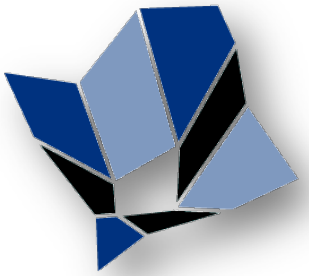

Projekt RADIESCHEN

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

Report „Synthese“

Entspricht dem Report D6.3
„Abschlussbericht des Projekts und Roadmap für eine Infrastruktur für Forschungsdaten in Deutschland“ nach Projektantrag

Inhalt

1. Einleitung

Weltweit sieht sich die Wissenschaft zunehmend mit der Herausforderung einer stetig ansteigenden Masse von Forschungsdaten konfrontiert. Diese Lawine der Daten ist ein Datenstrom, generiert aus Sensoren und wissenschaftlichen Instrumenten, aus digitalen Aufzeichnungen, aus sozialwissenschaftlichen Erhebungen oder auch bezogen aus dem World Wide Web.

Dieser rasante Zuwachs an Daten betrifft alle Wissenschaftszweige, ob nun die Archäologie mit der Protokollierung von Ausgrabungen, die Astrophysik mit Teleskopdaten aus der Beobachtung ferner Galaxien oder auch die Sozialwissenschaften mit Daten aus Umfragen und Erhebungen. Die Herausforderung für die Wissenschaft liegt nicht nur in der Bearbeitung der Daten durch Analyse, Reduktion und Visualisierung, sondern auch in dem Aufbau von Infrastrukturen zu deren Bereithaltung und Verwahrung.

Die nachfolgende Abbildung zeigt den Anstieg der Verwendung des Begriffs "Big Data" im Internet. "Big Data" wird zum Schlagwort ab ca. 2012 und ist ein hochaktuelles Thema.

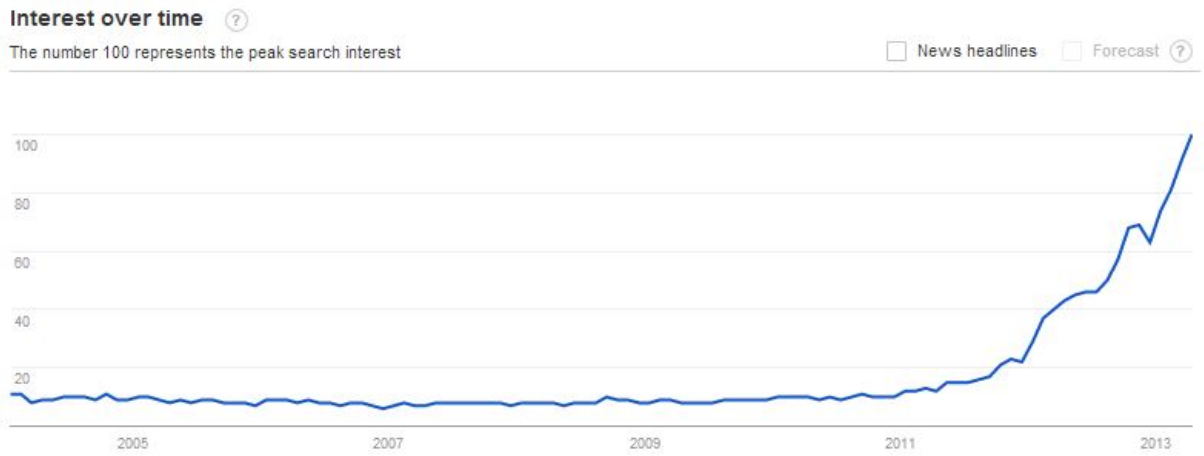

Abb. 1.: Google Trends – die Verwendung des Begriffs „Big Data“ bei Suchen im Internet vom Jahr 2004 bis heute. Die Kurve umfasst alle Anfragen weltweit und zeigt einen rasanten Anstieg ab Beginn des Jahres 2012[1].

Forschungsdaten sind jedoch nicht automatisch gleichzusetzen mit "Big Data". Kleinere Datensätze, wie z.B. die tägliche Beobachtung einer lokalen Wetterstation, sind ebenfalls Forschungsdaten, welche es verdienen, analysiert, verwahrt und im Hinblick auf eine mögliche Wiederverwendung annotiert zu werden. Diese kleinen Datensätze stellen sogar den größten Teil der verfügbaren Forschungsdaten dar. Forschungsdaten-Infrastrukturen und deren zugehörige Dienste und Werkzeuge sind demzufolge sehr unterschiedlich und auf die Behandlung dieser disziplin-spezifischen Datensätze ausgerichtet.

Das Projekt "Rahmenbedingungen einer disziplinübergreifenden Forschungsdaten-Infrastruktur (Radieschen)" stellt nun die Frage nach Anforderungen an generische Komponenten einer Infrastruktur und deren Vernetzung mit disziplin-spezifischen Bestandteilen. Grundlage bildet die Bestandsaufnahme bestehender Systeme und Infrastrukturen und deren Analyse im Hinblick auf Fragen der Organisationsstruktur, der eingesetzten Technologie und der entstehenden Kosten. Querschnittsthemen, wie das Wertesystem wissenschaftlicher Veröffentlichungen oder die Rolle der

[1] http://www.google.com/trends/

sozialen Medien in der Wissenschaft, spielen ebenso eine Rolle wie der Trend zur Auslagerung von Diensten an Service-Einrichtungen und Rechenzentren.

Der vorliegende Report gibt einen Überblick über die Ergebnisse des Projekts „Radieschen“. Zunächst erfolgt eine Einordnung des Projekts und der Projektziele in die deutsche und europäische Forschungslandschaft (Kapitel 2). Das darauffolgende Kapitel behandelt Zukunftsszenarien, in denen eine mögliche Entwicklung der Wissenschaftswelt in Deutschland im Jahre 2020 beschrieben wird. Kapitel 4 fasst die Ergebnisse der Arbeitspakete Kosten, Organisation und Technik zusammen, zeigt Handlungsempfehlungen auf und gibt einen Ausblick zu dem jeweiligen Thema.

Ein wichtiger Bestandteil des Radieschen-Projekts war die Interaktion mit der Forschungsdaten-Community. Kapitel 5 fasst die Erkenntnisse aus den während der Bestandsaufnahme durchgeführten Interviews zusammen und gibt einen Überblick über die Diskussionen, welche im Verlauf der Radieschen Workshops und des Forschungsdaten-Symposiums (FDI 2013) stattfanden. Kapitel 6 behandelt Querschnittsthemen, die zwar nicht Hauptbestandteil der Untersuchung waren, jedoch im Projektverlauf immer wieder zur Sprache kamen und diskutiert wurden. Der Report schließt mit einem Ausblick und gibt Empfehlungen für eine weitere Entwicklung von Forschungsdaten-Infrastrukturen.

2. Einordnung

eScience, e-Infrastructure, Research Data Management, Virtual Research Environments – diese Stichworte fallen oft im Kontext von Forschungsdaten und deren Infrastrukturen. Demzufolge gibt es eine Vielzahl von Projekten und Institutionen, die sich weltweit mit diesen Themen befassen.

Abbildung Abb. 2 zeigt eine Übersicht über **Large-Scale Research Infrastructures**, gefördert durch die EU ESFRI-Initiative[2]. Der Begriff "Research Infrastructures" ist hier sehr weit gefasst und beinhaltet Einrichtungen, Betriebsmittel und damit verbundene Dienste, die der wissenschaftlichen Community der verschiedensten Disziplinen zur Verfügung stehen. Als Beispiel sei hier Géant[3] genannt, welches als High-Speed Network die Kooperation, sowie das Teilen von Wissen und Ressourcen zwischen Forschern erleichtern soll. Géant zählt zu einem Projekt der e-Infrastructures Initiative der EU Kommission. Das Projekt Radieschen betrachtet einen Ausschnitt aus diesem recht großen Themengebiet und setzt seinen Fokus gezielt auf Forschungsdaten und den dazu benötigten Infrastrukturen in Deutschland. Damit ist ein genauerer Blick auf die tatsächlichen Gegebenheiten und Bedürfnisse der Forscher vor Ort in Deutschland möglich.

[2] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what
[3] http://www.geant.net/Pages/default.aspx

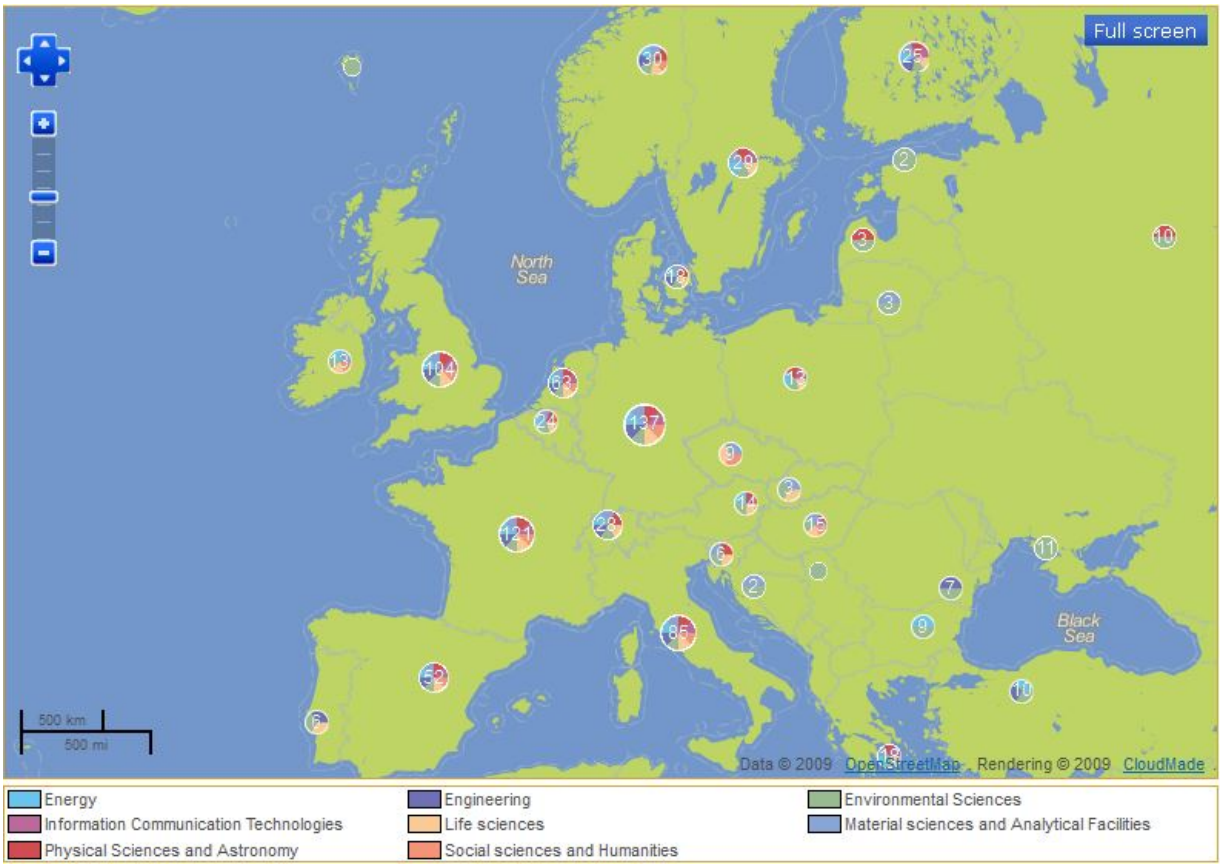

Abb. 2: Übersicht über aktuell bestehende Large-Scale Research Infrastructures in Europa[4]

Das **EU-Projekt EUDAT**[5] ähnelt in seiner Ausrichtung dem Projekt Radieschen. Auch hier steht die Herausforderung der Schaffung eines „cross-disciplinary data-services" im Vordergrund. Schwerpunkte der Forschung und Entwicklung bei EUDAT sind die Schaffung eines pan-europäischen Daten-Services, welches darauf abzielt, viele verschiedene wissenschaftliche Communities zu unterstützen.

EUDAT sieht die Heterogenität der Daten als Startpunkt, berücksichtigt jedoch gleichzeitig die Integration der Daten durch gemeinschaftlich nutzbare Lösungen und Dienste. Für die zu schaffende Collaborative Data Infrastructure (CDI) bedarf es einer abstrakten Architektur, die es erlaubt, bereits existierende Daten-Lösungen mit Datenzentren zu integrieren, welche gemeinsame Datendienste unterstützen. Ähnlich wie Radieschen befasst sich EUDAT mit der Wiederverwertung von Daten, Metadaten-Lösungen, Persistenten Identifikatoren und speziellen Lösungen für alle Arten von Daten, auch der sogenannten "Small Data".[6]

Im Unterschied zu EUDAT untersucht Radieschen nicht nur die technischen Möglichkeiten, sondern betrachtet auch wichtige Aspekte wie die organisatorischen Gegebenheiten und die Kosten für den Aufbau und Betrieb einer Forschungsdaten-Infrastruktur. Zudem orientiert sich Radieschen speziell an den Gegebenheiten der deutschen Forschungslandschaft.

