Das RADAR Projekt: Datenarchivierung und -publikation als Dienstleistung - disziplinübergreifend, nachhaltig, kostendeckend

Matthias Razum, FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur
Janna Neumann, Technische Informationsbibliothek Hannover

Zusammenfassung:
Die Nachvollziehbarkeit und Reproduzierbarkeit wissenschaftlicher Erkenntnisse basiert zunehmend auf digitalen Daten. Deren Publikation, Verfügbarkeit und Nachnutzung muss im Rahmen guter wissenschaftlicher Praxis gewährleistet werden. Das Projekt RADAR geht diese Herausforderung durch die Etablierung einer generischen Infrastruktur für die Archivierung und Publikation von Forschungsdaten an. Dafür vereinen fünf Projektpartner aus den Informations- und Naturwissenschaften (FIZ Karlsruhe, TIB in Hannover, KIT/SCC, LMU München und IPB Halle) ihre Kompetenzen. Durch enge Kooperation mit Wissenschaftler/innen, Datenzentren, Fachgesellschaften und Verlagen wird eine bedarfsgerechte Entwicklung der Infrastruktur sichergestellt. RADAR richtet sich an zwei Zielgruppen: Projekte (d. h. Forscher/innen) und Institutionen. Es verfolgt dabei einen zweistufigen Ansatz: ein disziplinübergreifendes Einstiegsangebot zur formatunabhängigen Datenarchivierung mit minimalem Metadatensatz und ein erweitertes Angebot mit integrierter Datenpublikation. Der thematische Schwerpunkt liegt bei den wissenschaftlichen Disziplinen im „long tail of science", in denen Forschungsdateninfrastrukturen meist noch fehlen. RADAR erlaubt eine temporäre oder – im Falle einer Datenpublikation – eine zeitlich unbegrenzte Datenarchivierung. Das angestrebte Geschäftsmodell zielt auf einen sich selbst tragenden Betrieb mit einer Kombination aus Einmalzahlungen und institutionellen Angeboten ab. RADAR ist als Baustein der internationalen Informationsinfrastruktur geplant, der sich über Schnittstellen auch in weitere Datenmanagement-Dienste Dritter integrieren lässt.

Summary:
The transparency and reproducibility of scientific results are increasingly based on digital data. In compliance with good scientific practice data need to be published, accessible, and re-usable. The RADAR project aims to establish a generic infrastructure, which will provide archiving and publication services for research data. Five partners from the information and natural sciences (FIZ Karlsruhe, TIB in Hanover, KIT/SCC, LMU Munich, and IPB Halle) have joined forces to address the challenges involved. By cooperating closely with researchers, data centers, scientific societies, as well as publishers, the partners ensure that the resulting infrastructure is designed to meet the requirements. Target groups are projects (e.g., researchers) and institutions (eg., libraries). Both groups are offered a two-stage approach with a cross-discipline starter package for format-independent data preservation with a minimum metadata set, and an enhanced package for preserving data with integrated data publication. RADAR focuses on the "long tail of science", which often lacks sufficient research data infrastructure. The repository will offer a temporary or - in case of data publication - long-term preservation of research data. A self-supporting business model will provide one-off payments and

institutional subscription services. As such, RADAR is intended to become an integral part of the international information infrastructure which also allows the integration of third-party services.

Zitierfähiger Link (DOI): 10.5282/o-bib/2014H1S30-44
Autorenidentifikation: Neumann, Janna: GND 139772863
Razum, Matthias: GND 1029295182

1. Motivation

Die Notwendigkeit, Forschungsdaten auch über das Ende eines Projekts hinaus verfügbar zu halten, ist inzwischen anerkannt. Vielfältige politische Vorhaben und Richtlinien (z.B. durch die OECD[1], UNESCO[2] , EU[3] oder die DFG[4]) unterstreichen dies. Erst die Bereitstellung der Daten erlaubt die Nachvollziehbarkeit und Reproduzierbarkeit wissenschaftlicher Ergebnisse (*reproducible research*, siehe Victoria Stodden[5]) und ermöglicht neuartige Ansätze des Erkenntnisgewinns, von Jim Gray als *Fourth Paradigm* bezeichnet.[6] Tatsächlich steht aber nur ein geringer Teil der produzierten Daten zur Verfügung. Fünf Problemfelder lassen sich identifizieren, die diesen Missstand begründen:

- das Fehlen einer dauerhaft angelegten und verlässlichen Infrastruktur zur Erschließung, Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten in vielen Fachdisziplinen und Einrichtungen,
- die fehlende oder unzureichende Integration der Datenarchivierung und -publikation in wissenschaftliche Arbeitsprozesse, was zu mangelhafter Bereitstellung von Daten und Metadaten führt und die publizierten Ergebnisse oft schlecht nachvollziehbar macht,
- eine fehlende Qualitätssicherung der Daten z.B. im Rahmen eines Peer-Review-Prozesses, die zu einer eingeschränkten Kommentierbarkeit sowie zur limitierten Zitierbarkeit der Daten führt,
- die Diskrepanz zwischen dem Aufwand für Datenaufbereitung/-archivierung und dem daraus resultierenden Mehrwert für den einzelnen Wissenschaftler,
- fehlende oder ungenügende Schnittstellen zum maschinellen Zugriff auf die Daten, um die Bereitstellung und Nachnutzung der Daten und Metadaten zu vereinfachen oder auch alternative Zugänge zu den in den Daten enthaltenen Informationen zu ermöglichen.

1 Vgl. Pilat, Dirk; Fukasaku, Yukiko: OECD Principles and Guidelines for Access to Research Data from Public Funding. In: Data Science Journal 6 (2007), S. OD4-OD11. https://www.jstage.jst.go.jp/article/dsj/6/0/6_0_OD4/_pdf (12.11.2014).

2 Charta zur Bewahrung des digitalen Kulturerbes: http://www.unesco.de/444.html (12.11.2014).

3 Vgl. Giaretta, David, u.a. (Hg.): Riding the wave – How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data.. 2010. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf (12.11.2014).

4 Vgl. Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten. Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Deutsche Forschungsgemeinschaft. 2009. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf (12.11.2014).

