Sag mir, wo die Daten sind…

Forschungsdaten sind der unverzichtbare und kostbare Rohstoff der Wissenschaft. Ihn zu sichern und weiter nutzbar zu halten, ist eine große Herausforderung.

Alexander Pawlak

Wir schreiben das 22. Jahrhundert: Auf dem Planeten Losannien ist die Forschung in einer Sackgasse gelandet. Zwar ist schon so gut wie alles Erdenkliche untersucht worden, aber keiner weiß, wo sich bestimmte Forschungsergebnisse befinden. Um dem Unwissen über das Wissen zu begegnen, rufen die Losannier eine eigene Wissenschaft namens Ignorantik ins Leben und durchforsten mit einem Spürcomputer die Datenspeicher des Planeten nach verschollenen Wissensschätzen. Doch im stetig wachsenden Netzwerk dauert die Suche bereits bis zu 16 Jahre, Tendenz steigend.

Dieses fiktive Dilemma hat der polnische Schriftsteller Stanislaw Lem bereits 1982 in seinem Science Fiction-Roman „Lokaltermin" ersonnen und legte damit den Finger in die richtige Wunde. Wie überall ist heutzutage auch in der Wissenschaft eine steigende Datenflut zu beobachten. So erzeugen allein die Detektoren des Large Hadron Colliders am CERN jährlich rund 15 Petabyte Daten. Dafür gibt es nicht nur ein eigenes Rechenzentrum vor Ort, sondern gleich ein global verteiltes Netzwerk: Für das „Worldwide LHC Computing Grid" stellen 170 Rechenzentren aus 34 Ländern über 100 000 Prozessoren zur Verfügung, um die Daten zu verarbeiten und für die Community verfügbar zu machen.

Doch abseits von der Großforschung, im Labor herkömmlicher Größe, sieht der Datenkreislauf meist wenig nachhaltig aus: Forscher gewinnen im Experiment Messdaten, werten sie aus, vergleichen sie mit anderen Beobachtungen oder Simulationen und veröffentlichen schließlich ihre Ergebnisse in den einschlägigen Fachzeitschriften. Dann müssen neue Daten her, die alten haben schließlich ihre Schuldigkeit getan, oder? Nein, sagte schon die Deutsche Forschungsgemeinschaft 1998 in ihrer Denkschrift zur „Sicherung guter wissenschaftlicher Praxis". Darin empfiehlt sie: „Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden." Ein Auslöser für die Empfehlung war ein schwerer Fall von Forschungsbetrug: Die Krebsforscher Friedhelm Herrmann und Marion Brach von der Universität Ulm hatten über viele Jahre hinweg ihre Daten manipuliert und fremde Ergebnisse kopiert und in den eigenen Publikationen verwendet. Ein verantwortungsvoller Umgang mit den Primärdaten ist daher ein entscheidender Schritt, um gute wissenschaftliche Praxis zu sichern und eventuelle Vorwürfe überprüfen zu können. Wie berechtigt die DFG-Empfehlung ist, rief der Fall des Physikers Jan Hendrik Schön im Jahr 2002 wieder eindrücklich in Erinnerung. Originaldaten konnte er nicht vorweisen.

Claudia Marcelloni, Maximilien Brice / CERN

Die Forschung – unendliche Datenmengen: ein Blick in die Magnetspeicher-Bibliothek des CERN für die Daten des Large Hadron Colliders.

Gespeicherte Forschungsdaten können aber noch viel mehr sein als nur ein Ausweis für wissenschaftliche Integrität, ist Stefan Winkler-Nees, Programmdirektor der DFG-Gruppe „Wissenschaftliche Literaturversorgungs- und Informationssysteme", überzeugt. „Die Denkschrift nimmt zur Aufbewahrung der Daten im Sinne einer Beweissicherung Stellung", erläutert er: „Das ist jedoch nur ein sehr kleiner Aspekt im komplexen Bereich Forschungsdaten." Was etwa, wenn sich aufgrund neuer Fragen oder Analysemethoden aus alten Daten neue Erkenntnisse herauskitzeln lassen, möglicherweise auch durch andere Forscher? Für eine solche „Nachnutzung" genügt ein Aufbewahren im Sinne der DFG-Richtlinien nicht. „Ein Wissenschaftler hat der Empfehlung ja Genüge getan, wenn er seine Rohdaten auf einer Floppy-Disk abspeichert und für zehn Jahre in irgendeine Schublade legt", erläutert

