Forschungsdaten sammeln und strukturieren:

Ein Thema auf dem „Kongress Bibliothek und Information“ in Leipzig 2013

Frank Scholze

Das Sammeln, Aufbereiten, Strukturieren, Verwalten und Zugänglichmachen, kurz das Managen von Forschungsdaten ist als Thema in der breiten bibliothekarischen Öffentlichkeit angekommen. Zwei Themenblöcke[1] und verschiedene Einzelvorträge[2] beschäftigten sich beim diesjährigen Kongress Bibliothek und Information in Leipzig mit Forschungsdaten. Während sich vor wenigen Jahren noch vorwiegend Wissenschaftler und einige wenige Datenkuratoren bzw. –bibliothekare an Spezialbibliotheken und außeruniversitären Forschungseinrichtungen mit sog. „Forschungsprimärdaten“ beschäftigten, sind heute zahlreiche Universitätsbibliotheken dabei, institutionelle oder fachliche Infrastrukturen für das Datenmanagement aufzubauen oder vorzubereiten. Monika Kuberek (TU Berlin), Johanna Vompras (Universität Bielefeld), Jana Porsche (IST Austria) und Elena Simukovic (HU Berlin) gaben in Leipzig exemplarisch jeweils einen Einblick in den Stand an ihrer Einrichtung. Um dieser allgemeinen Entwicklung Rechnung zu tragen und sie weiter zu befördern hat die Deutsche Forschungsgemeinschaft jüngst ein entsprechendes Förderprogramm aufgelegt[3].

Forschungsdaten sind zunächst einmal disziplinspezifisch. In allen experimentell und empirisch arbeitenden Wissenschaften werden Daten gemessen oder erhoben. Dies ist vielfach beschrieben worden und reicht von der Hochenergiephysik über die Astronomie, die Klima- und Geoforschung bis hin zu weiten Bereichen der Sozial und Wirtschaftswissenschaften, der Medizin und den Bio- und Lebenswissenschaften. Je aufwändiger die Experimente und Erhebungen sind, je größer die Menge der anfallenden Daten, desto stärker haben sich Infrastrukturen, Standards und Workflows zum Sammeln und Verwalten der Daten in den Disziplinen etabliert.

Sowohl international wie auch national gibt es erfolgreiche Beispiele für Forschungsdaten-Repositorien, etwa das World Data System der International Council of Science (ICSU)[4] mit seinen mehr als 50 weltweit verteilten disziplinspezifischen Datenzentren oder GESIS – Leibniz-Institut für Sozialwissenschaften[5] mit seinen archivierten Studien und empirischen Primärdaten aus den Sozialwissenschaften. Auch Fachdatenbanken wie z.B. die Protein Data Bank (PDB)[6] nehmen Forschungsdaten auf. Gerade letztere zeigt aber auch, dass diese existierenden Angebote oft nur einen Ausschnitt der insgesamt anfallenden Forschungsdaten erfassen.

Bibliotheken nehmen mehr und mehr die Disziplinen in den Fokus[7], die über weniger ausgeprägte Dateninfrastrukturen verfügen. In vielen Bereichen der Physik, der Chemie, der Bio- und der Ingenieurwissenschaften etc. werden kleinere oder mittelgroße Experimente durchgeführt, die erhobenen Daten analysiert, umgeformt oder verdichtet, so dass aus den Resultaten eine Publikation erstellt werden kann. Nach dem bzw. den Publikationen werden die Daten entweder gelöscht oder vergessen, meist auf lokalen (durchaus auch virtualisierten) Servern und Speichern. Eine oft zitierte These lautet nun, dass diese Daten zu wertvoll seien, um nach kurzer Zeit verloren zu gehen. Entweder könnten sie von anderen Wissenschaftlern mit ggf. anderen Methoden untersucht werden, um neue Erkenntnisse zu gewinnen oder sie könnten der eigenen Arbeitsgruppe ggf. erneute Messungen ersparen. Zudem sollten sie für Transparenz und Nachprüfbarkeit bzw. Nachvollziehbarkeit sorgen – ein Axiom der Wissenschaft.

Im Grundsatz ist dieser Forderung nach einem durchgängigen Datenmanagement für die Fachdisziplinen, in denen sich derartige Ansätze bislang nicht herausgebildet haben sinnvoll und richtig. Nur sind weitere Rahmenbedingungen zu berücksichtigen, um nicht

1 „Forschungsdaten-Repositorien - Infrastrukturen zur dauerhaften Zugänglichkeit von Forschungsdaten“ am 11.3. und „Forschungsdaten sammeln und strukturieren“ am 12.3.

2 U.a. im Themenblock Langzeitarchivierung.

3 http://www.dfg.de/foerderung/info_wissenschaft/info_wissenschaft_13_19/index.html