5 Vgl. Stodden, Victoria: Reproducible research for scientific computing: Tools and strategies for changing the culture. In: Computing in Science and Engineering 14,4 (2012), S. 13-17. http://dx.doi.org/10.1109/MCSE.2012.82 (12.11.2014).

6 Vgl. Hey, Tony; Tansley, Stewart; Tolle, Kristin (Hg.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, Washington: Microsoft Research, 2009. http://research.microsoft.com/en-us/collaboration/fourthparadigm/ (12.11.2014).

Das Projekt RADAR geht diese Herausforderungen durch die Etablierung einer generischen Infrastruktur für die Archivierung und Publikation von Forschungsdaten an.

2. Das Projekt RADAR

Im Rahmen der traditionellen wissenschaftlichen Grundlagenforschung kann der Aufbau, die Bereitstellung und ein dauerhafter Betrieb einer Datenarchivierungsinfrastruktur oftmals nicht realisiert werden. Sie muss vielmehr die Ziele, die wesentlichen Eigenschaften sowie ihre Anforderungen an ein nachhaltiges Datenmanagement definieren, um sich anschließend passenden Infrastrukturen zuzuwenden. Hier kommen die Träger der Informationsinfrastruktureinrichtungen wie z.B. Bibliotheken, Archive, Rechenzentren, spezialisierte Dienstleister oder auch Verlage ins Spiel.

Vor diesem Hintergrund haben drei Infrastruktureinrichtungen (FIZ Karlsruhe – Leibniz-Zentrum für Informationsinfrastruktur, die Technische Informationsbibliothek (TIB) in Hannover und das Steinbuch Centre for Computing (SCC) des KIT) gemeinsam mit zwei wissenschaftlichen Partnern (dem Department für Chemie der LMU München und dem Leibniz-Institut für Pflanzenbiochemie in Halle) das Projekt RADAR (Research Data Repository) im Themenfeld 4 („Forschungsnahe Informationsinfrastruktur") des 2012 veröffentlichen Förderprogramms der DFG zur „Neuausrichtung überregionaler Informationsservices" eingereicht. Ziel des Projekts ist die Etablierung eines Repositoriums zur Archivierung und Publikation von Forschungsdaten. Dieses soll als Basisdienstleistung für Forscher und wissenschaftliche Institutionen dienen. Das Projekt wurde offiziell im September 2013 gestartet und besitzt eine vorgesehene Laufzeit von drei Jahren mit einer Zwischenevaluation nach dem ersten Jahr.

Während der Projektlaufzeit wollen die Projektpartner eine Infrastruktur für die Datenarchivierung und -publikation (Datenzentrum) mit einem nachhaltigen und selbsttragenden Geschäftsmodell aufbauen. Der Aufbau und die Etablierung einer solchen Infrastruktur umfasst Werkzeuge und Prozesse, um Forschungsdaten:

- systematisch zu erschließen,
- dauerhaft in einem Datenarchiv zu bewahren und der Öffentlichkeit zugänglich zu machen,
- durch DOI-Vergabe verfügbar, zitierfähig und verlinkbar zu machen sowie
- für die Nachnutzung qualitätsgesichert bereitzustellen.[7]

RADAR verfolgt für sein vorgesehenes disziplinübergreifendes Dienstleistungsangebot einen zweistufigen Ansatz mit einem Einstiegsangebot zur Archivierung von Forschungsdaten und einem erweiterten Angebot mit integrierter Datenpublikation.

7 Vgl. Razum, Matthias; Neumann, Janna; Hahn, Matthias: RADAR – Ein Forschungsdaten-Repositorium als Dienstleistung für die Wissenschaft. In: Zeitschrift für Bibliothekswesen und Bibliographie 61,1 (2014), S. 18-27. http://dx.doi.org/10.3196/186429501461150 (12.11.2014). DOI: 10.3196/186429501461150.

Das Projekt gliedert sich in sieben Arbeitspakete (AP, siehe Abbildung 1). Das Projektmanagement in AP 1 nehmen das FIZ Karlsruhe und die TIB gemeinsam wahr. Die Anforderungsanalyse in AP 2 erfolgt durch die wissenschaftlichen Partner der LMU München und des Leibniz-Instituts für Pflanzenbiochemie (IPB) in Halle. Sie stellen sicher, dass der im Projekt konzipierte Dienst die Bedürfnisse der Forscher abdeckt. Gemeinsam mit der TIB erarbeiten die wissenschaftlichen Partner zudem allgemeine und fachspezifische Metadatenprofile (AP 3) für ausgesuchte Datentypen (NMR-Spektroskopie und 2D/DIGE-Bildern[8]). FIZ Karlsruhe und das Steinbuch Centre for Computing (SCC) des KIT konzipieren und implementieren die eigentliche Archivierungssoftware in AP 4. Die TIB kümmert sich in AP 5 um die Definition und Einführung notwendiger Prozesse für die Datenpublikation. Ein entscheidender Faktor für den Erfolg des Projekts ist die Etablierung eines auf Nachhaltigkeit angelegten Geschäftsmodells und der damit verbundenen Rahmenbedingungen. Diese Aufgabe übernehmen FIZ Karlsruhe und SCC in AP 6. Die Arbeitsergebnisse werden laufend durch die beiden wissenschaftlichen Partner in AP 7 evaluiert, um schon während der Projektlaufzeit die gewählten Ansätze auf ihre Tauglichkeit und Akzeptanz im Forschungsalltag zu prüfen.