  1617-9439/15/1111-24

Winkler-Nees. Doch damit sind die Daten für niemand anderen nachnutzbar, ja es ist noch nicht einmal garantiert, dass die Daten lesbar bleiben. Formate, Software und Computer unterliegen schließlich einem ständigen Wandel. „Digitale Daten halten eine Ewigkeit oder fünf Jahre. Je nachdem, was zuerst kommt", hatte es der Computerwissenschaftler Jeff Rothenberg von der RAND-Corporation 1995 auf den Punkt gebracht.[1)]

Stefan Winkler-Nees geht es zunächst weniger um die technischen Fragen, sondern um den nachhaltigen Umgang mit Forschungsdaten: „Dafür ein Bewusstsein zu schaffen, ist eine große Herausforderung. Daten zu teilen kann ein Vorteil sein – nicht nur für die Wissenschaft insgesamt, sondern auch für Forscher selbst." Ein gutes Beispiel dafür liefert der Blick in die Sterne.

Der Himmel auf der Festplatte

Die Astronomen haben eine gewisse Vorreiterrolle übernommen, wenn es darum geht, bei den Forschungsdaten von einer Ex-und-Hopp-Mentalität zu einer nachhaltigen Nutzung zu kommen. Der Heidelberger Astronom Joachim Wambsganß gibt dafür drei entscheidende Gründe an: „Bei astronomischen Daten handelt es sich erstens um sehr große Datenmengen, die zweitens keinen kommerziellen Wert haben und drittens nicht personenbezogen sind, anders als etwa in der Medizin oder den Sozialwissenschaften."

Die Astronomie-Community hat früh den bleibenden Wert ihrer Forschungsdaten und der Bedeutung einer internationalen Standardisierung erkannt und „baut" seit 2001 ein „virtuelles Observatorium" auf, d. h. eine digitale Infrastruktur, die Beobachtungsdaten vieler realer Observatorien unter einer einheitlichen Benutzeroberfläche online zur Verfügung stellt. Beobachtungen zu verschiedenen Zeiten und in unterschiedlichen Wellenlängenbereichen lassen sich so einfacher vergleichen und zusammenführen. 2002 wurde die „International Virtual Observatory Alliance" (IVOA) als Dachorganisation gegründet, um die internationale Zusammenarbeit zu koordinieren.[2)] Wambsganß, zu dessen Spezialgebieten die Erforschung von Quasaren und die Suche nach Exoplaneten gehören, ist verantwortlich für das „German Astrophysical Virtual Observatory" (GAVO). „Ein eigenes deutsches virtuelles Observatorium klingt zunächst absurd, aber aus förderpolitischen Gründen ist es sinnvoll, eine nationale Struktur zu etablieren", erläutert er.[3)]

Die gesammelten Daten ermöglichen immer wieder neue und überraschende Entdeckungen. Das allererste Paper, das allein auf „Beobachtungen" mit dem virtuellen Observatorium beruhte, erschien 2004. Ein europäisches Astronomen-Team konnte die Entdeckung von 30 supermassereichen Schwarzen Löchern in Kernen von aktiven Galaxien vermelden.

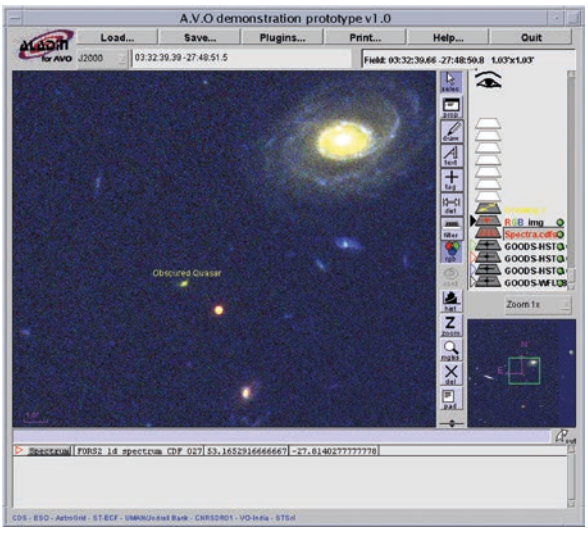
Ein weiteres gutes Beispiel für die Nachnutzung von Daten bietet das Karlsruher Luftschauerexperiment KASCADE, das von 1993 bis 2013 die Eigenschaften der kosmischen Strahlung untersuchte. Während seiner 20-jährigen Laufzeit nahm es mehr als 1,75 Milliarden Ereignisse auf, von denen etwa 425 Millionen die Qualitätsprüfung bestanden. Davon stehen derzeit etwa 160 Millionen online im „KASCADE Cosmic-ray Data Centre" zur Verfügung. Das Angebot richtet sich dabei nicht nur an die betreffenden Experten, sondern auch an eine breitere Öffentlichkeit. Für interessierte Studierende gibt es beispielsweise aufbereitete und detailliert erläuterte Lehrbeispiele. Auch das CERN macht seit November 2014 ausgewählte Datensätze für Lehrzwecke im Web zugänglich.

Mittlerweile haben sich disziplinübergreifende „Daten-Repositorien" im Web etabliert. Für Geowissenschaftler gibt es etwa PANGEA, wo Daten aus der Erdsystemforschung und den Umweltwissenschaften, in der Regel mit Zeitangaben und geografischen Koordinaten versehen, gesammelt und verfügbar gemacht werden.

Nutzen und Eigennutz

KASCADE, CERN und PANGEA berufen sich auf die „Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen", die 19 nationale und internationale Forschungsorganisationen 2003 beschlossen und die seitdem rund 500 wissenschaftliche Institutionen unterzeichnet haben. Eins der zentralen Ziele der Erklärung ist es, „auch die neuen Möglichkeiten der Wissensverbreitung über das Internet nach dem Prinzip des offenen Zugangs" zu fördern.

Sehr aufschlussreich ist eine Umfrage unter knapp 1600 Wissenschaftlerinnen und Wissenschaftlern. Über 73 Prozent von ihnen sind grundsätzlich überzeugt, dass öffentlich zugängliche Forschungsdaten zum Fortschritt der Wissenschaft beitragen. Wenn es aber darum geht, mit wem die Befragten ihre Daten teilen, ergibt sich ein deutlich anderes Bild. In den Naturwissenschaften teilen gerade einmal 17 Prozent ihre Daten öffentlich. Sicher schwingen dabei naheliegende Vorbehalte mit, etwa die Angst, dass Kollegen Dinge herausfinden könnten, die man selbst übersehen hat. Auch schreckt sicherlich der damit verbundene Aufwand ab, der dem jeweiligen Forscher selbst nicht zugutekommt. Anders wäre es vermutlich, wenn nicht nur die Veröffentlichung von Forschungsergebnissen, sondern



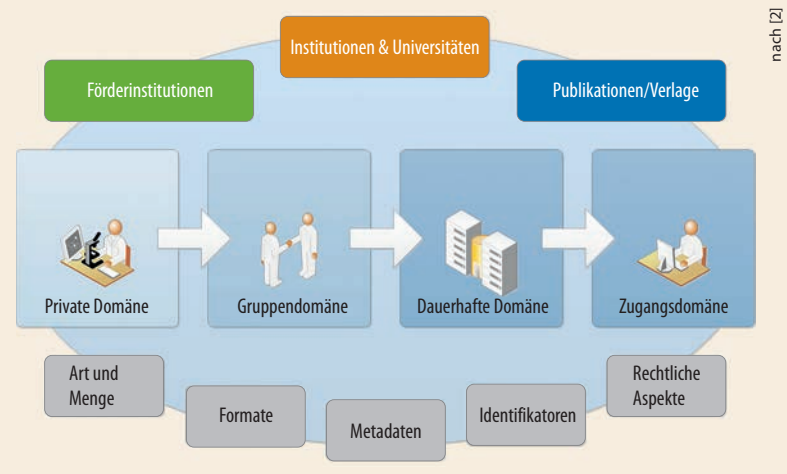
ESA/NASA, AVO Project, Paolo Padovani

Astronomen schauen längst nicht mehr selbst ins Teleskop, sondern können auch mit einem virtuellen Observatorium Entdeckungen machen.

1) Zum Thema Langzeitarchivierung von Forschungsdaten vgl. [1].

2) www.ivoa.net

3) In Deutschland geschieht das im Rahmen der projektorientierten Verbundforschung, mit der das BMBF beispielsweise die Entwicklung von astronomischen Instrumenten für Großteleskope fördert.



Das „Domänenmodell" etabliert eine gewisse Hierarchie des Forschungsdatenmanagements: Die Domänen unterscheiden sich in der Art des Datenaustauschs, den Kreis der Austauschpartner und in der Art der Nutzung. Weitere Unterschiede sind die Zuständigkeitsbereiche, in welche die Domänen fallen (oben) und welche Aspekte der Daten es beim Übergang zur nächsten Domäne abzustimmen gilt (unten).

auch die zugrundeliegenden Daten zur Reputation des Wissenschaftlers beitragen würden.

Wie erfolgreich eine solche Daten-Veröffentlichung sein kann, zeigt der Sloan Digital Sky Survey (SDSS), eine aufwändige Durchmusterung von ungefähr einem Viertel des Himmels. Der SDSS war ursprünglich ein privates Unternehmen der Sloan Foundation, bei dem die beteiligten Forscher alles allein machen und auswerten wollten. Als die Kosten aus dem Ruder liefen, kam die US National Science Foundation ins Spiel. Der öffentliche Geldgeber wollte das Projekt aber nur unterstützen, wenn die Daten nach einer gewissen Zeit öffentlich gemacht werden. Nach vielen Diskussionen einigte man sich darauf, die Daten nach einem Jahr „Karenzzeit", in der nur die Projektwissenschaftler damit arbeiten konnten, zu veröffentlichen. Ähnlich ist das auch bei den Weltraumteleskopen wie Hubble und Chandra üblich. „In so einem Survey steckt so viel drin, dass es unmöglich ist, alles in einem Wissenschaftlerleben auszuwerten", betont Wambsganß. Bei SDSS wurden die Daten nach und nach in neun Tranchen freigegeben. Dazu erschienen „Data Release Papers", die mittlerweile zu den meistzitierten Astronomie-Publikationen gehören und den Mitgliedern der Arbeitsgruppe entsprechend viel Reputation einbrachten.[4)]

Ein weites Datenfeld

Im Laufe der letzten Jahre sind viele Initiativen entstanden, die den Umgang mit Forschungsdaten auf unterschiedlichen Ebenen adressieren (**Infokasten**). Ende 2013 rief die Universität Bielefeld als erste deutsche Hochschule ihre Wissenschaftlerinnen und Wissenschaftler in einer Resolution des Rektorats dazu auf, Forschungsdaten besser auffindbar und möglichst nachnutzbar zu machen. Eine eigene Kontaktstelle informiert über Best Practices, aktuelle Förderrichtlinien und leistet Unterstützung bei der Erstellung von Datenmanagementplänen für zukünftige Forschungsvorhaben. Weitere deutsche Hochschulen sind diesem Beispiel gefolgt.

Auch in der Forschungsförderung findet dieses Thema Berücksichtigung. Bereits seit 2007 bietet die DFG im Rahmen ihrer Sonderforschungsbereiche die Möglichkeit, ein „Teilprojekt Informationsinfrastruktur" zu beantragen, z. B. zur Konzeption des Datenmanagements, Bereitstellung der erforderlichen Infrastruktur oder der Kooperation mit der Bibliothek oder dem Rechenzentrum vor Ort. In der Physik sind wieder astronomisch bzw. astrophysikalisch ausgerichtete Sonderforschungsbereiche Vorreiter.

Einen guten Eindruck von der komplexen Gemengelage des Forschungsdatenmanagements, bietet ein so genanntes Domänenmodell (**Grafik**): Zu Anfang steht der Wissenschaftler, der Forschungsdaten in der „Privaten Domäne" erzeugt und analysiert. Um diese in einem ausgewählten Kollegenkreis innerhalb und außerhalb seiner Institution zu diskutieren, muss er seine Forschungsdaten – meist in bereits bearbeiteter Form – über geeignete Systeme eingeschränkt zugreifbar machen (Gruppendomäne). Mit der Veröffentlichung der Daten gehen diese in die „Dauerhafte Domäne" über, die für die Archivierung und langfristige Erhaltung sorgt. Dabei ist die „curation boundary" zu überwinden. Dafür gilt es, die Daten geeignet zu erschließen, was teuer und arbeitsaufwändig sein kann. „Die wichtige Frage, die sich hier stellt, lautet: Was wollen wir aufheben und wie lange?", sagt Stefan Winkler-Nees: „Die Daten eines physikalischen Experiments, das sich problemlos überall wiederholen lässt, muss man sicher nicht für alle Ewigkeit konservieren. Anders ist das mit meteorologischen Messreihen, die sich schlecht wiederholen lassen." Hier kommt auch die Beschreibung der Primärdaten ins Spiel, die Metadaten. In der Astronomie bestehen diese etwa aus Ort und Zeitpunkt der Beobachtung, Wellenlängenbereich oder Informationen zu verwendeten Filtern und Zusatzgeräten. Allgemein sind Angaben über Formate, verwendete Software oder über Datenlizenzen nötig. „Diese Metadaten sind unabdingbar, denn sonst bleiben die Daten nur eine Ansammlung von Nullen und Einsen", erläutert Joachim Wambsganß.

Die vierte Domäne schließlich erlaubt den Zugang zu den archivierten Daten, z. B. über Fachportale oder virtuelle Forschungsumgebungen oder eigene „Data Journals". Hier haben sich mittlerweile auch wissenschaftliche Fachverlage positioniert. Mit dem einfachen „Ins-Web-Stellen" ist es bei solchen Datenveröffentlichungen nicht getan. Hier besteht die Herausforderung darin, Qualitätsstandards für Forschungsdaten zu definieren und ein entsprechendes Gutachterwe-

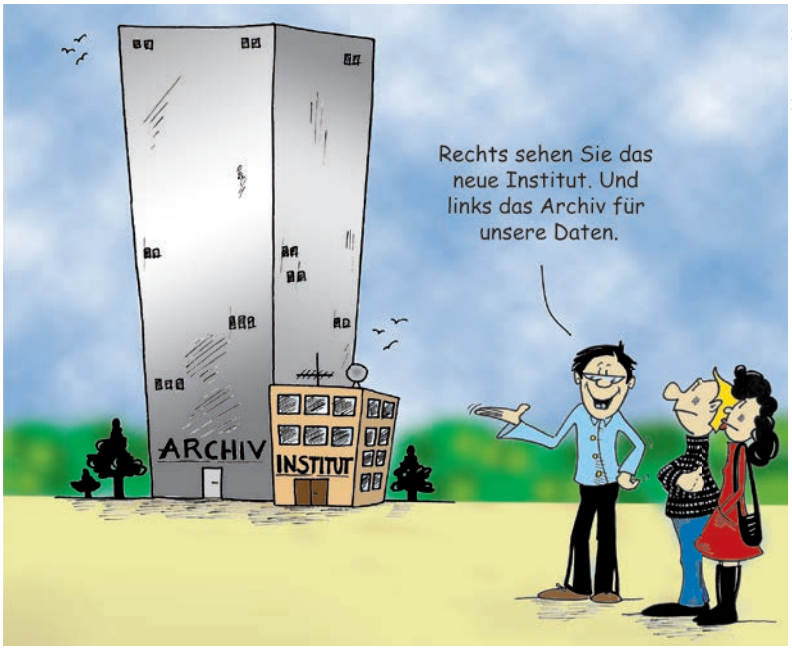
4) Mittlerweile stammt mehr als die Hälfte der Veröffentlichungen, die auf SDSS-Daten beruhen, von Astronomen, die nicht am Survey beteiligt waren.

sen („Peer Review“) zu etablieren, inklusive der dafür notwendigen Infrastruktur.

Die Vielfalt der Daten

Größere Forschungsprojekte wie Sonderforschungsbereiche oder besonders gut international vernetzte Communities wie in der Teilchenphysik und Astronomie haben es leichter, die „Domänenwände“ des Forschungsdatenmanagements zu überwinden. Doch was ist mit kleinen Forschungsgebieten oder Datenmengen, die noch bequem auf eine Festplatte passen? „Hier fehlt es meist noch am entsprechenden Problembewusstsein. Bereits die reine Archivierung gehören für Wissenschaftler nicht zum Kerngeschäft“, sagt Matthias Hahn vom Team des Projekts RADAR (Research Data Repositorium) am FIZ Karlsruhe, dem Leibniz-Institut für Informationsinfrastruktur [3].[5)]

Ziel ist es, einen fächerübergreifenden Service zu entwickeln, nicht zuletzt weil einzelne Institutionen allein mit Aufbau, Bereitstellung und dem dauerhaften Betrieb einer solchen Infrastruktur überfordert wären. Das Projekt versteht sich als komplementär zu bestehenden fachspezifischen Datenzentren. Zielgruppe sind Forscher, Einrichtungen und Disziplinen, denen entsprechende Infrastrukturen bisher fehlen. RADAR ist grundsätzlich als zweistufige Dienstleistung konzipiert. Die erste Stufe ist eine garantierte Datenarchivierung für 10 Jahre, was den Empfehlungen der DFG entspricht. Die zweite Stufe ist eine Archivierung ohne zeitliches Limit, die auch eine Option für eine Publikation der Daten bietet. Damit ist die Vergabe eines „Digital Object Identifiers“ (DOI) verbunden, mit dem der jeweilige Datensatz verfügbar, zitierfähig und verlinkbar ist. „Derjenige, der die Daten gibt, muss für die Nachnutzung auch eine Qualitätssicherung durchführen und vor allem die Daten ausreichend beschreiben“, betont Matthias Hahn. Die erste Projektphase von RADAR läuft bis Ende 2016. Wissenschaftliche Partner sind das Department für Chemie der LMU München und das Leibniz-Institut für Pflanzenbiochemie in Halle.



Wiebke Drenckhan

Auf dem Weg zu einem nachhaltigen Forschungsdatenmanagement sind noch viele Fragen zu beantworten, nicht nur fachinterne oder technische, sondern auch rechtliche, finanzielle und politische. Diese Problematik spricht auch die „Digitale Agenda 2014 – 2017“ der Bundesregierung an, welche die Vernetzung von Forschungsdatenbanken und Repositorien sowie virtuelle Forschungsumgebungen fördern und durch strategische Projekte unterstützen will. „In Bezug auf den Umgang mit Forschungsdaten ist ein Prozess im Gange, der in die richtige Richtung läuft, aber das noch sehr langsam. Vonseiten der DFG möchten wir dafür wichtige Anstöße geben“, bilanziert Stefan Winkler-Nees. Ob und wenn ja, auf welche Weise man Daten nachnutzen möchte, müsse jede Community letztlich für sich entscheiden. „In allen Disziplinen kann man von den bisherigen Ansätzen profitieren“, ist er überzeugt.

Literatur

[1] *H. Neuroth* et al., Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme (2012), http://bit.ly/1OE94nR
[2] *J. Klar* und *H. Enke*, Projekt RADIESCHEN: Report „Organisation und Struktur“ (2013), http://bit.ly/1kjkna3
[3] *M. Razum, J. Neumann* und *M. Hahn*, ZfBB **61**, 18 (2014), http://bit.ly/1Ox341C

LINKS ZU FORSCHUNGSDATEN

- **www.forschungsdaten.org**
Dieses Wiki sammelt Informationen rund um den Umgang mit digitalen Forschungsdaten und zu betreffenden Projekten und Initiativen.
- **www.langzeitarchivierung.de**
Deutsches Kompetenznetzwerk zur digitalen Langzeitarchivierung (nestor)
- **www.re3data.org**
Das Registry of Research Data Repositories (re3data) ist ein globales Verzeichnis von Forschungsdaten-Repositorien aus verschiedenen akademischen Disziplinen.
- **www.radar-projekt.org**
Das Ziel des Projekts RADAR ist der Aufbau einer fachübergreifender Forschungsdateninfrastruktur.
- **http://data.uni-bielefeld.de**
Auf dieser Seite finden sich Informationen rund um das Forschungsdatenmanagement wie Archivierung, Publikation oder Zitation von Forschungsdaten.
- **www.allianzinitiative.de**
Die Schwerpunktinitiative „Digitale Information“ wird von der Allianz der deutschen Wissenschaftsorganisationen getragen. Die Arbeitsgruppe Forschungsdaten veröffentlichte Anfang 2015 das Positionspapier „Research Data at Your Fingertips“, zu finden auf *http://bit.ly/1OC8HbQ*.
- **www.digitale-agenda.de**
Dei Digitale Agenda der Bundesregierung. Anmerkungen der AG Information der DPG: *http://bit.ly/1L9il4W*

5) Matthias Hahn ist seit Mitte des Jahres freiberuflich tätig.