[4] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=mapri
[5] http://www.eudat.eu/
[6] http://www.isgtw.org/feature/towards-collaborative-data-infrastructure-science

Ziel der **Research Data Alliance (RDA)**[7] ist es, auf internationaler Ebene Innovationen und Entdeckungen im Forschungsdatenkontext zu beschleunigen, deren Nutzen durch Wiederverwertung zu steigern, Standards zu harmonisieren und die Auffindbarkeit von Forschungsdaten zu erhöhen. Die Initiative zielt auf die Entwicklung und Förderung der Akzeptanz von Infrastrukturen, zugehöriger Policies, von Handlungsempfehlungen und die Entwicklung von Standards. Die RDA ist eine neu gegründete Organisation (August 2012). Die Alliance operiert weltweit mit Partnern in Europa, Australien und Amerika. Radieschen setzt seinen Schwerpunkt auf die Entwicklungen in Deutschland und liefert damit einen nationalen Beitrag zu den Arbeitszielen der Alliance.

Den Herausforderungen durch immer größer werdende Datenmengen für Wissenschaftler, Universitäten und Forschungsförderungseinrichtungen stellt sich die **DFG** durch ihre Ausschreibung „Informationsinfrastrukturen für Forschungsdaten“. In den insgesamt 27 durch diese Ausschreibung finanzierten Projekten, im Folgenden **„FD“-Projekte** genannt, entwickeln Wissenschaftler und Informationsspezialisten Infrastrukturen für Forschungsdaten, die auf das jeweilige Fach zugeschnitten sind (Stand Mai 2011). Dabei geht es nicht nur um die mittel- bis langfristige Archivierung, sondern u.a. auch um fachadäquate Metadaten sowie Fragen der Verknüpfung von Publikationen mit Forschungsdaten oder der Qualitätskontrolle[8]. Die 27 Projekte der Ausschreibung wurden im Rahmen der Radieschen-Bestandsaufnahme für Interviews angefragt.

Desweiteren besteht bei der DFG seit 2007 die Möglichkeit, im Rahmen von DFG-geförderten Sonderforschungsbereichen (SFB) Service-Projekte zu beantragen, "die sich mit dem Aufbau von Informationsinfrastruktur für das Forschungsvorhaben befassen. Auch diese Projekte bringen Wissenschaftler mit Informationsspezialisten aus Bibliotheken und Rechenzentren zusammen. Die DFG fördert bisher 27 solcher „INF“-Projekte im Rahmen der derzeit 232 aktiven SFBs"[9]. Im Rahmen der Bestandsaufnahme wurden auch diese Projekte durch Interviews befragt. Eine Rückkopplung mit der „INF“-Community erfolgte durch den von Radieschen veranstalteten Workshop der INF-Projekte.

Sowohl die FD-Projekte als auch die INF-Projekte haben eine praktische Ausrichtung und zielen auf die Realisierung oder den weiteren Ausbau von Forschungsdaten-Infrastrukturen. Radieschen agiert in diesem Rahmen als Beobachtungs- und Roadmap-Projekt und ist keiner der genannten Förderlinien zugeordnet.

Die **Helmholtz-Initiative “Large Scale Data Management and Analysis” (LSDMA)** bietet für die Forschungszentren der Helmholtz-Gemeinschaft in Deutschland einen Datenservice mit Community-spezifischen Data Life Cycle Laboratories (DLCL)[10].

The DLCLs arbeiten in enger Kooperation mit den Wissenschaftlern. Das Ziel ist die Verarbeitung, das Management und die Analyse der Daten während des gesamten Daten-Lebenszyklus. Die gemeinsamen Forschungsaktivitäten in den DLCLs resultieren in Community-spezifischen Werkzeugen und Methoden. Die DLCLs werden ergänzt durch ein Data Services Integration Team (DSIT). Dieses Team bietet generische Technologien und Infrastrukturen für den Einsatz in den

[7] http://rd-alliance.org/

[8] Making Scientific Research Data Accessible: Current Trends and Perspectives in Germany, Informationsworkshop in Washington DC, 21. Juni 2011, http://www.dfg.de/dfg_profil/geschaeftsstelle/dfg_praesenz_ausland/nordamerika/berichte/2011/110621_informationsworkshop_washington/index.jsp

[9] Vgl. Effertz und Schoch (2013)

[10] http://www.helmholtz-lsdma.de/

verschiedenen Forschungscommunities und basiert auf Forschung und Entwicklung in den Bereichen Daten-Management, Daten-Sicherheit, Storage Technologien und Daten-Langzeitarchivierung.

Die Helmholtz LSDMA setzen ihren Schwerpunkt auf das Handling großer bis sehr großer Datenmengen. Diese ist jedoch nicht in jeder Forschungsdisziplin gegeben. So umfasst z.B. das Forschungsdatenarchiv des EarthChem[11] nur 8 GB und enthält dennoch die Forschungsergebnisse der vergangenen Jahrzehnte in mehr als 300.000 Datensätzen. Ziel des Projekts Radieschen ist es, auch diese relativ kleinen Datensätze, die sogenannte "Small Data", in der Entwicklung von Forschungsdaten-Infrastrukturen und deren Werkzeugen zu berücksichtigen.

Forschungsdaten und der Aufbau von Forschungsdaten-Infrastrukturen sind keine lokalen Herausforderungen, sondern betreffen alle Wissenschaftsdisziplinen und Forschungseinrichtungen. Ob eine globale, europäische oder lokale Lösung gesucht wird, hängt vom Forschungsgegenstand ab. Nicht jede Disziplin arbeitet mit Petabytes an Daten und nicht jede Disziplin benötigt dauerhaften Zugriff auf die Daten über Landesgrenzen hinweg. Da Forscher und Wissenschaftler jedoch vorwiegend international agieren, wären internationale oder zumindest europäische Lösungen zu favorisieren. Dies gilt auch im Hinblick auf Wissenstransfer und die Erhöhung der Forschermobilität. Allen betrachteten Projekten gemein jedoch ist der Trend weg von der "Silo"-Lösung einzelner Datensammlungen hin zu einer Lösung, bei der die Dienste und Strukturen von speziellen Service-Anbietern, wie Rechenzentren oder speziellen Repositories in Anspruch genommen werden. Das Radieschen-Projekt liefert mit seinen Berichten hierzu die nötige Hintergrundinformation.

[11] http://www.earthchem.org/

3. Zukunftsszenarien

Das Aufkommen neuer Technologien und Entwicklungen stellt auch die Akteure im Bereich der Forschungsdaten-Infrastrukturen vor neue Herausforderungen. Bibliotheken als einer der Akteure ermöglichen Zugang zu digitalen Medien, unterstützen die Publikation von Forschungsdaten und deren Langzeitarchivierung. Digitale Medien und Forschungsdaten jedoch bringen neue Aspekte in das Tätigkeitsspektrum der Bibliotheken. Wie muss man sich die Bibliothek der Zukunft vorstellen? Die Bibliothek als Schnittstelle zu den Rechenzentren? Verschmelzen Bibliothek und Rechenzentrum zu einer neuen Serviceeinheit? Welche Rolle werden die wissenschaftlichen Verlage in Zukunft übernehmen? Momentan liegt die Gewichtung noch bei der traditionellen Form der Publikation in Form von Artikeln für Konferenzen und Journals. Aber wird das auch in Zukunft so bleiben? Neue Publikationsformen kündigen sich bereits an. Auch die Aufgaben der Rechenzentren können sich wandeln. Gestern war noch Bereitstellung von schneller Hardware im Fokus der Aufmerksamkeit, nun sind Daten das Thema, um das sich alles dreht.

Vor diesem Hintergrund stellt sich die Frage nach den Werkzeugen, um in einer vernetzten Welt, die sich ständig ändert, den richtigen Kurs zu finden und zu verfolgen. Ein Werkzeug aus dem Bereich des Innovation Managements ist die **Szenario-Technik**[12]. Nach Kurt Sontheimer[13] geht es bei der Szenario-Technik weniger um das Vorhersagen der Zukunft, sondern mehr um das Vorausdenken der Zukunft. Szenarien beschreiben mögliche künftige Situationen, beispielsweise die zukünftige Entwicklung des Wissenschaftsstandorts Deutschland, in die das Projekt zu positionieren wäre.

Zukunftsszenarien beruhen auf einem vernetzten System von Einflussfaktoren, wobei für jeden Einflussfaktor mehrere denkbare zukünftige Entwicklungsmöglichkeiten ins Kalkül gezogen werden können. Wesentliches Ziel der Szenario-Technik ist das Erkennen zukünftiger Chancen und Gefahren, um daraus strategische Entscheidungen abzuleiten.

Abb. 3 zeigt eine Projektion verschiedener, möglicher Zukunftsszenarien. Die aktuelle Ausgangslage ist durch den blauen Kreis links gekennzeichnet. Die X-Achse zeigt die Veränderungen über die Zeit. Die Y-Achse zeigt das Spektrum möglicher Szenarien in verschieden großen Kreisen an. Der Kreis um

- "Possible" beschreibt Entwicklungen, die möglicherweise eintreten können. Die Vorhersage basiert auf extrapoliertem Wissen, beispielsweise Hochrechnungen.
- "Plausible" beschreibt Szenarien, die eintreten könnten. Die Vorhersage basiert auf aktuellem Wissen.
- "Probable" stellt eine Situation dar, die wahrscheinlich eintreffen wird. Die Vorhersage basiert auf aktuellen Trends.
- "Preferable" beschreibt eine Situation, die man sich erhofft, basierend auf der fundierten Bewertungen der aktuellen Situation.

Die Methode erlaubt die Einbeziehung von Faktoren, die anderweitig schwierig zu erfassen sind, wie beispielsweise neue Erkenntnisse über Zukunft, einen tiefgreifenden Wandel der Werte oder bisher neue Regelungen und Innovationen.

[12] Gausemeier, J., Stoll, K., Wenzelmann, C. (2007)
[13] Sontheimer, K. (1970)

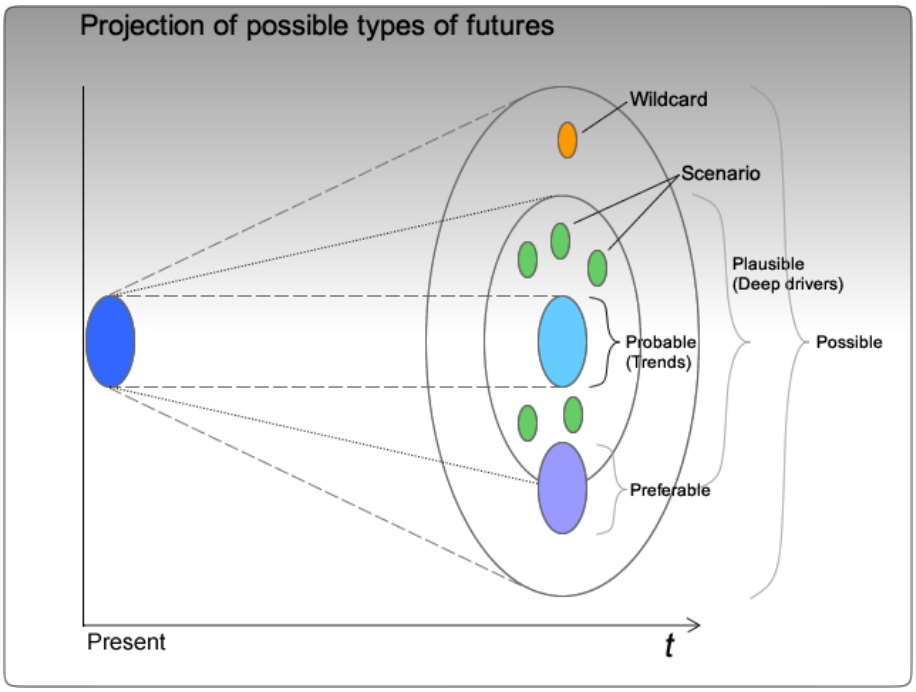

Abb. 3: Die Grafik[14] zeigt eine Projektion möglicher Zukunftsszenarien. Der blaue Punkt links stellt die Ausgangslage dar. Die grünen Punkte auf der rechten Seite zeigen mögliche Szenarien. Die Kreise des Trichters geben an, ob das jeweilige Szenario sich im Bereich des Wünschenswerten (Preferable), Wahrscheinlichen (Probable), Glaubwürdigen (Plausible) oder Möglichen (Plausible) befindet.

Die folgenden Zukunftsvisionen beschreiben mögliche Entwicklungen der Wissenschaftswelt in Deutschland im Jahre 2020 (oder später). Die Situationen sind überspitzt dargestellt, um Tendenzen zu verdeutlichen und mögliche Entwicklungsschritte ableiten zu können. Die Szenarien beschreiben Extremsituationen. Es ist nicht zu erwarten, dass die beschriebenen Situationen tatsächlich 1:1 so auftreten werden.

Die den Szenarien zugeordneten Grafiken verdeutlichen die Positionen der Akteure im Vergleich zum heutigen Zeitpunkt. Die Ausgangsposition, der heutige Zeitpunkt, liegt genau auf dem Kreuz in der Mitte. Die dargestellten Akteure sind die Wissenschaftler (W), die wissenschaftlichen Bibliotheken (B), die wissenschaftlichen Rechenzentren (R) und die Data Scientists (DS) als Verkörperung eines neuen Berufsbilds bei den Wissensarbeitern.

[14] Bild-Quelle: http://www.quesucede.com/page/show/id/scenario_planning

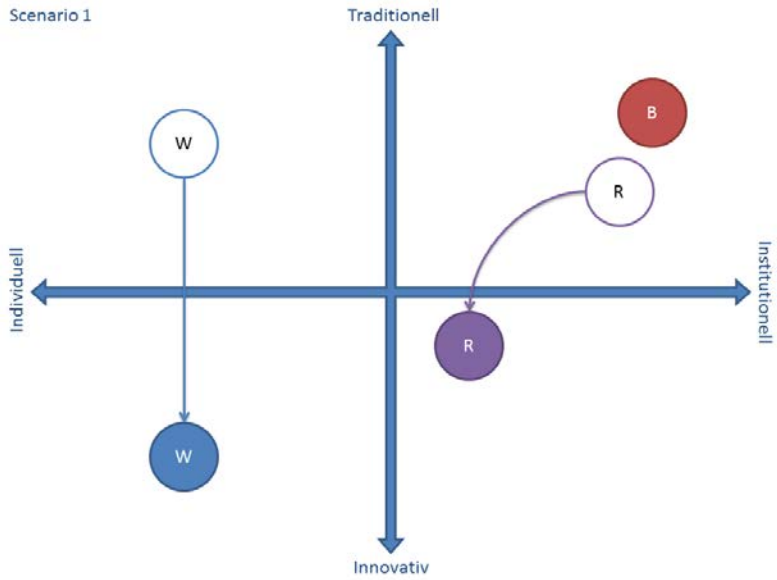

Abb. 4: Szenario 1 – Neue Leistungsidentikatoren in der Wissenschaft

Szenario 1 - Neue Leistungsindikatoren in der Wissenschaft

Lori ist eine erfolgreiche Wissenschaftlerin in den Geowissenschaften. Gerade ist sie von einer Vortragsreise aus Südafrika zurückgekehrt, da erhält sie die Nachricht, dass ihre Softwareveröffentlichung im Open Access Journal "Earth Science & Computing" angenommen wurde. Daten hat sie schon viele in Data Journals veröffentlicht, dies aber ist ihre erste Softwareveröffentlichung. Lori ist darüber besonders erfreut, denn diese Veröffentlichung ermöglicht ihr nun endlich, sich auf eine der besonders begehrten Positionen eines Leading Researcher in Australien zu bewerben. Eingangsvoraussetzung für diese heiß begehrten Positionen sind in datenintensiven Disziplinen wie den Geowissenschaften nicht mehr nur der Citation Index, sondern mittlerweile der Dreiklang aus Fachveröffentlichung, Datenveröffentlichung und Softwareveröffentlichung, denn erst in dieser Kombination werden Veröffentlichungen in den Naturwissenschaften als vollwertig und als maßgeblicher Beitrag zur Wissenschaft betrachtet.

Ihr Kollege Matthis betritt den Raum. Auch er freut sich, denn er als Co-Autor der Softwareveröffentlichung bekommt wertvolle European Research Credit Points (ERC Points) gutgeschrieben, die er nun einsetzen kann, um begehrte Messzeit an einem Massenspektrometer zu buchen. Ein Massenspektrometer steht zwar ganz in der Nähe in den Laboren des GFZ in Potsdam, jedoch bekommt man erst Messzeit genehmigt, nachdem man ein Mindestmaß an ERC Points erreicht hat. Die eingesetzten ERC Points amortisieren sich schnell, denn durch die geplanten Messungen werden sicher neue Erkenntnisse gewonnen, die Matthis für seine nächste Journal- und Datenveröffentlichung nutzen kann. So kann er nun weiter mit Lori und ihrem Team forschen und hoffentlich bald seine Doktorarbeit beenden.

Hauptaspekte des Szenarios:

- Einfaches Zählen von Publikationen und Zitaten zur Bewertung akademischer Leistungen wird abgelöst durch eine Kombination aus Fachveröffentlichungen, Datenveröffentlichungen und Softwareveröffentlichungen.
- Ein Scoring-System etabliert sich und regelt den Zugang zu Ressourcen.

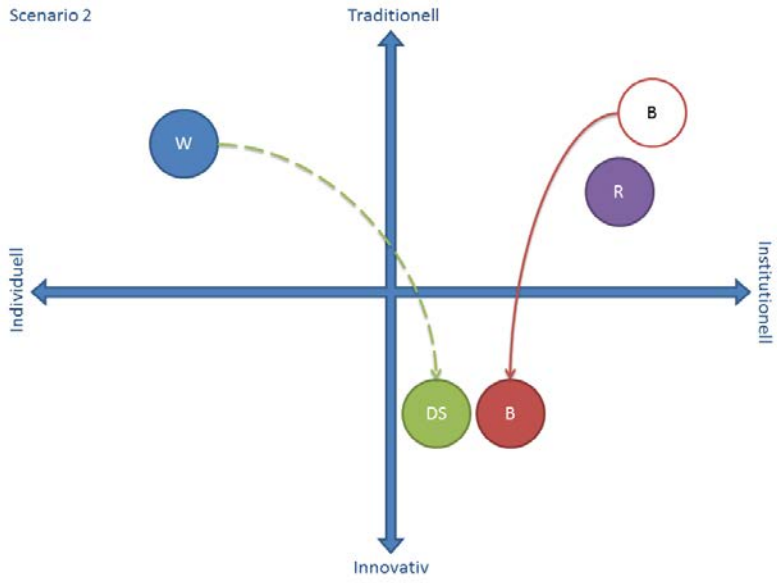

Abb. 5: Szenario 2 – Bibliotheken sind die Zukunft

Szenario 2 - Bibliotheken sind die Zukunft

Robert sieht sich um in seiner neugestalteten Bibliothek, einer Bibliothek im Verbund der Union of German Libraries for Science and Technology (UGL-ST). Hohe Aufenthaltsqualitäten, Besprechungsräume, Orte für Diskussionen, Orte für Schulung und Beratung prägen das Bild, und überall Displays mit Anzeigen der aktuellen Datenströme. Drahtloser Gigabit-Netzwerkzugang ist selbstverständlich. Die Datenbestände der Bibliothek sind landesweit mit den Beständen der anderen Universitäts- und Forschungsbibliotheken in der UGL-ST verbunden. Nach der Auflösung der traditionellen Bibliotheksverbünde ermöglichte die Gründung der UGL-ST es Deutschland mit der rasanten Entwicklung der Forschungsbibliotheken weltweit mitzuhalten. Lange vorbei sind die Zeiten, als einzelne institutionelle Bibliotheken noch Kataloge, Bücher und Zeitschriften vorhielten und Daten tief versteckt in den Rechenzentren und Arbeitsplatzrechnern lagen. Bücher und Zeitschriften gibt es noch immer, jedoch nur noch wenige in gedruckter Form. Die Bibliothek ist längst nicht mehr „Papiermuseum“, sondern sie hat sich zu einem Informationsdienstleister entwickelt, der Forscher mit Daten und Informationen versorgt, auch abseits textbasierter Medien.

Unter Roberts Mitarbeitern befinden sich einige hochausgebildete Data-Scientists. Diese sichten die eingehenden Datenströme, führen Qualitätschecks und erste Prüfungen auf eine mögliche Nachnutzung der Daten durch. Der Schwerpunkt der Tätigkeit der UGL-ST-Bibliotheken liegt jedoch nicht nur auf Archivierung und Katalogisierung von Daten. In der Folge der Zeitschriftenkrise zum Anfang des Jahrhunderts waren die Bibliotheken aktiv geworden und hatten die wissenschaftlichen Verlage, und mit ihnen das traditionelle Subskriptionsmodell für Zeitschriften, mit ihren eigenen Online-Publikationen vom Markt gedrängt. Ein global agierendes Open Access-Publikationshaus, die German Science Press unter dem Dach der UGL-ST schuf dafür die Voraussetzung. Die wissenschaftlichen Verlage waren zwar lange führend auf dem Gebiet der traditionellen Veröffentlichung von Zeitschriften und Büchern gewesen, auf dem Gebiet der Software- und Datenveröffentlichung jedoch waren sie mangels Kapazität und Reformwillen chancenlos. Im harten

Positionierungskampf waren sie untergegangen und führten nun ein Nischendasein oder waren von den großen wissenschaftlichen Bibliotheken übernommen worden.

Die neuen UGL-ST-Bibliotheken ähneln daher nur entfernt im Interieur den alten Universitätsbibliotheken, viel mehr sind sie heute Informations- und Kompetenzzentren mit den Kerneinheiten Bibliothek und Rechenzentrum, ein unverzichtbarer Teil der Forschungsinfrastruktur. Um auf der Höhe der Zeit zu bleiben und im harten internationalen Wettbewerb mithalten zu können, war es nötig geworden, selbständig neue Angebote zu entwickeln. Die UGL-ST hatte eine eigene Forschungsabteilung eingerichtet in der hochqualifizierte Data Scientists Software entwickelten und neue Techniken zur Datenanalyse und -visualisierung erforschten, um den schnellen Veränderungen durch neue wissenschaftliche Kommunikationsformen und der datengetriebenen Forschung gerecht zu werden. Gemeinsam und zentral Dienste für die Wissenschaft zu entwickeln, mit großer Verlässlichkeit anzubieten und gleichzeitig vor Ort kompetente Beratung und Dienste nah an den Bedürfnissen der Wissenschaftler zu vermitteln, darin liegt die Stärke der Bibliotheken in der Union of German Libraries for Science and Technology.

Hauptaspekte des Szenarios:

- Bibliotheken entwickeln sich weiter zu innovativen, vernetzten Informations- und Kompetenzzentren.
- Data Scientists, hochqualifizierte Experten im Umgang mit Daten, arbeiten in Bibliotheken in Bereichen wie der Kuratierung, Qualitätssicherung oder Archivierung.
- Bibliotheken übernehmen die Rolle der heutigen Wissenschaftsverlage.

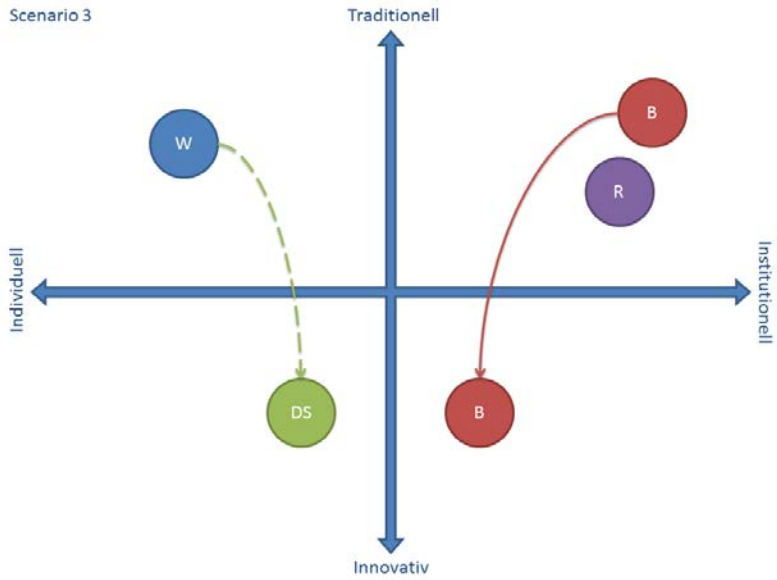

Abb. 6: Szenario 3 – Data Scientists, die Stars einer neuen Generation

Szenario 3 - Data Scientists, die Stars einer neuen Generation

Tom ist Data Scientist mit dem Spezialgebiet Forschungsdaten. Nach einem naturwissenschaftlichen Studium und dem Abschluss seiner Zusatzausbildung Forschungsdaten konnte er sich aus vielen Angeboten die für ihn interessanteste Position aussuchen. Trotz lukrativer Angebote aus Großbritannien und den USA entschied er sich für eine Position in Deutschland an der German

National Library for Science and Technology, kurz GNL-ST. Diese neugegründete Bibliothek war mit neuester Technologie ausgestattet und bot daher die besten Bedingungen für seinen Karrierestart. Zu den Aufgaben für Data Scientists an der GNL-ST gehört auch die Entwicklung von Algorithmen zur Klassifizierung und Annotierung von Daten, denn man hatte festgestellt, dass eine unstrukturierte Datenflut nicht sinnvoll auswertbar ist. Auch die Qualitätssicherung , Prüfung der Daten, sowie die Unterstützung bei einer möglichen Nachnutzung gehören zu ihren Aufgaben.

Der für Tom spannendste Teil aber ist die inhaltliche Sichtung der Daten und die Analyse möglicher Querverbindungen. Human-Computer-Interfaces ermöglichen den Data Scientists der GNL-ST über nicht-textbasierte Eingabegeräte (z.B. Datenhandschuhe, Gestensteuerung) und räumliche Displays mit den Daten im 3D-Raum zu interagieren. Die immersive Datenanalyse ist ein besonders beliebtes Verfahren für die datengetriebene Forschung in hochdimensionalen Datenräumen. So konnte er zum Beispiel bereits auf der Basis der Analyse von Satellitendaten über den Zustand der Ionosphäre zusammen mit der Analyse von ozeanographischen Daten auf eine mögliche Tsunamigefahr im Indischen Ozean hinweisen. Die Wissenschaftler vor Ort hatten seine Hinweise dankbar aufgenommen und sofort in ihr Warnsystem integriert. So kann nun schneller und präziser auf ein Warnsignal reagiert und die Bevölkerung, falls nötig, evakuiert werden.

Für die Zukunft wünscht sich Tom die Entwicklung weiterer innovativer Interaktionstechniken, die ihm erlauben, auch sehr große Datenmengen abstrahiert darzustellen und in Windeseile gesuchte Informationseinheiten zu sichten.

Hauptaspekte des Szenarios:

- Das Berufsbild des Data Scientists entwickelt und etabliert sich auch in der akademischen Welt.
- Data-Scientists arbeiten bei modernen, akademischen Informationsdienstleistern, die sich aus den traditionellen Wissenschafts-Bibliotheken entwickelt haben.
- Ihre Aufgaben umfassen Service für Wissenschaftler wie Ingest und Archivierung, aber auch Forschung im Bereich der Daten-Analyse.