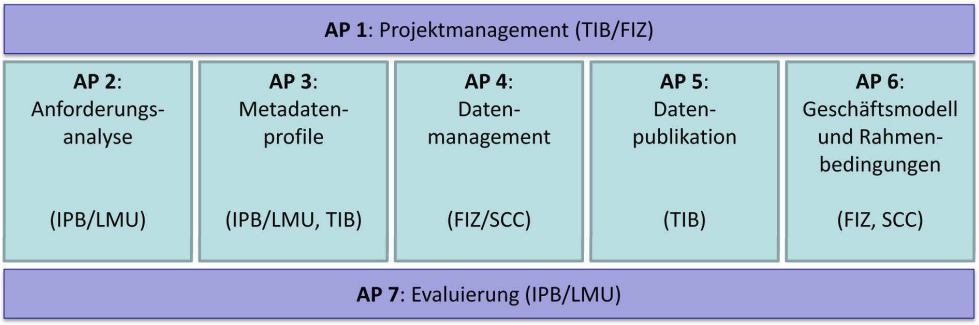

Abb. 1: Arbeitspakete im Projekt RADAR

3. Abgrenzung

RADAR versucht die genannten Herausforderungen in vielfältiger Weise anzugehen, kann und will aber nicht alle Probleme gleichzeitig lösen. Das Projekt fokussiert sich daher ganz bewusst auf einige Kernaufgaben und versucht sich ansonsten in die breitere und bereits etablierte (inter-)nationale Informationsinfrastruktur einzuordnen.

Forschungsdaten entstehen zu einem anderen Zeitpunkt als klassische Publikationen und durchlaufen einen eigenen „Lebenszyklus". Das Domänenmodell von Treloar und Harboe-Ree[9] sowie Klump[10] (siehe

8 NMR - "Nuclear Magnetic Resonance", eine spektroskopische Methode zur Untersuchung von Molekülen und Konzentrationsbestimmungen; 2D-/DIGE - "Differential in-Gel Electrophoresis", ein bildbasiertes Verfahren zur Proteinanalyse.

9 Vgl. Treloar, Andrew; Harboe-Ree, Cathrine: Data management and the curation continuum. How the Monash experience is informing repository relationships. Melbourne, 2008 (14th Victorian Association for Library Automation, Conference and Exhibition). http://arrow.monash.edu.au/hdl/1959.1/43940 (12.11.2014).

10 Vgl. Klump, Jens: Managing the Data Continuum, 2009. http://oa.helmholtz.de/fileadmin/user_upload/redakteur/Workshops/data_continuum_klump.pdf

Abbildung 2) beschreibt diesen Lebenszyklus und zeigt die mit den einzelnen Phasen verbundenen Prozesse im Forschungsdatenmanagement auf. Wissenschaftler/innen erzeugen und analysieren Forschungsdaten in der *privaten Domäne*. Zur Diskussion der Ergebnisse mit ausgewählten Kolleginnen und Kollegen innerhalb und außerhalb ihrer Institution machen sie diese – meist in bereits bearbeiteter Form – in der *kollaborativen Domäne* über geeignete Systeme eingeschränkt zugänglich. Mit der Veröffentlichung der Daten gehen diese in die öffentliche Domäne über, die für die Archivierung und langfristige Erhaltung sorgt. Damit verbunden ist die Überwindung der sogenannten *curation boundary*, bei der Daten selektiert und erschlossen werden müssen – intellektuelle Prozesse, die einen hohen Arbeits- und Kostenaufwand bedeuten. Die vierte Domäne schließlich erlaubt den Zugriff auf die archivierten Daten, z.B. über Fachportale oder virtuelle Forschungsumgebungen.[11]

Im Domänenmodell wird ersichtlich, dass mit einer zunehmenden Annäherung an den Arbeitsplatz der Wissenschaftlerinnen und Wissenschaftler (siehe Abbildung 2, kollaborative bzw. private Domäne), die Vielfalt der Forschungsprozesse und die Heterogenität der damit verbundenen Datentypen, Formate und Metadaten zunimmt. Daher ist es zielführend, RADAR zunächst in der dritten, öffentlichen Domäne anzusiedeln, da dies der einzige Bereich ist, in welchem sich eine generische, disziplinübergreifende Dienstleistung etablieren lässt. Gleichzeitig wirkt ein verlässliches und auf Dauerhaftigkeit ausgelegtes Datenarchiv dem Defizit einer fehlenden Infrastruktur entgegen. Mit RADAR soll somit eine wesentliche Grundlage für die Nachnutzung und Publikation von Forschungsdaten geschaffen werden, die einen erheblichen zusätzlichen Nutzen für Forschende (auch in der forschenden Industrie), Wissenschaft und Gesellschaft schafft.

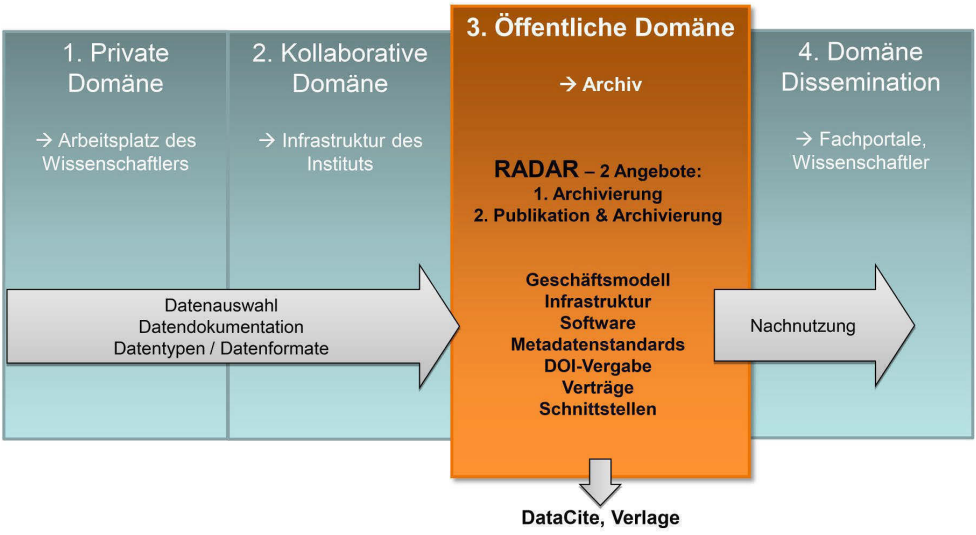

Abb. 2: Lebenszyklus von Forschungsdaten

11 Vgl. dazu Razum; Neumann; Hahn (wie Anm. 7).

Die sogenannten big data-Disziplinen wie Astronomie, Hochenergiephysik oder Erdwissenschaften mussten sich bereits frühzeitig mit der Etablierung und dem Betreiben von Dateninfrastrukturen befassen. Daraus resultierten sowohl international wie auch national etablierte Beispiele für Forschungsdaten-Repositorien, wie etwa das World Data System der International Council of Science (ICSU)[12] mit seinen mehr als 50 weltweit verteilten disziplinspezifischen Datenzentren oder GESIS – Leibniz-Institut für Sozialwissenschaften[13] mit seinen archivierten Studien und empirischen Primärdaten aus den Sozialwissenschaften. Bei anderen Disziplinen fehlte bisher – auch aufgrund der noch vergleichsweise geringen Datenmengen –der Druck, sich mit derartigen Infrastrukturen auseinanderzusetzen. Man spricht hier auch vom *long tail of science*.[14] Mit dem Fokus auf diesen Disziplinen will RADAR nicht in Konkurrenz zu etablierten Datenzentren treten, sondern vielmehr die verbleibenden Lücken schließen helfen.

