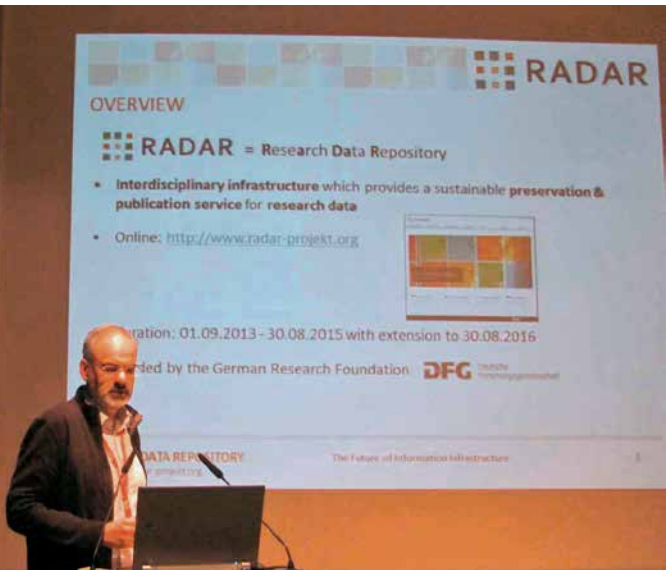
FIZ Karlsruhe setzt auf Forschungsdatenverwaltung

Marlies Ockenfeld

Auf dem Weg vom traditionellen deutschen Fachinformationszentrum zur international agierenden Informationsinfrastruktureinrichtung setzt das FIZ Karlsruhe seine Duftmarken, kürzlich mit einer von Matthias Razum geleiteten eintägigen Konferenz zur Zukunft der Informationsinfrastruktur am 15. September 2014 im ZKM | Zentrum für Kunst und Medientechnologie in Karlsruhe mit etwa hundert Teilnehmerinnen und Teilnehmern.



Dr. Matthias Hahn, Projektmanager im Bereich Langzeitarchivierung und Forschungsdatenmanagement bei FIZ Karlsruhe stellt das Projekt RADAR vor.

Als Infrastruktureinrichtung pflegt FIZ Karlsruhe sowohl Kontakt zu Bibliotheken und Forschungseinrichtungen wie auch zu Systemhäusern und Juristen. So war die Idee entstanden, gemeinsam einen Rundumblick auf das aktuelle Thema Forschungsdaten zu werfen. Dabei bilden Forschungsdaten ein weites Feld und selbstverständlich gibt es je nach Disziplin sehr unterschiedliche Herausforderungen und Regelungen für den Umgang mit den Daten, ihrer Speicherung, Verarbeitung und Nachnutzung. Und da es in aktuellen Veranstaltungen zumeist um digitale Daten geht, stellt sich zudem noch die Frage, ob und wie sichergestellt werden kann und soll, dass die einmal erhobenen, gemessenen oder berechneten Daten langfristig gespeichert und bei Bedarf – so es ihn tatsächlich geben sollte – trotz weiter entwickelter Speicher- und Verarbeitungstechnologien erneut genutzt werden können.

Konkret ging es dem Titel nach also um **Forschungsdatenmanagement: Organisatorische, technische und rechtliche Herausforderungen**. Veranstaltungssprache war wegen internationalen Besetzung Englisch, lediglich im ersten Block, in dem es um juristische Fragen ging, wurde wegen der spezifischen Rechtsfiguren, für die die englische Sprache keine eindeutige Übersetzung bietet, deutsch gesprochen.

Legal matter actually matter

Daniel Mietchen (Museum für Naturkunde Berlin), Andreas Wiebe (Universität Göttingen) und Paul C. Johannes (Universität Kassel) bestritten den ersten Block und erörterten juristische Fragestellungen, mit denen man etwa in Arbeitsverträgen, Autorenverträgen oder Lizenzvereinbarungen konfrontiert ist. Problematisch ist es, wenn Projekte nach Auslaufen der Förderung beendet werden oder wenn Start-ups in die Insolvenz gehen und niemand mehr zur Datenpflege oder zur Klärung von Lizenzbestimmungen zur Verfügung steht. Notwendig ist deshalb eine vollständige Dokumentation offener Daten und die saubere Anwendung üblicher Formate und Normen für ihre Speicherung, damit sie tatsächlich uneingeschränkt langfristig nachgenutzt werden können. Daneben sind allgemein akzeptierte Zitierregeln erforderlich. Bei der Datenpflege wären Empfehlungssysteme wünschenswert, die nachweisen, wer die Daten genutzt hat, etwa Journalisten, Bürgerwissenschaftler oder Roboter.

Uneinigkeit herrscht darüber, welches nationale Recht wann im weltweiten Internet anzuwenden ist. Sogenannte „Rohdaten“ sind juristisch gesehen nicht geschützt. Schutz genießen in Europa nur Datenbanken, in denen die Daten systematisch oder methodisch angeordnet sind, wobei dies nicht mit Kreativität oder Individualität, sondern in der getätigten Investition begründet ist. Das Datenbankherstellerrecht bietet Schutz gegen Entnahme der Daten, selbst wenn die Daten als solche frei sind. Während Referate (Abstracts) urheberrechtlichen Schutz gegen die Übernahme in ähnlicher Form genießen, gilt dies – eigentümlicherweise – nicht für Metadaten. Allerdings ist das Harvesting von Metadaten ein Rechtsverstoß.

Auf dünnem Eis bewegt man sich ggf. auch, wenn Forschungsdaten einen Personenbezug aufweisen, wenn also etwa die Metadaten Angaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person machen. Es gibt kein übergeordnetes Forschungsdatenschutzgesetz, sondern es gelten das Bundesdatenschutzgesetz (BDSG) sowie für die Hochschulen die jeweiligen Landesdatenschutzgesetze und darüber hi-

naus spezielle Gesetzgebungen, wie etwas das Krebsregistergesetz, das Statistikgesetz oder das Landesarchivgesetz. Für eine rechtmäßige Datennutzung ist regelmäßig eine Erlaubnis, entweder in Form einer Einwilligung oder einer Ermächtigung, notwendig. Um diese zu umgehen hilft eine Pseudonymisierung oder Anonymisierung, die allerdings bei der Anwendung von Big Data aufgehoben werden kann. Die Anlage 9 des BDSG macht Gestaltungsvorschläge für Forschungsdatenverwaltungssysteme; dazu gehören Zugangskontrolle, Zugriffskontrolle, Weitergabekontrolle, Eingabekontrolle, Verfügbarkeitskontrolle. Zur Beweissicherung und Beweiswerterhaltung und für das Einwilligungsmanagement ist eine logische oder physische Datentrennung erforderlich. Die Balance zwischen Wissenschaftsfreiheit auf der einen und persönlichem Datenschutz auf der anderen Seite zu finden bleibt eine Herausforderung.

Veränderte Einstellungen und Arbeitsabläufe notwendig

Tim Hasler (Zuse Institut Berlin), Armin Straube (nestor-Geschäftsstelle bei der DNB) und Fiona Murphy (Wiley) befassten sich mit den organisatorischen Rahmenbedingungen und Maßnahmen, um das Bewusstsein für den Wert von Forschungsdaten zu steigern und ihren Werterhalt zu sichern.

Im DFG-Projekt EWIG (ewig.gfz-potsdam.de), durchgeführt von GFZ, FU Berlin und ZIB, werden Komponenten zur Förderung von Langzeitarchivierung entwickelt, darunter die Entwicklung von Lehr- und Lernmodulen für einzelne Studiengänge, etwa den Bachelorstudiengang Meteorologie, wo im Statistik-Modul erarbeitet werden soll, was verlässliche Daten kennzeichnet. Dazu wurden u.a. Interviews mit wissenschaftlichem Personal und ein Workshop durchgeführt, die zeigten, dass es noch viele Barrieren in den Köpfen gibt, die eine Datenweitergabe behindern. Dies könnte sich ändern, wenn nicht mehr die auf den Forschungsdaten beruhenden Publikationen, sondern die Veröffentlichung der Daten selbst zur Reputation beitragen. Als Anregung für Dozenten, die sich bisher kaum mit Forschungsdatenverwaltung befasst haben, wurde im Projekt eine Handreichung erarbeitet, zu finden auf www.forschungsdaten.org. Mit dem Fokus auf die Sozialwissenschaften bietet der Wegweiser Forschungsdaten in den Sozial- und Wirtschaftswissenschaften www.data-archive.auffinden-zitieren-dokumentieren.de Anregungen.

Auf Grundlage der DIN 31644 „Kriterien für vertrauenswürdige digitale Archive“ entwickelte nestor ein Verfahren zur Selbsteinschätzung mit ausführlicher Dokumentation, die anschließend von zwei Auditoren überprüft wird. Besteht eine Einrichtung diese Zertifizierung, so erhält sie ein Nestor-Siegel. Die Kosten betragen 50 Euro. Das Verfahren ist nur für Einrichtungen anwendbar, DFG-Projekte können kein Nestor-Siegel bekommen, weil sie nicht auf Dauer angelegt sind. Das Nestor-Siegel ist ein eigenständiges Zertifikat, das mit seinen 34 Kriterien umfassender ist als das 16 Kriterien berücksichtigende in den Niederlanden entwickelte ältere Data Seal of Approval (www.datasealofapproval.org), doch auf diesem aufbaut. Eine Qualitätsbeurteilung der Daten ist Bestandteil der Selbsteinschätzung, wobei nachvollziehbar dokumentiert werden muss, ob lediglich ein Datenstrom gespeichert wird oder ob eine qualitativ anspruchsvolle Datenspeicherung erfolgt.

Wissenschaftsverlage machen bereits Gehversuche in Richtung Forschungsdatenpublikation, so Wiley mit dem Geoscience Data Journal, für das Fiona Murphy zuständig ist. Sie ist ebenso Gründungsmitglied des Projekts PREPARDE (Peer REview for Publication & Accreditation of Research Data in the Earth sciences). Es soll auch die Zusammenarbeit mit Bürgerwissenschaftlern ermöglichen und negative Ergebnisse, die meist verworfen oder verschwiegen werden, sammeln. Im Moment tut man sich noch schwer mit der Frage, was „Daten“ zugerechnet werden soll und was nicht, was also gepflegt, inhaltlich erschlossen und für die Nachnutzung zitierfähig bereit gestellt werden soll und wofür sich der Aufwand nicht lohnt bzw. keine Hosts, keine Finanzierung oder keine allgemeinen Verfahrensgrundsätze zur Verfügung stehen.

Fallstudien

Der dritte Block bot drei Fallstudien aus dem Forschungsumfeld. Die Chemikerin Nicole Jung (KIT Karlsruhe) ist in der organisch-chemischen Forschung tätig und berichtete von dem internen Projekt Chemotion (www.chemotion.net), bei dem mehr und mehr Eigenschaftsdaten von organischen Molekülen direkt bei ihrer Ermittlung mittels einer Open Source-Software vom Messinstrument in ein Laborinformationsmanagementsystem (LIMS) übertragen werden und jeweils mit einer zitierfähigen DOI versehen werden. Dabei wird weltweit sowohl mit Wissenschaftlergruppen als auch mit Geräteherstellern zusammengearbeitet. Die Plattform ist sowohl als Speicher für Forschungsdaten konzipiert als auch Kommunikationsplattform.

Klaus Rechert und Dennis Wehrle (Universität Freiburg) sind Informatiker und haben sich dem Problem der langfristigen Nutzung von Daten angenommen, die mit älteren Softwareversionen oder auf alten Rechnersystemen erzeugt worden sind. Mittels verfügbarer Open Source-Emulatoren haben sie eine Cloud-Anwendung entwickelt, mit der alte Datenbestände wiedererweckt, die seinerzeitigen Experimente nach-

vollzogen und die Daten begutachtet werden können, wie sie eindrucksvoll anhand einer HyperCard-Anwendung zeigten. Interessenten können mit den Daten arbeiten, sie aber nicht entnehmen. Informationen und eine Demoversion ihrer Lösung Emulation-as-a-Service finden sich unter http://bw-fla.uni-freiburg.de.

Dirk Fleischer (GEOMAR Kiel) knüpft an die Erfahrung an, dass Biologen im Regenwald, Meeresforscher auf hoher See, Geologen im Hochgebirge oder andere Feldforscher in der Regel mit Papier und Stift unterwegs sind, um Daten aufzuzeichnen. Sein Ansatz ist es, in die gewohnte Routinearbeit eine neue Technologie einzubringen, durch die Daten direkt bei der Entstehung erfasst und gespeichert werden. Seine Lösung basiert auf einem speziellen Stift (für 2000 Euro Jahreslizenz in Schweden zu mieten) und Papierformularen, auf die mit diesem Stift Eintragungen erfolgen. Jedes Formular repräsentiert zugleich eine Anwendung, die sowohl im Stift gespeichert als auch vom Stift einzeln an den Server übermittelt und von dort in einer Datenbasis abgelegt wird. Dabei werden etwa Zeit und Ort der Datenerhebung automatisch mit erfasst. Sollte die Datenübertragung und Datenspeicherung im Stift einmal nicht funktionieren, weil vergessen wurde, den Stift zu laden, so ist immerhin noch das beschriebene Blatt verfügbar.

Infrastrukturen für die Datenverwaltung

Um die Forschergruppen, die ihre Daten und Dokumente verarbeiten, anderen zur Verfügung stellen oder sie dauerhaft speichern wollen, konkurrieren verschiedene kommerzielle und gemeinnützige Einrichtungen mit technischen und organisatorischen Infrastrukturlösungen.

David Wilcox (DuraSpace) ist Product Manager des Open Source-Projekts Fedora (Flexible Extensible Durable Object Repository Architecture), dessen 52 Mitglieder abgesehen von FIZ Karlsruhe fast alle aus dem angloamerikanischen Raum stammen. Neu in Version 4 der Plattform ist die Unterstützung von Linked Open Data. Beispiele und die Dokumentation des Systems finden sich im Fedora-Wiki (https://wiki.duraspace.org/display/FF).

Alex Wade (Microsoft Research) präsentierte die Plattform Azure4 Research (www.azure4research.com) mit der cloud computing für die Forschung attraktiv gemacht werden soll. Die bereits eingeführte Plattform Microsoft Azure soll zur Zusammenarbeit im Forschungsumfeld und für datenintensive Rechenleistungen genutzt werden können. Ein Research Data Registry and Discovery Service Marketplace ist unter datamarket.azure.com erreichbar. Wer Daten hat, von denen er meint, sie könnten für andere interessant sein, kann sie einreichen und kann bei einer positiven Bewertung ein Jahr lang gratis die Plattform mit allen Programmen nutzen. Webinare und Trainings werden ebenso angeboten wie ein Azure Machine Learning Center.

Matthias Hahn (FIZ Karlsruhe) stellte abschließend die Ziele des DFG-Projekts RADAR – Research Data Archiving as a Service (www.radar-project.org) vor. In dessen Rahmen will FIZ Karlsruhe als Dienstleister die vertrauenswürdige Speicherung von Forschungsdaten in seinem Rechenzentrum anbieten. Alternativ soll die Wahl zwischen einer bis zu 15 Jahre währenden Datenhaltung ohne Veröffentlichung (2015 verfügbar) sowie einer unbegrenzten Datenspeicherung bei gleichzeitiger Veröffentlichung mit zugewiesenem DOI (2016 verfügbar) bestehen. Wie „unbegrenzt“ der Zeitraum bei digitalen Daten tatsächlich sein wird und welcher Aufwand dafür erforderlich sein wird, das sind offene Fragen. Dabei ist auch ein Zugriffsmanagement mitsamt Embargo- und Verlagsdienstleistungen vorgesehen. Zielgruppe sind abgeschlossene Projekte von Gruppen, die über keine eigene Infrastruktur zur Speicherung ihrer Daten verfügen. Inhaltlich erschlossen werden die Daten nach einem einheitlichen Erschließungsschema, das neun verbindliche und zwölf optionale Merkmale vorsieht.

Akzeptanz offen

Als Herausforderung angesichts all dieser Anstrengungen und Systeme bleibt die Frage nach der Akzeptanz durch die Forscher, die die Daten bereitstellen müssen. Um saubere Daten zu erhalten, ist es unumgänglich, dass die Datendokumentation zum Zeitpunkt der Erzeugung erfolgt. Was die Forscher dabei selbst übernehmen und was Aufgabe von Datendokumentaren sein wird, wird sich zeigen. Außerdem dürfte die starke Konkurrenz um Fördermittel im Wissenschaftsbetrieb in vielen Fällen gegen die Bereitstellung aktueller Daten sprechen, solange es keine weltweit geltenden rechtlichen Regelungen und anerkannte Regeln guter wissenschaftlicher Praxis bezüglich der Nachnutzung gibt.

Gefehlt haben bei dieser Konferenz diejenigen, die fremde Daten tatsächlich für eigene Forschungen oder Publikationen nutzen wollen. Auch über deren Akzeptanz lässt sich nur spekulieren. Doch plant FIZ Karlsruhe angesichts des großen Interesses 2015 eine Folgeveranstaltung, bei der dann vielleicht auch die konkreten Erfahrungen mit der Nachnutzung von Daten ein Thema sein könnten. ❙

Marlies Ockenfeld
Darmstadt
marlies.ockenfeld@gmx.net