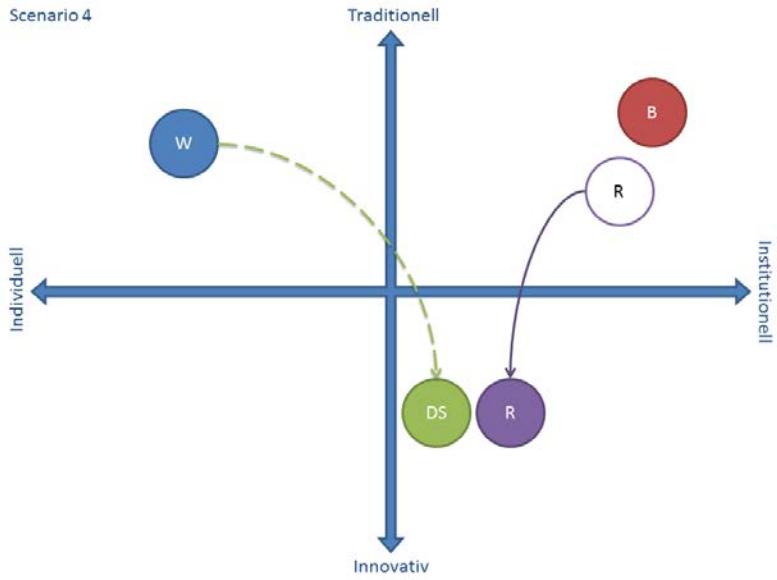

Abb. 7: Szenario 4 – Datenzentren übernehmen eine neue Rolle

Szenario 4 – Datenzentren übernehmen eine neue Rolle

Vorbei die Zeiten, als Rechenzentren in der Wissenschaft lediglich konservative Service Center waren, die auf Zuruf Speicherplatz und Server bereitstellten. Im Klischee des Rechenzentrums, wie Forscher es sich meist vorstellten, liefen seltsam gekleidete Gestalten in großen Hallen mit rauschendem Lüftungssystem zwischen Rechnern umher. Bewahren galt als das Gebot der Stunde, Neuerungen stand man in den Rechenzentren eher skeptisch gegenüber, denn irgendwo hatte gerade wieder ein Server seinen Geist aufgegeben. Die ambitionierten Rechenzentren engagierten sich im Hochleistungsrechnen. Doch irgendwann kam da ein Sprung und einige der akademischen Rechenzentren entwickelten sich zu Datenzentren.

Heute sind Data Center die natürlichen Ansprechpartner für Datenmanagement, Software-Services und auch klassische Veröffentlichungen. Letztere Aufgabe hatten die Data Center von den Bibliotheken und Verlagen übernommen, die mit der wachsenden Datenflut nicht mehr umgehen konnten, als der Ruf nach kombinierten Software- und Datenpublikationen immer lauter wurde. Und was lag da näher, als eigene Online-Journals zur Daten- und Softwareveröffentlichung anzubieten? Die Kapazitäten standen ja zur Verfügung. Nun war die ehemalige Bibliothek als Abteilung dem Data Center angegliedert und die wissenschaftlichen Verlage mit ihrem Papier- und Subskriptionsbasierten Angebot nahezu ausgestorben.

Data Scientists bevölkerten nun die Räume, welche mit modernster Interaktionstechnologie und Rechnern mit aktuellster Analyse-Software ausgestattet waren. Die Server und Hochleistungsrechner existierten natürlich noch - an einem zentralen Ort, gut abgesichert gegenüber unbefugtem Zugriff, Stromausfall und sonstigen Katastrophen. Serviceeinrichtungen waren die Data Centers geblieben, jedoch gingen ihre Aufgaben weit über das Niveau von Internet Hosting und Cloud-Angeboten hinaus. Angebote für Virtuelle Forschungsumgebungen (VFUs) und Research Data Engines (RDEs) bildeten nun den Schwerpunkt. Für den Wissenschaftler blieben die Details der Administration und der Softwareinstallation transparent verborgen, er konnte sich mit Hilfe einer Toolbox seine eigene Arbeitsumgebung für alle seine Projekte kombinieren und darin problemlos mit anderen Forschern seines Teams oder seiner Community zusammen arbeiten. Auch die Publikation von Forschungsergebnissen, Daten und Software waren nun aus dieser Umgebung heraus möglich.

Hauptaspekte des Szenarios:

- Rechenzentren entwickeln sich weiter zu Datenzentren, die den Forscherinnen und Forschern als primäre Ansprechpartner sowohl für Datenmanagement und Software-Services als auch für Publikationen aller Arten dienen.
- In den Datenzentren arbeiten Data-Scientists an der Bereitstellung der diversen Dienste (VFUs, RDE) für die Communities.

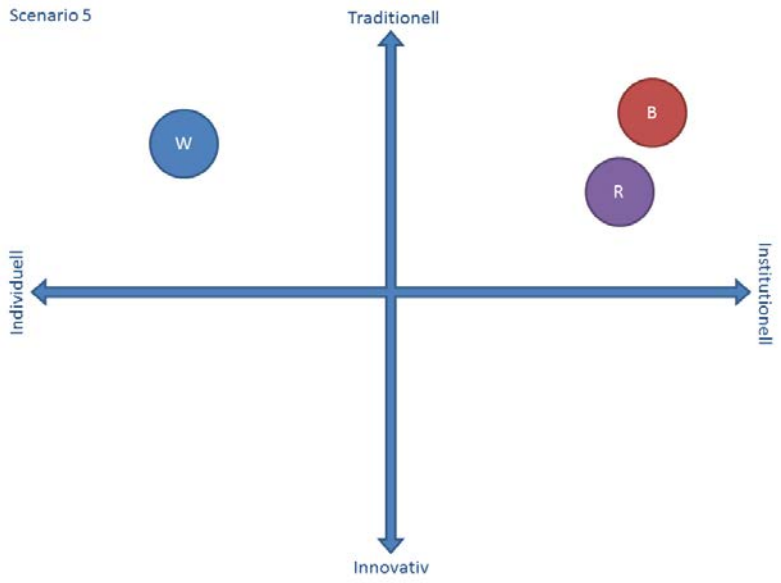

Abb. 8: Szenario 5 - Bewährtes bewahren

Szenario 5 - Bewährtes bewahren

Peter sitzt vor seinem Bildschirm und sortiert die Daten seines Projekts "Benedikt" in seine Datenbank. Es sind einzigartige Datensätze, die eine fast vollständige Analyse der Ikonenmalerei des 14.Jahrhunderts erlauben. Die Daten sind zum Teil nicht reproduzierbar und damit für ihn und seine Kollegen besonders wertvoll. Peter verwahrt seine Daten zunächst auf seiner externen Festplatte. Später will er sie in ein Repositorium für kunstgeschichtliche Daten transferieren, doch zunächst plant er seine Daten weiter auszuwerten und seine Ergebnisse zu veröffentlichen - in einem der noch verbliebenen traditionellen Wissenschaftsverlage. Er misstraut dem neuen Angebot der Online-Verlage, nachdem gegen einen seiner Kollegen ein Plagiatsvorwurf laut geworden war. Plagiatsforscher hatten behauptet, der Kollege hätte fremde Datensätze aus einer der Online-Datenbanken verwendet ohne ihre Herkunft eindeutig zu kennzeichnen. Damit ihm das nicht auch passiert, verwendet er nur eigene Datensätze und verwahrt diese sicher auf seiner Festplatte auf. Wo käme man denn da hin, wenn andere Wissenschaftler auf Basis seiner Arbeiten Ruhm und Ehre erlangten?

Mit dieser Position steht er in Deutschland nicht alleine da. Deutschland hat sich in den letzten Jahren mehr und mehr zu einer Enklave des Bewährten in einer Welt des Umbruchs entwickelt. Hier gibt es sie noch, die traditionellen Wissenschaftsverlage mit ihren Paper Publikationen, ebenso wie die Rechenzentren, die sich direkt um die Belange der Forscher vor Ort kümmern und ihnen ihre Arbeitsumgebungen nach Maß bereitstellen. In anderen europäischen Ländern ist bereits alles virtuell - Virtual Research Communities (VRCs), Research Data Engines, Pläne zum Daten-management. Lauter Dinge, mit denen man als Forscher lediglich seine wertvolle Zeit vertrödelte.

Nachteilig allerdings war, dass nun er und sein Team doch recht abgeschnitten waren von den weltweiten Forschungsaktivitäten in der Kunstgeschichte. Zusammenarbeit und Informations-austausch erfolgt mittlerweile international vielfach über VRCs. Datenaustausch über die großen Online-Datenbanken war zwar möglich, beruhte aber auf Gegenseitigkeit. Hinzu kam, dass seine sorgfältig gepflegte Publikationsliste weniger und weniger wert zu sein schien. Er hatte schon lange keine interessanten Jobangebote mehr aus dem Ausland erhalten. Dort wurde mehr und mehr Wert

auf den Dreiklang an Veröffentlichungen von Daten, Software und Methoden gelegt. Vielleicht sollte er diese Möglichkeiten doch mal näher untersuchen? Die Chancen auf ein Vorankommen als Wissenschaftler allein in Deutschland schienen ihm doch recht begrenzt. Er würde darüber später noch einmal intensiv nachdenken. Jetzt aber war erst einmal die Auswertung seiner Daten wichtiger. Der Fördermittelgeber verlangte nach einem Bericht und die Zeit bis zum Abgabetermin verging schnell.

Hauptaspekte des Szenarios:

- Bestrebungen nach Erneuerung werden aus den verschiedensten Gründen abgewehrt.
- Deutschland fällt im internationalen Vergleich zurück. Die Wissenschaftler sind zunehmend isoliert.

Fazit

Ein optimales Ergebnis kann nur erzielt werden, wenn die verschiedenen Akteure miteinander interagieren und bereit sind, ihre aktuelle Position zu überdenken und zu verändern.

Die Wissenschaftswelt ist dynamisch und verändert sich kontinuierlich. Welchen Weg diese Entwicklung nehmen wird, ist nicht vorhersehbar. Die vorgestellten Szenarien zeigen mögliche Entwicklungen - zum Positiven und zum Negativen. Es ist nun an den Akteuren selbst, die eigene Position in diesem Kontext zu definieren, diese zu überdenken und Schritte zu überlegen, mit denen eine möglichst positive Zukunftsentwicklung erzielt werden kann.

4. Synthese der Ergebnisse aus den Arbeitspaketen Kosten, Organisation und Technik

Kernstück des Projekts Radieschen ist die Erarbeitung dreier Reports zu den Themen Technologie, Organisation und Kosten. Die inhaltlichen Schwerpunkte der Reports liegen auf der Analyse der technischen Komponenten der Infrastruktur (Report Technik), der Analyse von Prozessen und Workflows im Lebenszyklus von Forschungsdaten sowie der Betrachtung von organisatorischen Aspekten (Report Organisation) und der Untersuchung von Kostenstrukturen für den Betrieb von Forschungsdaten-Infrastrukturen (Report Kosten). Dieses Kapitel zieht für die drei Reports ein Fazit, zeigt für die jeweiligen Themen Handlungsempfehlungen auf und gibt einen Ausblick auf die mögliche weitere Entwicklung in naher Zukunft.

Technologie

Die Auswertung der Interviews und die Analyse der untersuchten Materialien ergaben folgende markante Ergebnisse:

- In den Projekten und Fachdisziplinen werden vorzugsweise Eigenentwicklungen aufgebaut und genutzt. Dabei werden jedoch generische Werkzeuge der unteren technischen Ebene (z. B. Dateisysteme, Datenbanken) als Grundlage für die Eigenentwicklungen verwendet. Die Nutzung dieser grundlegenden Werkzeuge ist disziplinübergreifend weit verbreitet.
- In Bezug auf Hardware sind als Speichermedien vor allem Festplatten- und Bandsysteme im Einsatz.
- Falls bereits existierende Softwarelösungen zum Einsatz kommen, wird bei der Auswahl vor allem Wert auf Nachhaltigkeit gelegt, d. h. die Softwarelösungen müssen eine Perspektive bzgl. Support und Pflege bieten. Argumente gegen eine Nutzung generischer Komponenten sind häufig deren mangelnde Anpassbarkeit an aktuelle Arbeitsumgebungen und Anforderungen. Kommerzielle Dienste dienen dagegen häufig als Vorbilder in Bezug auf ihre Benutzerfreundlichkeit und ihre Integrationsmöglichkeiten in die privaten Arbeitsumgebungen.
- Die Analyse der Workflows der verschiedenen Disziplinen und befragten Institutionen ergab eine Reihe von datenbezogenen Arbeitsschritten, die fast überall auftreten. Dabei handelt es sich sowohl um disziplinspezifische als auch disziplinunabhängige Prozesse. Beispiele für disziplinspezifische Arbeitsschritte sind Datenerfassung, Qualitätskontrolle der Daten und disziplinspezifische Metadaten. Beispiele für disziplinunabhängige Prozesse sind die allgemeine Metadatenerzeugung/-ergänzung, Daten- und Metadatenspeicherung, Datentransfer, Datenreplizierung, (Langzeit-) Archivierung und der webbasierte Zugriff auf die Daten.

Aus der Analyse der technischen Systeme konnten folgende Handlungsempfehlungen extrahiert werden:

- Für eine Reihe von disziplinunabhängigen Arbeitsschritten können disziplinübergreifende Methoden und Werkzeuge genutzt werden. Dazu gehören: Datenformate, Metadatenerzeugung/-ergänzung für technische und kontextuelle Metadaten, Daten- und Metadatenspeicherung, Datentransfer, Datenreplikation, Backup, (Langzeit-)Archivierung, webbasierter Zugriff auf die Daten

- Werkzeuge zum Forschungsdatenmanagement sollten benutzerfreundlich sein und eine Integration in die wissenschaftliche Arbeitsumgebung erlauben.

- Die Entwicklung von Standards und Schnittstellen ist wichtig, damit es mit dem fortschreitenden technischen Wandel möglich ist, relativ einfach einzelne, veraltete Module durch aktuelle Technologien zu ersetzen.

- Persistente Identifikatoren sollten verstärkt eingesetzt werden. Einfach handhabbare Systeme mit dem erforderlichen technischen und organisatorischen Hintergrund existieren bereits.

- Der fachgebietsübergreifende Austausch zum Thema Forschungsdatenmanagement sollte intensiviert werden, da existierende und möglicherweise wiederverwertbare Lösungen oft nicht bekannt sind. Das Wissen über technische Lösungen und organisatorische Aspekte könnte z. B. über zu gründende thematische Kompetenzzentren verbreitet werden.

Generell geht der Trend zu einer Auslagerung von technischen Diensten an Service-Einrichtungen und Rechenzentren. Aufgrund der Flut von Daten und der kurzen Haltbarkeit elektronisch gespeicherter Informationen erscheint eine solche Auslagerung auch dringend erforderlich, insbesondere für eine professionelle Kuratierung. In Zukunft wird es immer wichtiger werden, professionelle Dienste zum Management von Forschungsdaten zur Verfügung zu haben und zu nutzen. Sicherlich sind einige Datenmanagement- und Datenverarbeitungsschritte spezifisch für eine Fachdisziplin. Für eine Vielzahl von Diensten können jedoch übergreifende Lösungen entwickelt werden, die allerdings eine langfristige Perspektive und Nachhaltigkeit vorweisen müssen.

Organisation

Der Fokus des Arbeitspakets 3 liegt auf der Untersuchung der Strukturen, in denen Forschungsdaten-management in Deutschland stattfindet, und den daran beteiligten Organisationen. Hierbei stehen insbesondere die von der DFG geförderten Projekte zum Thema im Vordergrund. Diese Fokussierung auf die DFG ist nicht zuletzt der Tatsache geschuldet, dass – bis vor wenigen Wochen – das Thema Digitales Forschungsdatenmanagement bei den anderen Fördergebern, trotz GWK Kommission Zukunft der Informationsinfrastruktur („KII“)[15], Allianz für Forschungsdaten und anderen Initiativen, kaum als herausragendes Förderthema sichtbar war. Auch wurde viel zu lange der wissenschaftliche Wert der Daten selbst unterschätzt und durch traditionelle Orientierung auf wissenschaftliche Publikationen das Thema nur im Bereich der der klassischen Gedächtnisinstitutionen und Bibliotheken verortet.

Es ist festzustellen, dass die Akzeptanz der Ergebnisse der bisherigen Projekte zum Datenmanagement in den einzelnen Communities (Formate, Standards, Workflows, Infrastruktur, Metadaten, …) verbesserungsbedürftig ist. Die Nutzung dieser Vorarbeiten durch neue Projekte sollte daher eine höhere Priorität im Förderungsprozess erhalten:

- Die Berücksichtigung existierender Standards sollte eine Anforderung bei der Bewilligung neuer Anträge sein.
- Im Bereich der Infrastruktur sollte die Kompetenz der diversen Gremien in den Fachdisziplinen stärker genutzt werden.

[15] http://www.gwk-bonn.de/index.php?id=205

Unter Bezugnahme auf das Domänen-Modell ist festzustellen, dass die wesentlichen Fortschritte für das Management von Forschungsdaten derzeit im Bereich der Übergänge zwischen Privater, Gruppen- und Dauerhafter Domäne erzielt werden müssen. Genau an diesem Punkt setzt das Instrument der INF-Projekte in den SFB der DFG an. Dieses Förderinstrument sollte daher gezielt weiterentwickelt werden:

- Jeder SFB sollte ein INF-Projekt[16] aufweisen. Die fortlaufende Beteiligung der Wissenschaftler an der INF-Konzeption und Durchführung ist anzustreben.
- Jedes INF-Projekt sollte durch gezielte Beratung befähigt werden, die Standards und Tools anzuwenden, die sich in den Fachdisziplinen etabliert haben. Ist dieser Konsens innerhalb der Disziplin nicht vorhanden, sollten (nicht vom INF selbst) gezielte Maßnahmen zu seiner Herstellung ergriffen werden.
- Für INF-Projekte sind nicht Kriterien wie Neuartigkeit oder Einzigartigkeit entscheidend, sondern Effizienz und Einbindung in einen infrastrukturellen Kontext.
- Die INF-Projekte sollten nicht den Versuch beinhalten, die Strukturen des gesamten Fachgebiets gleich mit zu erschaffen. Auch sollte die Verwendung von generischen Komponenten gegenüber der Entwicklung von speziellen Lösungen bevorzugt werden.
- Um eine langfristige Perspektive zu eröffnen, benötigen die INF-Teilprojekte korrespondierende Infrastrukturmaßnahmen der IT-Infrastruktur-Provider im Wissenschafts-bereich, in die sich ihre Arbeiten einbinden lassen.

Die Vielfalt der Ansätze für Forschungsdatenmanagement in den einzelnen Disziplinen zeigt, dass die Entwicklung von Tools und Standards für das Datenmanagement weiterhin dynamisch ist:

- Es fehlen klare Analysen der grundlegenden Workflows in den meisten Disziplinen. Nur durch Konzentration auf diese existierenden Workflows können effiziente Werkzeuge ausgewählt bzw. entwickelt werden, die dann auch tatsächlich durch die jeweilige Community akzeptiert werden.
- Bei der (Weiter-) Entwicklung von fachspezifischen Metadaten sollte ein klarer Fokus auf Auffindbarkeit und Nachvollziehbarkeit (Provenance) gelegt werden. Komplexe, darüber hinausgehende Systeme verlieren mit ihrem zu spezifischen Vokabular zu schnell Wiederverwendbarkeit und Aktualität. Metadaten sollten auch weitgehend durch die Werkzeuge selbst bereitgestellt werden.

Auch im Bereich der IT-Infrastruktur besteht weiterer Entwicklungsbedarf:

- IT-Infrastruktur und deren Provider konzentrieren sich zu stark auf das Scientific Computing und nehmen sich des Problems des Datenmanagements bislang nur unzureichend an. In Hinsicht auf eine bessere Unterstützung des Forschungsdatenmanagements ist jedoch eine Flexibilisierung der gegenwärtigen Förderungsstrukturen notwendig. Das D-Grid[17] war eine solche Initiative. Es hat sich jedoch gezeigt, dass es noch erhebliche Probleme gibt, die Bereitstellung von IT-Ressourcen zu flexibilisieren. Hierbei sind jedoch nicht technische, sondern rechtliche, förderpolitische und organisatorische Fragen entscheidend.

[16] Programmelement der Informationsmanagement und Informationsinfrastrukturin den DFG-Sonder-forschungsbereichen, kurz SFBs (siehe hierzu http://www.dfg.de/foerderung/programme/koordinierte_programme/sfb/programmelemente/programmelement_inf/index.html)
[17] http://www.d-grid.de/

Obwohl die drängendsten Probleme im Bereich des Forschungsdatenmanagements immer noch innerhalb der jeweiligen Disziplin zu verorten sind, sind auch im organisatorischen Bereich disziplinübergreifende Lösungsansätze erstrebenswert:

- Um die Ergebnisse der vielfältigen abgeschlossenen Projekte zu nutzen, sollten thematische Kompetenznetzwerke eingerichtet werden, die neue Projekte beratend unterstützen. Es hat sich gezeigt, dass ein großes Informationsbedürfnis über vorhandene Tools und Standards seitens der Forscherinnen und Forscher und der Projekte vorhanden ist, das zur Zeit jedoch nur unzureichend befriedigt werden kann.

Die wesentlichen Fortschritte für das nachhaltige Forschungsdatenmanagement werden in den nächsten Jahren immer noch innerhalb der Fachdisziplinen erzielt werden. Dies wird insbesondere durch fachspezifische Standardisierungsprozesse sowie Implementierung und Nutzung von fachspezifischen Workflows erreicht werden. Diese Entwicklungen müssen durch eine Flexibilisierung der IT-Infrastrukturen gestützt werden. Erst auf dieser Basis sind auch interdisziplinär entwickelte IT-Werkzeuge wie beispielsweise Tools zum Data-Mining erfolgreich einsetzbar. Nichtsdestotrotz ist eine Zusammenarbeit, insbesondere eine stärkere Kommunikation, zwischen den Disziplinen wünschenswert und anzustreben. Auch gemeinschaftliche organisatorische Strukturen sind sinnvoll, solange nicht entsprechende Strukturen innerhalb der Disziplinen vernachlässigt werden.

Kosten

Das Risiko eines Datenverlustes ist bei fehlender Archivierung sehr hoch. Selbst wenn die Daten noch irgendwo vorhanden sind, kostet die Wiederauffindung und Neu-Zusammenstellung viel Aufwand. Deshalb sollten zumindest alle wichtigen und nicht wiederherstellbaren Daten archiviert werden.

Primärarchive, d.h. solche Forschungsdatenarchive, die Daten direkt vom Forscher nehmen, haben den größten Aufwand an Arbeitszeit beim Ingest. Anders ist es bei Sekundärarchiven, d.h. solchen, die Daten nicht vom Forscher, sondern ausschließlich von anderen Archiven nehmen. Bei den beiden im Rahmen von Radieschen befragten Sekundärarchiven sind Auswahl und Ingest zusammengenommen der am wenigsten arbeitsaufwändige Schritt verglichen mit Speicherung plus Kuration und Bereitstellung.

Im Hinblick auf Kosten können folgende Handlungsempfehlungen gegeben werden:

- Eine weitergehende Automatisierung hilft Kosten zu senken. Das gilt vor allem für den Ingest als meist arbeitsaufwändigsten Schritt. Eine disziplinübergreifende Ingest-Software muss flexibel genug sein, um den Anforderungen unterschiedlicher Fächer gerecht zu werden, trotzdem leicht zu handhaben sein und dauerhaft gepflegt werden. Nur dann wird sich der gewünschte Effekt einer fächerübergreifenden und nachhaltigen Kostensenkung einstellen. Es ist Software-Projekten wie PubFlow[18] zu wünschen, dass durch Folgeprojekte Weiterentwicklung und Pflege ermöglicht werden.
- Einheitliche Datenstrukturen, Vokabulare und Metadatenstandards vereinfachen Software-Entwicklung, Qualitätskontrolle, Datenpflege und die Suche in Metadaten und Daten. Das hilft auch Kosten zu senken. Trotzdem sind die Bestrebungen zu Standards zu kommen in vielen Disziplinen noch unterentwickelt.

[18] http://www.pubflow.uni-kiel.de/

- Forschungsdaten-Dienstleistungen mit einem Preis zu versehen ist in jedem Falle problematisch. Ein kostenpflichtiger Zugang kann z.B. die Durchführung von Vorstudien erschweren.
- Falls Forschungsdaten-Dienstleistungen mit einem Preis versehen werden, müssen solche Preise frühzeitig bekannt sein, damit Forscher Mittel in der erforderlichen Höhe beantragen können.
- Die entstehenden Kosten sollten offen dar liegen und nachvollziehbar sein. Für den Anfang sei hier eine affine Preisfunktion empfohlen. Diese besteht aus einem konstanten Sockelbetrag, in dem mengenunabhängige Kosten berücksichtigt werden, und einer Summe linearer Teilfunktionen. Der Preis sollte linear mit der Zahl der Datensätze ansteigen. An die Stelle der Zahl der Datensätze kann auch eine andere Größe treten, die die Zahl der logischen Dateneinheiten wiedergibt. Wenn die Datenvolumina größer als der Durchschnitt sind, sollte als zweite Variable das Datenvolumen in die Preisfunktion aufgenommen werden, ebenfalls in Form einer linearen Teilfunktion. Eine affine Preisfunktion lässt sich leicht an erfasste oder geschätzte Kosten anpassen und besitzt darüber hinaus den Vorteil, transparent und nachvollziehbar für die Kunden zu sein.
- Ganz ohne eine bessere Förderung der Forschungsdateninfrastruktur wird es nicht gehen. Durch Effizienzsteigerung bei den Archiven allein sind schon die aktuellen Datenmengen nicht zu bewältigen.
- Zur genaueren Erforschung der Kosten wären betriebswirtschaftliche Begleitprojekte sinnvoll. Die Erfassung der Kosten ist genauer, wenn sie projektbegleitend und nicht nachträglich durchgeführt wird.
- Nur wenig bekannt ist über die Kosten der Pre-Ingest-Phase und der privaten Domäne. Hier besteht noch Forschungsbedarf.

Nach Daten des Statistischen Bundesamtes wurden 2011 vom Staat und von privaten Institutionen ohne Erwerbszweck 11 Mrd. € und von den Hochschulen noch einmal 13 Mrd. € für Forschung und Entwicklung ausgegeben. Wenn von den 11 Mrd. € nur 1 % zugunsten einer besseren Forschungsdateninfrastruktur umgewidmet würden, könnten ca. 25 weitere Forschungsdatenzentren von der Leistungsfähigkeit des DAS (Datenarchiv für Sozialwissenschaften beim GESIS) und ca. 10 weitere Virtuelle Forschungsumgebungen mit Datenzugang wie TextGrid oder C3Grid betrieben werden.