Die enorme Vielzahl der Daten- und Metadatenformate stellt für die langfristige Nachnutzbarkeit der Daten, insbesondere für die funktionale Langzeitarchivierung, eine große Herausforderung dar.[15] Die Beobachtung der sich entwickelnden Technologie der digitalen Langzeitarchivierung und der Anforderungen der relevanten Zielgruppen haben einen großen Einfluss auf notwendige *Preservation Policies* und Datenmanagementpläne.[16] Viele dieser Fragestellungen sind zurzeit noch Gegenstand der Forschung.[17] Um das Projekt nicht zu überfrachten, verzichtet RADAR vorerst bewusst auf eine funktionale Langzeitarchivierung der Daten, beobachtet aber relevante Projekte[18] und Initiativen[19] bzw. plant entsprechende Schnittstellen oder Kooperationsmöglichkeiten ein.

4. Zielgruppen

Auch aufgrund seines Angebotsmodells richtet sich RADAR an verschiedene Zielgruppen. *Wissenschaftler/innen* sollen mit RADAR die Ergebnisse ihrer projektbezogene Forschung einfach archivieren und publizieren können. Diese Zielgruppe gibt damit auch eine wichtige Anforderung für die Erarbeitung eines Geschäftsmodells (AP 6) vor: Da Projekte begrenzte Laufzeiten haben, muss es möglich sein, schon bei der Antragsstellung die Kosten für die dauerhafte oder auch zeitlich begrenzte (im Sinne von empfohlenen Haltefristen) Datenarchivierung zu kalkulieren und dann im Projektverlauf über eine Einmalzahlung abzugelten.

12 http://www.icsu-wds.org/
13 GESIS - Leibniz-Institut für Sozialwissenschaften, http://www.gesis.org/
14 Vgl. Borgman, Christine L.: The Conundrum of Sharing Research Data. In: Journal of the American Society for Information Science and Technology 63,6 (2012), S. 1059–1078.
15 Riley, Jenn: Seeing Standards. A Visualization of the Metadata Universe. http://www.dlib.indiana.edu/~jenlrile/metadatamap/ (12.11.2014).
16 Vgl. Neuroth, Heike, u.a. (Hg.): Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme. Göttingen, 2012. http://nbn-resolving.de/urn:nbn:de:0008-2012031401 (12.11.2014).
17 Vgl. Becker, Christoph: Vertrauenswürdige Planung in der digitalen Langzeitarchivierung. In: Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare 64,2 (2011), S. 233-246.
18 Z.B. SCAPE – Scalable Preservation Environments (http://www.scape-project.eu/) oder bwFLA (http://bw-fla.uni-freiburg.de/)
19 Z.B. NESTOR (http://www.langzeitarchivierung.de/)

Institutionelle Nutzer wie z.B. Bibliotheken dagegen sind eher an einer jährlichen Zahlungsweise interessiert, da sie sich nicht für sehr lange Zeit an den von ihnen beauftragten Dienstleister binden wollen. Auch sind so zu Beginn der Inanspruchnahme einer Archivierungsdienstleistung keine hohen Einmalzahlungen erforderlich, da sich die Kosten auf viele Jahre verteilen. Das kann insbesondere für Einrichtungen interessant sein, die bereits umfangreiche Datenbestände vorhalten. Wichtig für institutionelle Nutzer ist die Möglichkeit, RADAR „unsichtbar" in ihre eigenen Portale einzubinden und so weiterhin als Dienstleister und Ansprechpartner innerhalb ihrer Organisation wahrgenommen zu werden. Auch kann eine Bibliothek so die Funktionalität des Portals an die lokalen Gegebenheiten anpassen.

Auch *Kultureinrichtungen* wie Archive oder Museen können die Dienste von RADAR in Anspruch nehmen, um die im Rahmen von Digitalisierungsvorhaben entstehenden Digitalisate in Form von Masterdateien dauerhaft vorzuhalten und Zugriffskopien anzubieten, ohne die dafür notwendige Infrastruktur selbst vorhalten zu müssen.

Schließlich haben auch *wissenschaftliche Verlage* ein Interesse an der Verknüpfung traditioneller Publikationen mit den zugrundeliegenden Daten. Mit dem Georg Thieme Verlag wurde deshalb auch ein wissenschaftlicher Verlag in das Projektkonsortium als assoziierter Partner aufgenommen, um hier mögliche Kooperationsformen zu erarbeiten und zu evaluieren.

5. Bisher erreichte Ergebnisse

In den folgenden Abschnitten werden die im bisherigen Projektverlauf (im Zeitraum vom 01.09.2013 bis zum 31.05.2014) erzielten Arbeitsergebnisse der jeweiligen Arbeitspakete dargestellt. Auf die detaillierte Darstellung von AP 1: Projektmanagement und AP 7: Evaluierung wird verzichtet.

AP 2: Anforderungsanalyse

In einem ersten Schritt erfolgte eine Analyse bestehender Prozesse für Sammlung und Registrierung von Forschungsdaten bei den wissenschaftlichen Partnern. Diese wurde anschließend um fachwissenschaftliche Anforderungen angrenzender Disziplinen erweitert, um zu möglichst disziplinübergreifenden Anforderungen zu kommen. Als Teil dieser Analyse wurden auch bestehende Infrastrukturen (u. a. ZENODO[20], figshare[21], Dryad[22]) evaluiert. Dabei wurden nicht nur deren Leistungsumfang, sondern auch deren Geschäftsmodelle berücksichtigt und geprüft, inwieweit Teile davon für RADAR nachgenutzt werden können bzw. eine Anbindung oder Kooperation sinnvoll erscheint.

Im Rahmen der Analyse zeigte sich, dass RADAR durch seine generische Ausrichtung, die geplante zweistufige Servicestruktur und die Archivierung und Publikation innerhalb des deutschen Rechtsrahmens als wichtige Ergänzung eingestuft wird.

20 http://zenodo.org/
21 http://figshare.com/
22 http://datadryad.org/

AP 3: Metadatenprofile

Im Arbeitspaket 3 wurde in Zusammenarbeit mit der TIB von den wissenschaftlichen Partnern ein allgemeines sowie fachspezifisches Metadatenschema erarbeitet.

Das allgemeine Metadatenschema, für das als Basis das DataCite Metadata Schema v3.0[23] diente, soll den interdisziplinären, zentralen Nachweis der in RADAR archivierten und publizierten Forschungsdaten erlauben, wohingegen das fachspezifische Metadatenschema die disziplinspezifischen Anforderungen zur Suche und zur Nachnutzung von Forschungsdaten erfüllen soll. Exemplarisch ausgewählt wurden geeignete Metadatenparameter für fachspezifische NMR und 2D-/DIGE-Daten.

Das erstellte Schema umfasst die Definition von neun Pflichtfeldern, welche zusammen den allgemeinen, deskriptiven Teil des Metadatenprofils bilden, sowie die Definition, Abstimmung und Anpassung von zwölf optionalen Feldern, welche die fachspezifischen Beschreibungen der Datensätze abbilden. Die allgemeinen Pflichtfelder enthalten die Grundanforderungen für eine DOI-Registrierung nach dem DataCite Metadatenschema.[24]

Nach der Fertigstellung des entwickelten Metadatenprofils (Version 0.1) wurde die Umsetzung in eine XML Schema Definition (XSD) gestartet. Es wurde prototypisch eine Eingabemaske implementiert, die die Darstellung des Metadatenprofils für den RADAR-Nutzer veranschaulicht. Die Umsetzung dient zudem als Testservice für die wissenschaftlichen Partner zur Metadateneingabe und Datenspeicherung.

Die Entwicklung eines Glossars soll dazu dienen, den Wissenschaftler/innen exemplarische Beispiele (zunächst für NMR- und 2D-/DIGE- Analysen) für eine nachhaltige Beschreibung eines detaillierten Datensatzes aufzuzeigen. Das fachspezifische Glossar soll, gemäß dem generischen Grundsatz, sukzessive beim Ausbau von RADAR erstellt und auch um neu aufgenommene Fachgebiete (z.B. Materialwissenschaft und Werkstofftechnik) erweitert werden.

AP 4: Datenmanagement

Die Voraussetzungen für die Publikation von Forschungsdaten und deren Referenzierung sind eine verlässliche und dauerhafte Archivierung und persistente Identifizierung. Die Publikationsmöglichkeit ist ein wichtiges Element einer disziplinübergreifenden Informationsinfrastruktur. Damit dieses Element auch mit anderen Diensten (etwa zur Datenerfassung, zur (teil-)automatisierten Anreicherung mit Metadaten oder zur Dissemination) verwendet werden kann, muss die Systemarchitektur des Datenzentrums offen sein und über geeignete Programmierschnittstellen (*application programming interfaces*, API) verfügen.

Grundsätzlich unterscheidet die Systemarchitektur zwischen einem Verwaltungsteil und der eigentlichen Datenhaltung. Ersterer implementiert die Benutzungsoberfläche für die Anwender und

23 Vgl. DataCite e.V. (Hg.): DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.0. July 2013. http://schema.datacite.org/meta/kernel-3.0/doc/DataCite-MetadataKernel_v3.0.pdf (12.11.2014).

24 http://schema.datacite.org/

erlaubt die Steuerung des Gesamtsystems. Hier werden Forschungsdaten zu größeren Einheiten zusammengefasst und paketiert, um sie dann in Form eines SIP gemäß dem OAIS-Referenzmodell[25] an die Datenhaltungskomponente durchzureichen. Diese wird in der ersten Version von RADAR nur einmal existieren, ist aber so ausgelegt, dass zukünftig weitere Rechenzentren diese Aufgabe übernehmen und so eine geografische Replikation der archivierten Daten ermöglichen können.

Bisher wurden ca. 50 Anwendungsfälle (*use cases*) für das Gesamtsystem definiert, vollständig beschrieben und evaluiert. Für die zentralen Anwendungsfälle liegen Wireframes vor, also eine skizzenhafte Darstellung der späteren Benutzungsoberfläche. Als nächstes steht das Interaction Design und anschließend das Webdesign an, bevor die eigentliche Implementierung beginnt. Eine auf der NoSQL-Datenbank ElasticSearch[26] aufsetzende Infrastruktur zur Speicherung der Verwaltungsdaten ist inzwischen weitgehend fertig gestellt und befindet sich im Test.

Für die Verbindung zwischen der Verwaltungs- und Datenhaltungsschicht wurde eine Schnittstelle definiert und implementiert. Die Architektur für die eigentliche Datenhaltung ist inzwischen weit fortgeschritten (siehe Abbildung 3) und erste Tests der Datenübernahme zwischen Verwaltungs- und Datenhaltungsschicht konnten erfolgreich durchgeführt werden.

