Speichervolumen- und Kostenabschätzung für die Archivierung von Forschungsdaten

Auswertung einer Bedarfsumfrage an der Universität Osnabrück

Marco Gronwald

Das Thema Forschungsdaten und insbesondere das Forschungsdatenmanagement geraten zunehmend in den Fokus von Infrastrukturdienstleistern der Wissenschaft. Neben den zu etablierenden Servicedienstleistungen für die Forschenden, gewinnen auch vermehrt Speicherorte für die Langzeitarchivierung an Relevanz. Die vorliegende Arbeit leitete aus einer Umfrage an der Universität Osnabrück den zukünftigen Speicherbedarf ab und brachte ihn in Beziehung mit den entstehenden möglichen Kosten. Abhängig von der Speichermenge und Datenformaten, wächst der Speicherbedarf unterschiedlich schnell an und spiegelt sich in den anfallenden Kosten wider. Somit bleibt die Frage zu klären, welche Daten wo veröffentlicht oder archiviert werden (externes Repositorium vs. eigenes Rechenzentrum).

Research data and research data management is becoming increasingly be focused by infrastructure service providers in science. In addition to the services for researchers, storage for long-term archiving is also becoming more relevance in future. From a survey at the University of Osnabrück, the present work derived the future storage requirements and related them to the possible costs. Depending on the amount of memory and data formats, the storage requirements increase at different speeds and are reflected in the gained costs. Thus, the question remains to clarify which data is published or archived where (external repository vs. own data center).

1 Einleitung

Seit dem Jahr 2005 hat sich das weltweite Datenvolumen alle zwei Jahre verdoppelt (theatlas.com, 2015). Während die Menge an digitalen Daten im Jahr 2016 noch 16.1 Zettabyte (ZB) betrug, wird ein Anstieg auf 163 ZB (de.statistia.com, 2017) bis 180 ZB (whatsthebigdata.com, 2016) im Jahr 2025 prognostiziert. Welchen Anteil daran Forschungsdaten haben, ist derzeit noch ungeklärt.

Forschungsdaten sind Daten, die während des Forschungsprozesses entstehen oder Ergebnisse daraus sind (Kindling et al., 2013, S. 130). Die Deutsche Forschungsgemeinschaft (DFG) führte mit ihren „Leitlinien zum Umgang mit Forschungsdaten“ von 2015 aus, dass Forschungsdaten eine wesentliche Grundlage für das wissenschaftliche Arbeiten seien (DFG, 2015). Die Vielfalt solcher Daten entspricht hierbei der Vielfalt der unterschiedlichen wissenschaftlichen Disziplinen, Erkenntnisinteressen und Forschungsverfahren der MINT-Fächer als auch in den Geisteswissenschaften (Lindstädt, 2015, 2016; Neuroth, 2012).

Zu diesen Daten zählen u.a. Messdaten, Laborwerte, audiovisuelle Informationen, Texte, Surveydaten, Objekte aus Sammlungen oder Proben, die während der wissenschaftlichen Arbeit entstehen, entwickelt oder ausgewertet werden. Methodische Testverfahren, wie Fragebögen, Software, Modelle und Simulationen können ebenfalls zentrale Ergebnisse wissenschaftlicher Forschung darstellen und sollten daher ebenfalls unter dem Begriff Forschungsdaten gefasst werden (DFG, 2015).

Die von diversen Institutionen geforderte Transparenz des Forschungsprozesses stellt zugleich die Gewährleistung der Nachvollziehbarkeit von Forschung dar und ist grundlegend für die Qualitätssicherung der Wissenschaftskommunikation (Kindling, 2013, S. 138). Allerdings ist dies auf der Basis des konventionellen Publikationsverhaltens in der Wissenschaft nur in den seltensten Fällen möglich. Die Sicherung der Qualität, vor allem die weitere (Nach)-Nutzung von Forschungsdaten, wird dadurch erschwert. Erst wenn die Forschungsdaten zugänglich bzw. publiziert sind, wird die Möglichkeit geschaffen, die wahre, empirisch fundierte „Aussagekraft der auf dieser Grundlage ermittelten Erkenntnisse parallel zur methodischen Korrektheit und Logik in der Schlussfolgerung zu prüfen“ (Kindling, 2013, S. 138). Darüber hinaus erhöht die langfristige Bereitstellung von Forschungsdaten auch die Reichweite bzw. den Impact von Forschungsergebnissen oder gar die Zitationen der Wissenschaftler signifikant (Piwowar et al., 2013).

1.1 Wissenschaftspolitische Ebene

Das Management von Forschungsdaten benötigt allerdings auch die nötige Infrastruktur. Im Zuge wissenschaftlichen Arbeitens entstehen vermehrt digitale Daten, welche zugleich einer möglichen Belegfunktion dienen. In diesem Kontext werden nicht nur Geräte und Datenbanken angesprochen, sondern der Forschungsprozess selbst, da nahezu alle Forschungswerkzeuge eine relevante IT-Komponente mit

einschließen (RfII, 2016, S. 12). Internationale sowie nationale Akteure der Wissenschaftscommunity und wissenschaftspolitischer Entscheidungsträger (wie z.B. die Wissenschaftsminister der G8-Staaten) wurden an dieser Stelle auf das Thema aufmerksam und reagierten mit einer Vielzahl an Stellungnahmen, Studien und Positionspapieren zu Themen des digitalen Wandels (BMBF, 2017; European-Commission, 2016, 2017; G8, 2013; OECD, 2004). Diese Papiere beinhalten allesamt das Thema „Open-Data" und fordern neben dem systematischen Aufbau von Dateninfrastrukturen Anreize für einen Paradigmen- und Kulturwandel in der Wissenschaft sowie neue Berufsbilder (RfII, 2016, S. 12). Die Europäische Kommission postulierte bereits 2010 die „Collaborative Data Infrastructure (CDI)", welche einen Zugriff auf als vertrauensvoll eingestufte Daten sowie deren Nachnutzung unterstützt. Daran anschließend gründeten sich auch unter deutscher Beteiligung EU-geförderte Initiativen (Research Data Alliance (RDA), European Data Infrastructure (EUDAT) und European Open Science Cloud (EOSC), die als Maßgabe eine Infrastrukturentwicklung haben (RfII, 2016, S. 13).

Im nationalen Kontext verschärfte sich der Diskurs 2011 durch die „Kommission Zukunft der Informationsinfrastruktur (KII)" (KII, 2011) und den damit verknüpften Empfehlungen des Wissenschaftsrates von 2011 und 2012 (Wissenschaftsrat, 2012, 2011), Informationsinfrastrukturen auszubauen. Die Bundesregierung formulierte in ihrer „Digitalen Agenda 2014-2017" das Ziel, einen besseren Zugang zu Forschungsdaten als Teil einer Internationalisierungsstrategie für die Wissenschaft zu ermöglichen (Bundesregierung, 2014). Dem folgten Strategien des Bundesministeriums für Bildung und Forschung zum Thema Open Access (auch von Forschungsdaten) und Kompetenzerwerb im Umgang mit Forschungsdaten (BMBF, 2017, 2016). Die Hochschulrektorenkonferenz erkannte 2014 ebenfalls, dass das exponentielle Wachstum, die stetig steigende Komplexität und der vermehrte Gebrauch von digitalen Forschungsdaten einen maßgeblichen Einfluss auf den Forschungsprozess haben. Daher der Appell der HRK an die Hochschulleitungen, Leitlinien zum Umgang mit Forschungsdaten abzustimmen und zu vereinbaren sowie die Informationskompetenz der Wissenschaftler zu stärken und die strukturellen Voraussetzungen für ein effizientes, institutionelles Forschungsdatenmanagement zu schaffen (HRK, 2014). Ein Jahr später postulierte die HRK einen Orientierungspfad, sowie konkrete Handlungsoptionen und Szenarien, wie die Hochschulleitungen die Entwicklung eines Forschungsdatenmanagements steuern können (HRK, 2015).

1.2 Kostenentstehung

Für die Langzeitarchivierung von Forschungsdaten fallen durch mehrere Komponenten Kosten an. Einerseits durch die für den Betrieb des Archivs benötigte technische Infrastruktur und andererseits durch die umfangreichen organisatorischen sowie administrativen Tätigkeiten. Beucke (2010) sowie Hagel et al. (2009) untersuchten in ihren Arbeiten, wie Geschäftsmodelle für die Langzeitarchivierung beschaffen sein sollten und welche Komponenten dazu notwendig sind. Hagel et al. (2009, S. 102-114) stellen 15 Kostenmodelle vor und vergleichen diese. Jedoch stellen diese eher theoretische bzw. schematische Modelle dar, mit denen sich keine konkrete Kosten errechnen lassen können. Dies ist zunächst auch nicht nötig, da diese der Abschätzung und/oder Planung von Sollkostenanteilen dienen, die durch eine spätere Istkostenrechnung nicht übertroffen werden sollen (Hagel et al., 2009, S. 103). Nach Wollschläger (2008, S. 1-2) lassen sich die Kosten, die durch eine Langzeitarchivierung entstehen, in drei Arten gliedern:

- Initiale Kosten: Entstehung bei der Informationsbeschaffung und Projektplanung zu einem Langzeitarchivierungssystem.
- Beschaffungskosten: Entstehung von Kosten bei der Beschaffung von Hard- und Software, aber auch bei dem Personalrekruiting.
- Betriebskosten: Kosten, die beim dauerhaften Betrieb des Archivierungssystem entstehen.

Wobei die konkreten Kosten dabei jeweils abhängig sind von: Der Menge der digitalen Daten, der Heterogenität der Datenformate, der Qualität der Metadaten und der gewünschten Zugriffsmöglichkeiten und Schnittstellen, sowie den Anforderungen Dritter an die archivierende Institution bzw. Verpflichtungen der Institution gegenüber Dritten.

1.3 Ziel der Arbeit

Wollschläger (2008, S. 2) stellt dar, dass die konkreten Kosten der Langzeitarchivierung abhängig sind von ihrer Menge und der Heterogenität der Datenformate. Hierzu soll mit einer ersten Prognose ermittelt werden, welcher Speicherbedarf durch die derzeit ‚gespeicherten' Forschungsdaten auf den Dienst-PCs der Forschenden der Universität Osnabrück besteht und wie dieser sich über die kommenden Jahre entwickeln könnte. Neben der Archivierung der Forschungsdaten auf einem (noch zu etablierendem) institutionellen Repositorium an der Universität, bietet das disziplinunabhängige Repositiorium RADAR (Research Data Repository) des FIZ Karlsruhe die Möglichkeit, Forschungsdaten zu archivieren und nach Bedarf zu publizieren. Hierzu soll die Frage geklärt

werden, wie hoch die Kosten einer Datenarchivierung bei RADAR über einen Zeitraum von zehn Jahren wären. Abhängig vom Datenvolumen und der Option der Datenpublikation, kann daraus abgeleitet werden, ab welchem Datenvolumen RADAR als Dienstleistung genutzt werden könnte.

2 Methodik und Datengrundlage

Um den Status Quo zum Umgang mit Forschungsdaten an der Universität Osnabrück zu ermitteln, wurde im dritten Quartal 2017 eine Online-Befragung durchgeführt, die sich an alle Forschenden der Universität Osnabrück richtete (Gronwald, 2018). Als Plattform für die Online-Befragung wurde das vorkonfigurierte Online-Befragungssystem des Fachbereichs 1 (FB Kultur- und Sozialwissenschaften), welches auf dem System LimeSurvey (Version 2.64.0+170307, LimeSurvey GmbH, Hamburg) basiert, verwendet. Der Fragebogen bestand aus 16 Einzelfragen, die in vier Fragegruppen untergliedert waren: Allgemeine Fragen, Kenntnisfragen zu Forschungsdaten, Fragen zu Regelungen sowie Fragen zur Selbsteinschätzung und zum Schulungsbedarf. Wobei für die Abschätzung des Speicherbedarfs nur eine Frage herangezogen wurde und in diesem Artikel dargestellt wird. Der Online-Fragebogen wurde im Vorfeld an vier Fachreferenten der Universitätsbibliothek zu Testzwecken geschickt. Nach einigen Anpassungen wurde die finale Version erstellt.

In Zusammenarbeit mit der Universitätsleitung, dem Personaldezernat und der Vizepräsidentin für Forschung und Nachwuchsförderung wurden zunächst die Emailadressen aller Professorinnen und Professoren zusammengetragen. Am 11.9.2017 wurde die E-Mail mit der Bitte um Teilnahme sowie dem Link zur Online-Umfrage über den Emailverteiler der Hochschulleitung an alle 217 Professorinnen und Professoren (Stand: 31.12.2017) der Universität Osnabrück verschickt. Die Umfrage war vom 11.9. bis zum 31.12.2017 freigeschaltet. Dieser Email war ein Begleitschreiben angehängt, welches die Unterstützung der Umfrage durch die Vizepräsidentin für Forschung und Nachwuchsförderung versicherte. Darin wurde auch gebeten, die Umfrage an die wissenschaftlichen Mitarbeiterinnen und Mitarbeiter, Doktorandinnen und Doktoraden, Bachelor- und Masterkandidaten sowie an studentische Hilfskräfte weiterzuleiten.

2.1 Teilnahme und Rücklauf

Die Anzahl der im wissenschaftlichen Dienst beschäftigten Mitarbeiter beträgt 968 (Stand: 31.12.2017). Ab dem 24.10.2017 wurden die Fachreferentinnen und Fachreferenten der Universitätsbibliothek darum gebeten in ihren Fachbereichen regelmäßig an die

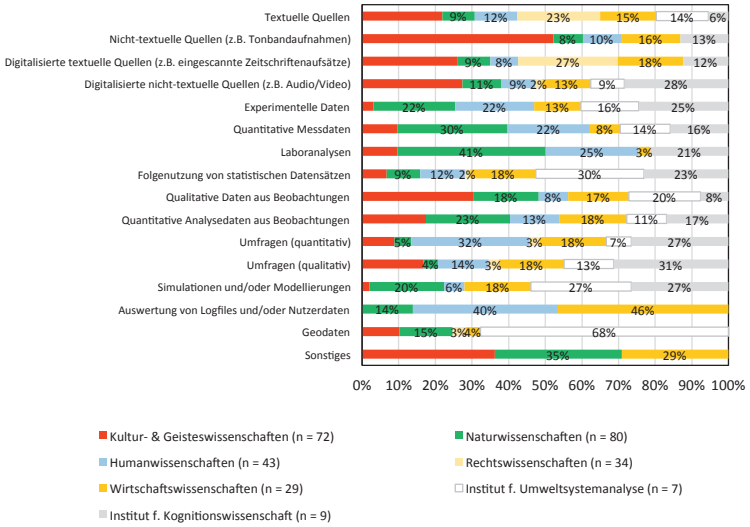


Abbildung 1: Relative Verteilung der Datenformate auf die Fachbereiche. Abgebrochene Fragebögen wurden herausgerechnet (Nennungen = 1037; n = 258).

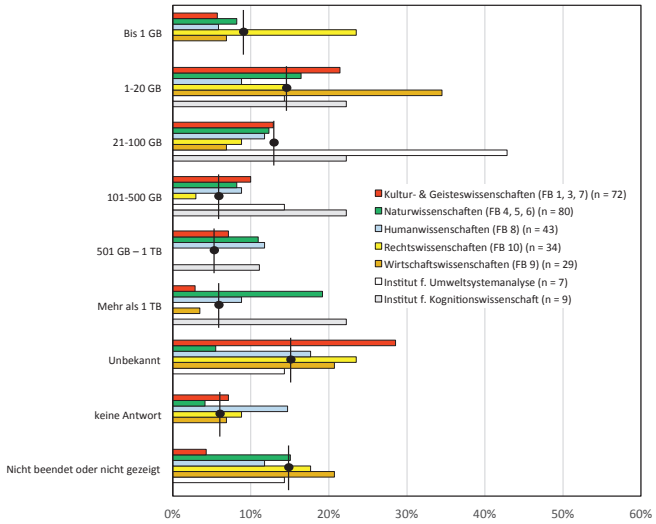


Abbildung 2: Gesamtgröße der Forschungsdaten aus bisherigen oder derzeit laufenden Projekten (bzw. Abschlussarbeit) ungefähr? (• = Mittelwert aller Fachbereiche; n = 258).

Teilnahme an der Umfrage zu erinnern. Zum Ende der Umfrage haben 278 Personen an der Umfrage teilgenommen, wobei 221 Fragebögen vollständig ausgefüllt wurden (ca. 80%). Bei einer Grundgesamtheit von 968 Personen und einem Konfidenzniveau von 95% läge die minimale Stichprobengröße bei 276 Teilnehmern. Der Stichprobenfehler der durchgeführten Umfrage liegt somit bei 4,7%. Da lediglich der Fachbereich, in dem die Teilnehmer tätig sind, sowie die Karrierestufe (Professur, Habilitation, Doktorand etc.) abgefragt wurden, ist die Umfrage als anonym anzusehen.

Werden die Personen berücksichtigt, die einem Fachbereich zuzuordnen waren (n=258), beträgt die Rücklaufquote der Zielgruppe 27% bzw. der Stichprobenfehler 5,2%. Die Verteilung auf die Fachbereiche und Institute zeigt, dass der Anteil der Teilnehmer

Tabelle 1: Beteiligung nach Fachbereichen.

Fachbereich / Institute	Grundgesamtheit (wiss. Personal)		Stichprobe*		Rücklaufquote Fachbereiche in %	Min. Stichproben-größe für einen Fehler von 5%	Stichproben—fehler in %
	Absolut	*%*	*Absolut*	*%*			
FB 1 Kultur- und Sozialwissenschaften	90	*9%*	31	*12%*	34%	74	**14,3%**
FB 3 Erziehungs- und Kulturwissenschaften	138	*14%*	21	*8%*	15%	102	**19,8%**
FB 4 Physik	54	*6%*	18	*7%*	33%	48	**19,0%**
FB 5 Biologie / Chemie	150	*15%*	29	*11%*	19%	109	**16,4%**
FB 6 Mathematik / Informatik	78	*8%*	26	*10%*	33%	65	**15,8%**
FB 7 Sprach- und Literaturwissenschaften	87	*9%*	18	*7%*	21%	72	**20,7%**
FB 8 Humanwissenschaften	98	*10%*	34	*13%*	35%	79	**13,7%**
FB 9 Wirtschaftswissenschaften	62	*6%*	29	*11%*	47%	54	**13,4%**
FB 10 Rechtswissenschaften	85	*9%*	34	*13%*	40%	70	**13,1%**
IKFN (Institut für Kulturgeschichte der frühen Neuzeit)	7	*1%*	0	*0%*	0%	7	**∞**
IKW (Institut für Kognitionswissenschaften)	36	*4%*	9	*3%*	25%	33	**28,7%**
IMIS (Institut für Migragionsforschung und Interkulturelle Studien)	17	*2%*	2	*1%*	12%	17	**67,1%**
IUSF (Institut für Umweltsystemanalyse)	35	*4%*	7	*3%*	20%	33	**33,6%**
SFB 944 "Physiology and dynamics of cellular microcompartments"	31	*3%*	-	-	-	-	**-**
Summe	**968**	***100%***	**258* (278)**	***100%***	**27%**	**276**	**5,2% (4,7%)**

*Insgesamt haben 278 Personen teilgenommen. 20 Personen haben die Umfrage vorzeitig beendet oder gaben bei der Zugehörigkeit zum Fachbereich keine Antwort ab.

Tabelle 2: Beteiligung nach Fächerclustern.

Fächercluster	Grundgesamtheit (wiss. Personal)		Stichprobe		Rücklaufquote Fachbereiche in %	Min. Stichprobengröße für einen Fehler von 5%	Stichproben—fehler in %
	Absolut	*%*	*Absolut*	*%*			
Kultur- & Geisteswissenschaften (FB 1, 3, 7; IKFN, IMIS)	339	*35%*	72	*28%*	21%	181	**10,3%**
Naturwissenschaften (FB 4, 5, 6; IUSF; SFB 944)	348	*36%*	80	*31%*	23%	183	**9,6%**
Humanwissenschaften (FB 8; IKW)	134	*14%*	43	*17%*	32%	100	**12,4%**
Rechtswissenschaften (FB 10)	85	*9%*	34	*13%*	40%	85	**13,1%**
Wirtschaftswissenschaften (FB 9)	62	*6%*	29	*11%*	47%	54	**13,4%**
Summe	**968**	***100%***	**258* (278)**	***100%***	**27%**	**276**	**5,2% (4,7%)**

*Insgesamt haben 278 Personen teilgenommen. 20 Personen haben die Umfrage vorzeitig beendet oder gaben bei der Zugehörigkeit zum Fachbereich keine Antwort ab.

an der Umfrage die Verteilung auf die Fachbereiche repräsentiert. Lediglich die Institute IKFN, IKW, IMIS und IUSF sind deutlich unterrepräsentiert. Der SFB 944 konnte für die Auswertung der Verteilung der Befragten als eigene Institution nicht berücksichtigt werden, da einige Mitarbeiter auch im FB 5 (Biologie / Chemie) tätig sind. Obwohl der Stichprobenfehler aller an der Umfrage teilgenommenen Personen bei 4,7% liegt, verteilt sich dieser sehr heterogen über die Fachbereiche. So ist ein Stichprobenfehler von <15% lediglich für die Fachbereiche 1, 8, 9 und 10 anzunehmen. Ein Stichprobenfehler zwischen 15% und 25% liegt bei den Fachbereichen 3, 4, 5 und 7 vor. Aufgrund der geringen Rücklaufquote liegt der Stichprobenfehler für die Institute IKW, IMIS sowie IUSF bei über 25%. Da die Fächerverteilung in den erhobenen Fachbereichen sehr heterogen ist, wurde geprüft, ob die Teilnehmer in den Fächerclustern ebenfalls repräsentativ verteilt sind (Tab. 2). Dabei zeigt sich, dass die Kultur- & Geisteswissenschaften, sowie die Naturwissenschaften im Vergleich zu den Human-, Rechts- und Wirtschaftswissenschaften überrepräsentiert sind. Insgesamt liegen die Stichprobenfehler bei den Fächerclustern unterhalb von 15%. Aufgrund der höheren Grundgesamtheit und Rücklaufquote, liegen die Stichprobenfehler in den Kultur- & Geisteswissenschaften sowie Naturwissenschaften bei 10,3% bzw. 9,6%. Der Fehler der Human-, Rechts-, und Wirtschaftswissenschaften liegen zwischen 12,4% und 13,4%.

2.2 Datengrundlage für die Berechnung des zukünftigen Speicherbedarfs & Kosten

Ausgehend von der Frage über den Umfang des bisherigen Speicherbedarfs für Forschungsdaten an der Universität Osnabrück (Abb. 2) werden In einem zweiten Schritt, abhängig vom derzeitigen Datenvolumen, die für die Langzeitarchivierung anfallenden Kosten geschätzt. Dieses Kostenmodell basiert im Wesentlichen auf dem Geschäftsmodell des Archivierungsdienstes RADAR (Research Data Repository) des FIZ Karlsruhe. RADAR ist ein disziplinübergreifendes Forschungsdatenrepositorium für die Archivierung und Publikation von Forschungsdaten abgeschlossener wissenschaftlicher Studien und Projekte (RADAR, 2017). Mit einer antwortabhängigen Datenauswahl wurde die in der Umfrage ermittelte Verteilung der Speichervolumina in Beziehung gesetzt mit den zuvor abgefragten Datenformaten. Die Darstellung erfolgt als relative Verteilung der anteiligen Datenformate an den Datenvolumina (Abb. 3).

Für die Berechnung des zukünftigen Speicherbedarfs werden die aus den Antworten ermittelten Kategoriegrenzen (von – bis) herangezogen, um daraus die Gesamtdatenvolumen zu errechnen (Tab. 3). Mit den Stichprobenfehlern aus Tab. 1 lässt sich durch Fehlerfortpflanzungsrechnung der Gesamtfehler ableiten. Dieser beträgt 26% und wird in den folgenden Abbildungen auf die jeweilige obere und untere Grenze hinzugerechnet (rote Linien).

Die 164 Teilnehmer, die diese Frage beantwortet haben, belegen derzeit ein Gesamtdatenvolumen von 44 ± 10 TB. Das durchschnittliche Speichervolumen pro Wissenschaftler (basierend auf den 164 Teilnehmern, die diese Frage beantwortet haben) beträgt somit 0,27 ± 0,06 TB. Allerdings spiegelt die Teilnehmerzahl nur 16,9% der 968 Wissenschaftler wider.

Mit $\emptyset TB_{AnteilWiss} = \frac{\%Wiss \times \Sigma Wiss \times \emptyset TB_{Wiss}}{100}$ wurde der Speicherbedarf ermittelt, den ein bestimmter Prozent-

satz an Wissenschaftlern benötigen würde, um all ihre derzeitigen Forschungsdaten zu archivieren (Abb. 4 A). Wobei gilt: %Wiss = Anteil der Wissenschaftler; ∑Wiss = 968 Wissenschaftler und $ØTB_{Wiss}$ = Das durchschnittliche Speichervolumen pro Wissenschaftler.

Die jährliche Steigerungsrate wurde in einer Umfrage an der Universität Trier ermittelt und diente als Faktor für die Berechnung (Lemaire et al., 2016, S. 25). Die Befragten sollten dabei ihr Datenvolumen der nachfolgenden drei Jahre nach dem Umfragezeitpunkt abschätzen. Da die Größe der Universität Osnabrück (Anzahl der wissenschaftlichen Mitarbeiter und Anzahl der Studierenden) der Universität Trier ungefähr der Anzahl der Universität Osnabrück entspricht, wurden diese Daten (Tab. 5) genutzt und dazu verwendet, das zu erwartende Speichervolumen der nächsten Jahre der Universität Osnabrück zu ermitteln (Abb. 4 B & C).Die durchschnittliche Wachstumsrate beträgt 5,6 % pro Jahr und entspricht in etwa der Rate, die Strodl & Rauber (2011) annahmen (5 %).

Für eine Kostenabschätzung der Archivierung von Forschungsdaten wurde das Geschäftsmodell des Archivierungsdienstes RADAR (Research Data Repository) des FIZ Karlsruhes herangezogen (Tab. 5) (RADAR, 2017). Dabei wird für das Kostenmodell einerseits die Größe der zu archivierenden Datenmenge (<50 TB und >50 TB) unterschieden. Andererseits wird unterschieden, ob die Forschungsdaten archiviert und veröffentlich werden sollen oder nur archiviert („Dark Archive“).

Im Fall „ohne Publikation“ kann RADAR als reines ‚Dark Archive‘ verwendet werden, bei dem die Forschungsdaten nicht veröffentlicht werden. Die Forschungsdaten können hierbei bis zu 15 Jahre unveröffentlicht archiviert werden. Der Bereitsteller der Daten kann über die Zugriffsrechte entscheiden. Werden über RADAR die Forschungsdaten zusätzlich veröffentlicht, so können diese unbegrenzt aber mindestens 25 Jahre archiviert werden. Den Forschungsdaten wird u.a. eine DataCite-DOI zugewiesen und darüber hinaus wird eine Schnittstelle für Peer-Review-Verfahren angeboten. Die Berechnung des Kostenmodells berücksichtigt allerdings nur den von der DFG geforderten Archivierungszeitraum von 10 Jahren (Abb. 5).

3 Ergebnisse & Diskussion

Die Umfrage ergab, dass unabhängig vom Speichervolumen die drei am häufigsten genannten Datenformate Text-, Bild- und Tabellenformate waren (Abb. 3). Es zeigt sich jedoch deutlich, dass der Anteil an Textformaten in dem Speichervolumenbereich bis zu 1 GB mit 54% deutlich höher ist als im Vergleich zu den Datenvolumenbereichen >1 GB - >1 TB. So sinkt der Anteil an Textdokumenten bei einem Speichervolumen von mehr als 1 TB auf 16% ab. Mit steigendem Speichervolumen wächst der Anteil der Videoformate von 4% (bis zu 1 GB) auf 14% an (mehr als 1 TB). Ein Anstieg ist ebenfalls für gerätespezifische Formate (6% bei bis zu 1 GB zu 12% bei mehr als 1 TB Speichervolumen) und Metadatenformate (0% bei bis zu 1 GB zu 4% bei mehr als 1 TB Speichervolumen) zu beobachten.

Wird der Speicherbedarf der Wissenschaftler betrachtet, so wächst dieser an, sobald ein höherer Anteil der Wissenschaftler der Universität Osnabrück ihre

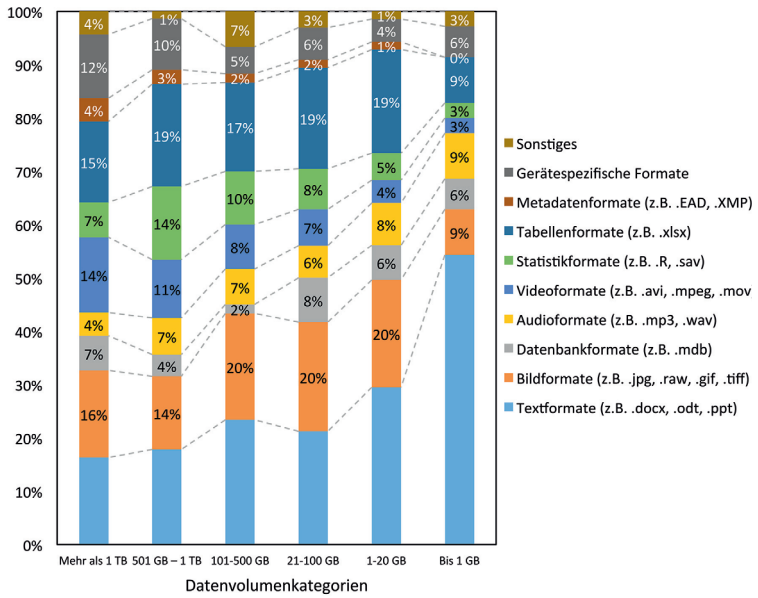


Abbildung 3: Relative Verteilung der Datenformate nach Datenvolumenkategorien.

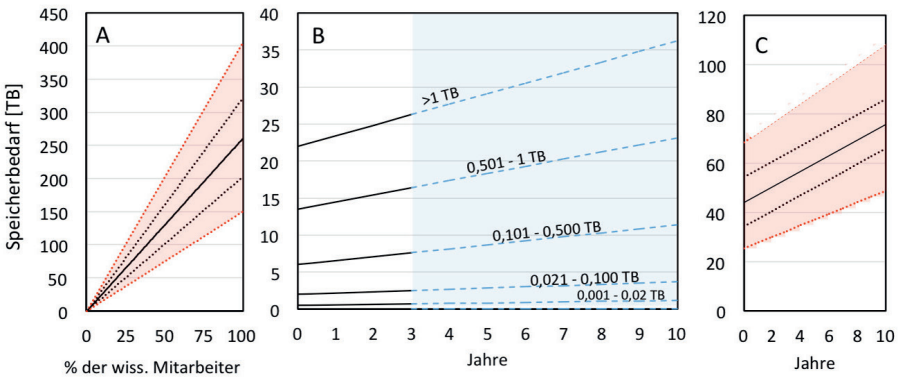


*Abbildung 4: **A:** Derzeitiger Speicherbedarf, wenn ein gewisser Anteil an wiss. Mitarbeiter all ihre Daten archivieren würden. Entwicklung des theoretischen Speicherbedarfs über die nächsten 10 Jahre: **B:** Entwicklung pro Datenvolumenkategorien. **C:** Gesamtentwicklung des Speichervolumens beginnend mit dem derzeitigen Datenvolumen von ca. 44 TB. (– = mittlerer Wert; ··· = untere/obere Grenze; ··· = 26% Stichprobenfehlerbereich; --- = Hochrechnung).*

Daten archivieren würden (Abb. 4 A). Jedoch ist hier zu beachten, dass die Wissenschaftler nur die Forschungsdaten in bzw. über ein Repositorium archivieren und/oder publizieren würden, wenn diese zu einer Veröffentlichung oder gar zu einem Projekt gehören. Daher kann davon ausgegangen werden, dass nur ein geringerer Anteil dieser Daten für eine Archivierung oder Publikation in Frage kommt. Da die herangezogenen Werte lediglich auf der durchgeführten Umfrage basieren, sei der Unsicherheitsbereich der Prognose an dieser Stelle deutlich zu betonen. Herausfordernd bei der Ermittlung zukünftigen Speicherbedarfs von Wissenschaftlern ist die derzeit noch nicht eindeutig abzuwägende Frage nach den anfallenden zukünftigen Forschungsfragen und deren Bewilligung bzw. Realisierungsmaßnahmen oder gar die Wahl der Methodik. Für die Planung der zu erwartenden Datenmengen

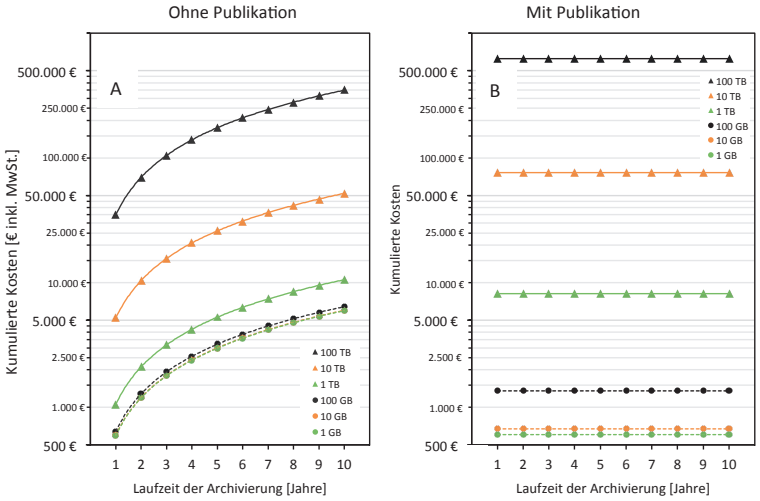


*Abbildung 5: Kumulierte Archivierungskosten bei dem Repositorium RADAR für Speichervolumina von 1GB - 100 TB bei Laufzeiten von 1-10 Jahren mit **(A)** und ohne **(B)** Datenpublikation.*

spielt aber nicht nur die Gesamtspeichermenge eine Rolle, sondern auch, welche Speichervolumenkategorie in welchem Maße anwächst. So zeigt sich generell, dass nicht nur das gesamte Speichervolumen in den nächsten Jahren wachsen wird (von derzeit insgesamt geschätzten 44 TB auf ca. 55 TB in den kommenden 3 Jahren) (Abb. 4 C), sondern auch, dass die Volumenkategorien unterschiedlich ansteigen werden (Abb. 4 B). Hierbei zeigt sich v.a., dass die Größe der Volumenkategorie von der Dauer ihres Wachstums abhängig ist. In der Volumenkategorie >500 GB werden vermutlich die gerätespezifischen Datenformate sowie die Bild- und Videoformate zunehmen, da diese von ihrer Qualität besser werden bzw., je nach Methodik, die erzeugte Datenmenge pro Zeiteinheit größer wird. In der Volumenkategorie <500 GB werden die Text- und Tabellenformate einen großen Anteil der Daten ausmachen. Es muss allerdings darauf hingewiesen werden, dass eine sichere Aussage ab dem Jahr 3 nicht mehr möglich ist (roter Bereich), da die Werte zunächst auf den Zuwachsraten von Lemaire et al. (2016) basieren. Dort wurde in der Fragestellung nach dem Datenvolumen der kommenden drei Jahren gefragt. Die Entwicklung des Speichervolumens setzt zudem ein lineares Wachstum voraus, mit dem auch das Kostenmodell von Strodl & Rauber (2011) kalkuliert hat. Dies erscheint unrealistisch, da von einem exponentiellen Wachstum ausgegangen werden muss (Ackerman et al., 2011; Zaslavsky et al., 2012).Werden Forschungsdaten weder über ein institutionelles Repositorium der Universität, noch über ein fachliches Repositorium der in einer Fachdisziplin erzeugten Forschungsdaten archiviert, sondern hierfür das Repositorium RADAR genutzt, so fallen je nach Speichervolumen verschiedene Kosten an (Abb. 5 A). Vor allem in den Speichervolumenbereichen von 1-100 GB liegen die Kosten allesamt in einem ähnlichen Bereich und beginnen für eine Laufzeit von einem Jahr bei ca. 600 € (inkl. MwSt.), wobei hier der Anteil der jährlichen Grundgebühr den größten Anteil ausmacht. Nach zehn Jahren Laufzeit liegen die kumulierten Kosten bei 6.000 – 6.400 €. Ab einem Speichervolumen von 10 TB erhöhen sich die Kosten um den Faktor 10 und für 100 TB um den Faktor 100. So fallen bei der reinen Archivierung (ohne Datenpublikation) eines Speichervolumens von 100 TB nach zehn Jahren kumulierte Archivierungskosten von 350.000 € an. Ab diesem Speichervolumen ist die höhere jährliche Grundgebühr von 11.900 € zu berücksichtigen. Werden die Daten zusätzlich zur reinen Archivierung auch veröffentlicht (Abb. 5 B), sind einmalige Kosten zu zahlen. Die Kosten lägen für ein Datenvolumen von z. B. 100 GB bei rund 1.400 €. Das Archivieren und Publizieren eines Datenpakets mit einem Volumen von 1 TB würde rund 8.000 € kosten. Ausgehend von der verwendeten Methode und Datengrundlage kann davon ausgegangen werden, dass die Nutzung des Archivierungsdienstes RADAR sich für ein Datenvolumen von bis zu 10 TB (ohne Publikation) und von bis zu 1 TB (mit Publikation) in einem angemessenen Kostenrahmen für eine Universität oder einem Projekt bleibt. Eine Langzeitarchivierung von Datenvolumina von mehr als 10 TB, ob mit oder ohne Publikation, bedeutet, dass z.B. in einem drittmittelfinanzierten Projekt (parallel zu Personal- oder Sachkosten), zu einem hohen Kostenfaktor wird. Ab welchem Speichervolumen lohnt sich nun die Langzeitarchivierung von Forschungsdaten bei dem Archivierungsdienst RADAR und ab wann sollten die universitären Rechenzentren diese Aufgabe übernehmen? Hierzu muss eindeutig die Motivation

dahinter definiert werden, ob es sich um eine reine Langzeitarchivierung handelt oder ob ausgewählte Forschungsdaten zu einer Veröffentlichung publiziert werden sollen (sollte kein geeignetes fachliches Repositorium existieren).

Um die These zu diskutieren seien in Tabelle 6 die Speicher- bzw. Storagekosten von zwei Universitäts-Rechenzentren dargestellt. Die jährlichen Kosten und die Kosten, die nach zehn Jahren Langzeitarchivierung entstehen, wurden aus den angegebenen Preisen umgerechnet. So können diese besser mit dem Archivierungsdienst RADAR verglichen werden. Hierbei ist jedoch zu beachten, dass die Rechenzentren nicht immer persistente Identifikatoren, wie z. B. DOI oder URN auf die archivierten Daten vergeben und die Daten nicht als „Publikation" verstanden werden können. Auch wurden die Kostenentwicklungen für Speicher (auch induziert durch die technische Weiterentwicklung) sowie die Kosten für einen etwaigen Ausbau der Rechenzentren nicht berücksichtig. Kalkuliert wurde mit den auf den Webseiten ermittelten Preisen. Forschungsdaten, die einer Publikation oder einem Projekt zugehörig veröffentlicht werden sollen, unterliegen anderen Voraussetzungen als Rohdaten, die lediglich der Langzeitarchivierung dienen (persistente Identifikatoren, Embargo-Fristen, sensible Daten usw.). Teilweise werden die Rohdaten für eine Publikation auch soweit vereinfacht, dass ihr Speichervolumen geringer wird. Somit käme, aufgrund der angebotenen Services (unbegrenzte Speicherzeit, optionale Sperrfristen, Zuweisung eines DataCite-DOI, Schnittstelle für Peer-Review, Lizenzierung, Indexierung der Metadaten), RADAR wieder in Frage, um ebendiese Forschungsdaten evtl. bis zu einem Volumen von 1 TB dort zu veröffentlichen. Auch wenn die Kosten hierbei über denen der Rechenzentren liegen. An dieser Stelle sei die Frage zu stellen, ob aus Kostengründen nicht jedes Rechenzentrum ähnliche Services wie RADAR in ihr Portfolio aufnehmen sollte, um so zugleich dem Aufbau der Nationalen Forschungsdateninfrastruktur (NFDI) (RfII, 2017) entgegenzukommen.

Schlussfolgerung

Forschungsdaten geraten mehr und mehr in den Fokus der wissenschaftspolitischen Ebene. Neben den Forschern selbst, hat auch die wissenschaftlich interessierte Öffentlichkeit hat ein Interesse am Zugang zu Forschungsdaten. Die Zugänglichmachung und Archivierung von digitalen Forschungsdaten in dafür geschaffene Repositorien wird zunehmend genutzt. Dem stehen Speicherkosten gegenüber, die bei der Archivierung entstehen. So muss trennscharf definiert werden, welche Forschungsdaten zu publizieren und welche zu archivieren sind. Je nach Bedarf stehen den Forschenden hierzu externe Dienstleistungen, wie das betrachtete Repositorium RADAR, zur Verfügung. Aufgrund der anfallenden Kosten ist zu empfehlen, dort lediglich die für Publikationen aufbereiteten Daten zu veröffentlichen. Für die Archivierung von Roh- oder Long-Tail-Data sollte zunächst geprüft werden, ob die lokalen Rechenzentren hierfür genutzt werden können oder ob es günstigere Alternativen gibt. Allerdings lässt sich der Wert von Forschungsdaten nicht ausschließlich in einer monetären Zahl darstellen.

Tabelle 3: Speichervolumenbereiche, die als Grundlage für die Berechnung des benötigten Speichervolumens herangezogen werden.

Antwortbereiche	Nennungen	Anteil	Volumenbereiche [TB]			Gesamtdatenvolumen [TB]		
			Untere Grenze	Mittlerer Wert	Obere Grenze	Untere Grenze	Mittlerer Wert	Obere Grenze
Mehr als 1 TB	22	13%	1	1	1	22	22	22
501 GB – 1 TB	18	11%	0,501	0,75	0,999	9,02	13,5	17,98
101-500 GB	20	12%	0,101	0,3	0,5	2,02	6	10
21-100 GB	33	20%	0,021	0,06	0,1	0,64	1,98	3,3
1-20 GB	49	30%	0,01	0,01	0,02	0,49	0,49	0,98
Bis 1 GB	22	13%	0,001	0,001	0,001	0,02	0,02	0,02
Teilnehmer	164					34,24	43,99	54,28
			Ø Speichervolumen pro Wissenschaftler:			**0,21**	**0,27**	**0,33**

Tabelle 4: Geschätztes zu archivierendes Forschungsdatenvolumen der nächsten 3 Jahre (Übernommen aus: Lemaire et al., 2016, S. 25, Tab. 5)

Geschätztes zu archivierendes Forschungsdatenvolumen der nächsten 3 Jahre	
> 1 TB	19,4%
101 GB – 1 TB	16,1%
21 – 100 GB	15,5%
1 – 20 GB	21,3%
< 1 GB	9,0%
Kann ich nicht einschätzen	17,4%
Keine Angabe	1,3%

Tabelle 5: Jährliche Archivierungskosten (inkl. MwSt.) von RADAR (RADAR, 2017; Stand der Preise: 22.12.2017).

Speichervolumen	Archivierungskosten		Jährliche Grundgebühr
	ohne Publikation	**mit** Publikation (einmalig)	
< 50 TB	€ 0,46 pro GB / Jahr	€ 7,56 pro GB	€ 595,00
> 50 TB	€ 0,23 pro GB / Jahr	€ 6,12 pro GB	€ 11.900,00

Tabelle 6: Preise für die Langzeitarchivierung von digitalen Daten an zwei beispielhaften Rechenzentren von Universitäten.

Institution	Preis pro GB und Monat (ohne MwSt.)	Speicher-volumen	Preis Speichervolumen pro Monat	Preis Speichervolumen pro Jahr	Kosten nach 10 Jahren
TU Berlin	--	1 TB	29,17 €	350,00 €[1]	3.500,00 €[2]
		10 TB	291,67 €	3.500,00 €	35.000,00 €
		100 TB	2.916,70 €	35.000,00 €	350.000,00 €
RZ Ruhr Universität Bochum	0,05 €[3]	1 TB	5,00 €	60,00 €	600,00 €
		10 TB	50,00 €	600,00 €	6.000,00 €
		100 TB	500,00 €	6.000,00 €	60.000,00 €

[1] „1 TB mit Datensicherung jährlich € 350,00, Mindestlaufzeit 3 Jahre"

[2] „1 TB mit Datensicherung für DepositOnce € 3500,00, Laufzeit 10 Jahre." https://www.campusmanagement.tu-berlin.de/menue/dienste/daten_server/speicherplatzsan [05. Februar 2019].

[3] „Option 1: Netzwerkspeicher inkl. Snapshots und Notfallbackup zum UARuhr-Backupdienst" https://it-services.ruhr-uni-bochum.de/sites/default/files/entgelttabelle_zum_servicekatalog.pdf [05. Februar 2019].

Literaturverzeichnis

Ackerman, Elise/ Guizzo, Erico: 5 Technologies That Will Shape the Web - Innovations that will make the web smarter and sleeker and irresistibly more social, too, in: IEEE Spectrum's special report on the battle for the future of the social Web, 48(6) (2011) S. 40-45.

Beucke, Daniel: Geschäftsmodelle für die digitale Langzeitarchivierung : Das Beispiel Forschungsdaten, in: Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft (278) (2010). doi: 10.18452/2025.

BMBF: Digitale Innovationen – Neue Dimensionen von Bildung und Wissenschaft erschließen (2017) URL: https://www.bmbf.de/pub/Digitale_Innovationen.pdf [14. Juni 2018].

BMBF: Open Access in Deutschland – Die Strategie des Bundesministeriums für Bildung und Forschung (2016) URL: https://www.bmbf.de/pub/Open_Access_in_Deutschland.pdf [14. Juni 2018].

Bundesregierung: Digitale Agenda 2014 – 2017. Bundesministerium für Wirtschaft und Energie, Bundesministerium des Innern, Bundesministerium für Verkehr und digitale Infrastruktur (2014) URL: https://www.digitale-agenda.de/Content/DE/_Anlagen/2014/08/2014-08-20-digitale-agenda.pdf?__blob=publicationFile&v=6 [22. Mai 2018].

de.statistia.com: Prognose zum Volumen der jährlich generierten digitalen Datenmenge weltweit in den Jahren 2016 und 2025 (in Zettabyte) (2017) URL: https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen [03. Mai 2018].

DFG: Leitlinien zum Umgang mit Forschungsdaten (2015) URL: http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf [17. Juni 2018].

European-Commission: COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS : European Cloud Initiative – Building a competitive data and knowledge economy in Europe (2016) URL: http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52016DC0178 [17. Juni 2018].

European-Commission: European Open Science Cloud – New Research & Innovation Opportunities (2017) URL: http://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none [17. Juni 2018].

G8: G8 Science Ministers Statment (2013) URL: https://www.gov.uk/government/news/g8-science-ministers-statement [18. Juni 2018].

Gronwald, M.: Rohdaten zur Umfrage zum Umgang mit Forschungsdaten an der Universität Osnabrück (Dataset) (2018) URL: http://nbn-resolving.org/urn:nbn:de:gbv:700-2018050417137 [11. Mai 2018]

Borghoff, Uwe M./ Brantl, Markus/ Hagel, Harald/ Minkus, Michael/ Rödig, Peter/ Beinert, Tobias/ Lang, Susanne/ Schoger, Astrid: Entwicklung von Organisations- und Geschäftsmodellen zur Langzeitarchivierung digitaler Objekte aus DFG-geförderten Digitalisierungsprojekten : Studie im Auftrag der Deutschen Forschungsgemeinschaft – April 2009 (2009) URL: https://www.digitale-sammlungen.de/content/dokumente/2009_04_Studie_Organisations_und-Geschaeftsmodelle.pdf [10. August 2018].

HRK: Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können. Orientierungspfade, Handlungsoptionen, Szenarien. Empfehlung der 19. Mitgliederversammlung der HRK am 10. November 2015 in Kiel (2015) URL: https://www.hrk.de/uploads/tx_szconvention/Empfehlung_Forschungsdatenmanagement__final_Stand_11.11.2015.pdf [14. Juni 2018].

HRK: Management von Forschungsdaten – eine zentrale strategische Herausforderung für Hochschulleitungen. In: 16. Mitgliederversammlung der Hochschulrektorenkonferenz am 13. Mai 2014, Frankfurt am Main 2014.

KII: Gesamtkonzept für die Informationsinfrastruktur in Deutschland – Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder (2011) URL: https://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Infrastruktur/KII_Gesamtkonzept.pdf [15. Juni 2018].

Kindling, Maxi: Qualitätssicherung im Umgang mit digitalen Forschungsdaten / Quality assurance of digital research data / La garantie de la qualité des données numériques de recherche, in: Information - Wissenschaft & Praxis, 64(2-3) (2013) 137. doi: 10.1515/iwp-2013-0020.

Kindling, Maxi/ Schirmbacher, Peter: „Die digitale Forschungswelt" als Gegenstand der Forschung / Research on Digital Research / Recherche dans la domaine de la recherche numérique, in: Information - Wissenschaft & Praxis, 64(2-3) (2013) 127. doi: 10.1515/iwp-2013-0017.

Lemaire, Marina/ Rommelfanger, Yvonne/ Ludwig, Jan/ Lürken-Uhl, Alexander/ Merkler, Benjamin/ Sturm, Peter: Umgang mit Forschungsdaten und deren Archivierung. Bericht zur Online-Bedarfserhebung an der Universität Trier. Trier: Universitätsbibliothek Trier 2016.

Lindstädt, Birte: Forschungsdatenmanagement als überregionale Aufgabe der Informationsversorgung – was kann eine Zentrale Fachbibliothek wie ZB MED Leibniz-Informationszentrum Lebenswissenschaften leisten?, in: 2015, 2(4) (2015) 5. doi: 10.5282/o-bib/2015H4S73-77.

Lindstädt, Birte: Management und Publikation von Forschungsdaten – Serviceleistungen einer wissenschaftlichen Bibliothek, in: Bibliotheksdienst, 50(7) (2016). doi: 10.1515/bd-2016-0078.

Neuroth, Heike: Langzeitarchivierung von Forschungsdaten : eine Bestandsaufnahme / Nestor. Heike Neuroth (Hrsg.). Göttingen: Göttingen Univ.-Verl. 2012.

OECD: Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communiqué, Communiqué. Organisation for Economic Co-operation and Development. Paris, France 2004.

Piwowar, Heather A./ Vision, Todd J.: Data reuse and the open data citation advantage, in: PeerJ, 1 (2013) e175. doi: 10.7717/peerj.175

RADAR. (2017): Preise für Forschungseinrichtungen und Hochschulen. Retrieved from https://www.radar-service.eu/de/preise [10. Januar 2019].

RfII: Ein Diskussionsimpuls zu Zielstellung und Voraussetzungen für den Einstieg in die Nationale Forschungsdateninfrastruktur (NFDI) (2017) URL: http://www.rfii.de/download/rfii-diskussionspapier-2017 [18. Dezember 2018].

RfII: Leistung aus Vielfalt – Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland (2016) URL: http://www.rfii.de/?wpdmdl=1998 [18. Dezember 2018].

Strodl, Stephan/ Rauber, Andreas: A cost model for small scale automated digital preservation archives. iPRES2011, November 1-4. National Library Board Singapore & Nanyang Technological University. URL: https://fedora.phaidra.univie.ac.at/fedora/get/o:294219/bdef:Content/get [28. Februar 2019].

theatlas.com: The dramatic rise of data creation and replication (2015) URL: https://www.theatlas.com/charts/E1Wxox0c [25. April 2018].

whatsthebigdata.com: Amount of Data Created Annually to Reach 180 Zettabytes in 2025 (2016) URL: https://whatsthebigdata.com/2016/03/07/amount-of-data-created-annually-to-reach-180-zettabytes-in-2025 [25. April 2018].

Wissenschaftsrat: Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. (2359-12). Berlin 2012. URL: https://www.wissenschaftsrat.de/download/archiv/2359-12.pdf [18. Mai 2018].

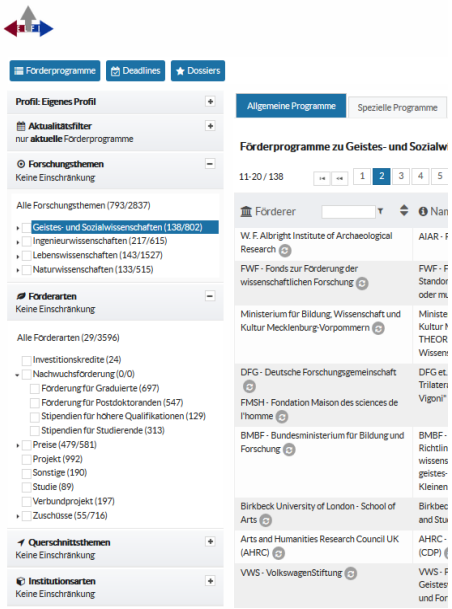
Wissenschaftsrat: Übergreifende Empfehlungen zu Informationsinfrastrukturen. (10466-11). Berlin 2011. URL: https://www.wissenschaftsrat.de/download/archiv/10466-11.pdf [18. Mai 2018].

Wollschläger, Thomas: „Kapitel 5: Geschäftsmodelle“, in: nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. (Vol. 1.5) 2008.

Zaslavsky, Arkady/ Perera, Charith/ Georgakopoulos, Dimitrios: Sensing as a Service and Big Data. Paper presented at the International Conference on Advances in Cloud Computing (ACC), Bangalore, India 2012. URL: https://arxiv.org/abs/1301.0159 [13. Juli 2018].

Korrespondierender Autor
Dr. Marco Gronwald
Referent eScience und Fachreferent für Geographie und Geowissenschaften
Universitätsbibliothek
Alte Münze 16
49074 Osnabrück
marco.gronwald@uni-osnabrueck.de
ORCID: 0000-0002-5081-4390

www.elfi.info



Preise, Stipendi
eisekosten, Tagu
Doktorandenförd
Postdoktoranden
pendium, Druckko
Anschubfinanzie
Verbundprojekt,
Tagungsorganisa
Studie, Reisekos
ommerschule, **Exi**
Essay-Wettbewer
Dissertationspr
Studienförderun
ation, Auslandsf
osten, bilateral

In unserer Datenbank finden Sie die Finanzierung für Ihr Forschungsprojekt!

ELFI Gesellschaft für Forschungsdienstleistungen mbH
Postfach 25 02 07
D-44740 Bochum
Tel. +49 (0)234 / 32-22940