4 http://www.icsu-wds.org/

5 GESIS - Leibniz-Institut für Sozialwissenschaften, http://www.gesis.org/

6 Die PDB wird vom Research Collaboratory for Structural Bioinformatics (RCSB) unterhalten, http://www.rcsb.org

7 Z.B. im Projekt RADAR – Research Data Repositorium http://www.tib-hannover.de/en/the-tib/news/news/id/409/

in einem pauschalen „jeder Messwert muss bewahrt werden“ stecken zu bleiben. Die Daten müssen mit entsprechenden Kontextinformationen versehen sein, um „lesbar“, d.h. nachnutzbar, zu bleiben. Dies bedeutet einen gewissen Aufwand, der jedoch leistbar und zum Teil auch automatisierbar ist, wie best practice Beispiele aus verschiedenen Disziplinen belegen.[8] Erneute Messungen können jedoch - wie die Re-Analyse von Daten - ebenfalls zu Erkenntnisgewinnen beitragen, zudem können sie günstiger sein als ein dauerhaftes Vorhalten bereits erhobener Daten. Darüber hinaus kann die Entwicklung neuer Messinstrumente sowie neuer Messmethoden ältere Daten obsolet machen. Setzt man sich jedoch konstruktiv mit all diesen und weiteren Fragen auseinander, ist man bereits dabei, einen Datenmanagementplan aufzustellen. Denn auch die bewusste Entscheidung, Daten nur kurz oder mittelfristig aufzubewahren, kann Teil einer Gesamtstrategie sein. Und nur so gelingt es, das Datenmanagement in die Arbeitsprozesse der Wissenschaft zu integrieren. Bibliotheken können hier als Partner der Wissenschaft das Bewusstsein stärken, dass digitale Infrastrukturen ein untrennbarer Teil der Forschung geworden sind, die nicht ausschließlich von zufälligen Entscheidungen oder evolutionären Entwicklungen abhängig sein sollten. Zudem bleibt die zentrale Forderung nach Nachprüfbarkeit bzw. Nachvollziehbarkeit der Forschung.

Der gemeinsame Bezugspunkt der unterschiedlichen Vorträge in Leipzig war daher, Bibliotheken als Berater und Begleiter zu sehen, die den Fokus auf eine mittel- bis langfristige Verfügbarkeit von Forschungsdaten legen, wobei die Klärung von Rahmenbedingungen und Selektionskriterien ein notwendiges Merkmal ist. Dies wird z.B. an der persistenten Identifizierung von Forschungsdaten deutlich. Diese skaliert nur, wenn sie in Forschungsprozesse je nach fachlichen Gegebenheiten integriert ist, ebenso wie die Verbindung zwischen bzw. Integration von Daten und Publikationen. Geteilt waren die Meinungen darüber, ob durch diese Fokussierung Bibliotheken wieder näher an die Forschung „heranrücken“ oder ob sie nicht nur als integraler Bestandteil der Wissenschaft deren digitalen Wandel mit vollziehen. Diesen Gedanken unterstrich vor allem Thomas Bürger mit seinem Vortrag „Die Bibliothek als Forschungsinfrastruktur“, der auch die Perspektive der Geisteswissenschaften einnahm. Hier stellen sich andere grundsätzliche Fragen im Bereich der Forschungsdaten als bei den experimentierenden, messenden oder simulierenden Disziplinen, nicht zuletzt hinsichtlich der Definition „Was sind eigentlich (digitale) Forschungsdaten für unsere Disziplin?“

„Der Vorhang zu und alle Fragen offen“ könnte daher das Fazit der Vorträge zu Forschungsdaten beim Kongress Bibliothek und Information in Leipzig 2013 lauten, aber gerade die offenen Fragen treiben nicht nur die Wissenschaft, sondern auch das Bibliothekswesen an. Insofern war es bezeichnend, dass eine der beiden Vortragsveranstaltungen vom DFG-Projekt re3data (Registry of Research Data Repositories)[9] organisiert wurde. Dieses Projekt versucht eine Bestandsaufnahme dessen, was sich als Forschungsdatenrepositorium bezeichnet und gleichzeitig eine strukturierte Beschreibung und Darstellung der Funktionen dieser Forschungsinfrastrukturen. Weder das eine noch das andere ist abgeschlossen, jedoch haben sich bereits jetzt aus beinahe allen Disziplinen best practice Beispiele herauskristallisiert, die hinsichtlich Funktionen und Transparenz Vorbildcharakter besitzen.[10] Dies verbreitert und vertieft den Diskurs über Forschungsdateninfrastrukturen und hilft Bibliotheken, ihre Position zu bestimmen, so wie dies gegenwärtig auch Verlage und andere Firmen tun. Der Vorhang ist dabei so offen wie die Fragen, das Thema wird auch weitere Kongresse und Bibliothekartage beschäftigen.

Spannend für die Wissenschaft ist dabei weniger die Frage der institutionellen Zuständigkeit – auch wenn diese sicherlich immer wieder intensiv diskutiert werden wird, da auch primär fachlich orientierte Infrastrukturen eine institutionelle Verankerung benötigen – sondern die Frage wie digitales Forschungsdatenmanagement zurück wirkt auf Methoden, Kommunikation und Erkenntnisgewinn in den einzelnen Disziplinen – kurz wie sich die digitale Wissenschaft im 21. Jahrhundert entwickeln wird.[11]

Frank Scholze
KIT-Bibliothek
Straße am Forum 2
76131 Karlsruhe
www.bibliothek.kit.edu
scholze@kit.edu

8 Siehe http://www.re3data.org/ bzw. http://blogs.plos.org/mfenner/2013/06/01/re3data-org-registry-of-research-data-repositories-launched/

9 Pampel et al. (2013) Making research data repositories visible: the re3data.org registry. In: PeerJ PrePrints 1:e21v1 http://dx.doi.org/10.7287/peerj.preprints.21v1

10 Clarin-Eric, Archaeology Data Service, EASY (DANS), IQSS Dataverse network, CrystalEye (beta), Durham HepData Project, figshare, Global Change Master Directory (NASA), NeuroMorpho, Neuroscience Information Framework, ClinicalTrials.gov, Ecological Archives (ESA), PANGAEA, WDC for Remote Sensing of the Atmosphere

11 Michael Nielsen (2011) Reinventing Discovery: The new era of networked science. Princeton.
The Royal Society. (2012). Science as an open enterprise. The Royal Society Science Policy Centre report 02/12. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf