Forschungsdaten in der digitalen Bibliothek

Author(s):
Töwe, Matthias

Publication Date:
2014-08

Permanent Link:

https://doi.org/10.3929/ethz-a-010337865 →

Forschungsdaten in der digitalen Bibliothek

Matthias Töwe*

Zusammenfassung
Die digitale Bibliothek schreibt in Hochschule und Forschung in vielen ihrer Funktionen die Rolle der physischen Bibliotheken fort. Der folgende Beitrag führt aus, welche darüber hinaus gehenden neuen Aufgaben die Hochschulbibliotheken im Zusammenhang mit der Organisation, Publikation und dauerhaften Erhaltung von Forschungsdaten beschäftigen und stellt Angebote der ETH Zürich vor, die derzeit aufgebaut werden. Von Interesse sind dabei sowohl die befristete Aufbewahrung gemäss den Vorgaben der guten wissenschaftlichen Praxis als auch die dauerhafte Publikation und Langzeitverfügbarkeit von Daten für die Nachnutzung.

Abstract
In many of its functions, the digital library perpetuates the role of the physical library in university and research. The following article describes new tasks for university libraries beyond this in the context of the organization, publication and long-term preservation of research data. It also presents some evolving services at ETH Zurich. Both the safeguarding of data for limited periods of time according to good scientific practice as well as the permanent publication and long-term availability of data for re-use are discussed.

Forschungsrelevante Informationen und Ergebnisse wissenschaftlichen Arbeitens zu sammeln, zugänglich zu machen und zu bewahren – dies sind seit jeher Aufgaben wissenschaftlicher Bibliotheken. Im Mittelpunkt auch der digitalen Bibliothek steht die wissenschaftliche Informationsversorgung, die geprägt ist durch formale Publikationen wie Zeitschriftenartikel oder Monographien. Diese werden ergänzt durch Dokumente und Quellen, die als Teil von Vor- und Nachlässen in Bibliotheken gelangen und ursprünglich nicht publiziert wurden. Voraussetzungen für den einfachen Zugang zu diesen Inhalten und für ihre wissenschaftliche Nutzung sind eine aussagekräftige Erschliessung und die Digitalisierung in hoher Qualität. Eine Plattform für diesen Zweck ist etwa e-manuscripta.ch[1].

Die Entwicklung hin zur digitalen Bibliothek widerspiegelt die in den vergangenen Jahrzehnten gewachsene digitale Durchdringung des gesamten Forschungsprozesses. Diese wird einerseits getrieben durch die Bedürfnisse, die sich aus wissenschaftlichen Fragestellungen ergeben, andererseits ermöglichen qualitative und quantitative Fortschritte bei der Verarbeitung digitaler Daten überhaupt erst bestimmte wissenschaftliche Aktivitäten. In welchem Mass digitale Methoden zum Einsatz kommen und in welchem Umfang digitale Daten produziert werden, hängt sehr stark vom jeweiligen Fach ab. Bestimmte Wissenschaftszweige sind ohne digitale Daten und die Methoden zu ihrer Gewinnung und Analyse nicht denkbar, wie etwa die experimentelle Hochenergiephysik, die allerdings auch in dieser Hinsicht als Extremfall anzusehen ist. Daneben gibt es weitere Wissenschaftsdisziplinen, die heute durch die Produktion bzw. Sammlung von «Big Data» und/oder durch deren Nutzung geprägt sind. Dabei sind «Big Data» nicht anhand der reinen Datenmenge definierbar. Wichtiger ist die Perspektive, aus einer mehr oder weniger unstrukturierten Datensammlung mit geeigneten statistischen Methoden und Algorithmen Informationen zu gewinnen, die andernfalls nicht erreichbar sind. Diese Methoden sind Gegenstand intensiver Forschung und um sie sinnvoll anwenden zu können, müssen die Daten in geeigneter Form verfügbar sein. Bereits aus dieser knappen Beschreibung kann abgeleitet werden, dass der Umgang mit «Big Data» hohe Anforderungen an die Rechenzentrumsinfrastruktur stellt und selbst noch Gegenstand der Forschung ist. «Big Data» stehen daher in der Regel nicht im Mittelpunkt der Aktivitäten von Bibliotheken, einige Dienstleistungen können jedoch auch für die Verwaltung dieser Daten attraktiv sein.

* ETH-Bibliothek, ETH Zürich, Rämistrasse 101, 8092 Zürich.

E-Mail: *matthias.toewe@library.ethz.ch*
http://www.library.ethz.ch/Digitaler-Datenerhalt

Matthias Töwe, Dr. phil. nat., ist in Hamburg geboren und aufgewachsen. Nach dem Studium der Chemie in Hamburg und dem Doktorat an der Universität Basel absolvierte er an der Universitätsbibliothek Basel die Ausbildung zum Wissenschaftlichen Bibliothekar. Seit 2003 arbeitet Matthias Töwe an der ETH-Bibliothek, zunächst für das Konsortium der Schweizer Hochschulbibliotheken, unter anderem als Leiter des Moduls E-Archiving. Von 2008 bis 2010 koordinierte er das landesweite Projekt Elektronische Bibliothek Schweiz: e lib.ch. Seit Ende 2010 leitet er die Fachstelle Digitaler Datenerhalt der ETH-Bibliothek.

1 *http://www.e-manuscripta.ch/*, Zugriff am 10. Juni 2014

Nahezu flächendeckend ist dagegen die Entwicklung, dass digitale Verfahren die zur Verfügung stehenden Möglichkeiten in quantitativer und/oder qualitativer Hinsicht erweitern. Ohne die Möglichkeit einer scharfen Abgrenzung kann man hier von «Small Data» sprechen, die in jedem Forschungsvorhaben anfallen, und zwar auch dort, wo der eigentliche Forschungsgegenstand im Bereich von «Big Data» liegt. «Small Data» sind dabei unter anderem durch die Zusammensetzung aus eigenständigen, definierten Einheiten gekennzeichnet. In vielen Fällen handelt es sich um abgeschlossene Dokumente oder andere Dateien, die für den Menschen lesbar sind, aber auch abgeschlossene und für die maschinelle Auswertung bestimmte Dateien gehören dazu. Aufgrund ihres Charakters und ihres Umfangs stehen solche Objekte im Mittelpunkt der Dienstleistungen von Hochschulbibliotheken für ihre Kunden.