5. Analyse der Diskussion mit der Community

Erkenntnisse aus den Interviews

Die Projektergebnisse von Radieschen basieren zu einem erheblichen Teil auf den im Rahmen des Projekts durchgeführten Interviews. Die Interviews wurden mit Repräsentanten einer Vielzahl wissenschaftlicher Communities geführt. Das Radieschen-Projekt-Team führte insgesamt mehr als 28 Interviews durch. Die Mehrzahl der Interviews fand als persönliche Befragung statt. Einige Interviews wurden aus Termingründen telefonisch geführt. Der umfangreiche Fragenkatalog deckte neben einem allgemeinen Teil Fragen aus den Bereichen Technologie, Organisation und Kosten ab.

Die Essenz der Interviews in Worte zu fassen ist kein einfaches Unterfangen, einige klare Trends sind jedoch deutlich erkennbar. Das personalisierte Interview-Format trug erheblich zum Verständnis der Daten, Workflows und generellen Besonderheiten der einzelnen Projekte bei.

Die Anforderungen an die Hardware stellen wahrscheinlich den größten gemeinsamen Nenner der verschiedenen Projekte dar. Alle Projekte nannten einen relativ **standardisierten Hardware-Setup**: Blade-Server für Web und Datenbanken, High End Harddisk, jedoch ansonsten keine sehr exotischen Konfigurationen. Mehrere Teilnehmer pflegten ihre eigenen Systeme, während andere auf die Hosting Einrichtungen ihrer eigenen Datencenter zurückgreifen. Virtualization wird oft eingesetzt und allgemein als eine positive Entwicklung betrachtet. Einige der Befragten berichteten vom Einsatz eines Low-End Clusters (Hadoop[19]).

Viele Projekte nutzen extensiv **Freie und Open-Source Software**, insbesondere auf Seiten des Servers. Falls spezifische Software entwickelt wurde, wird diese oft unter GPL oder einer ähnlichen freien Lizenz veröffentlicht.

Resource Repositories sind eine populäre Wahl, um Daten und Literatur zu verwahren. Diese sind jedoch oft nicht ausreichend oder spezifisch genug, um ganze Datensätze zu speichern. Alternativ werden dann **periphere Datenspeichersysteme** eingesetzt.

Beim Thema **Metadaten** dagegen zeigen sich viele Variationen: Diese reichen von abstrakten Text-Fragmenten in den Lebenswissenschaften zu detaillierten Schemata (DDI in den Sozial-wissenschaften, ABC in Biologie). Die meisten Interviewpartner gaben an, dass es einer ernsthaften Anstrengung bedarf, Metadaten von hoher Qualität zusammen zu stellen.

Die Mehrzahl der Befragten sympathisiert mit **Open Access Policies** und versucht, so offen wie möglich mit ihren Daten umzugehen. Ausnahmen bestehen in Bezug auf Privacy (z.B. Daten einzelner Haushalte), durch juristische Einschränkungen auf Seiten der Verleger, risiko-behaftetes Material (z.B. die Standorte gefährdeter Arten). In einigen Fällen kann jedoch durch Anonymisierung oder die Hinzunahme sogenannter "Noise Data" dennoch eine Veröffentlichung gewährleistet werden.

Ein oft erwähntes Konzept speziell im sozio-ökonomischen Bereich ist es, direkt die **datenspeichernde Einrichtung zu besuchen**, um dort mit sensitiven Daten zu arbeiten. Einige Versuche allerdings wurden bereits unternommen, um eine orts-unabhängigere Lösung zu entwickeln.

Die **Workflows** für die Datenarchivierung sind spezifisch für jede wissenschaftliche Community. Abgesehen von sehr abstrakten Schritten sind Gemeinsamkeiten daher sehr schwierig zu definieren.

Die Relevanz von **Langzeit-Archivierung** (von Daten und Software) als Thema war allen Befragten bewusst. Das Thema wurde jedoch nicht immer explizit adressiert, da dies eine Kostenabschätzung und eine Einbettung in die Organisation und deren Arbeitsabläufe bedingt.

Generell kann man sagen, dass **es keinen durchschnittlichen Nutzer** gibt. Sogar innerhalb einer Community variiert die IT-Affinität der Nutzer beträchtlich. Jedoch sind in manchen Bereichen, wie z.B. in der Astronomie, IT-Kenntnisse weiter verbreitet als in anderen Bereichen, wie z.B. den Sozialwissenschaften.

Die **Anzahl der Personen**, die an einem Projekt mitarbeiten, ist generell **nicht einfach zu definieren**. Die meisten Organisationen setzen Mitarbeiter in mehreren Projekten ein, um so Synergie-Effekte zu

[19] http://en.wikipedia.org/wiki/Apache_Hadoop

nutzen. Auch gaben viele der Befragten an, dass es schwierig bis unmöglich sei, die Kosten für den Ingest eines Datensatzes zu kalkulieren.

Generell gibt es **signifikante Unterschiede zwischen den befragten Projekten** in Größe, Entwicklungsstand, Zukunftsperspektive, Anzahl der Kooperationen, etc. Da ein Ziel bei der Durchführung der Interviews war, eine möglichst große Bandbreite abzudecken, stellt dies kein ungewöhnliches Ergebnis dar. Die Interviews sind daher eher als **Indikatoren für die Breite des Themengebiets** zu betrachten und sollten weniger auf mögliche Generalisierungen hin untersucht werden.

Workshop- und Symposium-Ergebnisse

Im Rahmen des Radieschen-Projekts wurden ein Experten-Workshop und ein Symposium veranstaltet. Am Workshop (April 2012) nahmen ca. 85 Personen teil, am Symposium 135 Teilnehmer (Januar 2013). Die Teilnehmer kamen zumeist aus dem deutschsprachigen Raum und vertraten die unterschiedlichsten Forschungsdisziplinen und Fachrichtungen – von Tiermedizin und Geowissenschaften bis hin zu Rechenzentren und wissenschaftlichen Bibliotheken. Der Experten-Workshop unterteilte sich in Arbeitsgruppen mit den Themen

- WS1: Policies und Anreize: Was sind sinnvolle und notwendige Richtlinien im Umgang mit Forschungsdaten?
- WS2: Einbindung in den Forschungsprozess: Kommen die Daten zur Infrastruktur oder die Infrastruktur zu den Daten?
- WS3: Generische vs. Disziplinspezifische Dienste: Was sind die Erfolgskriterien disziplinübergreifender Dienste?
- WS4: Möglichkeiten und Grenzen der Auslagerung und Zentralisierung von Diensten

Das Symposium bestand aus einen Vortragsteil und kleineren Expertenrunden im Rahmenprogramm. Die Themen der Expertenrunden reichten von Daten-Management über Policy bis hin zu Virtuellen Forschungsumgebungen. Beiden Veranstaltungen gemein war das Ziel, die Forschungsdaten-Community zusammen zu bringen, zu vernetzen, sowie den Erfahrungsaustausch und die Diskussion der Teilnehmer untereinander zu fördern.

Im Rahmen des **Experten-Workshops** kamen die Wissenschaftler überein, dass die Frage des zentralen oder dezentralen Angebots von Diensten stark von Volumen der Daten bestimmt wird. Im Vordergrund sollte in erster Linie die Diskussion um „Heterogene vs. Homogene Daten“ stehen und weniger die Frage, ob es sich um ‚Big Data‘ oder ‚Small Data‘ handelt. Die vordringlichste Frage sei, ob Infrastrukturen existieren, um heterogene, hochkomplexe Daten einfach darzustellen und ob Infrastrukturen existieren, um aus sehr großen Datenmengen Informationen zu extrahieren. Bislang existiert eine Vielfalt von Formaten für sowohl große als auch kleine, heterogene als auch homogene Datensätze, welche schwer zu verwalten ist.

Eine weitere Schlussfolgerung der Diskussion des Experten-Workshops war, dass Werkzeuge für das Forschungsdaten-Management ein Niveau erlangen müssen, welches mit kommerziellen Werkzeugen vergleichbar ist. Desweiteren sollten Policies, Infrastrukturen und Anreizsysteme im Gleichklang behandelt und weiterentwickelt werden.

In der Diskussion wurde deutlich, dass auch ein kultureller Wandel in der Bewertung eines systematischeren Umgangs mit Forschungsdaten nötig ist und nachhaltige, nutzerfreundliche Anwendungen entwickelt werden müssen, welche sich nahtlos in die wissenschaftlichen Arbeitsabläufe einfügen.

Die große Zahl der Teilnehmer an dem **Symposium zu Forschungsdaten-Infrastrukturen (FDI 2013)** im Januar 2013 zeigte, dass das Thema Forschungsdaten als wichtiges Thema wahrgenommen wird. In den Diskussionen war zu beobachten, dass die Entwicklung von Forschungsdaten-Infrastrukturen nach wie vor heterogen verläuft. Die Diskussion zeigte jedoch auch, dass das Thema Forschungsdaten-Infrastrukturen eine konzeptionelle Reife erreicht hat, die mit Konzepten in anderen europäischen Staaten, den USA oder Australien vergleichbar ist.

Insbesondere in der technischen und der konzeptionellen Entwicklung ist eine gewisse Konvergenz zu beobachten. Daneben zeigen sich in anderen Bereichen konzeptionelle Unschärfen, bei denen noch Forschungsbedarf besteht, so z.B. bei den Themen „Qualität“, „Vertrauen“ und „Kosten- und Preismodelle“. Lücken bestehen auch noch bei Datenmanagement-Werkzeugen und deren Integration in die Arbeitsabläufe der Wissenschaftler. Die Teilnehmer bemängelten, dass viele Akteure im Datenmanagement nicht weit genug bekannt seien und auch die Vernetzung innerhalb der Community optimiert werden könnte. Auch sei der Bedarf an Angeboten in den Bereichen Qualifizierung und Beratung besonders ausgeprägt.

Bei der Veranstaltung wurde deutlich, dass eine Verbesserung des Umgangs mit Forschungsdaten nicht nur eine Frage der unterstützenden technischen Infrastrukturen ist, sondern auch einen kulturellen Wandel in der Wissenschaft erfordert. Der kulturelle Wandel im Hinblick auf Forschungsdaten bewegt sich in Richtung eines offeneren Umgangs mit diesem Teil der wissenschaftlichen Überlieferung, wie es im Bericht „Science as an Open Enterprise“ der Royal Society (2012) vorgeschlagen wird. Der Grad der Offenheit wird bestimmt durch die Spannung zwischen Vertrauen in die „Peers“ und Kontrolle über das eigene „Werk“. Neben den sozialen Normen muss auch noch der rechtliche Rahmen für Forschungsdaten weiterentwickelt werden.

Für eine Verbesserung der Situation sei es notwendig, die Erstellung von Daten, Software und Infrastrukturen neben den bisher üblichen Literaturveröffentlichungen als Beitrag im wissenschaftlichen Wertesystem zu verankern. Dies bedingt, dass der Nutzen einer Forschungsdateninfrastruktur für Forscher offensichtlicher werden muss. Datenpolicies und die Berücksichtigung dieser Leistungen in den institutionellen Bewertungssystemen könnten diesen Wandel unterstützen.

Positiv bewerteten die Teilnehmer die Möglichkeit, neue und vorbildliche Lösungen für den Umgang mit Forschungsdaten kennenzulernen. Einige Projekte und vielversprechende Ergebnisse wurden vorgestellt. Durch Evaluation der Ergebnisse und Erfahrungsaustausch kann hier eine Entwicklung angestoßen werden, welche über das Stadium des Experimentierens hinausgeht. Wichtig ist, dass der Austausch zwischen den Akteuren weitergeht und auch Praxisvermittlung einschließt, um den Kreis der Akteure zu erweitern.

Sowohl die Teilnehmer als auch die Organisatoren des Symposiums bewerteten die Veranstaltung als äußerst hilfreich für den Austausch von Ideen, der Generierung neuer Impulse und für die Vernetzung der Akteure untereinander. Über Folgeveranstaltungen wird bereits nachgedacht.

Ergebnisse der Diskussion des INF-Workshops

Am 11. April 2013 wurde vom Radieschen-Projekt in der SUB Göttingen ein Community-bildender Workshop der SFB-INF Projekte durchgeführt. Der Workshop war mit 40 Teilnehmern aus 21 verschiedenen SFB-INF Projekten sehr gut besucht.

Die SFBs erhalten mit 561 Mio € etwa 20% der von der DFG bewilligten Forschungsförderung. Das ist nach der Einzelförderung (ca. 35%) der zweitgrößte Anteil im Förderprogramm der DFG[20]. Derzeit befinden sich 232 SFBs in der Förderung, darunter 27 SFBs mit einem INF Teilprojekt.

Der Workshop traf auf ein großes Interesse im Kreis der SFB-INF-Projekte. Es konnte eine sehr weitgehende Abdeckung erzielt werden. Während der Veranstaltung zeigte sich dann ein hoher Diskussionsbedarf. Eine Vielzahl verschiedener Fragen wurde aufgeworfen und diskutiert.

Es hat sich gezeigt, dass die SFB-INF-Projekte entweder bei den ***Bibliotheken*** oder den ***Rechenzentren*** der jeweiligen Standorte angesiedelt sind. Mehrere Standorte, wie die Universitäten Bielefeld, Freiburg, Trier, Kiel etc. nutzen die SFB-INF-Projekte um standortweite Lösungen für eine Forschungsdaten-Infrastruktur aufzubauen.

Die typischen Aktivitäten der INF-Projekte, die sich im Vorfeld des Workshops durch eine Befragung der Teilnehmer herauskristallisierten, wurden im Workshop noch einmal bestätigt. In dieser Befragung haben 18 SFB-INF-Projekte als typische Aktivitäten benannt:

- Bereitstellung einer Plattform zur zentralen Speicherung und zum Austausch der Daten, z.B. Datenbank, Repository, Fileserver (13 Nennungen)
- Bereitstellung einer kollaborativen Arbeitsumgebung, z.B. Projektmanagement- oder Portalsoftware, selbstentwickelte Webportale mit integrierten Tools (10 Nennungen)
- Beratung und Unterstützung, z.T. auch Schulung, z.B. Datenaufbereitung, Datenanalyse, Metadaten, Policy-Entwicklung (7 Nennungen)
- Entwicklung, Implementierung und Bereitstellung von Tools, z.B. computerlinguistische Werkzeuge (7 Nennungen)
- Publikation (6 Nennungen)
- 4 Nennungen und weniger: Archivierung, Administration, Entwicklung von Standards und Formaten etc.

Ein besonders intensiv diskutierter Punkt waren die Fragen der Langzeitarchivierung und Nachhaltigkeit. Einige der SFB-INF-Projekte zielen darauf ab, die produzierten Daten langfristig vorzuhalten, während andere Teilnehmer davor warnten, die SFB-INF Aktivitäten mit solchen Ansprüchen zu überfrachten. Der Aufbau einer geeigneten und langfristig/nachhaltig verfügbaren Forschungsdaten-Infrastruktur sollte besser außerhalb der Projekte stattfinden. Die INF-Projekte können nicht dauerhaft und umfassend eine fehlende Forschungsdaten-Infrastruktur ersetzen. Die anwesende DFG-Expertin verwies hier klar auf die Verantwortung der Hochschulen , da diese den SFB beantragen und damit auch die nötige (digitale) Infrastruktur bereit stellen müssen. Dies schließe die Langzeitarchivierung und -verfügbarkeit mit ein.

Ein weiterer intensiv diskutierter Punkt war die Frage der Akzeptanz. Die Erfahrungen in Hinsicht auf die Akzeptanz des INF-Projektes durch die anderen Teilprojekte sind sehr unterschiedlich. In einem

[20] Die genannten Zahlen beziehen sich auf das Jahr 2011 (vgl. Effertz und Schoch 2013)

Fall (SFB 649) ist der PI des INF-Projektes identisch mit dem Sprecher des Gesamt-SFBs; dieses SFB-INF-Projekt konnte von keinen Herausforderungen hinsichtlich seiner Akzeptanz berichten. Aus den anderen SFB-INF-Projekten wurden verschiedene Erfahrungen berichtet. Die Akzeptanz der anderen Teilprojekte zu erlangen ist ein langwieriger und zeitaufwendiger Prozess, der vor allem dann gelingt, wenn die SFB-INF Projekte sehr genau auf die Bedürfnisse und die Prozesse im Arbeitsablauf der anderen Wissenschaftler eingehen.

Beim Ausblick wurden einige konkrete Wünsche für künftige Aktivitäten formuliert. Insbesondere wurde der Wunsch geäußert, auch künftig weitere SFB-INF-Workshops durchzuführen. Betont wurde, dass insbesondere ein sehr fokussierter Workshop-Ansatz entlang spezifischer Themenaspekte erwünscht ist. Zu den genannten Themen gehören beispielsweise:

- Akzeptanz bzw. wie diese erhöht werden kann
- eingesetzte Tools, Technologien, kollaborative Arbeitsumgebungen (Nachnutzung)
- Heterogenität (von Daten, Formaten, Anforderungen, Technologien etc.)
- Policies, z.B. die Forschungsdaten-Policiy eines SFBs, aber auch das Forschungsdaten-Management der Teilprojekte

6. Querschnittsthemen

Die **Kostenaufteilung zwischen den Akteuren** innerhalb einer Organisation ist nach wie vor ungeklärt. Zwar ist eine grobe Kostenschätzung möglich, welche Kosten durch Ingest, Aufbereitung der Daten und deren Verwahrung entstehen, jedoch ist es zumeist nicht möglich, diese Kosten den einzelnen Akteuren, wie z.B. Mitarbeitern oder Abteilungen eines Instituts, zu zuordnen (siehe hierzu auch Report „Kosten“). Eine eindeutige Zuordnung zu einer Kostenstelle ist zumeist nicht gegeben.

Eine offene Frage bleibt ebenso die Aufteilung der Kosten zwischen dem daten-produzierenden Projekt (z.B. das Projekt „Radieschen“), den Instituten, an denen das Projekt angesiedelt ist (z.B. der Abteilung GFZ-CeGIT) und der Institution (hier das Helmholtz Zentrum Potsdam Deutsches GeoForschungsZentrum GFZ), welche das Projekt leitet. Hierzu bestehen bislang bei den großen Wissenschaftseinrichtungen keine Richtlinien. Als eine Folge bleiben somit auch das Fortbestehen und die Instandhaltung bestehender Forschungsdaten-Repositorien im Ungewissen. Fördermittelgeber fördern Projekte zur Lösung aktueller Herausforderungen. Bestehende Repositorien werden als Bestand, und somit in das Aufgabengebiet der übergreifenden Institution fallend, gesehen. Die übergreifende Institution wiederum verweist auf die Einwerbung von Fördermitteln zum Weiterbetrieb solcher Repositorien. Die Einführung einer Sonderabgabe, ähnlich des Solidaritätszuschlags, bei der Bewilligung von Projekten wäre hier eine Möglichkeit, die aktuelle Lage zu verbessern. Dieser Zuschlag wäre zu gleichen Teilen von der übergreifenden Institution und dem Fördermittelgeber zu tragen. Zudem sollte das Thema Nachhaltigkeit bei der Beantragung zukünftiger Projekte besser berücksichtigt werden, etwa durch die Ausarbeitung eines Nachhaltigkeitsplans als Teil der Projektarbeit.

Ein immer wiederkehrendes Thema der Diskussion ist das **Wertesystem zur Anerkennung der wissenschaftlichen Arbeit in ihren verschiedenen Formen**. (Siehe hierzu auch den Report von Knowledge Exchange zu „The Value of Research Data - Metrics for datasets from a cultural and technical point of view“ [21]). Bislang in den Metriken berücksichtigt werden die traditionellen

[21] http://www.knowledge-exchange.info/datametrics

Formen der Publikationen (H-Index[22], Journal Impact Factor (JIF)[23]). Daten- oder Software-Publikationen haben bei diesen Indices bislang keine bis wenig Relevanz. Eine Anerkennung solcher Veröffentlichungen würde das Bewusstsein für die Relevanz von Forschungsdaten und deren Publikation deutlich stärken und somit auch die Weiterentwicklung von Forschungsdaten-Infrastrukturen vorantreiben. Ähnliches gilt auch für die Veröffentlichung von Software. Hilfreich in diesem Kontext wäre die Entwicklung neuer Kategorien bei der Generierung von Kennzahlen in wissenschaftlichen Evaluationssystemen, welche verschiedene Formen von Publikationen gleichberechtigt berücksichtigen, sowie eine verstärkte Unterstützung und Werbung für die neuen Formen der (Open Access) Journals zu Daten und Software-Publikation durch die Verleger, Forscher und auch durch die Fördermittelgeber.

Modernen Kommunikationsformen wie z.B. den **Sozialen Medien** stehen viele Wissenschaftler eher zurückhaltend gegenüber. Dies gilt auch für Wissenschaftler der Generation Y, die den wissenschaftlichen Nachwuchs darstellen, jedoch noch nicht zu den sogenannten „Digital Natives[24]“ zählen. Social Media und Online Foren in der Forschung werden zumeist nicht als legitime Werkzeuge der Forschung akzeptiert[25].

Neue web-basierte Werkzeuge und andere, neuartige Anwendungen jedoch können zur internationalen Vernetzung der Wissenschaftler beitragen, zum Datenaustausch und auch zur weiteren Datenerhebung eingesetzt werden. So nutzt zum Beispiel bereits eine kleine Gruppe von Wissenschaftlern verstärkt Open Data, exploriert Crowd Sourcing Mechanismen und setzt Citizen Science[26] als Unterstützung ihrer Arbeit ein (ESA – Projekt AstroDrone[27]). Diese unkonventionelle Art der Herangehensweise birgt viel Potential und sollte eine entsprechende Unterstützung finden.

7. Ausblick und Empfehlungen

Der vorliegende Synthese-Report beschreibt den Status Quo der Entwicklung von Forschungsdaten-Infrastrukturen in Deutschland. Es wird die aktuell verwendete Technologie beschrieben, ein Überblick über Organisationsstrukturen gegeben und die Kostenverteilung untersucht. Desweiteren zeigen die einzelnen Reports bestehende Lücken in der Entwicklung auf und geben Hinweise auf Entwicklungsbedarf in der nahen Zukunft.

Wie jedoch sieht die Entwicklung von Forschungsdaten-Infrastrukturen in der weiteren Zukunft aus? Wird die Entwicklung geradlinig verlaufen, d.h. ist es absehbar, welche Neuerungen im Verlauf der Zeit eintreten werden? Sind bahnbrechende Umwälzungen zu erwarten? Lassen sich dafür bereits Anzeichen erkennen?

Die Technikgeschichte lehrt uns, dass Innovationen mit zunächst gering eingeschätzten Auswirkungen durchaus das Potential entwickelten, bereits bewährte Technologien vom Markt zu verdrängen und deren Platz einzunehmen. Beispiele sind der Siegeszug des Telefons, welches die Telegrafie ablöste oder Wikipedia, welches als Online-Enzyklopädie führend wurde und damit

[22] http://de.wikipedia.org/wiki/H-Index
[23] http://de.wikipedia.org/wiki/Impact-Faktor
[24] http://de.wikipedia.org/wiki/Digital_Native
[25] The British Library and JISC (2012)
[26] The Royal Society (2012)
[27] ESA Projekt “AstroDrone“ http://www.esa.int/ger/ESA_in_your_country/Germany/Smartphone-App_verwandelt_Spielzeug-Drohne_in_Raumsonde

langjährig etablierte Werke wie z.B. die Encyclopaedia Britannica von ihrer Position verdrängte. Einiges, was uns heute schon selbstverständlich vorkommt, existierte vor zehn Jahren noch nicht einmal in unserer Vorstellung. Im Zusammenhang mit Forschungsdaten illustrieren die Begriffe „Grid“ und „Cloud“ gut die Dynamik dieser Entwicklung. Ende der 1990er kam der Begriff „Grid“ auf als das Versprechen, unbegrenzte IT-Ressourcen quasi „aus der Steckdose“ beziehen zu können. Für den privaten Nutzer schien das ohne Relevanz zu sein, da sich das Grid-Konzept in erster Linie an Nutzergemeinschaften orientierte. Gewerbliche Anwendungen, wie sie in der Förderung des BMBF auch gedacht waren, wurden nie etabliert, da die angesprochenen Firmen kein Vertrauen in die Sicherheit der Grid-Anwendungen hatten.[28]

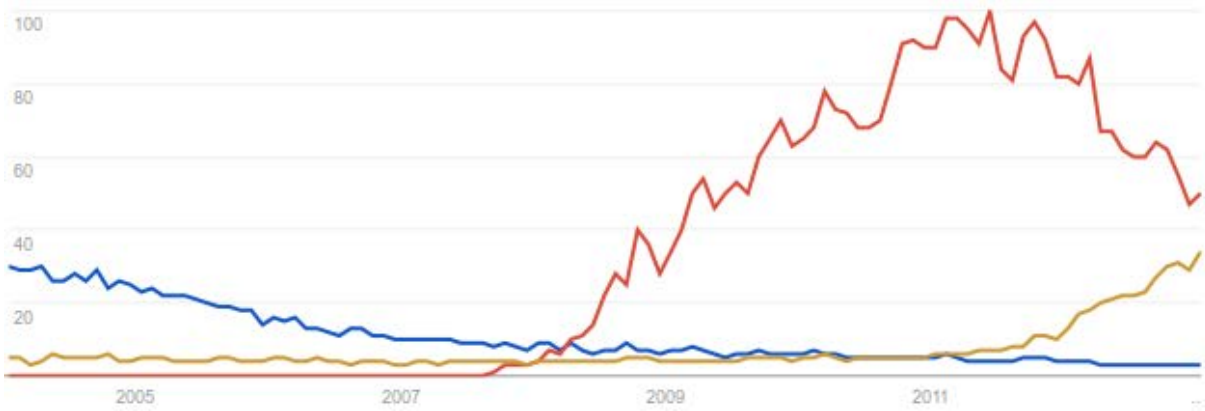

Abb. 9 Histogramm der Anfragen bei Google zu den Begriffen „Grid Computing“ (blau), „Cloud Computing“ (rot) und „Big Data“ (gelb) im Januar 2013. Quelle: Google Trends.

Mit dem Aufkommen des Begriffs „Cloud“ verschwand der Begriff „Grid“ praktisch in der Bedeutungslosigkeit (Abb. 9).

