Forschungsdaten-Management

von Jens Klump und Jens Ludwig

Blick zurück: Entwicklung des Forschungsdaten-Managements in den letzten zehn Jahren

Das Deutsche GeoForschungsZentrum (GFZ) in Potsdam und das Zentrum für marine Umweltwissenschaften (MARUM) der Universität Bremen hatten nach erfolgreichen Pilotstudien 2004 zusammen mit der Technischen Informationsbibliothek Hannover (TIB) und dem Deutschen Klimarechenzentrum (DKRZ) das DFG-Projekt „Publikation und Zitierbarkeit wissenschaftlicher Primärdaten“ (STD-DOI) begonnen. Mit den bereits gesammelten Erfahrungen im Management geowissenschaftlicher Datenbestände am GFZ und am MARUM wollte man dieses Wissen mit der Expertise ausgewiesener Bibliotheken in einem Leistungszentrum für Forschungsinformation in den Geowissenschaften zusammenführen, dem „Zentrum für geowissenschaftliches Informationsmanagement“ (CeGIM). Am von der DFG geförderten Pilotprojekt waren das GFZ, das MARUM, die SUB Göttingen, die UB der TU Bergakademie Freiberg und die Bibliothek des Wissenschaftsparks Albert Einstein in Potsdam beteiligt. Dieses Projekt war der Ausgangspunkt der seitdem andauernden Zusammenarbeit der Abteilung Forschung & Entwicklung der SUB Göttingen und des Zentrums für Geoinformationstechnologie des GFZ bzw. seiner Vorgängereinheit, dem Daten- und Rechenzentrum des GFZ.

Schon sehr früh hatte man in den Geowissenschaften erkannt, dass Strukturen für die Bereitstellung und Langzeiterhaltung der Daten notwendig sind. Bereits für das Internationale Geophysikalische Jahr (IGY, 1957 bis 1958) wurden die World Data Center (WDC) eingerichtet, um die Daten des IGY bereitzustellen und zu erhalten (vgl. Dittert/Diepenbroek/Grobe 2001; Pfeiffenberger 2007). Darüber hinaus sollte das System der WDC in den Zeiten des Kalten Kriegs helfen, durch Zusammenarbeit und Transparenz zum Vertrauen zwischen Wissenschaftlern der beiden Machtblöcke beizutragen und dadurch zur politischen Entspannung beitragen. Mit dem Ende des „Kalten Krieges“ und der Entstehung des Internets verlor das System der

WDC in der bis dahin bestehenden Form seinen Sinn, sodass 2008 der International Council for Science (ICSU) beschloss, das System der WDC zu reformieren und in das World Data System (WDS) zu überführen.

Das entstehende Internet erlaubte einen immer einfacheren und schnelleren Austausch von Forschungsdaten, sodass bald die kulturelle Weiterentwicklung des Umgangs mit Forschungsdaten hinter den technischen Möglichkeiten zurückblieb (vgl. Klump et al. 2006). Ein Ergebnis der Unzufriedenheit mit dieser Entwicklung war die Veröffentlichung der „Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen“ vom 22. Oktober 2003 (Berlin Declaration 2003). Die „Berliner Erklärung“ wird zwar in erster Linie als Meilenstein der Open-Access-Bewegung gesehen, jedoch beinhaltet sie auch die Forderung nach einem offenem Zugang zu Forschungsdaten und geht damit über die „Regeln für eine gute wissenschaftliche Praxis“ der DFG (1998) hinaus, die in erster Linie die Beweissicherung zur Verhinderung und Prüfung wissenschaftlichen Fehlverhaltens behandeln (vgl. Klump et al. 2006). Das Potenzial, das im offenen Zugang zu wissenschaftlichem Wissen steckt, wurde auch von der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) erkannt, weshalb deren Rat 2006 eine Empfehlung über den Zugang zu Daten aus öffentlich geförderter Forschung verabschiedete (vgl. OECD 2009). Es wird üblicherweise von den Mitgliedsstaaten der OECD erwartet, dass diese eine Empfehlung des OECD-Rates in nationale Gesetzgebung überführen. Im Falle der Empfehlung zum Umgang mit Forschungsdaten wurde dieser Prozess auch weiter von der OECD begleitet. Ein Ergebnis ist, dass die DFG seit 2010 in ihrem Leitfaden für Antragsteller eine Erklärung verlangt, wie mit den im Projekt gewonnenen Forschungsdaten umgegangen werden soll (vgl. DFG 2010), und das Thema des Umgangs mit Forschungsdaten auch systematisch weiterentwickelt. Noch weiter geht der im Februar 2013 veröffentlichte Erlass der US-Bundesregierung (vgl. Holdren 2013), die alle US-Bundesbehörden dazu verpflichtet, Ergebnisse öffentlich geförderter Forschung – dazu zählen Literatur und Daten – binnen zwölf Monate nach Veröffentlichung der Arbeiten kostenfrei zugänglich zu machen.

In der Praxis ist die Entwicklung eines systematischen und auch offenen Umgangs mit Forschungsdaten unter den Fachdisziplinen ungleich ausgeprägt. Ein offener Umgang mit Forschungsdaten ist vor allem in den Feldern zu beobachten, in denen der einzelne Forscher von einer Zusammenarbeit mit anderen Gruppen in seinem Feld profitiert (vgl. Neuroth et al. 2012). Dies drückt sich auch in dem Ungleichgewicht aus, dass die Bereitschaft, Daten

nachzunutzen überwiegt gegenüber der Bereitschaft, Daten mit anderen zu teilen (vgl. Borgman 2012). Die fehlende Bereitschaft dazu liegt wenigstens teilweise darin begründet, dass es heute wenig Anreize für den Wissenschaftler gibt, die zusätzliche Arbeit, die für das Bereitstellen von Daten für andere notwendig wäre, zu leisten. Darüber hinaus muss jedoch auch untersucht werden, wie die Akteure und ihre Rollen im Umgang mit Forschungsdaten verteilt sind und an welchen Institutionen sie angesiedelt sind.

Eine besondere Bedeutung in der Entwicklung neuer Praktiken im Umgang mit Forschungsdaten kommt den Bibliotheken zu. Insbesondere an akademischen Einrichtungen sind sie Teil der Infrastruktur und haben dabei unter anderem die Funktion der Gedächtnisorganisation und des Informationsdienstleisters. Bibliotheken könnten somit eine Rolle bei der Übernahme von digitalen Forschungsergebnissen spielen, um ihren Vertrieb, ihre Veröffentlichung und ihre langfristige Aufbewahrung zu organisieren. In der Ausbildung des Personals kommt das Thema Forschungsdaten gerade erst an und die konzeptionelle und technische Entwicklung in diesem Feld ist immer noch so schnell, dass Lehrmaterialien binnen kurzer Zeit veralten.

Bibliotheken sind gegenüber den Fachwissenschaften „fachfremd“. Das heißt, dass sie digitale Forschungsdaten nur übernehmen können und begrenzt zu deren systematischen fachlichen Beschreibung beitragen können. Auch die Entwicklung und der Betrieb der informationstechnischen Infrastrukturen, wie sie hier in Zukunft benötigt werden, ist bisher kein typischer Tätigkeitsbereich einer Bibliothek. Insofern müssen in Zukunft möglicherweise auch kooperative Modelle der Zusammenarbeit von Bibliotheken, Rechenzentren und disziplinären Kompetenzzentren für Forschungsdaten entwickelt werden. Diese Fragen, wie der Umgang mit Forschungsdaten verbessert werden kann und welche Rolle dabei Bibliotheken spielen könnten, wollen wir in diesem Kapitel erörtern.

Die Gegenwart: Aufgaben und Zuständigkeiten im Forschungsdaten-Management

Auch wenn Forschungsdaten-Management kein neues Thema ist, so wird es derzeit stärker als je zuvor diskutiert. Im Bereich der Bibliotheken wird in dem Thema oftmals eine Chance gesehen, sich angesichts der unklaren Rolle der Bibliotheken in der Informationsinfrastruktur der Zukunft ein neues Thema zu erschließen. Manchmal erfolgt dies mit einer Selbstverständlich-

keit und vielleicht auch Naivität, die bei Forschungsinstitutionen, die sich schon lange um Forschungsdatenbestände kümmern, die Skepsis gegenüber der Popularität des Themas bekräftigt. Um zu verstehen, welche Institutionen welche Rolle im Forschungsdaten-Management sinnvoll spielen können, ist es notwendig, den unscharfen Begriff des Forschungsdaten-Managements genauer aufzuschlüsseln.

Es kommt vor, dass Vertreter von Bibliotheken behaupten, dass sie bereits sehr lange Forschungsdaten-Management betreiben. Diese überraschende Aussage wird verständlich, wenn man sich vor Augen führt, wie breit der Begriff Forschungsdaten verwendet wird. Oft wird darunter alles verstanden, was in der Forschung benutzt wird und digital vorliegt, sodass auch z.B. digitale Artikel oder digitalisierte Bücher und Nachlässe als Forschungsdaten gelten. Um die eigentlichen Besonderheiten des Forschungsdaten-Managements in den Blick zu bekommen, ist es aber hilfreich, Forschungsdaten in einem engeren Sinne als die Daten zu verstehen, die im Forschungsprozess für die Untersuchung eines Forschungsgegenstands benutzt werden oder wurden. Der paradigmatische Fall sind Daten aus Messungen, aber auch Ergebnisse aus Simulationen oder Umfragen können als typisch gelten.

Forschungsdaten werden nicht nur begrifflich mit Publikationen assoziiert, sondern auch viele Aktivitäten des Forschungsdaten-Managements hängen mit Publikationen oder Publizieren zusammen. Die bereits erwähnte Empfehlung Nummer 7 der „Regeln für eine gute wissenschaftliche Praxis“, dass „Primärdaten als Grundlagen für Veröffentlichungen [...] auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden [sollen]“ (DFG 1998: 12), begründet beispielhaft die Notwendigkeit des Forschungsdaten-Managements als Dokumentation für Publikationen. Es gibt auch noch weitere Fälle, in denen sich Forschungsdaten-Management aus einer Pflicht zur Dokumentationen begründet, bei denen aber nicht unbedingt die Forschung selbst, sondern auf ihr basierende Entscheidungen oder Produkte überprüfbar sein sollen, wie z.B. Patientenbehandlungen in der Medizin oder kommerzielle Produkte in der Wirtschaft.

Eine etwas anders ausgerichtete Art von Forschungsdaten-Management zielt nicht darauf, Forschungsdaten als Dokumentation für Publikationen zu erhalten, sondern Forschungsdaten als Publikationen zur Nachnutzung bereitzustellen. Dadurch, dass das Veröffentlichen von Daten jeder Art durch das Internet viel einfacher geworden ist, ist es in großem Maßstab möglich geworden, Forschungsdaten zu publizieren. Insbesondere bei aufwendig erzeugten Forschungsdaten soll eine höhere Effizienz des Wissenschaftssys-

tems dadurch erreicht werden, dass Daten für neue Forschung nachgenutzt werden.

Bibliotheken werden traditionell als Sammler und Bewahrer von Publikationen verstanden und vor diesem Hintergrund scheint es naheliegend, sie auch als Akteure im Bereich der beiden eben beschriebenen Arten von Forschungsdaten-Management zu sehen. Typische Fragestellungen sind dann, wie Forschungsdaten in die bestehenden Repositorien übernommen und in den Katalogen nachgewiesen werden können. Für den Anwendungsfall der Dokumentation kann die Bibliothek wahrscheinlich auch ihre traditionellen Kompetenzen nutzbringend anbringen und ein wichtiger Akteur sein. Da auf solche Daten allerdings nur sehr selten zugegriffen wird und sie nicht mehr unmittelbar produktiv genutzt werden, handelt es sich aber um keine besonders prestigeträchtige Aktivität. Im Fall der Forschungsdaten-Publikationen für die Nachnutzung wird zwar wahrscheinlich auch selten auf den einzelnen Datensatz zugegriffen, aber durch die Nutzung für neue Forschung erhält das Forschungsdaten-Management ein positiveres Ansehen.

Ob Bibliotheken dafür die geeignetsten Institutionen sind, ist allerdings zweifelhaft, denn das Management von Forschungsdaten als Publikation für die Nachnutzung ist eine sehr voraussetzungsreiche Aktivität. Es existieren nicht ohne Grund viele disziplinspezifische Forschungsdatenzentren, da eine hohe Disziplinverankerung und großes Fachwissen notwendig sind. Typische Aktivitäten sind die aufwendige Aufbereitung und Dokumentation von Datensätzen, die Qualitätskontrolle von Forschungsdaten, die Pflege langfristiger Zeitreihen, die Interaktion mit der Fach-Community und die Begleitung der Entwicklung der Disziplin, um z.B. über Veränderungen der Anforderungen oder der Terminologie auf dem Laufenden zu bleiben. In einigen Fachdisziplinen wird die Veröffentlichung von Forschungsdaten in ganz bestimmten Zentren vorausgesetzt, wie z. B. der Protein Data Bank. Damit ist nicht ausgeschlossen, dass auch Bibliotheken Forschungsdaten veröffentlichen und pflegen oder gar die Rolle eines Forschungsdatenzentrums übernehmen, aber es erfordert eine fachliche Spezialisierung, die wohl nur sehr wenige Institutionen vorweisen können. Der Unterschied drückt sich auch in der Zielgruppe des Forschungsdaten-Managements aus: Während die Dokumentation von Forschungsdaten nur erfordert, einen lokalen Auftrag zu erfüllen, meist auf dem kleinsten gemeinsamen Nenner verschiedener Disziplinen, geht man mit der Pflege publizierter Forschungsdaten eine Verpflichtung gegenüber einer globalen und ausdifferenzierten Fachdisziplin ein.

Der Fokus auf den Zusammenhang von Forschungsdaten und Publikationen vernachlässigt aber noch eine dritte Art des Forschungsdaten-Managements – das wissenschaftliche Arbeiten mit Forschungsdaten in einem Forschungsprojekt selbst zu unterstützen. In diesem Fall geht es nicht um das Management von Publikationen, sondern primär darum, die wissenschaftliche Untersuchung selbst zu verbessern, zu vereinfachen oder durch neue Forschungsmethoden zu erweitern (wie z.B. Data Mining). Die meisten Aufgaben und Aktivitäten des Forschungsdaten-Managements, wie z.B. die adäquate Beschreibung der Forschungsdaten und ihres Entstehungskontextes durch Metadaten oder die Verwaltung einer Speicherinfrastruktur, können unabhängig von ihrer Positionierung in Forschungs- und Datenlebenszyklen nützlich sein (s. Abb. 1). Sie können sowohl für die Dokumentation als auch für die Nachnutzung und auch während der Projektphase nützlich sein. Das Forschungsdaten-Management während eines Projekts ist damit sozusagen der Zwilling von Konzepten wie E-Science, E-Research, Digital Humanities oder Virtuellen Forschungsumgebungen, die stärker neuartige, kollaborative Werkzeuge in den Vordergrund stellen – aber, um zu funktionieren, auch entsprechend aufbereitete Forschungsdaten benötigen. Ein Beispiel für diese Art des Forschungsdaten-Managements sind die INF-Teilprojekte, die im Programm Sonderforschungsbereiche der DFG zur Unterstützung eines Gesamtprojekts beantragt werden können.

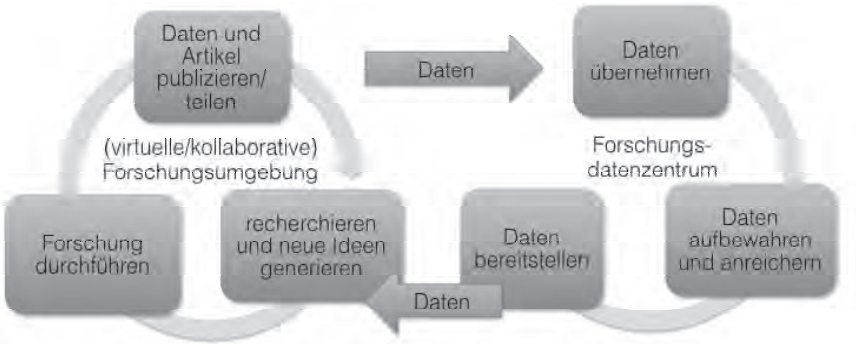

Abb. 1 Idealisierter Forschungs- und Datenzyklus

Das Forschungsdaten-Management im Projekt erweist sich wahrscheinlich auch für eine spätere Nachnutzung als sehr sinnvoll, da Forschungsdaten, die schon während der Projektzeit gut gemanagt wurden, voraussichtlich mit weniger Aufwand in die Nachnutzung überführt werden können. Eine Reihe

von Forschungsdatenzentren bietet auch aus diesem Grund disziplinspezifische Forschungswerkzeuge an, die z.B. durch eine standardkonforme Datenerzeugung sowohl den Wissenschaftlern während des Projekts die Arbeit erleichtern, als auch den Aufwand für die Publikation und für das langfristige Datenmanagement für Forschungsdatenzentren reduzieren.

Ob sich damit der Gesamtaufwand für das Forschungsdaten-Management reduziert, wird sich zeigen müssen. Es ist aber davon auszugehen, dass das Forschungsdaten-Management im Projekt die Akzeptanz für das Datenmanagement verbessert, da die Wissenschaftler unmittelbar einen Nutzen haben. Hinter dieser Akzeptanzfrage verbirgt sich eine grundsätzlichere Schwierigkeit für das Wissenschaftssystem in Bezug auf Forschungsdaten: Die Wahrnehmung ist häufig, dass die Datenerzeugung die eigentlichen Kosten sind und die Datenmanagementkosten nur etwas Zusätzliches sind, auf das man auch verzichten könnte. Wenn man es aber für den gesamten Nutzungszeitraum der Daten betrachtet, der deutlich über das erzeugende Projekt hinausgehen kann, dann sind sie allerdings genauso ein Teil der Nutzungskosten, wie es die Erzeugungskosten sind. Datenmanagement im Projekt kann vielleicht den Aufwand für die Datenpublikation und langfristige Datenarchivierung zu einem gewissen Grad in das Projekt verschieben, das die Daten erzeugt, und damit hoffentlich eine realistischere Wahrnehmung und Kalkulation des Aufwands befördern.

Die Rolle von Infrastruktur im Forschungsdaten-Management

Für das Forschungsdaten-Management als Dokumentation sind die Institutionen verantwortlich, an denen die Wissenschaftler arbeiten. Dies ergibt sich relativ klar daraus, dass es sich um die Aufbewahrung der Daten für Verantwortungszwecke handelt, und die Institutionen diejenigen sind, die die Rahmenbedingungen für die Arbeit der Wissenschaftler organisieren. Es handelt sich um eine relativ klar definierte und zu einem gewissen Grad standardisierbare Aufgabe, für die die lokalen Infrastruktureinrichtungen oder -abteilungen die richtigen Akteure sind. Für die Nachnutzung wiederum wurde oben ausgeführt, dass es sich um eine komplexe, disziplinspezifische und höchstens teilweise standardisierbare Aufgabe handelt, die wahrscheinlich am besten von überregionalen oder internationalen, spezialisierten und in den Disziplinen verankerten Zentren übernommen wird. Wer ist aber der richtige Akteur, um das Forschungsdaten-Management in Projekten zu übernehmen?

Dies kann sich je nach Kontext unterschiedlich darstellen. Wenn es sich um relativ technik-affine Disziplinen oder um geringe Aufwände handelt, dann übernehmen die Wissenschaftler das gegebenenfalls selbst – einige Disziplinen haben auch verwandte Zweige der Informatik, die in diesen Fällen kompetente Ansprechpartner sein können. Die zentralen Forschungsdatenzentren, die die Publikationen von Forschungsdaten und ihre langfristige Pflege übernehmen, bringen sicherlich ebenfalls die inhaltliche Kompetenz mit, würden aber als zentrale Einrichtung vor der Herausforderung stehen, sehr viele verteilte Forschungsprojekte unterstützen zu müssen. Zusätzlich zu dem Skalierungsproblem wäre es schwierig, in die jeweiligen Projektgegebenheiten eingebunden zu sein und ihren oftmals lokalen Besonderheiten gerecht zu werden, sodass Forschungsdatenzentren wahrscheinlich nur in Fällen besonderer Bedeutung Projekte direkt beim Forschungsdaten-Management unterstützen können (wie z. B. die langfristig angelegten Umfrageprojekte in den Sozialwissenschaften) und sich eher auf die Entwicklung zentraler Werkzeuge und Standards im Sinne von Referenz-Sammlungen (vgl. National Science Board 2005) konzentrieren.

Wie ist aber die Situation, wenn das Datenmanagement in Projekten nicht effizient von den Wissenschaftlern selbst übernommen werden kann, in interdisziplinären Projekten oder wenn keine fachbezogene Informatik ausgeprägt wurde oder lokal vorhanden ist? Im Prinzip handelt es sich um komplexe und sehr projektspezifische Aufgaben, die aber nicht unbedingt vollständig disziplinspezifisch sein müssen. Einrichtungen wie Bibliotheken oder Rechenzentren können dafür die richtigen Institutionen sein und kompetentes Personal haben, wenn sie die Tätigkeit des Forschungsdaten-Managements nicht nur unter dem Vorzeichen der Publikation betrachten, sondern sich auf die Unterstützung des Forschungsprozesses einlassen. Sie können direkt im Projekt integriert sein als „embedded data managers“ und für die gesamte Institution als lokale Forschungsdaten-Support-Teams dienen. Wenn sie über die Projekte, in denen sich diese Investition lohnt, finanziert sind, dann kann diese Form des Forschungsdaten-Managements auch skalieren.

Es ist klar, dass diese lokalen Support-Teams nicht die Fachkompetenz zentraler Forschungsdatenzentren ersetzen können, so wie die Zentren nicht die Integration im Projekt vor Ort ersetzen können. Die Support-Teams aber können wissen, wann sie welche spezialisierten Zentren hinzuziehen müssen. Es bietet sich hier die Möglichkeit für ein sinnvolles Zusammenspiel von Infrastruktureinrichtungen. Und erst dieses Zusammenspiel wird das sein können, was wir eigentlich Forschungsdaten-Infrastruktur nennen sollten.

Tabelle 1:
Übersicht über verschiedene Formen des Forschungsdaten-Managements

	Dokumentation für Publikationen	Nachnutzung als Publikation	Datenmanagement im Projekt
Ziel	Nachvollziehbarkeit für Verantwortungszwecke	bessere und effizientere Forschung durch Nutzung bereits vorhandener Daten	bessere und effizientere Forschung durch besseres Datenmanagement in Projekten
Auswahlkriterium	Daten, die Grundlage einer Publikation sind, oder entsprechend Auflagen	fachwissenschaftliche Qualität der Daten und Kosten-Nutzen-Abschätzung	Relevanz der Projekte und Kosten-Nutzen-Abschätzung
Preismaßstab, Kostenfaktor	Volumen	Komplexität oder Anzahl Datensätze	Aufwand, Abschätzung anhand von Personenmonaten
Zielgruppe	Universität/Institution (weniger Wissenschaftler)	Disziplin, Forschungs-Community	Forschergruppe/ Forschungsprojekt
Dienstleister	Lokale Service-Infrastuktureinrichtungen. z.B. Rechenzentren und Bibliotheken	spezialisierte Datenzentren mit disziplinärer Verankerung und Kompetenz	lokale Zentren mit Support-Teams für Forschungsdaten-Management
Aufbewahrungsdauer	lang, aber begrenzt (zehn Jahre oder entsprechend der Auflagen)	lang und unbestimmt (potenziell unbegrenzt)	kurz und begrenzt (Projektlaufzeit)
Zugriffshäufigkeit	sehr selten	selten	häufig
Volumen	hoch	mittel	mittel
Aufgaben-Charakteristik	standardisierbar, teilweise automatisierbar	komplex, begrenzt automatisierbar	je Projekt individuell angepasste Maßnahmen, personalintensiv, kaum automatisierbar

Bisher wurde in diesem Text der Begriff Infrastruktur nicht weiter erläutert – aber anhand der Behauptung, dass erst das Zusammenspiel von lokalen und zentralen Einrichtungen die Forschungsdaten-Infrastruktur darstellt, sol-

len zwei Aspekte hervorgehoben werden: 1. Infrastruktur ist nicht nur Technik; 2. Infrastruktur besteht nicht nur aus einzelnen Komponenten.

- Zu häufig wird explizit oder implizit ein technischer Service oder eine Technologie als die Lösung einer Aufgabe im Bereich des Forschungsdaten-Managements betrachtet, anstelle die Organisationen und personellen Dienstleistungen als das Fundamentale anzusehen. “When dealing with infrastructures, we need to look to the whole array of organizational forms, practices, and institutions which accompany, make possible, and inflect the development of new technology [...] People, routines, forms, and classification systems are as integral to information handling as computers, Ethernet cables, and Web protocols” (Edwards et al. 2007: 3). Oder – um einen pointierten Artikeltitel der California Digital Library zu zitieren: “Preservation Is Not a Place” (Abrams et al. 2009). Ein Repositorium alleine löst keine Aufgaben, sondern das wird nur durch Prozesse, in denen Personen Technik benutzen, erreicht. Und wenn das Ziel eine Infrastruktur für die öffentliche und langfristige Verfügbarkeit von Forschungsdaten ist, dann ist dies nur durch nachhaltige Organisations- und Personalressourcen zu erreichen, die die Bereitstellung und Pflege von Forschungsdaten fortwährend betreiben, obwohl es zu Veränderungen in allen Bereichen wie z.B. der benutzten Technologie, dem Hintergrundwissen der Nutzergruppe und ihren Erwartungen kommt. Zu den größten Herausforderungen im Bereich des Forschungsdaten-Managements zählt sicherlich dieser Umgang mit Veränderungen und die Einplanung eines eigenen Lebenszyklus für Infrastrukturkomponenten, die entwickelt, genutzt, irgendwann nicht mehr die funktionellen oder ökonomischen Anforderungen erfüllen und erneuert, abgewickelt oder ersetzt werden müssen.
- Zudem ist Infrastruktur nicht nur eine einzelne Einrichtung, ein einzelner Service oder eine einzelne Entwicklung. Einzelne Komponenten sind immer auf ihren Kontext ausgerichtet und erst im Zusammenspiel mit anderen Komponenten werden sie für die verschiedensten Situationen und Kontexte nutzbar. “[T]rue infrastructures only begin to form when locally constructed, centrally controlled systems are linked into networks and internetworks governed by distributed control and coordination processes” (Edwards et al. 2007: 7). Für das Forschungsdaten-Management ist dafür eine zentrale Frage, wie die verschiedenen Institutionen und Disziplinen zusammenarbeiten und wie disziplinspezifisch und wie generisch sowohl die einzelnen Dienste als auch die Institutionen sein müssen

und können. Und selbst die Disziplinen sind unscharfe Konzepte, die beliebig ausdifferenziert werden können. Betrachten wir die Physik als eigene Disziplin? Wenn wir sie als zu groß ansehen, wäre die Astrophysik oder erst die Radioastronomie hinreichend eingegrenzt? Oder sind die einzelnen Großprojekte schon so spezifisch geworden, dass man sie als eigenständig betrachten muss? Gleichzeitig finden wir im Datenmanagement nicht nur disziplinspezifische Anforderungen, sondern auch viele methodenspezifische Aufgaben, die sich in einer Vielzahl von Disziplinen finden und für die generische Dienste Synergiepotenzial versprechen.

Es existieren bereits eine Reihe von einzelnen Infrastrukturkomponenten, die sich in einigen Disziplinen oder Teildisziplinen gut zusammenfügen, und auch althergebrachte Netzwerk wie das oben beschriebene WDC oder WDS sowie neue Initiativen wie die Research Data Alliance[1]. Es ist wahrscheinlich aber relativ unumstritten, dass es noch keine Forschungsdaten-Infrastruktur in dem Sinne gibt, wie man Infrastruktur gerne hätte: überall verfügbar, robust und transparent. Die hier beschriebenen Arten von Forschungsdaten-Management und die sich daraus ergebenden Anforderungen an Infrastruktureinrichtungen umreißen einen Rahmen, aber lassen noch viele Fragen unbeantwortet.

Mögliche Entwicklungen: Wie kann und wie sollte die Landschaft in zehn Jahren aussehen?

Die technische Entwicklung schreitet weiter schnell voran. 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. Anfang der 2000er-Jahre 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 (vgl. Klump 2008).

1 http://rd-alliance.org

Zeitgleich mit dem Aufkommen der Smartphones nach Vorbild des iPhone im Jahre 2007 und den neuen, damit verbundenen Online-Diensten kamen zwei neue Begriffe auf: Cloud Computing und App.

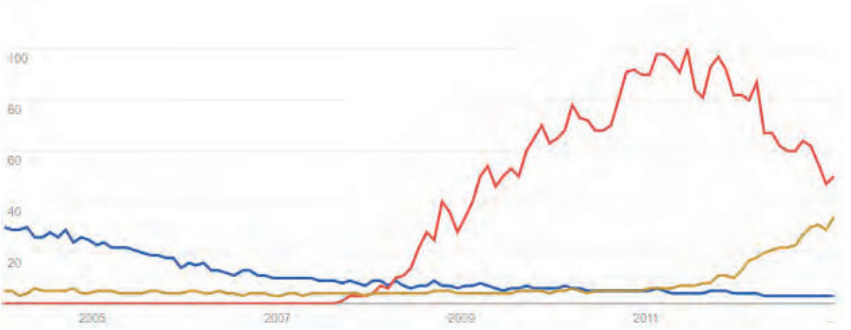

Abb. 2 Histogramm der Anfragen bei Google zu den Begriffen „Grid Computing“ (blau), „Cloud Computing“ (rot) und „Big Data“ (gelb) im Januar 2013.
Quelle: Google Trends

Mit dem Aufkommen des Begriffs „Cloud“ verschwand der Begriff „Grid“ praktisch in der Bedeutungslosigkeit (Abb. 2). Selbst der Begriff „Cloud“ ist unbedeutend verglichen mit dem Suchbegriff „App“, der sich als Synonym für Anwendungen (applications) auf mobilen Geräten etabliert hat (Abb. 3). Soziale Netzwerke spielten vor 2004 noch eine untergeordnete Rolle, heute hat Facebook rund eine Milliarde Nutzer.

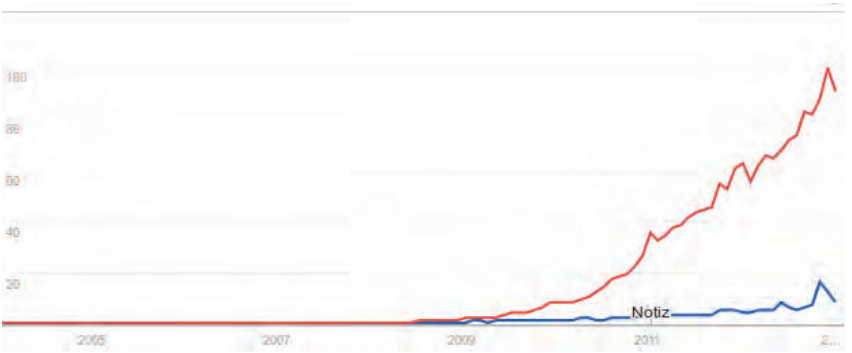

Abb. 3 Histogramm der Anfragen bei Google zu den Begriffen „Cloud Computing“ (blau) und „App“ (rot). Diese beiden Begriffe sind heute auch im Management von Forschungsdaten präsent, spielten aber noch vor 2010 fast keine Rolle.
Quelle: Google Trends

Was bedeutet das für Technologien und Dienste im Hinblick auf 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 Beitrag „e-Science and its implications" von Hey und Trefethen (2003), mit dem der Begriff der „Datenflut" (data deluge) in die Diskussion eingeführt wurde. Die 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 (vgl. McNally et al. 2012). Der Informatik-Pionier Jim Gray sprach auch von einem Paradigmenwechsel (vgl. Hey/Tansley/Tolle 2009). Aber ist dieser Paradigmenwechsel eingetreten?

Nimmt man die bewilligten Projektmittel als Näherungswert für die Größe eines Projekts und das zu erwartete Datenvolumens, so zeigt sich, dass auch die Verteilung von Forschungsprojekten mit einer „long-tail economy" (Anderson 2004) beschrieben werden kann. Nur relativ wenige Projekte erzeugen sehr große Datenmengen, die allermeisten Projekte erzeugen relativ kleine Datenmengen. Bei diesen Daten handelt es sich jedoch überwiegend um intellektuell vernetzte Objekte, die einen hohen Grad an semantischer und struktureller Komplexität besitzen und oft auch zwischen Projekten miteinander vernetzt sind (vgl. Heidorn 2008).

Eine weitere bedeutende Entwicklung ist die Entstehung sozialer Netzwerke und einer Vielzahl von Projekten mit Inhalten, die – teils anonym –

von Nutzern erstellt werden. Das herausragendste Beispiel von nutzergenerierten Inhalten ist die Online-Enzyklopädie Wikipedia. Mit dem Erscheinen dieser vielen Projekte wurde erwartet, dass dies auch Auswirkungen auf die Wissenschaft haben wird, insbesondere durch die „digital natives“, die eine Welt ohne Internet gar nicht kennen. Von ihnen wurde erwartet, dass sie die neuen Werkzeuge, wie zum Beispiel Wikis, für eine offenere wissenschaftliche Kommunikation nutzen. Der Bericht „Researchers of Tomorrow“, der von der British Library, JISC und HEFCE (2012) herausgegeben wurde, untersuchte durch Befragung von etwa 6000 graduierten Studenten, wie sich das Verhalten der „digital natives“ als Wissenschaftler von dem der vorangegangenen Generation ihrer wissenschaftlichen Betreuer unterscheidet. Zur Überraschung der Autoren stellte sich heraus, dass die „digital natives“ im Privatleben zwar die neuen Möglichkeiten des Internets und der Informationstechnologie intensiv nutzen, in ihrer wissenschaftlichen Praxis jedoch das Nutzerverhalten ihrer betreuenden Wissenschaftler kopierten, da sie diese als erfolgreiche Vorbilder ansehen. Es ist daher nicht zu erwarten, dass sich allein durch das Heranwachsen der „digital natives“ die Bereitschaft zur Zusammenarbeit im Netz, der Umgang mit Forschungsdaten, und die Bereitschaft, diese mit anderen zu teilen, rasch ändern – solange dies nicht als kulturell verankerte Wissenschaftspraxis vermittelt wird. Hinzu kommt, dass sich wissenschaftliche Arbeit hinsichtlich der Motivation der Autoren grundlegend von anderen „Open Initiatives“ unterscheidet (vgl. Klump 2012).

Herausforderungen an die Organisation

Die Anwendung der modernen Informations- und Kommunikationstechnik im Wissenschaftsbereich hat es deutlich einfacher gemacht, institutionsübergreifend zu kooperieren. Insgesamt nehmen kooperative Strukturen in der Forschung zu und der Umgang mit Forschungsdaten ist in den Bereichen bereits besser ausgebildet, in denen kooperative Strukturen vorherrschen (vgl. Neuroth et al. 2012). Der Zugewinn an Reputation durch vernetzte Projekte lässt Kooperation inzwischen oftmals als notwendige Arbeitsweise in der modernen Wissenschaft erscheinen. Je effizienter man einzelne Ergebnisse und technische Möglichkeiten ausnutzen will, desto flexibler und sachbezogener muss die Kooperation sein. Allerdings muss hier eine gewisse Spannung konstatiert werden: Der einzelne Wissenschaftler ist weiterhin an einer Organisation angestellt, die für ihn grundsätzliche Rahmenbedingungen be-

reitstellt, aber die Kooperation findet nicht mehr nur innerhalb einer Organisation statt.

Formen kooperativer Forschung stellen die Organisationen, in denen die daran teilnehmenden Wissenschaftler beheimatet sind, vor neue Herausforderungen. Es muss ein Rahmen geschaffen werden, der diese extremen Anforderungen an flexibler Kooperationen – geradezu volatiler, nur noch aufgabenbezogener Kooperation – erfüllen kann. Wünschenswert wäre, wenn Wissenschaftler sich allein mit ihrer speziellen Expertise in diese Projekte einbringen könnten, in einer Art Matrixorganisation. Dies umzusetzen ist jedoch auf der Ebene der Individuen schwierig, da Forschung nicht in Geschäftsprozessen denkt und auch die Struktur der Forschungsförderung und der Einwerbung von Mitarbeiterstellen durch Drittmittel diese Art von Organisation nicht vorsieht.

Die Anforderungen einer flexiblen „Matrixorgansiation“ stehen in einem gewissen Widerspruch zu den Erwartungen der Heimatorganisationen, die bei der Zuteilung von Ressourcen fragen, was Kooperationen ihnen für Nutzen bringen. Konflikte entstehen dann notwendigerweise, wenn diese Anforderungen sich schneller ändern als die Ziele der beheimatenden Organisation. So fördert der technische Fortschritt zwar globale Kooperation, aber nicht notwendigerweise auch lokale. Die Dynamik der Forschung erfordert zum Teil spontane Organisation, wie sie durch neue Technologien möglich geworden ist – diese steht aber oft im Widerspruch zu lokaler Organisation, in der zum Beispiel eine Abteilung den Nutzen eines Projekts für eine übergeordnete Einrichtung demonstrieren muss, der für die Projektpartner evident ist.

Herausforderungen an die Infrastruktur

Infrastruktureinrichtungen stehen im Bezug auf die rasche Entwicklung der Informationstechnik und ihrer Nutzungsmöglichkeiten vor einem Grunddilemma: Es ist ihre Aufgabe, stabile Dienste zu betreiben und diese stetig zu verbessern. Für diese Situation einer „sustaining innovation“ sind Infrastruktureinrichtungen im Allgemeinen gut vorbereitet. Auch für diese Dienste ist es notwendig, ihren Lebenszyklus zu planen und sie an technische Innovationen anzupassen. Immer wieder tauchen jedoch neue Technologien auf, deren Bedeutung für die Arbeitsweise der Forschung nicht von Anfang an klar zu erkennen ist. Aufgrund ihrer Zielsetzung, stabile Dienste anzubieten,

fällt es Infrastruktureinrichtungen schwer, sich auf neue Entwicklungen einzulassen, deren langfristiger Erfolg noch nicht abzuschätzen ist. Im schlimmsten Fall kann dies dazu führen, dass sich eine Institution auf eine fundamentale Neuerung nicht einlassen kann und letztlich scheitert. Dieses Innovationsmuster wird „disruptive innovation“ genannt (vgl. Christensen 2003). Insbesondere wissenschaftliche Bibliotheken werden als gefährdet angesehen (vgl. Lewis 2004).

Zudem sind auch Neuerungen in der Informationstechnologie oft kurzlebig. Viele der, zum Beispiel, von Google angebotenen Dienste haben eine durchschnittliche Lebensdauer von etwa drei Jahren (vgl. Arthur 2013), 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 ausbleibendem Erfolg wieder abgeschaltet werden, ohne den Betrieb der Plattform als Ganzes zu beeinträchtigen.

Eine erfolgreiche Strategie für Infrastruktureinrichtungen könnte daher 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 der Widerspruch zwischen dem Anspruch der Infrastruktur an Stabilität und der Anforderung nach flexiblen, möglicherweise kurzlebigen Anwendungen überbrücken.

Ein Maßstab, an dem neue Entwicklungen in der Informationstechnik gemessen werden können, ist, inwieweit sie die Prozesse der Forschung und Kommunikation unterstützen. Zu schnelle Änderungen in den technischen Rahmenbedingungen erschweren es potenziellen Nutzern, ihre Entwicklungen auf die neuen Dienste aufzubauen. Im Rahmen des D-Grid wurde, zum Beispiel, kritisiert, dass sich die Grid-Schnittstellen zu schnell änderten, um Anwendungen zu entwickeln, die sich in die Arbeitsabläufe der Forschung einpassen. Lieber hätte man mit älteren Versionen der Grid-Middleware gearbeitet, als die eigenen Anwendungen ständig auf neue Versionen der Grid-Middleware anpassen zu müssen.

Für Infrastruktureinrichtungen der akademischen Informationsversorgung ist die Einschätzung neuer Technologien schwer, da man sich dort nicht in der Lage sieht, technologische „Versuchsballons“ zu starten und gegebenen-

falls auch nach kurzer Zeit wieder aufzugeben. Zudem wurden Informationsinfrastrukturen meist als monolithische, in sich geschlossene Anwendungen gebaut. Um eine bessere Anpassung der Dienste an die Bedürfnisse der Forschung zu erreichen, wurden meist Nutzerbeiräte eingerichtet, um einen direkteren Austausch mit den Nutzern zu erreichen. Doch auch etablierte Nutzergruppen sind vor den Herausforderungen einer „disruptive innovation" nicht gefeit (vgl. Christensen 2003).

Literaturverzeichnis

Abrams, S.; P. Cruse; J. Kunze (2009): Preservation Is Not a Place. In: *International Journal of Digital Curation*, 4 (1): 8–21, doi:10.2218/ijdc.v4i1.72.

Anderson, C. (2004): The Long Tail, In: *Wired*. Online: http://www.wired.com/wired/archive/12.10/tail_pr.html.

Arthur, C. (2013): Google Keep? It'll probably be with us until March 2017 – on average. In: *The Guardian*, 22. März 2013. Online: http://www.guardian.co.uk/technology/2013/mar/22/google-keep-services-closed.

Berlin Declaration (2003): Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/.

Borgman, C. L. (2012): The Conundrum of Sharing Research Data. In: *Journal of the American Society for Information Science and Technology*, 63 (6): 1059–1078, doi:10.1002/asi.22634.

British Library; HEFC; JISC (2012): Researchers of Tomorrow – The research behaviour of Generation Y doctoral students. London: JISC. Online: http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow.

Christensen, C. M. (2003): *The Innovator's Dilemma*. New York, NY: HarperCollins Publishers.

DFG (1998): *Sicherung guter wissenschaftlicher Praxis*. Bonn: Deutsche Forschungsgemeinschaft.

DFG (2010): Merkblatt für Anträge auf Sachbeihilfen mit Leitfaden für die Antragstellung und ergänzenden Leitfäden für die Antragstellung für Projekte mit Verwertungspotenzial, für die Antragstellung für Projekte im Rahmen einer Kooperation mit Entwicklungsländern, Deutsche Forschungsgemeinschaft (DFG), Bonn. http://www.dfg.de/download/formulare/1_02/1_02.pdf.

Dittert, N.; M. Diepenbroek; H. Grobe (2001): Scientific data must be made available to all. In: *Nature*, 414 (6862): 393, doi:10.1038/35106716.

Edwards, P.; S. Jackson; G. Bowker; C. Knobel (2007): Understanding Infrastructure: Dynamics, Tensions, and Design. http://hdl.handle.net/2027.42/49353.

Heidorn, P. B. (2008): Shedding Light on the Dark Data in the Long Tail of Science. In: *Lib. Trends*, 57 (2), 280–299, doi:10.1353/lib.0.0036.

Hey, T.; S. Tansley; K. Tolle (Hrsg.) (2009): *The Fourth Paradigm: Data-Intensive Scientific Discovery* (v 1.1). Redmond, WA: Microsoft Research. Online: http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

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

Holdren, J. P. (2013): Increasing Access to the Results of Federally Funded Scientific Research. The White House, Office of Science and Technology Policy. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

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

Klump, J. (2012): Offener Zugang zu Forschungsdaten: Open Data und Open Access to Data – Die ungleichen Geschwister. In: U. Herb (Hrsg.): *Open Initiatives: Offenheit in der digitalen Welt und Wissenschaft*. Saarbrücken: Universaar, S. 45–53. Online: http://nbn-resolving.de/urn:nbn:de:bsz:291-universaar-873.

Klump, J.; R. Bertelmann; J. Brase; M. Diepenbroek; H. Grobe; H. Höck; M. Lautenschlager; U. Schindler; I. Sens; J. Wächter (2006): Data publication in the Open Access Initiative. In: *Data Science Journal*, 5, 79–83, doi:10.2481/dsj.5.79.

Lewis, D. W. (2004): The Innovator's Dilemma: Disruptive Change and Academic Libraries. In: *Library Administration & Management*, 18 (2), 68–74. Online: http://hdl.handle.net/1805/173.

National Science Board (2005): Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Technical report, National Science Foundation, Washington, DC. http://www.nsf.gov/pubs/2005/nsb0540/.

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

Neuroth, H.; S. Strathmann; A. Oßwald; R. Scheffel; J. Klump; J. Ludwig (Hrsg.) (2012): *Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme*.

Boizenburg: Hülsbusch. Online: http://nestor.sub.uni-goettingen.de/bestandsaufnahme.

OECD (2006): Recommendation of the Council concerning Access to Research Data from Public Funding. Organisation for Economic Co-operation and Development, Paris.

Pfeiffenberger, H. (2007): Offener Zugang zu wissenschaftlichen Primärdaten. In: *Zeitschrift für Bibliothekswesen und Bibliographie*, 54 (4–5): 207–210. Online: http://hdl.handle.net/10013/epic.28454.d001.