Die zunehmende Nutzung digitaler Methoden ist selbstverständlich kein Thema nur der Natur- oder Ingenieurwissenschaften, sondern betrifft in steigendem Masse auch die Geisteswissenschaften. Man kann es als Zeichen einer bewussten Auseinandersetzung mit dem stattfinden Wandel interpretieren, dass mit den «Digital Humanities» ein eigener Begriff geschaffen wurde, der nach und nach und auf sehr unterschiedliche Arten mit Leben gefüllt wird.

Der Suche nach einer scharfen Definition von Forschungsdaten sind durch die Heterogenität der Methoden und Anforderungen der einzelnen Fächer Grenzen gesetzt. Die OECD verwendet die folgende Definition: «»[…] factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.[2]»

Der Fokus der OECD-Guidelines liegt auf dem verbesserten Zugang zu Forschungsdaten. Daher klammert die Definition die folgenden Inhalte aus: «This term does not cover the following: laboratory notebooks, preliminary analyses, and drafts of scientific papers, plans for future research, peer reviews, or personal communications with colleagues or physical objects (e.g. laboratory samples, strains of bacteria and test animals such as mice). Access to all of these products or outcomes of research is governed by different considerations than those dealt with here.[3]»

Die Deutsche Forschungsgemeinschaft bezog sich 2009 in ihren Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten auf folgende Definition: «Forschungsprimärdaten sind Daten, die im Verlauf von Quellenforschungen, Experimenten, Messungen, Erhebungen oder Umfragen entstanden sind. Sie stellen die Grundlagen für die wissenschaftlichen Publikationen dar.[4]»

Die weitere Differenzierung wird den Fachgemeinschaften überlassen: «In Abhängigkeit von der jeweiligen Fachzugehörigkeit sind die Forschungsprimärdaten unterschiedlich zu definieren. Die Wissenschaftler sollen in ihren Fachcommunities selber entscheiden, ob bereits Rohdaten hierzu zählen oder ab welchem Grad der Aggregierung die Daten langfristig aufzubewahren sind. Des weiteren soll die Granularität in groben Umrissen vereinbart sein: wie viele Daten ergeben einen Datensatz, der mit einer stabilen Adresse (persistent identifier) ausgestattet wird?[5]»

Das EU-Forschungsrahmenprogramm Horizon 2020 unterscheidet in seinen Guidelines on Open Access to Scientific Publications and Research Data[6] für seine Open Research Pilotphase zwei Arten von Forschungsdaten: «1) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible[7]», mit der Erläuterung «Associated metadata refers to the metadata describing the research data deposited[8].» «2) other data, including associated metadata, as specified and within the deadlines laid down in the data management plan[9]», mit der Erläuterung «[f]or instance curated data not directly attributable to a publication, or raw data[10].»

Bereits die ersten Erkenntnisse aus unserem eigenen Projekt bestätigten, dass eine wirklich aussagekräftige Vorab-Definition von «Forschungsdaten» nicht möglich sein würde. Konsequenterweise wird die Entscheidung darüber, was ihre Forschungsdaten sind, weiterhin den Forschenden überlassen, wobei

2 OECD PRINCIPLES AND GUIDELINES FOR ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING, OECD 2007, *http://www.oecd.org/science/sci-tech/38500813.pdf* , Zugriff am 09. Juni 2014

3 Ebd.

4 Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, Deutsche Forschungsgemeinschaft 2009, *http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf*, Zugriff am 09. Juni 2014

5 Ebd.

6 Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, European Commission, Version 1.0, 11. Dezember 2013, *http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf*, Zugriff am 10. Juni 2014

7 Ebd.

8 Ebd., Fussnote 18

9 Ebd.

10 Ebd., Fussnote 19

es sich bisher noch um individuelle Entscheidungen der einzelnen Verantwortlichen handelt. Eine breite Diskussion innerhalb der Fachcommunities ist aus unserer Sicht bisher die Ausnahme und es stellt sich auch die Frage, wie eng eine Fachcommunity gefasst sein müsste, um zu einem akzeptierten und in der Praxis anwendbaren Konsens zu gelangen.

Mit dem Fortschreiten der Nutzung und letztlich auch mit der wachsenden Abhängigkeit von digitalen Daten und Methoden hat sich das Bewusstsein verbreitet, dass die klassische und in ihrer Form eingeschränkte wissenschaftliche Publikation mehr und mehr nur noch die Spitze des Eisbergs des dahinterstehenden Erkenntnisprozesses adäquat abbilden kann. Gleichzeitig bleiben riesige Mengen an Rohdaten oder in unterschiedlicher Art und Weise bearbeitete Daten meist für Dritte unsichtbar, aus denen die publizierten wissenschaftlichen Erkenntnisse gezogen wurden.

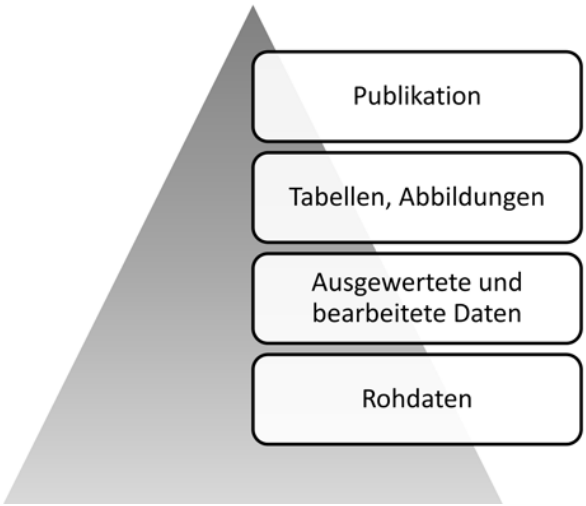

Abb.1. *Schematische Darstellung der Informationsverdichtung im Forschungsprozess.*

Dies wird von Forschenden in ihrer Rolle als Autorinnen und Autoren ebenso wie in ihrer Eigenschaft als Rezipientinnen und Rezipienten wissenschaftlicher Publikationen zunehmend als Einschränkung empfunden. Dass die Publikation die verdichtete Form zur Dokumentation einer unter Umständen jahrelangen Arbeit ist, hat sich als zentraler Faktor der wissenschaftlichen Kommunikation über Jahrhunderte grundsätzlich bewährt – was Kritik an den herrschenden Geschäftsmodellen im Übrigen nicht ausschliesst. In der Vergangenheit war es schlicht nicht möglich, grössere Datenmengen zusammen mit Publikationen oder auch auf anderem Wege zu verbreiten und die konzentrierte Verbreitung von Wissen und Informationen war notwendig. Im Zuge der digitalen Transformation und der umfassenden informationstechnologischen Durchdringung wissenschaftlicher Prozesse hat sich dies fundamental geändert und zunehmend besteht die Erwartung, dass begleitend zur weiterhin erforderlichen intellektuellen Ausarbeitung auch die Daten, die dem Inhalt einer wissenschaftlichen Publikation zugrunde liegen, mit ähnlicher Leichtigkeit erreichbar sein sollten wie die Publikation selbst. Denn durch die technischen Möglichkeit digitale Datenmengen zu sammeln und nach Bedarf auszutauschen, bietet sich die Chance, publizierte Resultate transparenter zu dokumentieren und nachvollziehbarer zu machen. Das wiederum begünstigt wissenschaftliche Forschungs- und Innovationsprozesse. Die entsprechenden Zusammenhänge wurden inzwischen auf verschiedenen Ebenen untersucht. Meilensteine waren etwa der Bericht «Riding the wave – How Europe can gain from the rising tide of scientific data[11]» zuhanden der Europäischen Kommission sowie der «Report on Integration of Data and Publications[12]» des Projekts «Opportunities for Data Exchange - ODE[13]».

Bei genauerer Betrachtung sind mehrere unterschiedliche Aufgaben beim Umgang mit Forschungsdaten zu lösen. Im Sinne der guten wissenschaftlichen Praxis ist es erforderlich, Roh- und bearbeitete Forschungsdaten für eine bestimmte, im jeweiligen Fach übliche Frist aufzubewahren. In manchen Fällen gibt es weitergehende gesetzliche Regelungen und in zunehmendem Masse machen die Institutionen der Forschungsförderung Vorgaben in diesem Sinne. Wo keine weitergehende Regelung besteht, scheint sich eine Mindestfrist von zehn Jahren zu etablieren. Die betreffenden Daten müssen nicht zwingend weltweit frei zugänglich gemacht werden. Es kann allerdings sinnvoll sein und wird teilweise auch von den Herausgebern von Zeitschriften verlangt, zumindest Teile davon bereits zusammen mit dem Manuskript einer Veröffentlichung für die Gutachter zugänglich zu machen. Die befristete Aufbewahrung von Forschungsdaten in diesem Sinne unterstützt und ermöglicht also einen Teil der Qualitätssicherung im Forschungs- und Publikationsprozess.

Transparenz und Nachvollziehbarkeit sind für Forschende selbstverständlich keine neuen Forderungen. Sie sind Grundpfeiler der wissenschaftlichen Arbeitsweise, die per definitionem überprüfbar sein muss. Verändert haben sich vor allem die Menge und

11 Riding the wave – How Europe can gain from the rising tide of scientific data, Final Report of the High level Expert Group on Scientific Data, European Union 2010, *http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf*, Zugriff am 10. Juni 2014

12 *http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf*, Zugriff am 10. Juni 2014

13 *http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/*, Zugriff am 10. Juni 2014

die Bedeutung digitaler Daten. Die Rohdaten müssen zumindest für eine gewisse Zeit greifbar bleiben – so sehen es z. B. seit Jahren auch die Richtlinien der ETH Zürich für die gute wissenschaftliche Praxis[14] vor.

Es liegt jedoch nahe, über die reine Aufbewahrung für den Fall einer Anfechtung hinaus einen Schritt weiter in Richtung einer wissenschaftlichen Nachnutzung zu gehen: Wenn einmal gemessene oder erhobene Daten publiziert, dokumentiert und langfristig verfügbar gehalten werden, besteht grundsätzlich die Möglichkeit, sie zur Beantwortung neuer Fragestellungen heranzuziehen oder sie mit neuen, vielleicht effektiveren Methoden erneut auszuwerten.

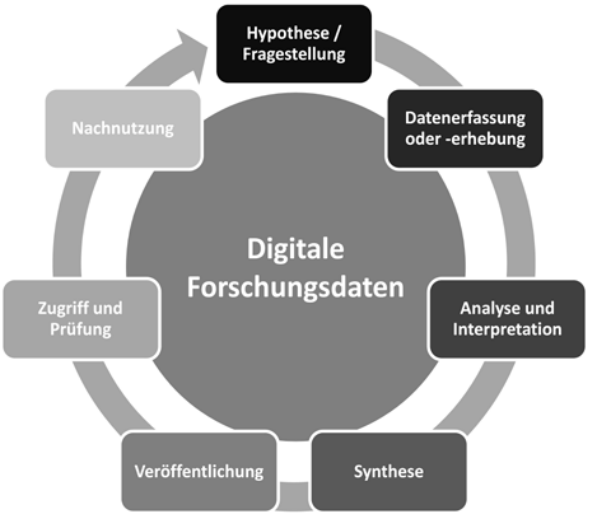

Abb. 2. *Vereinfachte schematische Darstellung des Lebenszyklus digitaler Forschungsdaten.*

Es gilt demnach, frühzeitig innerhalb eines Forschungsvorhabens die Voraussetzungen für die Nachnutzung der dort gewonnenen Forschungsdaten durch Dritte zu schaffen. Als Endergebnis wird zunehmend die eigenständige Publikation von Daten angestrebt, die unabhängig von einem bestimmten Artikel erfolgt. Voraussetzungen hierfür sind die dauerhafte Zitierbarkeit und eine umfassende Dokumentation der Daten und ihres wissenschaftlichen Kontexts. Dieses Vorgehen ist bisher nur in wenigen wissenschaftlichen Disziplinen üblich, während der freizügige Austausch von Forschungsdaten in anderen Fächern aus unterschiedlichen Gründen keine Tradition hat.

Für die Nachnutzung kommen vor allem Daten in Frage, die unter bestimmten Gesichtspunkten und mit einer Reihe von Methoden ausgewertet wurden, die aber das Potential haben, zu einem späteren Zeitpunkt, mit anderen Methoden oder auch in einem anderen fachlichen Kontext neu analysiert zu werden. Vorreiter im Hinblick auf die qualifizierte Nachnutzung einmal gewonnener Daten sind neben den Geowissenschaften die Sozialwissenschaften. Letztere verfügen in vielen Ländern seit Jahrzehnten über Datenzentren, die je nach Ausrichtung nicht nur wichtige Teile der wissenschaftlichen Infrastruktur sind, sondern auch selbst wissenschaftlich aktiv sind. Die Zentren aus aktuell 13 europäischen Staaten sind im Consortium of European Social Science Data Archives (CESSDA)[15] zusammengeschlossen.

Oft ist es erst die Aufbereitung der Daten durch diese Datenzentren, die eine Nachnutzung ermöglicht. Erhebungsdaten der Sozialwissenschaften und Beobachtungsdaten der Geowissenschaften teilen das Merkmal der Zeitgebundenheit: Sie stellen meist eine Momentaufnahme dar, die sich zu einem späteren Zeitpunkt grundsätzlich nicht noch einmal erzeugen lässt: Weder Erdbeben und Wetterereignisse noch die Befragung einer Personengruppe in einem bestimmten Kontext sind wiederholbar.

In anderen naturwissenschaftlichen oder technischen Fächern sind Daten zwar prinzipiell wieder zu beschaffen – und in vielen Fällen wird man diesem Weg und damit der Gewinnung besserer, z.B. höher aufgelöster Daten den Vorzug geben. Der Aufwand für eine solche Reproduktion kann jedoch gross sein, und zwar auch wenn man nicht die Grossforschungsanlagen der Hochenergiephysik heranzieht. Selbst für die Ausführung von Modellrechnungen kann es so langwierig und teuer sein, die benötigte Rechenzeit zu erhalten, dass es sinnvoller sein kann, die Ergebnisse von Modellrechnungen aufzubewahren und nicht nur die eigentlich für die Reproduktion ausreichenden Parameter und die Beschreibung des Modells.

Die beiden Aspekte Nachvollziehbarkeit und Nachnutzung sind Anforderungen, die Forschungsförderinstitutionen verstärkt an die Verantwortlichen der von ihnen unterstützten Forschungsvorhaben stellen. Wegen ihres Einflusses auf die internationale Diskussion seien hier die Vorgaben des EU-Forschungsrahmenprogramms Horizon 2020 an die Teilnehmenden an seinem Open Research Pilot zitiert:

«1) Step 1: participating projects are required to deposit the research data described above, preferably into a research data repository. ‹Research data repositories› are online archives for research data. They can be subject-based/thematic, institutional or centralised. […] In addition, it is expected that the

14 Revidierte Version vom 25. Oktober 2011, aufbereitet als Broschüre: *https://www.ethz.ch/content/dam/ethz/main/research/pdf/forschungsethik/Broschure.pdf*, Zugriff am 10. Juni 2014

15 *http://www.cessda.net/*, Zugriff am 10. Juni 2014

Open Access Infrastructure for Research in Europe (OpenAIRE) will become an entry point for linking publications to underlying research data.

2) Step 2: as far as possible, projects must then take measures to enable for third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data. One straightforward and effective way of doing this is to attach Creative Commons Licence (CC-BY or CC0 tool) to the data deposited (*http://creativecommons.org/licenses/*, *http://creativecommons.org/about/cc0*).

At the same time, projects should provide information via the chosen repository about tools and instruments at the disposal of the beneficiaries and necessary for validating the results, for instance specialised software or software code, algorithms, analysis protocols, etc. Where possible, they should provide the tools and instruments themselves.[16]»

Neben den Förderorganisationen anderer Länder macht auch der Schweizerische Nationalfonds bereits seit einigen Jahren gewisse Vorgaben für die von ihm geförderten Projekte.[17] Hochschulen und Forschungseinrichtungen unterstützen dieses Anliegen ebenfalls, wie zuletzt die im Mai 2014 von der deutschen Hochschulrektorenkonferenz publizierten Empfehlungen bestätigen.[18] Sie stellt unter anderem fest: «Die Hochschulleitungen sind gefordert, die strukturellen Voraussetzungen für ein effizientes, den gesamten Lebenszyklus der Daten (Erzeugung, Verarbeitung, Speicherung, Erschließung und Archivierung) umfassendes Forschungsdatenmanagement zu schaffen. Dabei geht es nicht nur darum, die technischen Voraussetzungen bereit zu stellen. Ebenso wichtig ist es, die Abläufe und die Rollenverteilung an der Hochschule zu organisieren und transparent zu machen.[19]»

Wissenschaftlerinnen und Wissenschaftler sind somit von verschiedenen Seiten mit höheren Erwartungen konfrontiert, die sie sinnvoll beantworten müssen. Diese Erwartungen sind zum Teil wissenschaftlicher Natur, etwa wenn Forschungsdaten von den Forschenden selbst so beschrieben und dokumentiert werden müssen, dass eine wissenschaftlich seriöse Nachnutzung möglich wird. Es fallen aber auch in grösserem Umfang organisatorische oder IT-Aufgaben an, die von Forschenden zwar durchaus bewältigt werden können, deren Erfüllung aber in den meisten Fällen zu Lasten ihres «Kerngeschäfts» in Forschung und Lehre geht. Es ist daher Aufgabe von Infrastruktureinrichtungen für Forschung und Lehre, Wissenschaftlerinnen und Wissenschaftler von diesen Aufgaben zu entlasten oder sie zumindest bei der Erfüllung zu unterstützen. Die Entscheidung darüber, welche Daten für welche Zeiträume aufbewahrt werden sollen, muss dabei innerhalb gesetzlicher Grenzen oder innerhalb des von den Richtlinien der jeweiligen Institution gesteckten Rahmens bei den Forschenden bleiben.

Doch warum sollen sich überhaupt die einzelnen Institutionen selbst mit der Erhaltung der Daten ihrer Angehörigen beschäftigen? Wäre es nicht sinnvoller, internationale Datenarchive zu nutzen, am besten solche, die fachlich spezialisiert sind? Tatsächlich sind etablierte und akzeptierte fachspezifische Angebote der bevorzugte Ort für die Sammlung eines Teils der Daten. Es gibt eine Reihe leistungsfähiger und organisatorisch stabiler Dienste – aber bei genauerer Betrachtung nur für erstaunlich wenige Wissenschaftszweige.[20] Typischerweise sind dies Fächer, die eine längere Tradition der Veröffentlichung und des Austausches von Forschungsdaten haben. Es ist daher kein Zufall, dass beispielsweise der innerhalb der Fachgemeinschaft breit abgestützte Dienst Pangaea[21] für die Geowissenschaften sehr häufig zitiert wird. Selbst die Abdeckung der Fächer des naturwissenschaftlich-technischen Spektrums der ETH Zürich ist insgesamt ausgesprochen lückenhaft. Zusätzlich machen die existierenden Dienste einschränkende Vorgaben zur Art der abzulegenden Daten, so dass es nicht möglich ist, zusätzliche und als wichtig erachtete Materialien mitzuliefern. Und schliesslich sind auch eine befristete Aufbewahrung von Daten oder die Ablage ohne Veröffentlichung der Daten in der Regel nicht vorgesehen.

Die Angebote von Hochschulen zur Aufbewahrung von Daten werden sich bei den grundsätzlich ver-

16 Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, European Commission, Version 1.0, 11. Dezember 2013, *http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf*, Zugriff am 10. Juni 2014

17 Beitragsreglement vom 14. Dezember 2007 (Stand 1. Juli 2012, *http://www.snf.ch/SiteCollectionDocuments/allg_reglement_d.pdf*) sowie Reglement über die Information, die Valorisierung und die Rechte an Forschungsresultaten vom 17. Juni 2008 (Fassung vom 01. Mai 2014, *http://www.snf.ch/SiteCollectionDocuments/allg_reglement_valorisierung_d.pdf*) , Zugriff am 10. Juni 2014

18 Empfehlung der 16. Mitgliederversammlung der HRK am 13. Mai 2014 in Frankfurt am Main: Management von Forschungsdaten – eine zentrale strategische Herausforderung für Hochschulleitungen, *http://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_13052014_01.pdf*, Zugriff am 10. Juni 2014

19 Ebd.

20 Nachweise solcher Dienste werden inzwischen systematisch gepflegt im Registry of Research Data Repositories (*www.re3data.org*) und in Databib (*http://databib.org*). Beide werden auch explizit in den Guidelines zu Horizon 2020 genannt.

21 PANGAEA – Data Publisher for Earth & Environmental Science, *http://www.pangaea.de/*, Zugriff am 10. Juni 2014

fügbaren Funktionen mit solchen fachspezifischen Diensten überschneiden, etwa bei der Registrierung von persistenten Identifikatoren (z.B. Digital Object Identifier, DOI), der Veröffentlichung und der Erhaltung auf Bit-Ebene. Sobald spezielle Funktionen und Methoden einzelner Fächer abgebildet werden müssen, stossen allgemeine, lokal betriebene Daten Archive an Grenzen. Dagegen können und müssen sie offener sein hinsichtlich der abzuliefernden Inhalte und der gewünschten Zugriffsregelungen.

Es erscheint zunächst möglicherweise nicht als zwingend, dass wissenschaftliche Bibliotheken als Betreiber solcher Datenarchive in den Hochschulen auftreten. Es gibt aber einleuchtende Gründe, warum sie eine wichtige Rolle spielen sollten, auch wenn diese je nach Struktur und Aufgabenteilung innerhalb einer Hochschule unterschiedlich aussehen kann. So kommt etwa der Dokumentation des Kontextes von Forschungsdaten eine entscheidende Bedeutung beim Auffinden relevanter Daten und bei ihrer qualifizierten Nachnutzung zu. Die Beschreibung von inhaltlichen, administrativen und technischen Eigenschaften mit Hilfe von standardisierten Metadaten gehört zu den Kernkompetenzen von Bibliotheken und Archiven. Diese Kompetenzen können und sollten auch für den Umgang mit Forschungsdaten genutzt werden und können Forschende bei der Erfüllung ihrer Aufgaben unterstützen. Es ist ganz klar nicht das Ziel, Forschende dabei zu Bibliothekaren zu machen und ebenso klar ist, dass starre Metadatenschemata zur wissenschaftlichen Dokumentation beitragen können, dass aber darüber hinaus in aller Regel eine ausformulierte Dokumentation nötig ist, z.B. in Form der zu einem Datenpaket gehörenden Dissertation.

Auch haben Bibliotheken in den vergangenen Jahren massgeblich den Aufbau von Dienstleistungen vorangetrieben, die nun als Bausteine für eine Forschungsdateninfrastruktur zur Verfügung stehen. Dies betrifft zum Beispiel die Registrierung von Digital Object Identifiern (DOI) als persistente Identifikatoren, die die Möglichkeit für die Zitierung von Datenobjekten schaffen. Für die Schweiz hat die ETH Zürich als Mitglied im internationalen Konsortium DataCite[22] die Registrierung aufgebaut. Die Dienstleistungen des DOI-Desk[23] der ETH-Bibliothek können von allen Hochschulen der Schweiz genutzt werden. In Erweiterung dieser Dienstleistung beteiligen sich Bibliotheken wie die ETH-Bibliothek via DataCite auch daran, die Verbreitung von eindeutigen Identitäten für wissenschaftliche Autorinnen und Autoren voranzutreiben. Sie nutzen dafür ORCID[24] (Open Researcher and Contributor ID). Eine solche eindeutige Identifikation ermöglicht die zweifelsfreie Zuordnung von Publikationen oder Datensätzen zu einer Person, unabhängig davon, in welcher Schreibweise ihr Name oder ihre Institution erscheinen und wer ihr aktueller Arbeitgeber ist.

Viele Hochschulbibliotheken unterstützen bereits heute die Abläufe rund um das elektronische Publizieren, etwa die Erfassung von Publikationen der Hochschulangehörigen für die Bibliographie und das Reporting der Institution oder die Förderung von Open Access-Veröffentlichungen auch in eigenen institutionellen Repositorien. Anforderungen an den Umgang mit Forschungsdaten zielen nicht zuletzt darauf, Datensätze entweder als Begleitmaterial zu formalen Veröffentlichungen bereitzustellen oder auch als eigenständig, zitierbare und beschriebene Objekte. Eine enge Verbindung der entsprechenden Abläufe für Publikationen und Forschungsdaten ist daher sinnvoll, damit Forschende nicht mit unnötig vielen unterschiedlichen Plattformen konfrontiert sind.