Was bedeutet das für Technologien und Dienste für den Umgang mit Forschungsdaten? Ein Blick auf die oben skizzierten Trends zeigt die Dynamik der Entwicklung und macht deutlich, wie schwer es ist, die Entwicklung des Umgangs mit Forschungsdaten für die nächsten zehn Jahren vorherzusagen. Es ist unmöglich vorherzusagen, welche technischen Lösungen zur Verfügung stehen werden. Auch Trends lassen sich nur in begrenztem Maße identifizieren, denn die Entwicklung wird weiterhin stark von *disruptive innovation* Mustern beeinflusst, was für sich selbst wiederum einen Trend in der weiteren Entwicklung darstellt.

Was ist nun wirklich ein Trend und was ist nur aufgebauscht? Ist „Big Data“ tatsächlich ein bedeutender Trend in der Wissenschaft? Werden die „Digital Natives“, die eine Welt ohne Internet nicht kennen, als die Wissenschaftler von morgen diese Technologien anders und freier nutzen? Hier hilft es, von den Technologien zu abstrahieren und zu fragen, welche Prozesse technisch unterstützt werden sollen.

Im Jahr 2003 erschienen der einflussreiche Artikel „e-Science and its implications“[29] von Hey und Trefethen, mit dem der Begriff der „Datenflut“ (*data deluge*) in die Diskussion eingeführt wurde. Die

[28] Klump, 2008.
[29] Hey und Trefethen, (2003).

Diskussion darüber war damals noch sehr mit den erwarteten Datenmengen befasst. Mit dem technischen Fortschritt kamen jedoch auch andere Aspekte mit ins Blickfeld, nämlich die Möglichkeit durch explorative Analyse der Daten neue Hypothesen zu formulieren und zu prüfen. Hier hängt der wissenschaftliche Fortschritt unmittelbar an der Verfügbarkeit der Daten und der Möglichkeit sie zu verarbeiten, was auch mit dem Begriff „*data intensive science*“ bezeichnet wird.[30]

Disruptive Innovation[31] läßt sich an vielen weiteren Beispielen der Technikgeschichte nachverfolgen. Allen gemein ist der Beginn der Innovation in einem Nischenmarkt unbemerkt oder unbeachtet von den Marktführern. Eine Disruptive Innovation kann aus einer neuen Technologie, einem neuartigen Produkt oder auch einer innovativen Dienstleistung bestehen. Abb. 10 zeigt den Verlauf einer Disruptiven Innovation, hier bezeichnet als Disruptive Technology. Die Technologie ist zunächst den etablierten Produkten unterlegen. Nach Akzeptanz im unteren Marktsegment (Low Quality Use) verbessert sich die Technologie und höherwertige Märkte werden erobert (Medium Quality Use – High Quality Use) bis schließlich ein voll ausgereiftes Produkt angestammte Marktführer von ihrer Position verdrängt (Most Demanding Use).

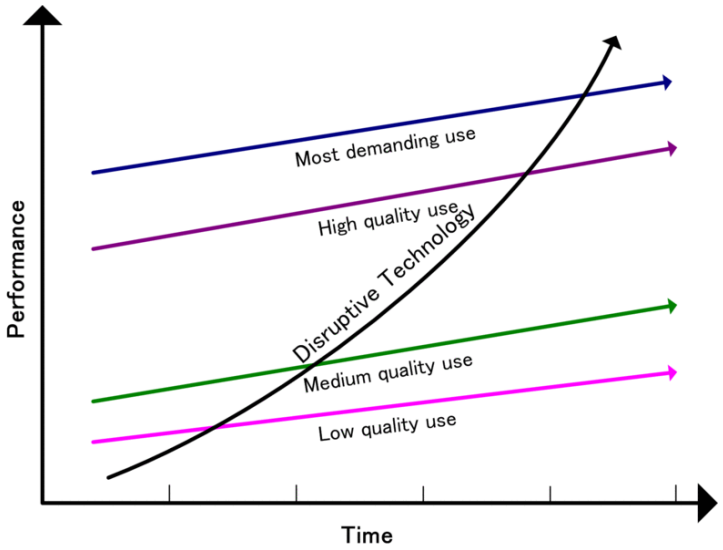

Abb. 10: Verlauf einer Disruptiven Innovation (Quelle: Wikipedia)

Auch im Bereich Forschungsdaten und Forschungsdaten-Infrastrukturen sind weitere Innovationen zu erwarten, sowohl auf technologischer Ebene (Academic Cloud Computing, Data-Driven Research), als auch auf sozialer Ebene (Wertesystem, Publikationsmetriken). Nicht alle diese Neuerungen lassen sich als Disruptive Innovationen bezeichnen. Jedoch sollte man auch im Bereich Forschungsdaten-Infrastrukturen offen sein für Innovationen, Trends beobachten und neue Entwicklungen fördern. Abb. 11 zeigt einen Blick auf den Gartner Hype Cycle zu Emerging Technologies. Der Hype Cycle zeigt Neue Technologien von ihrem ersten Aufkommen im Bereich der „Technology Trigger“, dem Beginn des Hypes über das „Tal der Desillusionierung“ bis hin zur weitreichenden Akzeptanz der Technologie und Umsetzung in marktfähige Produkte.

[30] McNally u.a., (2012)
[31] http://en.wikipedia.org/wiki/Disruptive_innovation

Als Schlussfolgerung könnte es für Fördermittelgeber daher interessant sein, **in der Programmsteuerung agile Komponenten einzubinden**, um schnell und flexibel auf neue Entwicklungen reagieren zu können. Auch wäre vorstellbar, Projekte mit Think-Tank Charakter oder Projekte als eine Art Versuchsballon für neue Technologien und Entwicklungen zu fördern.

Wünschenswert im Bereich Forschungsdaten-Infrastrukturen wäre auch ein **Projekt, welches das Zusammenwirken der Forschungsprojekte dieses Förderbereichs koordiniert**. Ein solches Projekt sollte auf die Durchführung von Networking-Aktivitäten, dem Austausch und der Verbreitung der Ergebnisse und dem Aufbau einer gemeinsamen Wissensbasis der Projekte fokussieren.

Forschungsdaten-Infrastrukturen befinden sich, von einigen Ausnahmen abgesehen, in der generell eher konservativ ausgerichteten Umgebung von Wissenschaftlichen Bibliotheken und den klassischen Wissenschaften. Die Erforschung von Sachverhalten und das Bereitstellen von Diensten steht hier im Vordergrund und weniger die Suche nach Innovationen. Ein Projekt zum Thema Innovation Management & Forschungsdaten-Infrastrukturen oder **auch Open Innovation im Forschungsdaten-Kontext** könnte der weiteren Entwicklung von Forschungsdaten-Infrastrukturen und deren Öffnung gegenüber neuen Trends und Innovationen förderlich sein.

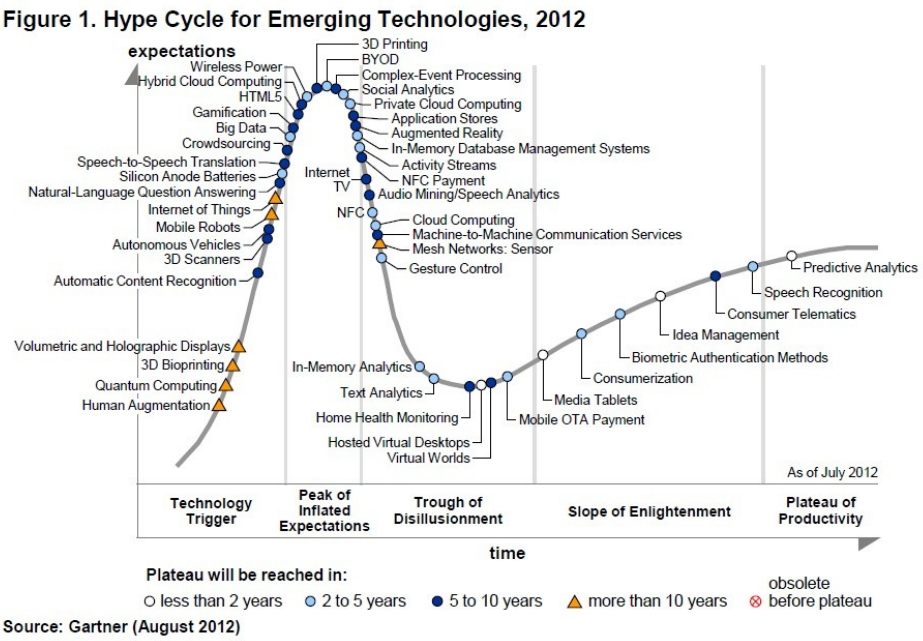

Abb. 11 Der Gartner Hype Cycle bieten eine grafische Repräsentation über den Status der Reife und der Annahme neuer Technologien und Anwendungen und zeigen deren potentielle Relevanz bei der Lösung realer Geschäftsprobleme und der Exploration neuer Möglichkeiten. Der Hype Cycle ‚Emerging Technologies' fokussiert auf Entwicklungen mit breiter, industrie-übergreifender Bedeutung und einem hohen Transformationspotential.

Jedoch sind auch Neuerungen in der Informationstechnologie oft kurzlebig. Viele der beispielsweise von Google angebotenen Dienste haben eine durchschnittliche Lebensdauer von etwa drei Jahren[32], was in etwa der Dauer von Innovationszyklen in der Informationstechnik entspricht. Das Beispiel Google zeigt, dass es auch für einen Weltkonzern nicht möglich ist, den Erfolg eines seiner Dienste vorherzusehen und deshalb begegnet man dieser Herausforderung, in dem man ein Portfolio von Diensten auf einer gemeinsamen Plattform entwickelt. Einzelne Dienste können dann bei

[32] Arthur, C. (2013)

ausbleibendem Erfolg wieder abgeschaltet werden, ohne den Betrieb der Plattform als Ganzes zu beeinträchtigen. Eine erfolgreiche Strategie für Infrastruktureinrichtungen könnte sein, ein **modularisiertes Portfolio von Diensten** zu entwickeln, das auf einer gemeinsamen Plattform aufbaut. Diese Strategie würde es den Einrichtungen erlauben, die Dienste flexibel den sich stets ändernden Bedürfnissen der Forschung anzupassen, während die darunter liegende Plattform stetig weiterentwickelt werden kann. Auf diese Weise ließe sich dem Widerspruch zwischen dem Anspruch der Infrastruktur an Stabilität mit den Anforderungen flexibler, und möglicherweise kurzlebiger, Anwendungen überbrücken.

Generell ist die Entwicklung von Forschungsdaten-Infrastrukturen in Deutschland geprägt von einer durchaus aktiven Forschungsdaten-Community. Förderlich wäre jedoch eine **verstärkte Zusammenarbeit mit Projekten auf europäischer Ebene**, sowie eine Öffnung gegenüber anderen Fachrichtungen wie dem bereits genannten Innovation Management. Beachtung finden sollte auch die **parallel verlaufende Entwicklung in der Wirtschaft** im Umgang mit großen Datenmengen („Big Data“) und, in diesem Kontext, in der **Herausbildung des Berufs des Data Scientists**. Trotz unterschiedlicher Ausrichtungen und Ziele gibt es hier deutliche Parallelen[33]. Es wäre durchaus denkbar, diese Entwicklungen zu korrelieren oder zumindest zu beobachten und daraus Anregungen für die weitere Entwicklung von Forschungsdaten-Infrastrukturen zu übernehmen.

8. Literaturverzeichnis

Arthur, C. (2013), Google Keep? It’ll probably be with us until March 2017 - on average, The Guardian, 22. März. [online] Available from: http://www.guardian.co.uk/technology/2013/mar/22/google-keep-services-closed

Biesdorf, S.; Court, D. and Willmott, P. (2013), “Big Data: What’s your plan?”, McKinsey & Company, McKinsey Quarterly - Insights & Publications, March 2013, http://www.mckinsey.com/insights/business_technology/big_data_whats_your_plan

Effertz, E.; Schoch, K. (2013), Teilprojekte zur Informationsinfrastruktur in Sonderforschungsbereichen. „INF-Projekte“. (Präsentation anlässlich des Gemeinsamen Workshops der SFB-INF-Projekte am 11. April 2013 in Göttingen)

Gausemeier, J., Stoll, K., Wenzelmann, C. (2007), Szenario-Technik und Wissensmanagement in der strategischen Planung.
In: Gausemeier, J. (Hrsg.) Vorausschau und Technologieplanung, 1.Aufl. ,S.3-30, HNI, Paderborn, 2007

Sontheimer, K. (1970), Voraussage als Ziel und Problem moderner Sozialwissenschaft. In: Klages, H.: Möglichkeiten und Grenzen der Zukunftsforschung. Herder, Wien, Freiburg, 1970

Hey, T., und A. Trefethen (2003), e-Science and its implications, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 361(1809), 1809–1825, doi:10.1098/rsta.2003.1224.

Klump, J. (2008), Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten, nestor-Materialien, Kompetenznetzwerk Langzeitarchivierung (nestor), Frankfurt (Main), Germany. [online] Available from: http://nbn-resolving.de/urn:nbn:de:0008-2008040103.

[33] Biesdorf, S. Court, D. and Willmott, P. (2013)

McNally, R., A. Mackenzie, A. Hui, und J. Tomomitsu (2012), Understanding the “Intensive” in “Data Intensive Research”: Data Flows in Next Generation Sequencing and Environmental Networked Sensors, IJDC, 7(1), 81–94, doi:10.2218/ijdc.v7i1.216.

The British Library and JISC (2012), Researchers of Tomorrow: the research behavior of Generation Y students, June 2012, http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow.aspx

The Royal Society (2012), Science as an open enterprise, June 2012, http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf