine Beschreibung der Datentypen und Funktionsabläufe wird das Referenzmodell als vom Einsatzbereich unabhängig angesehen und bildet so eine wesentliche Grundlage für Archivsysteme.

Sowohl die technischen als auch die organisatorischen Prozesse einer Infrastruktur zur digitalen Langzeitarchivierung müssen so entworfen sein, dass dem System Vertrauen entgegengebracht werden kann. Insbesondere ist dabei die Nachhaltigkeit zu adressieren.

2 Nachhaltigkeit und Vertrauenswürdigkeit

Um das Vertrauen in digitale Archive zu stärken, wurden im Rahmen des Projekts ArchiSafe, das im Rahmen der eGovernment-Initiative BundOnline 2005 an der Physikalisch-Technischen Bundesanstalt (PTB) durchgeführt wurde, Konzepte für ein vertrauenswürdiges elektronisches Archiv entworfen. Adressiert werden fünf Kernanforderungen: Vollständigkeit, Integrität und Authentizität, Lesbarkeit, Verfügbarkeit und Verkehrsfähigkeit [HR08]. Auch für den Bereich der Forschung wurden im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts "Beweissicheres elektronisches Laborbuch" (BeLab) Konzepte zur beweiswerterhaltenden Archivierung entwickelt [Jo13]. Dabei wurden neben den technischen Möglichkeiten insbesondere juristische Fragestellungen berücksichtigt. Neben diesen Projekten wurden in den letzten Jahren Kriterienkataloge entworfen, anhand derer ein umgesetztes Archiv bewertet und zertifiziert werden kann, um das Vertrauen in die Infrastruktur zu fördern. Basierend auf [CR07] lassen sich die Anforderungen in drei Kategorien unterteilen: Organisatorische Strukturen, Datenmanagement und technische Aspekte.

2.1 Organisatorische Strukturen

Das *Anwendungsgebiet* des Systems und die Nutzergruppen sind zu definieren. Es müssen Leitlinien, die sowohl inhaltliche Kriterien als auch Betriebskriterien darlegen, zur allgemeinen Verfügung gestellt [SHH08] und Rechte und Pflichten des Anbieters (auch bezüglich Archivierungszeiträume) sowie der Urheber der Daten festgeschrieben werden [Ne10]. Das Ziel der *langfristigen Erhaltung* von Daten mit Möglichkeit der Verwaltung und des Zugriffs muss klar erkennbar sein [Mc12]. Dazu gehört auch die Definition von Strategien für das vorzeitige Ende des Betriebs oder das substantielle Ändern des Anwendungsbereiches [Do09]. Zur Erhaltung des Repositoriums müssen kurz- und langfristige Geschäftspläne (*Geschäftsmodell*) verfasst sein, die in einem zyklischen Prozess erarbeitet werden. Die Pläne müssen dabei mindestens jährlich überprüft und gegebenenfalls angepasst werden. Das Modell muss Einnahmen, Investitionen, Ausgaben und Risiken enthalten und dabei von Dritten transparent nach gesetzlichen Vorgaben überprüft werden können [RL02]. Die formale Prüfung und die Beurteilung des Repositoriums müssen regelmäßig erfolgen, so dass technologische

Entwicklungen und wachsende Anforderungen rechtzeitig erkannt werden. Daher ist eine regelmäßige *Zertifizierung* sinnvoll [CR07].

2.2 Datenmanagement und technische Aspekte

Das Gesamtsystem sollte *OAIS-konform* aufgebaut sein [SHH08]. OAIS bietet dabei ein allgemeines Framework aus Terminologien und Konzepten, um Architekturen und Operationen digitaler Archive zu beschreiben und zu vergleichen [Mc12]. Das Repositorium muss festlegen, welche *Eigenschaften eines digitalen Objekts* erhalten bleiben und in welcher Form es gespeichert wird. Beispielsweise kann zu einem Dokument nur der Text oder auch die Darstellung gesichert werden. Für die gängigsten Typen angenommener digitaler Objekte sollten Kriterien festgelegt sein, die genau beschreiben, um welche Objekte es sich handelt. Dem Objekt sollte dann ein *Persistent Identifier* zugewiesen werden [Do09]. Dabei gelten inhaltlich veränderte Dokumente als neue Dokumente. Die *Mindestzeit der Verfügbarkeit* eines Objekts für Dokumenten- und Publikationsservices darf beispielsweise 5 Jahre nicht unterschreiten und sollte in den Leitlinien festgehalten werden. Dabei sollte stets die Verbindung von Metadaten mit deren Objekte gewährleistet sein [Ne10].

Um alle Anfragen der Nutzer nach hinterlegten Daten erfüllen zu können, muss für jedes Objekt ein *Minimum an Metadaten*, die zur Identifizierung des Objekts dienen, hinterlegt sein. Hierzu muss das Repositorium schon von Beginn an eine Mindestmenge an Metadaten fordern [SHH08]. Das Repositorium sollte zudem aufzeigen, wie Metadaten gewonnen werden [CR07]. Es sollte die Möglichkeit bestehen über eine *Schnittstelle* standardisierte Metadaten ausgeben zu können [Ne10]. Die *Erreichbarkeit* des gesamten Angebots sollte über eine Webseite gewährleistet sein. Generell sollten aber verschiedene Möglichkeiten des Zugriffs vorhanden und frei zugänglich sein [CR07].

Das *Betriebskonzept* muss eine angemessene Verfügbarkeit des Systems garantieren und ausreichend dokumentiert sein. Die Dokumentation sollte Angaben über Komponenten, Zugangsregeln (personelle Verantwortlichkeit und Vertretung) und die pflichtmäßig durchzuführenden regelmäßigen Wartungen enthalten. Dabei sollten Technologien zur Sicherung und Wiederherstellung der Server-Software, der Metadaten und der Objekte eingesetzt werden [Ne10]. Das Repositorium muss dazu in der Lage sein, die Anzahl vorhandener Kopien digitaler Objekte und deren Ablageort zu definieren [Bl10].

3 Archivierungsinfrastrukturen

National aber auch international gibt es eine Reihe von Beispielen für Forschungsdaten-Repositorien, wie z. B. das World Data System der International Council of Science (ICSU)[1], das Studien und empirischen Primärdaten aus den Sozialwissenschaften archiviert und zur Verfügung stellt. Im Bereich der Geowissenschaften ist ein weiteres Beispiel das Repositorium PANGAEA[2]. Aber auch Fachdatenbanken, wie z. B. die

[1] http://www.icsu-wds.org/
[2] http://pangaea.de

Protein Data Bank (PDB)[3], die Informationen über experimentell gewonnene Daten von Proteinstrukturen und Nukleinsäuren zur Verfügung stellt, werden genutzt.

Die Verlässlichkeit der Technologie ist eine der wichtigsten Faktoren bei der Archivierung von Daten, die über mehrere Jahrzehnte gespeichert, lesbar und zugänglich sein müssen. In Rechenzentren werden Daten vor allem auf magnetischen Datenträgern in Form von Festplatten oder auf bis zu 700 Metern langen und 0,5 Zoll breiten Magnetbändern gespeichert. Beide Speichertechnologien haben eine vergleichbare Nutzungsdauer von vier bis sechs Jahren. Diese ist jedoch abhängig von der Anzahl der Lese- und Schreib-Vorgängen. Obwohl diese in einem Datenarchiv erwartungsgemäß gering sind, müssen die Speichermedien regelmäßig zur Überprüfung der Datenintegrität (Bitstream Preservation) gelesen werden. Am Ende der Lebensdauer einer Festplatte oder eines Bandes müssen die darauf befindlichen Daten ein letztes Mal zur Migration auf ein neues Gerät oder Medium gelesen werden.

Für große Archive und Langzeitarchive werden aus Kostengründen Magnetbänder zur Datenspeicherung eingesetzt. Bandspeichermedien sind nicht verbunden mit den eigentlichen Lese-/Schreibgeräten. Dadurch können sie mit einem niedrigeren Energieverbrauch betrieben werden und bieten so gegenüber Festplatten einen Kostenvorteil. Mit modernen Bandlaufwerken wird eine Abwärtskompatibilität bis zu drei Generationen gewährleistet, so dass in der Praxis eine Lebensdauer von acht bis zwölf Jahren erreicht werden kann. Ein weiterer Vorteil ist die in den letzten 10 Jahren angestiegene Speicherdichte und die damit um das 30- bis 50-fach angestiegene Speicherkapazität.

Systeme, auf denen Repositorien für eine verlässliche Archivierung aufbauen, bestehen aus zwei Kernkomponenten. Eine Datenbank, in der der Speicherort der archivierten Daten nach dem Ingest vermerkt wird, und ein System, das den Zugriff auf unterschiedliche Speichermedien abstrahiert. Zusätzlich sorgt es für die Migration der gespeicherten Daten zwischen den Speicherebenen oder Hierarchien auf der Grundlage von dem Benutzer definierter Regeln. Die Datenbank erlaubt des Weiteren die Ablage von Metadaten in unterschiedlichen Formaten. Kommerzielle Systeme, wie HPSS[4], SAM-QFS[5] oder DMF[6], sind in der Lage Millionen Objekte und mehrere Petabyte auf unterschiedlichen Speichersystemen, von Solid-State-Drive (SSD) bis Band, zu verwalten. Neben der Softwareentwicklung zum Datenmanagement in der Teilchenphysik entstanden die mit den kommerziellen Produkten vergleichbaren Open Source Pakete, wie zum Beispiel dCache und xrootd, die ebenfalls mit Petabyte-großen Speichersysteme umgehen und die eine automatische Migration zwischen Speicherebenen/-hierarchien unterstützen. iRODS, ein weiteres Open Source Paket, bietet ein regelbasiertes Datenmanagement mit dem eine automatische Migration in Abhängigkeit von den gespeicherten Metadaten, wie Größe oder Eigentümer, gesteuert werden kann.

---

[3] http://www.rcsb.org/
[4] http://www.hpss-collaboration.org/
[5] http://www.oracle.com
[6] http://www.sgi.com

4 Disziplinübergreifendes Forschungsdaten-Repositorium

Die Nachvollziehbarkeit und Reproduzierbarkeit von wissenschaftlichen Ergebnissen, etwa in [LMS12] unter dem Begriff *reproducible research* beschrieben, sowie neuartige Ansätze des Erkenntnisgewinns, die von Jim Gray als Fourth Paradigm bezeichnet wurden [HTT09], führen zu einer steigenden Bedeutung der Forschungsdaten im Rahmen von Veröffentlichungen. Verbunden damit ist der Bedarf, Forschungsdaten, auf denen eigene Ergebnisse beruhen, zu referenzieren. Das Referenzieren – in Form von Zitaten ein Kernelement wissenschaftlicher Kommunikation und Reputationsbildung – bedarf aber ebenso wie die Reproduzierbarkeit von wissenschaftlichen Erkenntnissen einen zuverlässigen und dauerhaften Zugriff auf die zugrunde liegenden Forschungsdaten und damit einer entsprechenden Infrastruktur.

4.1 Anwendungsbereich

Das von der DFG geförderte Projekt Research Data Repositorium (RADAR) stellt eine solche Infrastruktur für die Archivierung und Publikation von Forschungsdaten bereit. Partner des Projekts sind das FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur, die Technische Informationsbibliothek (TIB) in Hannover, das Steinbuch Centre for Computing (SCC) des KIT und zwei wissenschaftliche Partner (das Department für Chemie der LMU München und dem Leibniz-Institut für Pflanzenbiochemie in Halle).

Das Projekt richtet sich an Wissenschaftler (z. B. in Projekten) und Institutionen (z. B. Bibliotheken) mit zwei Angeboten: Ein disziplinübergreifendes Einstiegsangebot zur formatunabhängigen Datenarchivierung mit minimalen Metadatensatz und einem darauf aufbauenden Angebot mit integrierter Datenpublikation. Das Angebot der Archivierung (Einstiegsangebot) bietet eine formatunabhängige Archivierung mit Bitstream Preservation und einem minimalen Metadatensatz gemäß dem DataCite Metadata Kernel aus. Darüber hinaus erlaubt es eine Verknüpfung von Daten mit Metadaten.. Im Einstiegsangebot behält der Datengeber das alleinige Zugriffsrecht auf seine in RADAR archivierten Forschungsdaten, welches auf einen ausgewählten Benutzerkreis erweiterbar ist. Die so archivierten Forschungsdaten erhalten einen persistenten Identifier in Form eines Handles. Dieses Angebot richtet sich an Kunden, die primär an einer Datenarchivierung unter Einhaltung der von der DFG empfohlenen Haltefristen von 10 Jahren interessiert sind. Daneben können vom Benutzer weitere Haltefristen gewählt werden, die eine Mindesthaltefrist von 5 Jahren, siehe Abschnitt 2.2, nicht unterschreiten. Darüber hinaus eignet es sich aber auch für andere Daten wie z. B. Negativdaten[7], die man nicht als Teil einer Publikation verwenden möchte. Das zweite Angebot unterstützt die dauerhafte Datenarchivierung mit integrierter Datenpublikation. Über die Leistungen des Einstiegsangebots hinaus ist hier die Vergabe von format- und disziplinspezifischen Metadaten sowie die Vergabe von dauerhaften DOI-Namen im Rahmen eines Publikationsprozesses vorgesehen. Durch diese Maßnahmen ist das Referenzieren von Daten, die zu einer Veröffentlichung im klassischen Sinne gehören,

[7] Daten, die zu keinem Ergebnis geführt haben.

aber auch eine reine Datenpublikation mit eigener DOI-Registrierung möglich. Dies bedingt auch eine aufwändigere Beschreibung der so publizierten und archivierten Forschungsdaten mit disziplinspezifischen Metadaten.

Gegenüber kommerziellen Diensten wie z. B. Dropbox[8] oder figshare[9] setzt sich RADAR in drei wesentlichen Punkten ab: Hinter RADAR steckt ein Betreiberkonsortium aus wissenschaftsnahen Gedächtnisorganisationen und Infrastruktureinrichtungen, die dem deutschem Recht unterliegen und keine kommerziellen Interessen verfolgen. Gegenüber kommerziellen Anbietern wird z. B. durch die Überwachung von Haltefristen und die Vergabe von persistenten Identifikatoren (Handles und DOIs) ein Mehrwert geboten. Drittens unterscheidet sich RADAR durch die zugesicherte Bitstream Preservation.

Als Entwicklungsgrundlage der RADAR Infrastruktur dient eine zuvor durchgeführte Anforderungsanalyse, in der insbesondere der Forschungsprozess und der zugehörige Workflow durch zwei an dem Projekt beteiligte Partner betrachtet wird. Dabei soll herausgestellt werden, welche Daten als erhaltungswürdig eingestuft werden können, welche Software in dem entsprechenden Forschungsbereich zum Einsatz kommt, und welche Datenformate daraus resultieren. Das Ziel der Analyse ist es, einen prototypischen Workflow zu genieren, anhand dessen die Anforderungen an die Archivierung im Allgemeinen, aber auch an einen forschungsnahen Daten Ingest, herausgestellt werden können. Basierend auf den beschriebenen zwei Angeboten sollen zudem im Rahmen der Anforderungsanalyse allgemeine und fachspezifische Metadatenfelder identifiziert werden. Bei der Ausgestaltung dieser Metadatenprofile kann über eine Kooperation mit dem Projekt Large Scale Data Management & Analysis (LSDMA)[10] am KIT auf umfangreiche Vorarbeiten zurückgegriffen werden. Thematisch legt RADAR den Schwerpunkt auf die *small sciences*, in denen Forschungsdateninfrastrukturen meist noch fehlen. RADAR erlaubt eine temporäre oder im Falle einer Datenpublikation zeitlich unbegrenzte Datenarchivierung. Das Projekt will nicht in Konkurrenz zu anderen, wie die in Abschnitt 3 genannten Beispiele, treten. Wie zuvor dargestellt sind die in den bisherigen Datenzentren vorhandenen Lösungen meist stark auf eine Zielgruppe ausgerichtet oder stellen kleinere Fachanwendungen für eine begrenzte Nutzergruppe dar. Hier versucht RADAR über das Angebot eines generischen, d. h. disziplinunabhängigen Repositoriums, die Lücke zu schließen und gleichzeitig über seine offenen Schnittstellen eine Basis für die Etablierung weiterer Fachanwendungen zu bilden.

4.2 Technische Aspekte

Als Frontend wird dem Nutzer eine intuitiv zu bedienende Web Oberfläche geboten. Hierüber lassen sich sowohl Daten und Metadaten erfassen als auch Regelungen, wie Servicelevel und Zugriffsrechte definieren. Zur schnelleren Übertragung von großen Datenmengen können des Weiteren direkte Speicherschnittstellen, die entsprechend

[8] https://www.dropbox.com/
[9] http://figshare.com/
[10] http://www.helmholtz-lsdma.de/

geeignete Protokolle unterstützen, genutzt werden. Die Gesamtarchitektur ist in Abbildung 1 dargestellt. Sowohl die im Abschnitt 3 beschriebenen kommerziellen als auch die Open Source Systeme bieten Funktionen zur Zugriffskontrolle. Des Weiteren unterstützen sie die Überprüfung der Datenintegrität auf der Grundlage von Prüfsummen. Die Absicherung der Datenintegrität kann durch das Anlegen von mehreren Kopien erhöht werden. So besteht die Möglichkeit mehrere Kopien der Daten anzufertigen und bei Bedarf und nach Ausstattung der Anlage auf unterschiedliche geografische Standorte zu verteilen. Dabei – insbesondere durch die Nutzung von Bändern als Speichermedium – muss mit einer längeren Zugriffszeit gerechnet werden. Möglichkeiten der Migration und schnellere Zugriffszeiten können durch den Betreiber durch unterschiedliche Servicelevel definiert werden, so dass der Nutzer selbst über Sicherheit der Datenintegrität und Zugriffszeiten in Abhängigkeit von entstehenden Kosten entscheiden kann. Die Bitstream Preservation ist nur eine Herausforderung der digitalen Langzeitarchivierung, auf die im Projekt RADAR besonderen Fokus gelegt wird. Teilweise existieren für weitere Aufgaben der Langzeitarchivierung bereits Lösungen oder müssen noch erforscht werden, z. B. im Bereich der Interpretation von Daten. Für diese soll das System über geeignete Schnittstellen flexibel sein.

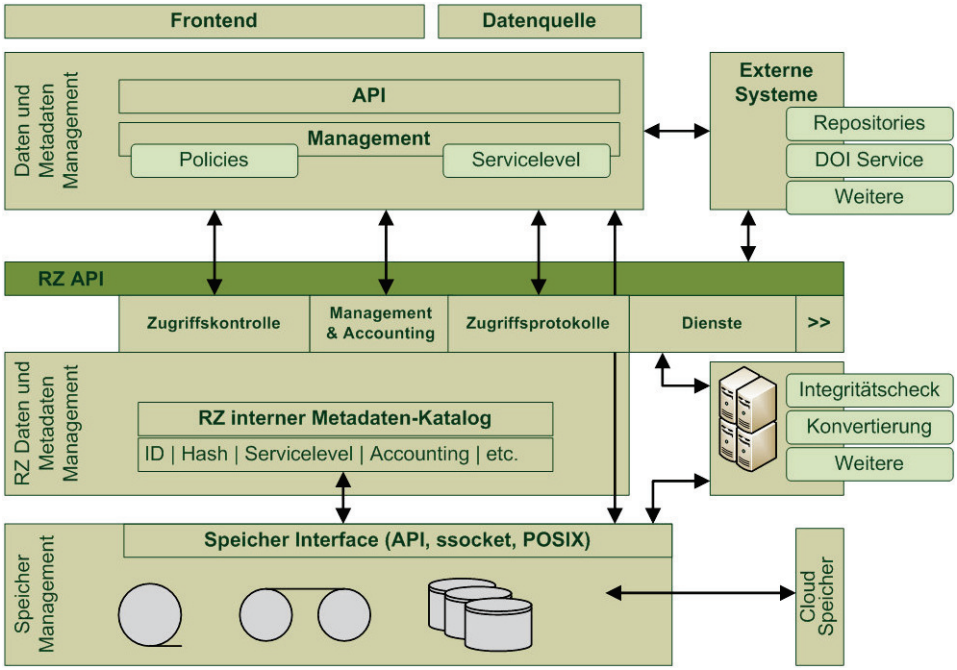

Abbildung 1: RADAR Architekturentwurf

Wie in Abbildung 1 dargestellt, ist der Kern der RADAR Architektur eine erweiterbare Schnittstelle für die Archiv Storage Services (Rechenzentrum Programmierschnittstelle – RZ API), die von einem Rechenzentrum angeboten werden. Die Schnittstelle verbirgt die Nutzung verschiedener Storage Services vor dem Frontend. Die RZ API stellt neben Schnittstellen für die Speicherung von Dateien Mechanismen für Zugangskontrolle,

Management, Accounting und zukünftig weitere inhaltsabhängige Services zur Verfügung. Da auf dieser Ebene auch die Bitstream Preservation sichergestellt werden muss, bietet sie die Möglichkeit der Neuberechnung von Prüfsummen. Dies ist aufgrund ansteigender Datenmengen nur über einen automatisierten Workflow möglich. Ein Automatismus könnte beispielsweise auch für Formatkonvertierungen umgesetzt werden. Vorgesehen ist, dass die RZ API auch Zugriffe auf das Metadaten System des Frontend ermöglicht und damit schnell Informationen über die gespeicherten Daten abfragen kann. Die Schicht oberhalb der RZ API übernimmt das Management von Kopien der Daten an unterschiedlichen Standorten und deren Erhalt. Außerdem übernimmt sie Koordinierungsaufgaben, die z. B. für die Vergabe von DOIs notwendig sind. Ein wichtiges Feature ist aber auch, dass die eingestellten Daten ohne Frontend abgerufen werden können. Dazu ist ein selbstbeschreibendes Dateiformat vorgesehen aus dem die Metadaten wieder hergestellt werden können.

Die Definition der RZ API soll es des Weiteren ermöglichen weitere Rechenzentren in das System einbinden zu können. Die RZ API Schicht trennt so Art der Speicherstrukturen des Rechenzentrums und Arbeitsweise des Daten und Metadaten Managements. Über die Anbindung weiterer Rechenzentren lässt sich der Erhalt der Datenintegrität verteilt über geografisch unterschiedliche Standorte realisieren. Des Weiteren können Benutzer die Standorte, an denen die Daten hinterlegt werden sollen, wählen. Neben der Anbindung von universitären Rechenzentren soll es möglich sein, auch kommerzielle Anbieter mit einzubinden. Für diese Möglichkeiten müssen entsprechende Services definiert und entsprechende Kosten im Rahmen des Geschäftsmodells kalkuliert werden.

Um das Vertrauen in das System zu stärken, soll eine Zertifizierung durchgeführt werden. Da eine große Anzahl von Kriterienkatalogen und Zertifizierungen existieren, muss in einem ersten Schritt eine Bewertung der Zertifikate und eine Eignungsprüfung für das angestrebte System durchgeführt werden. Die Zertifizierung wird sowohl das Datenzentrum als auch das darauf aufsetzende System (Management-Schicht) umfassen. Darüber hinaus soll auch das Verfahren der Datenpublikation zertifiziert werden.

4.3 Geschäftsmodell

Ein wesentlicher Punkt zur Nachhaltigkeit des Projekts ist das selbsttragende Geschäftsmodell, das im Laufe des Projekts ausgearbeitet werden soll. Als Zielgruppe werden neben Forschungsprojekten (d. h. Wissenschaftlern) auch institutionelle Nutzer, wie z. B. Bibliotheken, die das Forschungsdatenmanagement in ihrer Institution selbst übernehmen, die Aufgabe der Bitstream Preservation aber an einen Dienstleister auslagern möchten, adressiert. Insbesondere Wissenschaftler sind jedoch nicht in der Lage nach Ende ihres Projekts dauerhaft für die Archivierung ihrer Daten zu bezahlen. Daher sieht das Geschäftsmodell Einmalzahlungen für die Archivierung und optionale Publikation der Forschungsdaten in Abhängigkeit von Datenumfang und Haltedauer vor. So können schon während der Antragstellung Kosten zur Archivierung und Datenpublikation abgeschätzt und mit beim Fördergeber beantragt werden.

Da RADAR (vorerst) nur die Bitstream Preservation sicherstellt und die funktionale Langzeitarchivierung (Content Preservation) mittels Formatüberwachungen, Migration und Emulation außen vor lässt, lassen sich die zu erwartenden Kosten für die Zukunft recht gut abschätzen. Die höchsten Kosten fallen beim Ingest der Daten (also einer initialen Aufbereitung der Daten mit Beratung, gegebenenfalls einmaliger Formatkonvertierung in ein geeignetes Archivformat, Anreicherung mit Metadaten usw.) einmalig an. Die Kosten für die Speicherung sind abhängig von der Datenmenge, nicht deren Format oder Komplexität. Die Speicherung von Daten wird – bei gleichbleibender Menge – seit Jahrzehnten günstiger. Das weltweit rasant ansteigende Datenvolumen [Ec10] lässt erwarten, dass sich dieser Trend auch in Zukunft fortsetzt, auch wenn dies gegebenenfalls auch einen Technologiewechsel erfordern wird. Ist der grundlegende Betrieb des Rechenzentrums, in dem die zu archivierenden Daten gespeichert sind, durch andere Aufgaben gewährleistet, werden die Kosten für die weitere Archivierung mit zunehmender Zeit marginalisiert [BCL08]. Durch den Entwurf der RZ API, siehe Abschnitt 4.2, soll gewährleistet werden, dass auch weitere Rechenzentren, die diese API zur Verfügung stellen, eine Archivierung der Daten durchführen können. Dadurch können Verantwortlichkeiten auf andere Anbieter ausgelagert werden.

5 Fazit und Ausblick

Wie dargestellt, existiert eine Vielzahl von Richtlinien, Empfehlungen und Kriterienkatalogen, die den Aufbau eines vertrauenswürdigen Repositoriums darstellen und anhand derer sich eine Entwicklung bewerten lässt. Aufgabe wird es sein geeignete Schnittstellen zwischen Speicherschicht und Anwendungsschicht zu definieren. Die Herausforderung ist dabei insbesondere den Aufwand zur Implementierung der Schnittstelle so gering wie möglich zu halten, so dass Rechenzentren mit unterschiedlichsten Speicherstrukturen – aus Langzeitarchivierungssicht auch erwünscht – diese implementieren können. Bei der darüber liegenden Management-Schicht soll ebenfalls das Ziel einer möglichst flexiblen Schnittstelle verfolgt werden, um auch hier die Möglichkeit der Anbindung unterschiedlichster Frontends zu gewährleisten.

Datenspeicherung und vor allem die Datenarchivierung von Forschungsdaten ist eine auf Dauerhaftigkeit angelegte Dienstleistung. Der Großteil der dabei entstehenden Kosten soll auf die wissenschaftlichen Nutzer des Datenzentrums umgelegt werden, wobei bei der Entwicklung des Geschäftsmodells die aktuellen Gegebenheiten des Wissenschaftsbetriebs berücksichtigt werden müssen. Hier existiert eine starke Ausrichtung auf geförderte Projekte mit begrenzter Laufzeit. Nach dem Ende des Projekts besteht daher keine Möglichkeit weiterhin für die Datenarchivierung zu bezahlen. Eine direkte Förderung des Datenzentrums wird bei dem angedachten Geschäftsmodell weitgehend vermieden. Die Ausarbeitung eines Kostenmodells, das die benutzten Speichermodalitäten und zukünftige technische Entwicklungen berücksichtigt, muss erarbeitet werden. Darüber hinaus müssen vertragliche und rechtliche Regelungen für den Betrieb und das Dienstleistungsangebot des Datenzentrums erarbeitet werden. Wichtige Aspekte sind dabei die Weiterverwertung der Daten durch Dritte, Haftungsfragen und die Dauer der Datenspeicherung.

Literaturverzeichnis

[BCL08] Beagrie, N.; Chruszcz, J.; Lavoie, B.: Keeping Research Data Safe - A Cost Model and Guidance for UK Universities. Final Report. s.l: JISC, 2008, S. 91-92, http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf, abgerufen am 10.04.2014.

[Bl10] Blue Ribbon Task Force: Sustainable economics for a digital planet: Ensuring long-term access to digital information. In: Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf, abgerufen am 10.04.2014.

[CC12] CCSDS: Reference Model for an Open Archival Information System (OAIS). Magenta book. Technischer Bericht, CCSDS, 2012.

[CR07] Center for Research Libraries (CRL): Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), Chicago, 2007, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf, abgerufen am 10.04.2014.

[Do09] Dobratz, S. et al.: Catalogue of Criteria for Trusted Digital Repositories, nestor materials, Deutsche Nationalbibliothek, Frankfurt (Main), Germany, http://nbn-resolving.de/urn:nbn:de:0008-2010030806, abgerufen am 10.04.2014.

[Ec10] The Economist Newspaper Limited: The Data Deluge. In: The Economist, 2010. http://www.economist.com/node/15579717, abgerufen am 10.04.2014.

[HR08] Hackel, S.; Roßnagel, A.: Langfristige Aufbewahrung elektronischer Dokumente: In: Klumpp (Hrsg.), Informationelles Vertrauen für die Informationsgesellschaft, Berlin 2008.

[HTT09] Hey, T.; Tansley, S.; Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, Washington : Microsoft Research, 2009.

[Jo13] Johannes, P.C.; Potthoff, J.; Roßnagel, A.; Neumair, B.; Madiesh, M.; Hackel, S.: Beweissicheres elektronisches Laborbuch – Anforderungen, Konzepte und Umsetzung zur langfristigen, beweiswerterhaltenden Archivierung elektronischer Forschungsdaten und –dokumentation. In: Roßnagel, A. (Hrsg.): Der Elektronische Rechtsverkehr, Band 29, NOMOS, 2013.

[Mc12] McGovern, N.Y.: Aligning National Approaches to Digital Preservation. Atlanta, USA, 2012, http://educopia.org/sites/educopia.org/files/ANADP_Educopia_2012.pdf, abgerufen am 10.04.2014.

[Ne10] Netzwerkinformation e.v., Deutsche Initiative für und Arbeitsgruppe Elektronisches Publizieren: DINI-Zertifikat Dokumente- und Publikationservice 2010, http://nbn-resolving.de/urn:nbn:de:kobv:11-100182794, abgerufen am 10.04.2014.

[Ne12] Neuroth, H.; Oßwald, A.; Scheffel, R.; Strathmann, S.; Ludwig, J.: Langzeitarchivierung von Forschungsdaten: eine Bestandsaufnahme. Hülsbusch, Boizenburg, 2012. http://nestor.sub.uni-goettingen.de/bestandsaufnahme/index.php, abgerufen am 10.04.2014.

[RL02] Research Libraries Group (RLG): Trusted Digital Repositories: Attributes and Responsibilities, Report, 2002. http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf, abgerufen am 10.04.2014.

[SHH08] Sesink, L.; van Horik, R.; Harmsen, H.: Data Seal of Approval, Data Archiving and Networked Services (DANS), Den Haag, 2008, http://www.datasealofapproval.org/, abgerufen am 10.04.2014.

[LMS12] LeVeque, R.J.; Mitchell, I.A.; Stodden, V.: Reproducible research for scientific computing: Tools and strategies for changing the culture. In: Computing in Science and Engineering. July/August 2012, Bd. 14, 4, S. 13-17.

[Th02] Thibodeau, K.: Overview of technological approaches to digital preservation and challenges in coming years. The state of digital preservation: an international perspective, Seiten 4–31, 2002.