Wegen des engen Zusammenhangs mit dem elektronischen Publizieren sind es meist die Hochschulbibliotheken, die auch die Hochschulbibliographie pflegen und – je nach Hochschule - mehr oder weniger systematisch auch Dienstleistungen zu bibliometrischen Auswertungen wie Zitationsanalysen anbieten. Es ist unumstritten, dass für eine konsequente Umsetzung der Ziele der Forschungsförderer im Sinne der Nachnutzung von Daten zusätzliche Anreize nötig sind. Dazu gehört insbesondere die verstärkte Anerkennung der Veröffentlichung von dokumentierten Daten als Publikation. Um diese zu ermöglichen, kann auf die bestehenden Strukturen des wissenschaftlichen Publizierens zurückgegriffen werden. Vereinzelt haben sich auch wissenschaftliche Zeitschriften etabliert, die ausschliesslich die Beschreibung von wissenschaftlichen Datensätzen als Artikel veröffentlichen und sie so zitierbar und als eigenständige Publikation sichtbar machen. Als Beispiel sei die Zeitschrift Earth System Science Data (ESSD)[25] aus den Geowissenschaften genannt.

Implizit ist klar, dass Daten, die einmal publiziert und mit einem persistenten Identifikator versehen wurden, auf Dauer nicht nur auffindbar, sondern auch nutzbar bleiben sollten. Als eher langlebige Institutionen sind Hochschulbibliotheken aus organisatorischer Sicht in einer guten Position, um die Erhaltung

22 http://www.datacite.org, Zugriff am 10. Juni 2014
23 http://www.library.ethz.ch/DOI-Desk, Zugriff am 10. Juni 2014
24 http://orcid.org/, Zugriff am 10. Juni 2014
25 http://www.earth-system-science-data.net/, Zugriff am 10. Juni 2014

von Daten über längere Zeiträume zu gewährleisten. Aber auch inhaltlich beschäftigen sich Bibliotheken wie die ETH-Bibliothek teilweise schon seit Jahren mit der langfristigen Erhaltung der Nutzbarkeit von digitalen Daten. Bei Forschungsdaten sind diesen Bemühungen zwar Grenzen gesetzt durch die Vielfalt der verwendeten Formate, die häufig nicht ausreichend dokumentiert sind. Die grundsätzlichen Erwägungen gelten aber auch hier und es ist zu hoffen, dass durch einen Erfahrungszuwachs über längere Zeiträume weitere Verbesserungen erzielt werden können.

Zum Teil bieten Verlage die Möglichkeit an, Daten als so genannte «Supplementary Material» mit einem Artikel zusammen zu hinterlegen. Nicht wenige Verlage sehen dies allerdings angesichts des Umfangs und der langfristigen Herausforderungen nicht als ihre Aufgabe an. Auch spricht die Erfahrung mit teilweise kurzlebigen Angeboten dafür, dass die Aufgabe allenfalls bei gewinnorientierten Unternehmen nicht optimal aufgehoben ist. Auch lässt sich ein gewisses Unbehagen nicht leugnen, nun auch noch die Forschungsdaten aus der Hand zu geben: In der Diskussion um Open Access zu Publikationen ist deutlich geworden, dass es kaum zu vermitteln ist, warum Hochschulen den Zugang zu öffentlich finanzierten Publikationen mit öffentlichen Mitteln von Verlagen zurückkaufen müssen. Für die Forschungsdaten soll eine solche Situation verhindert werden. Neben internationalen fachspezifischen Repositorien sind daher auch institutionelle Lösungen der Hochschulen zum Thema geworden.

Im Folgenden soll ausgeführt werden, wie die Bewahrung von Forschungsdaten den praktischen Nutzen der digitalen Bibliothek für die Wissenschaft wesentlich bereichern kann und welche konkreten Schritte die ETH-Bibliothek dabei geht. Soweit es um die dauerhafte Bewahrung von Forschungsdaten geht, ist die Fachstelle Digitaler Datenerhalt[26] die Hauptakteurin innerhalb der ETH-Bibliothek. Sie wurde 2010 zunächst als Projektteam für den Aufbau des ETH Data Archive installiert und hat 2014 den produktiven Betrieb aufgenommen. Wichtige Partner sind die Gruppe IT-Services der ETH-Bibliothek, das Team Elektronisches Publizieren sowie Hochschularchiv und Informatikdienste der ETH Zürich. Eine eigentliche Daten-Policy gibt es bisher noch nicht. Sie soll in Zusammenarbeit der verschiedenen betroffenen Stellen erarbeitet und insbesondere mit den Forschenden abgestimmt werden. Aus Sicht der ETH-Bibliothek ist ein wesentliches Ziel dabei die Klärung der Verantwortlichkeiten und die Präzisierung bestehender Regelungen, weniger die Festlegung neuer Verpflichtungen für die Forschenden.

Als Orientierungspunkte dienen im Folgenden grob die Stationen des Lebenszyklus.

– Hypothese / Fragestellung:
  In dieser wesentlichen Phase des wissenschaftlichen Erkenntnisprozesses können Bibliotheken und andere Dienstleister Forschende unterstützen, indem sie das Auffinden relevanter Publikationen, Informationen und Daten erleichtern und den Zugriff darauf vermitteln. Dazu tauschen sie Metadaten freizügig aus und pflegen persistente Referenzen auf Datensätze. Die ETH-Bibliothek nutzt Digital Object Identifier (DOI). Sie betreibt dazu im Rahmen des DataCite-Konsortiums das DOI-Desk für die Schweiz. Im Idealfall erhalten Forschende so die Möglichkeit, existierende Daten, die für sie von Interesse sind, für ihre Arbeit zu berücksichtigen. Zwingende Voraussetzung dafür ist eine umfassende Dokumentation der Daten und ihres Entstehungskontexts. Wo immer möglich, dürfte der direkte Kontakt mit den Datenproduzenten wichtig sein.

– Antragstellung:
  Wenn eine Projektfinanzierung beantragt wird, um die Fragestellung untersuchen zu können, müssen Forschende immer häufiger einen Datenmanagementplan (DMP) mitliefern, der beschreibt, welche Daten erzeugt oder erhoben werden sollen und wie mit ihnen umgegangen werden soll (Veröffentlichung, Möglichkeit der Nachnutzung, langfristige Erhaltung, Anonymisierung….). Es gibt international mehrere Versuche, die Erstellung von Datenmanagementplänen mit Online-Werkzeugen zu unterstützen. In der Praxis gibt es allerdings nicht «den» Datenmanagementplan, da jeweils unterschiedliche fachliche Gegebenheiten zu berücksichtigen sind. Da sich einige Fragen zu diesem frühen Zeitpunkt noch nicht oder noch nicht abschliessend beantworten lassen, wird auch vorgesehen, den Plan während der Projektdauer nachzuführen. Hier können Bibliotheken vor allem dann unterstützen, wenn es im jeweiligen Fach keine klar etablierte Praxis mit anerkannten fachspezifischen Repositorien gibt. Sie können Kriterien für die Bewertung und Auswahl möglicher externer und interner Repositorien liefern und falls nötig bei der Einschätzung von deren Eignung helfen. Grosse Bedeutung hat auch die Diskussion der geplanten Dateiformate und ihrer Vor- und Nachteile für die spätere Weiterverwendung. Aber wie die Erfahrung lehrt, werden auch ganz praktische Fragen gestellt, z.B. zu sinnvollen Datei- und Ordnernamen oder Identifikationssystemen.

26 *http://www.library.ethz.ch/Digitaler-Datenerhalt*, Zugriff am 10. Juni 2014

Es steht ausser Frage, dass die Bibliotheken wie auch die Forschenden das notwendige Wissen zunächst selbst aufbauen müssen. Der Vorteil einer Einbindung der Bibliotheken liegt nicht zuletzt in der Möglichkeit, die Erfahrungen aus verschiedensten Anwendungsfällen zu bündeln und nachnutzbar zu machen.

– Datenerfassung und -erhebung sowie Analyse und Interpretation:
Die hier verwendeten Methoden sind stark fachabhängig. Es ist sinnvoll, die verwendeten Formate frühzeitig auf ihre Eignung für die langfristige Aufbewahrung und Veröffentlichung zu prüfen und auch zu klären, wie die verwendeten Verfahren, Algorithmen usw. zusammen mit den Datensätzen dokumentiert und verfügbar gehalten werden sollen. Auf diese Weise könnte z.B. zu einem frühen Zeitpunkt entschieden werden, zusätzlich zu einem herstellerspezifischen Ausgabeformat auch noch ein offenes Austauschformat zu verwenden, in dem eine bessere Chance für eine spätere Nachnutzung besteht. Das Beispiel zeigt allerdings auch die Grenzen, die der Langzeitarchivierung gesetzt sind: Es dürfte eher selten möglich sein, ein solches Austauschformat zu finden, das tatsächlich alle Eigenschaften des ursprünglichen Formats übernimmt und nutzbar erhält. Darum ist die fachliche Abwägung der Forschenden notwendig, welchen Informations- oder Funktionsverlust sie für welchen Zweck in Kauf zu nehmen bereit sind.

– Synthese und Veröffentlichung:
Datensätze können entweder aus einer formalen Publikation heraus zitiert werden, bevorzugt mittels DOI (siehe oben) oder sie können als eigenständige Objekte behandelt und ebenfalls mit einem DOI versehen und veröffentlicht werden. Beides geschieht mit steigender Tendenz. Als Anreiz für die sorgfältige Dokumentation von Datensätzen ist allerdings eine erhöhte Anerkennung dieser Beiträge als wissenschaftliche Leistung nötig, damit die Beschreibung und Bereitstellung von gepflegten Datensätzen langfristig neben die heute fast ausschliesslich betrachteten formalen Publikationen (Artikel, Buchkapitel, Bücher) treten kann. Dabei helfen neben den persistenten Objektreferenzen wie DOI auch eindeutige Identifikatoren für die Autorinnen und Autoren. Beispiel ist hier ORCID, ein ebenfalls vor allem von Bibliotheken gepflegtes und vorangetriebenes System, das Zuschreibungsfehler z.B. aufgrund inkonsistenter Adress- und Zugehörigkeitsdaten vermeidet. Die Metadaten von veröffentlichten Forschungsdaten werden ab Mitte 2014 auch im Wissensportal der ETH-Bibliothek[27] zugänglich gemacht, das die über die ETH-Bibliothek zugänglichen Informationsressourcen bündelt und für die ETH-Angehörigen zugänglich macht.

– Zugriff und Prüfung:
Bibliotheken stellen mit ihren Partnern – häufig Rechenzentren – zunehmend Infrastrukturen bereit, die die Ablieferung von Daten für eine befristete oder dauerhafte Aufbewahrung ermöglichen und den Zugriff gemäss vereinbarten Vorgaben steuern. Solche institutionellen Lösungen werden nötig, weil es nur in vergleichsweise wenigen Fächern etablierte und allgemein anerkannte fachspezifische Repositorien für die Datenarchivierung gibt. Und auch in Fächern, in denen sie existieren, machen sie teilweise enge Vorgaben für die Art und Nutzung der Inhalte, so dass ein grösserer Teil von Daten nicht dort deponiert werden kann. Eine kontinuierliche Herausforderung liegt darin, die Ablieferung von Daten für die Forschenden einfach zu halten und dennoch die erforderliche Qualität der Beschreibung sicherzustellen. Die ETH-Bibliothek bietet neben der direkten Ablieferung über einen Web-Dialog auch einen Viewer und Editor für die lokale Nutzung an, mit dem Forschende ihre Daten strukturieren, organisieren und nach Bedarf in der gewünschten Granularität mit Metadaten versehen können, bevor sie diese als Archivpakete an das ETH Data Archive übergeben. Dieses Werkzeug wird auch für die an das Hochschularchiv der ETH Zürich abliefernden Stellen eingerichtet.

– Erhaltung:
Neben der offensichtlichen Notwendigkeit der Erhaltung von Daten auf Bit-Ebene müssen Daten auch nutzbar erhalten werden. Die Chancen dafür steigen, wenn offene, gut dokumentierte Standardformate zum Einsatz kommen oder zumindest zusätzlich zu allenfalls proprietären Dateiformaten erzeugt und aufbewahrt werden (siehe oben). Selbst wenn solche Formate gefunden werden, muss in der Folge kontinuierlich beobachtet werden, ob die vorhandenen Formate weiterhin nutzbar bleiben und unterstützt werden können oder ob Alternativen ins Auge gefasst werden müssen. Über lange Zeiträume hinweg kann diese Aufgabe nur von institutionellen Dienstleistern wie den Bibliotheken erbracht werden. Selbstverständlich ist die Rückkopplung mit den potentiellen Nutzerinnen und Nutzern, also den Mitgliedern der Fachgemeinschaft anzustreben. Deren Engagement dürfte allerdings direkt davon abhängen, ob sie

27 *http://www.library.ethz.ch*, Zugriff am 10. Juni 2014

selber in irgendeiner Form von den anstehenden Fragen betroffen sind.

– Nachnutzung:
  Die oben beschriebenen Massnahmen schaffen die Voraussetzung dafür, dass eine Nachnutzung von Daten technisch möglich bleibt. Eine Schlüsselrolle bei der inhaltlich sinnvollen Verwendung und auch beim blossen Auffinden relevanter Daten kommt den Metadaten zu. Die Beschreibung von inhaltlichen, administrativen und technischen Eigenschaften mit Hilfe von standardisierten Metadaten gehört zu den Kernkompetenzen von Bibliotheken und Archiven. Dies Kompetenzen können und sollen auch für den Umgang mit Forschungsdaten genutzt werden. Die bisherige Erfahrung zeigt, dass das Interesse für eine aussagekräftige Beschreibung bei den Forschenden gering ist, wenn sie lediglich die befristete Aufbewahrung ihrer Daten als Ziel haben. Es ist denkbar, dass die Bereitschaft wächst, hier einen gewissen Aufwand zu treiben, wenn erste schlechte Erfahrungen mit zu wenig klaren Metadaten gemacht wurden. Bei den Daten, die zur Veröffentlichung bestimmt sind, müssen die Mindestanforderungen von DataCite für Metadaten zur DOI-Registrierung erfüllt sein.[28] Da diese lediglich fünf Pflichtfelder vorsehen, muss zur aussagekräftigen Dokumentation in der Regel auf eine Veröffentlichung verwiesen werden.

Die ETH-Bibliothek versucht gemeinsam mit engagierten Pilotpartnern aus der Forschung praktikable Lösungen zu erarbeiten. Da die Anwendungsfälle in der Praxis sehr unterschiedlich sind, wird es leider kaum eine universelle Lösung für alle Datenproduzentinnen und –produzenten an der ETH Zürich geben. Ziel ist es, langfristig über einen Werkzeugkasten zu verfügen, mit dessen Hilfe im Zusammenwirken von Forschenden und ETH-Bibliothek eine grössere Bandbreite von Anforderungen an die Aufbewahrung, Bereitstellung und digitale Langzeitarchivierung abgedeckt werden kann. Selbstverständlich dupliziert die ETH-Bibliothek nicht die bestehende Infrastruktur der Informatikdienste. Die Zusammenarbeit mit den Informatikdiensten der ETH Zürich ist daher eine wichtige Säule für die Erhaltung von Daten. Die Gruppe ID Speicher stellt einerseits die Infrastruktur bereit, in der die Daten des ETH Data Archive abgelegt werden, und andererseits kann sie unkomplizierte Lösungen zur reinen Datenspeicherung anbieten, wenn eine Ablieferung an das ETH Data Archive nicht sinnvoll ist. Dies ist insbesondere der Fall, wenn eine Aufbewahrung für weniger als zehn Jahre gewünscht ist oder wenn keine Beschreibung der Inhalte mit Metadaten (mehr) möglich ist und auch keine Absicht besteht, Daten mit einem DOI zu versehen und sie für Dritte zugänglich zu machen.

Welche Angebote für die Wissenschaftlerinnen und Wissenschaftler überhaupt interessant sind, hängt ausgesprochen stark vom jeweiligen Fach ab. Dies beginnt bereits bei der Menge der produzierten Daten. Während «Big Data» bereits zum Schlagwort avanciert ist, fallen dennoch in vielen Projekten eher kleinteilige Daten in mehr oder weniger strukturierter Form an. Die Herausforderungen unterscheiden sich und gerade in wissenschaftlichen Disziplinen wie den Geowissenschaften, die sich gewissermassen alle mit dem gleichen «Untersuchungsgegenstand» beschäftigen, ist das Bewusstsein für den Wert einmal erhobener Daten und für den Nutzen ihres offenen Austauschs vielleicht am ausgeprägtesten. Folgerichtig gibt es hier auch eine entsprechende Kultur der Nachnutzung von Daten und etablierte Plattformen zur Unterstützung dieses Austauschs. Es wäre vermessen, solchen gut verankerten Diensten Konkurrenz machen zu wollen. Viel sinnvoller ist es, von diesen Diensten zu lernen, um eine eigenen Lösung anbieten zu können für Fächer, in denen solche Plattformen fehlen sowie für Daten, die aus verschiedenen Gründen bei solchen Diensten keinen Platz finden, z.B. weil sie (noch) nicht frei zugänglich gemacht werden können oder schlicht nicht in das Profil passen.

Ausblick

Der Umgang mit Forschungsdaten ist immer noch ein vergleichsweise junges Thema, an dem verschiedene Anspruchsgruppen Interesse haben. Dazu gehören Forschende, Forschungsförderer, Bibliotheken, Rechenzentren und andere Dienstleister sowie Hochschulleitungen, Verlage und weitere. Gewisse Rahmenbedingungen werden auf nationaler oder internationaler Ebene gesetzt, während es aufgrund der unterschiedlichen lokalen Gegebenheiten ganz unterschiedliche praktische Umsetzungen und Aufgabenteilungen innerhalb der Hochschulen geben kann. - Als sicher erscheint bisher, dass Forschungsdaten alle Beteiligten in den kommenden Jahren intensiv und auf unterschiedliche Art und Weise beschäftigen werden. ■

28 Vgl. DataCite Metadata Schema v 3.0, *http://schema.datacite.org/meta/kernel-3/index.html*, Zugriff am 10. Juni 2014