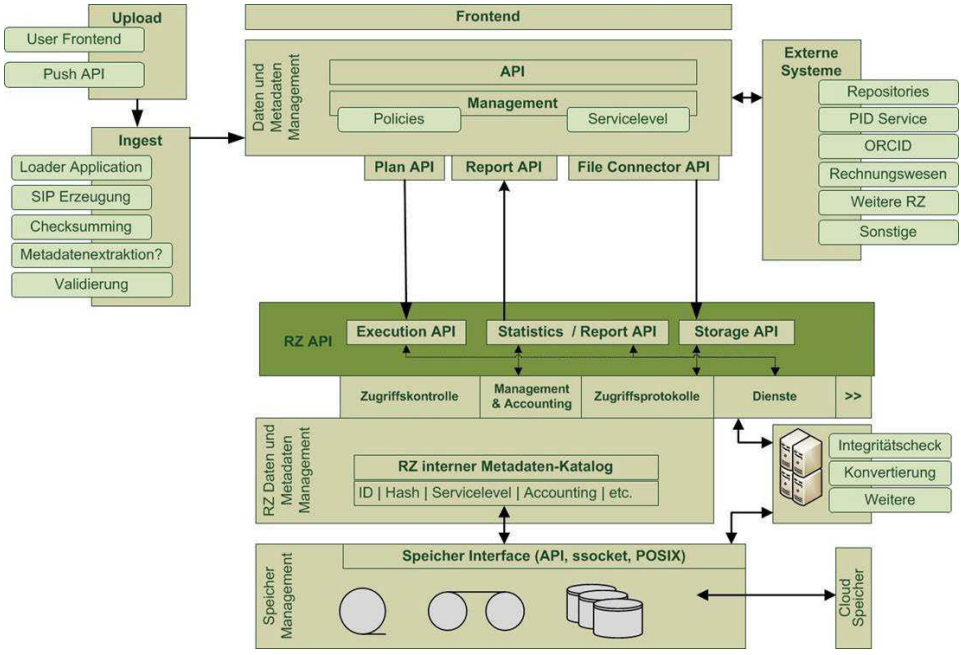

Abb. 3: Schematische Architektur von RADAR mit verteilter Datenhaltung und Schnittstellen (Quelle: Jan Potthoff, KIT/SCC)

25 Vgl. Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System (OAIS). Washington, DC: National Aeronautics and Space Administration, 2002.

26 http://www.elasticsearch.org/

AP 5: Datenpublikation

Die Veröffentlichung von Forschungsdaten sollte ein integraler Bestandteil des Forschungsprozesses sein[27], jedoch existieren derzeit kaum Standards, die die Publikationsworkflows von Forschungsdaten eindeutig beschreiben. Zwei elementare Anforderungen müssen hierfür jedoch erfüllt sein: Die eindeutige und persistente Identifizierbarkeit und ein verlässlicher dauerhafter Zugriff.

Die an das RADAR Forschungsdatenarchiv angeschlossene DOI-Registrierung ermöglicht die Vergabe von eindeutigen und persistenten Identifikatoren (DOI-Namen) für Forschungsdaten und damit eine eindeutige Referenzierbarkeit. Die so eröffnete Möglichkeit zur Zitierung von Forschungsdaten erhöht damit nicht nur die Anerkennung der wissenschaftlichen Datenproduzenten in ihrer Fachcommunity, sondern auch die Sichtbarkeit und damit die Verfügbarkeit ihrer Forschungsergebnisse.[28]

In diesem Arbeitspaket werden Workflows zu den verschiedenen Angebotsmodellen formuliert. Das Einstiegsangebot zeichnet sich durch formatunabhängige Archivierung, Bitstream Preservation sowie dem generischen Pflichtmetadatensatz aus. Dieses Angebot richtet sich an Kunden, die primär an der Einhaltung von empfohlenen Haltefristen[29] interessiert sind. Darüber hinaus eignet es sich aber auch für andere Daten wie z.B. Negativdaten[30], die etwa im Rahmen von weiterführenden Analysen von hohem Interesse sein können.

Die zweite, *erweiterte Angebotsstufe* zielt auf eine dauerhafte Datenarchivierung mit Datenpublikation ab. Hier ist die Vergabe von format- und disziplinspezifischen Metadaten sowie von dauerhaften DOI-Namen in den Publikationsprozess implementiert. In einem ersten Entwurf wurden für die beiden Angebotsstufen drei Workflows mit angepassten Kundenprofilen definiert:

- Der Entscheidungsprozess zwischen Basisangeboten Archivierung oder Publikation.
- Die angebotenen Varianten der Datenpublikation (direkte Publikation, Publikation mit zeitlichem Embargo, Publikation im Rahmen einer Verlagskooperation).
- Die für 2015/16 vorgesehene Ausbaustufe mit der Übertragung bereits archivierter Daten (Basisangebot) in die erweiterte Angebotsstufe zur Datenpublikation.
  - Anbindung DOI-Service durch Schnittstellenbeschreibung
  - DOI-Zuweisung im Angebot Datenpublikation
  - Entwurf Nutzungs- und Datenschutzbedingungen für RADAR

Neben Workflows, die in den Forschungsalltag wissenschaftlicher Institutionen integriert werden können, ist für die Akzeptanz des Datenarchivs in den Fachcommunities weiterhin eine eindeutige und transparente Definition der Verantwortlichkeiten der verschiedenen Akteure unabdingbar. Nur

27 Vgl. FIZ Chemie; TIB Hannover; Universität Paderborn: Konzeptstudie Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie. 2010. http://www.tib-hannover.de/fileadmin/projekte/primaer-chemie/Konzeptstudie_Forschungsdaten_Chemie.pdf (12.11.2014).

28 Vgl. Kotarski, Rachael, u.a.: Report on Best Practices for Citability of Data and Evolving Roles in Scholarly Communication. 2012. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-ReportBestPracticesCitabilityDataEvolvingRolesScholarlyCommunication.pdf (12.11.2014).

29 Vgl. Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten (wie Anm. 4)

30 Daten, die zu keinem oder nicht dem gewünschten Ergebnis geführt haben.

so kann die Qualität der archivierten Daten und damit auch die des Datenzentrums sichergestellt werden. Dies beinhaltet vor allem die kooperative Entwicklung von Policies zur Datenerhaltung zwischen den Akteuren der Datenproduktion und Datenarchivierung. Daher wurde in diesem AP ebenfalls ein Entwurf von Autorenrichtlinien für das geplante Produkt erstellt, welcher z.B. konkrete Formatempfehlungen, Lizenzmodelle und Zitierstandards umfasst. Potentiellen RADAR-Kunden soll so die Archivierung und Publikation ihrer Daten erleichtert werden. Gleichzeitig wird Wissenschaftler/inne/n mit solchen Richtlinien und Empfehlungen das Fachwissen vermittelt, um archivierungswürdige, qualitativ hochwertige Daten auszuwählen und somit als nationales Kulturgut langfristig zu erhalten und nachhaltig auch über institutionelle Grenzen hinweg verfügbar zu machen.[31]

AP 6: Geschäftsmodell

Die Erarbeitung eines sich selbst tragenden, nachhaltigen Geschäftsmodells für den Betrieb von RADAR ist eine zentrale Aufgabe des Projekts. Dabei sind unterschiedliche Aspekte zu berücksichtigen:

- die Festlegung der Zielgruppen und Analyse ihrer Anforderungen,
- eine genaue Beschreibung der angebotenen Dienstleistungen,
- die Analyse der Kostenfaktoren für den Aufbau des Repositoriums und insbesondere für den laufenden Betrieb,
- die Untersuchung möglicher Einnahmequellen (z.B. Unterstützung durch Drittmittel),
- die Definition des Betreibermodells (z.B. Anbieter der Dienstleistung, Rechtsform, usw.),
- eine Festlegung unterstützter Zahlungsmodelle und Preisfindung.

Die vorgesehenen Zielgruppen wurden gegenüber dem Antrag ausgeweitet, um ein zukünftiges Geschäftsmodell auf eine breitere Basis zu stellen und weitere Einnahmequellen zu erschließen. Sie umfassen neben Wissenschaftler/innen auch institutionelle Nutzer, Kultureinrichtungen sowie wissenschaftliche Verlage und wurden bereits weiter oben im Abschnitt Zielgruppen näher mit ihren spezifischen Bedürfnissen beschrieben. Dabei wurde auch die Abgeltung der erbrachten Archivierungsdienstleistung in Form einer Einmalzahlung erwähnt. Für die Kalkulation eines solchen Angebots muss man die notwendigen Aufgaben und Verfahren vollständig verstanden haben. Während das im Fall der funktionalen Langzeitarchivierung im Bereich von Forschungsdaten mit ihrer großen Formatvielfalt und Heterogenität noch nicht der Fall ist, sieht es bei der reinen Bitstream Preservation anders aus.

Die retrospektiven Betrachtung der Kosten für Speichersysteme[32] über die letzten zehn Jahre zeigt, dass die Hardware-Kosten für notwendige Speicher- und Rechenkapazität pro Speichereinheit (also z.B. pro Terabyte) kontinuierlich sinken, die Administrationskosten für diese Systeme hingegen weitgehend stabil blieben und die Kosten für Wartungsverträge und Software-Lizenzen moderat steigen. Die regelmäßige Überprüfung der Datenintegrität und das zugehörige Reporting kann weitgehend automatisiert werden. Als relatives Restrisiko bleiben die steigenden Energiekosten,

31 Vgl. Sustainable Economics for a Digital Planet - Ensuring Long-Term Access to Digital Information. Final Report. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (12.11.2014).

32 Dies betrifft sowohl Disk- als auch Tape-basierte Speichersysteme. Eine Differenzierung nach Serviceklassen (etwa Verlässlichkeit, Geschwindigkeit o.ä.) fand dabei nicht statt.

die jedoch pro Speichereinheit über die Jahre hinweg aufgrund von Effizienzgewinnen bei neueren Hardwaregenerationen abnehmen.

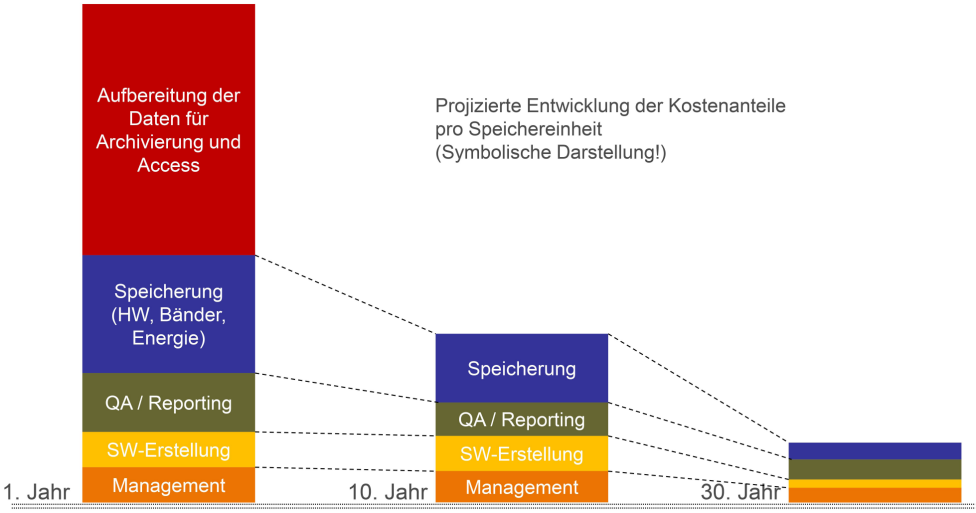

Abb. 4: Schematische Darstellung der Kostenentwicklung für Bitstream Presevation

Die meisten Kosten entstehen bei der Überwindung der von Treloar und Harboe-Ree beschriebene „Curation Boundary“[33] zur Aufnahme neuer Daten in das Archiv (siehe roten Block in Abbildung 4), die jedoch nur einmalig anfallen. Damit hängen die Kosten der Bitstream Preservation weitgehend von der gespeicherten Datenmenge ab. Durch Extrapolation der bekannten Kostenstrukturen lässt sich eine relativ zuverlässige Prognose für die Kostenentwicklung der nächsten zehn oder zwanzig Jahre abgeben.[34] Bei einer darüber hinausgehenden Speicherdauer werden die Kosten weitgehend marginalisiert. Die benötigte Rechenzentrumsleistung für jetzt eingestellte Datenmengen wird in dreißig bis fünfzig Jahren nur noch geringe, im Vergleich zu heute kaum mehr ins Gewicht fallende Kosten verursachen.

Aktuell wird an der Erfassung der Kostenstrukturen gearbeitet. Im geplanten RADAR-Webservice soll es dem späteren Kundenkreis ermöglicht werden, auf der Grundlage der voraussichtlichen Datenmenge und der gewünschten Serviceleistungen die anfallenden Datenmanagement- und Speicherkosten im Rahmen eines Kostenvoranschlags darzustellen. Diese Kostengrundlage kann dann bei der weiteren Planung des Datenmanagements berücksichtigt und beispielsweise im Rahmen von Projekt- und Drittmittelanträgen aufgeführt werden. Weiterhin steht in diesem AP die Analyse zu möglichen Betriebsmodellen und Rechtsformen im Fokus.

33 Vgl. Treloar, Andrew; Harboe-Ree, Cathrine (wie Anm. 9).

34 Vgl. Beagrie, Neil; Chruszcz, Julia; Lavoie, Brian: Keeping Research Data Safe - A Cost Model and Guidance for UK Universities. Final Report. s.l. : JISC, 2008, S. 4-6. http://webarchive.nationalarchives.gov.uk/20131202191249/http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf(12.11.2014).

6. Zusammenfassung und Ausblick

RADAR zielt auf den Aufbau und die Etablierung eines Basisangebots zur nachhaltigen Forschungsdatenarchivierung und eines erweiterten Angebots für zitierfähige Datenpublikationen und deckt damit einen essentiellen Teilaspekt zur Unterstützung im überregionalen Forschungsdatenmanagement ab. Die erste Ausbaustufe beschränkt sich inhaltlich ganz bewusst auf diesen pragmatischen Ansatz, um eine verlässliche Grundlage in der (inter-)nationalen Informationsinfrastruktur zu entwickeln und bereitzustellen. Die Projektpartner mit ihren spezifischen Kompetenzen tragen dabei auch zur Entwicklung und Etablierung eines nachhaltigen Dienstes bei.

Die offene Architektur des Systems ermöglicht über modifizierbare Schnittstellen und durch Andockung von Systemen Dritter eine breite Verteilung der generischen Infrastruktur. Die transparente Kostenstruktur erlaubt ein kalkulierbares und damit auch zu (re-)finanzierendes Forschungsdatenmanagement der Wissenschaftler/innen und/oder Institutionen.

RADAR hilft somit Ressourcen zu sparen, über Skaleneffekte die Kosten für die Datenarchivierung zu verringern und die umfangreichen und komplexen Herausforderungen des Forschungsdatenmanagements anzugehen.

Literaturverzeichnis

- Beagrie, Neil; Chruszcz, Julia; Lavoie, Brian. Keeping Research Data Safe - A Cost Model and Guidance for UK Universities. Final Report. s.l. : JISC, 2008. http://webarchive.nationalarchives.gov.uk/20131202191249/http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf (12.11.2014).

- Becker, Christoph: Vertrauenswürdige Planung in der digitalen Langzeitarchivierung. In: Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare 64,2 (2011), S. 233-246.

- Borgman, Christine L.: The Conundrum of Sharing Research Data. In: Journal of the American Society for Information Science and Technology 63,6 (2012), S. 1059–1078.

- Consultative Committee for Space Data Systems (CCSDS): Reference Model for an Open Archival Information System (OAIS). Washington, DC: National Aeronautics and Space Administration, 2002.

- DataCite e.V. (Hg.): DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.0. July 2013. http://schema.datacite.org/meta/kernel-3.0/doc/DataCite-MetadataKernel_v3.0.pdf (12.11.2014).

– Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten. Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Deutsche Forschungsgemeinschaft. 2009. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf (12.11.2014).

– FIZ Chemie; TIB Hannover; Universität Paderborn: Konzeptstudie Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie. 2010. http://www.tib-hannover.de/fileadmin/projekte/primaer-chemie/Konzeptstudie_Forschungsdaten_Chemie.pdf (12.11.2014).

– Giaretta, David, u.a. (Hg.): Riding the wave – How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. 2010. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf (12.11.2014).

– Hey, Tony; Tansley, Stewart; Tolle, Kristin (Hg.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, Washington: Microsoft Research, 2009. http://research.microsoft.com/en-us/collaboration/fourthparadigm/ (12.11.2014).

– Klump, Jens: Managing the Data Continuum, 2009. http://oa.helmholtz.de/fileadmin/user_upload/redakteur/Workshops/data_continuum_klump.pdf (12.11.2014).

– Kotarski, Rachael, u.a.: Report on Best Practices for Citability of Data and Evolving Roles in Scholarly Communication. 2012. http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-ReportBestPracticesCitabilityDataEvolvingRolesScholarlyCommunication.pdf (12.11.2014).

– Neuroth, Heike, u.a. (Hg.): Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme. Göttingen, 2012. http://nbn-resolving.de/urn:nbn:de:0008-2012031401 (12.11.2014).

– Pilat, Dirk; Fukasaku, Yukiko: OECD Principles and Guidelines for Access to Research Data from Public Funding. In: Data Science Journal 6 (2007), S. OD4-OD11. https://www.jstage.jst.go.jp/article/dsj/6/0/6_0_OD4/_pdf (12.11.2014).

– Razum, Matthias; Neumann, Janna; Hahn, Matthias: RADAR – Ein Forschungsdaten-Repositorium als Dienstleistung für die Wissenschaft. In: Zeitschrift für Bibliothekswesen und Bibliographie 61,1 (2014), S. 18-27. http://dx.doi.org/10.3196/186429501461150 (12.11.2014).

– Riley, Jenn: Seeing Standards. A Visualization of the Metadata Universe. http://www.dlib.indiana.edu/~jenlrile/metadatamap/ (12.11.2014).

– Stodden, Victoria: Reproducible research for scientific computing: Tools and strategies for changing the culture. In: Computing in Science and Engineering 14,4 (2012), S. 13-17. http://dx.doi.org/10.1109/MCSE.2012.82 (12.11.2014).

– Sustainable Economics for a Digital Planet - Ensuring Long-Term Access to Digital Information. Final Report. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (12.11.2014).

– Treloar, Andrew; Harboe-Ree, Cathrine: Data management and the curation continuum. How the Monash experience is informing repository relationships. Melbourne, 2008 (14th Victorian Association for Library Automation, Conference and Exhibition). http://arrow.monash.edu.au/hdl/1959.1/43940 (12.11.2014).