

Bericht über das DFG-Rundgespräch „Forschungsprimärdaten“ am 17.01.2008

In der wissenschaftlichen Forschung, z.B. bei Messprozessen oder Experimenten, gewonnene Primärdaten sind von erheblichem Wert für die aktuelle und künftige Forschung. Auch deshalb haben sich alle wesentlichen Forschungseinrichtungen entsprechend den Vorschlägen zur Sicherung guter wissenschaftlicher Praxis aus dem Jahr 1998 verpflichtet, Primärdaten als Grundlagen von Veröffentlichungen auf haltbaren und gesicherten Trägern für zehn Jahre aufzubewahren. Trotzdem ist ihre Existenz oft nur wenigen Wissenschaftlern bekannt, die Daten werden meist nicht publiziert und ihre langfristige Sicherung ist oft nicht geregelt.¹ Die Deutsche Forschungsgemeinschaft (DFG) lud vor diesem Hintergrund im Januar 2008 zu einem Rundgespräch zur „Speicherung und Bereitstellung von Forschungsprimärdaten“ ein, um darauf aufbauend Fördermaßnahmen in der Informationsinfrastruktur zu spezifizieren.

An dem eintägigen Gespräch nahmen neben den Mitgliedern des DFG-Unterausschusses für Informationsmanagement rund 20 Wissenschaftler und Vertreter von Forschungseinrichtungen verschiedener Fachdisziplinen teil, wobei das Spektrum der vertretenen Einrichtungen vom Seminar für Sprachwissenschaft der Universität Tübingen bis zum Jülich Supercomputing Centre / Forschungszentrum Jülich reichte. Einleitend führte Sigrun Eckelmann (DFG) in die Thematik ein und nannte als Ziel des Rundgesprächs die Konkretisierung der geplanten Förderinitiative der DFG. Nach einer Vorstellungsrunde folgten am Vormittag vier kurze Referate als Grundlage für die jeweils folgende Diskussion.

Unter dem Schlagwort „Visionen und Träume“ entwarf zunächst Gerhard Schneider (Rechenzentrum der Universität Freiburg) eine Vorstellung von der zukünftigen Verfügbarkeit von Primärdaten und nannte vier Visionen: 1. einen Verbund von technischen Speichereinrichtungen, die nach gemeinsamen Standards die Daten vorhalten und auffindbar machen, 2. allgemein akzeptierte Standards, um Softwareumgebungen inklusive der notwendigen Funktionalität lebendig halten zu können sowie Standards, die eine verlustlose Migration von „lebendigen“ Daten zwischen akzeptierten Softwaresystemen erlauben, 3. die Anerkennung der wissenschaftlichen Community, dass Ergebnisse nur dann aufgehoben werden können, wenn sie nach (einschränkenden, bewertenden) Standards gespeichert werden und 4. die Klärung und Berücksichtigung rechtlicher Fragen.

Nach der lebhaften Diskussion referierte Malte Dreyer (Max Planck digital library) zu „Typen von Primärdaten“ und schlug vor, Primärdaten als „Artefakte des wissenschaftlichen Prozesses“ zu verstehen und den Begriff damit weiter zu fassen, als in der bisherigen Diskussion geschehen. Dreyer betonte angesichts des Umfangs der entstehenden Datenmengen die Notwendigkeit eines Lebenszyklusmanagements und stellte die Bedeutung des jeweiligen fachspezifischen Kontextes der Datenentstehung heraus. Nach den Erfahrungen seiner Einrichtung gibt es bereits innerhalb einzelner Fachdisziplinen eine große Heterogenität der Daten und zugehörigen Metadaten.

¹ Zum Sachstand vgl. umfassend: nestor - materialien 6 - Langzeitarchivierung von Rohdaten / Thomas Severiens, Eberhard R. Hilf, Institute for Science Networking Oldenburg GmbH an der Carl von Ossietzky Universität Oldenburg, Frankfurt am Main : nestor c/o Die Deutsche Bibliothek, 2006.

Das dritte Referat war unter dem Titel „Die Player“ den an der Publikation und Archivierung von Primärdaten beteiligten Personen und Einrichtungen gewidmet. Ulrich Niess (Stadtarchiv Mannheim) plädierte für eine Anlehnung an die Vorgehensweise der öffentlichen Archive und unterschied zwischen einer primären und sekundären Nutzung der Daten. In der Phase der primären Nutzung läge die Hoheit über die Daten bei dem sie erhebenden Wissenschaftler bzw. der Forschungseinrichtung, dies schließe den Zeitraum der zehnjährigen Aufbewahrungsfrist nach den Regeln guter wissenschaftlicher Praxis ein. Nach Ablauf der - fachspezifisch sinnvoll festzulegenden - Aufbewahrungsfrist sollten die Daten dem zuständigen Archiv zur Übernahme angeboten und von diesem im Benehmen mit der anbietenden Einrichtung bewertet werden. Die Daten, denen ein sekundärer Nutzen, d. h. ein bleibender Wert für weitere Forschungsvorhaben zugesprochen werden könne, seien durch das zuständige Archiv dauerhaft zu archivieren und damit auch der Öffentlichkeit zur Verfügung zu stellen.

Im Anschluss an die kontroverse Diskussion folgte das letzte Referat, das sich den „Organisationsstrukturen“ widmete. Hans Pfeiffenberger (Alfred Wegener Institute) verdeutlichte einleitend am Beispiel des Internationalen Polarjahres Komplexität und Kosten eines langfristigen Datenmanagements, stellte wichtige Kriterien für eine angemessene Organisationsstruktur vor und plädierte abschließend für Centers of Excellence für ausgewählte Disziplinen und die Selbstorganisation der Wissenschaft als Grundlage einer funktionsfähigen Infrastruktur.

Nach einer Mittagspause folgte die Präsentation von zwei bereits durch die DFG geförderten Projekten, die sich mit der Publikation und Archivierung von Primärdaten befassen. Zunächst stellte Jan Brase (Technische Informationsbibliothek und Universitätsbibliothek Hannover) das Projekt „Datenpublikation und die Registrierungsagentur für wissenschaftliche Daten an der Technischen Informationsbibliothek - Etablierungsphase mit dem Schwerpunkt Erdsystemforschung“ als Nachfolger des Projekts "Publikation und Zitierfähigkeit wissenschaftlicher Primärdaten - CoData“ vor (siehe www.tib-hannover.de/de/die-tib/projekte/codata/). Anschließend präsentierte Erich Weichselgartner (Zentrum für Psychologische Information und Dokumentation) das Projekt „PsychData: Dokumentation und Archivierung von Rohdatensätzen aus der psychologischen Forschung“ (siehe <http://psychdata.zpid.de/>) und betonte besonders den hohen Aufwand einer retrospektiven Datendokumentation, die ein Vielfaches der Ressourcen binde, die bei einer zeitnahen Dokumentation der Daten im Kontext der Erhebung benötigt worden wären.

Abschließend fasste Thekla Kluttig (Sächsisches Staatsarchiv) die Diskussionen und Ergebnisse des Rundgesprächs zusammen. Weitgehender Konsens bestehe zu fünf Punkten: 1. Derzeit ist der unkontrollierte Verlust von Primärdaten als Fakt festzustellen, 2. es besteht Handlungsbedarf, die Selbstorganisation der Wissenschaft hat nicht überall funktioniert, 3. zwischen den Fachdisziplinen gibt es eine erhebliche Heterogenität der Daten, der Metadaten, der Forschungseinrichtungen und der bestehenden Standards, die es zu berücksichtigen gilt, 4. die Wissenschaftler spielen eine zentrale Rolle für die Qualität der Daten und die Verbesserung der Dokumentation der notwendigen Kontextinformationen, für sie sind daher sinnvolle Anreize zu schaffen sowie 5. die Informationstechnik selbst ist nicht das zentrale Problem, auch wenn nicht alle technischen Fragen bereits beantwortet sind. Erheblicher Dissens be-

stehe vor allem zu zwei wichtigen Punkten: Liegt die Hoheit über die Daten stets bei den Wissenschaftlern bzw. den Forschungseinrichtungen oder sind angesichts der mit öffentlichen Mitteln geförderten Forschung die dabei entstandenen Daten durch „zuverlässige Einrichtungen mit einem Archivierungsmandat“ (Schneider) für die Öffentlichkeit nutzbar zu machen? Und: Sollen alle entstandenen Daten langfristig archiviert werden (Totalarchivierung) oder ist nach einer - oft zehnjährigen - Aufbewahrungsfrist eine fachspezifisch differenzierte Bewertung der Daten sinnvoll, um nur die dauerhaft archivwürdigen Daten zu sichern? Für letzteres plädierten die teilnehmenden Vertreter von Gedächtnisorganisationen (Bibliotheken, Archive). Kluttig nannte zusammenfassend als Ergebnisse des Rundgesprächs: Angesichts der unbefriedigenden Situation bei der Publikation und Archivierung von Forschungsprimärdaten seien Anreize für Veränderungen notwendig. Adressaten seien vor allem die Forschungseinrichtungen und (auch künftigen) Wissenschaftler. Dabei seien fachspezifische Lösungen anzustreben und auch nur solche durchsetzbar. Voraussetzung für eine positive Entwicklung sei, das Bewusstsein für den Nutzen und die Kosten einer Datenarchivierung zu wecken. Ein wichtiges unmittelbares Ziel bestünde in der Verbesserung der Aufbereitung von Datenbeständen, d. h. vor allem ihre Dokumentation bereits durch den die Daten erhebenden Wissenschaftler. Eine Totalarchivierung könne zwar im Einzelfall sinnvoll sein, grundsätzlich sollte aber die Erarbeitung von fachspezifischen Auswahlkriterien angestrebt werden. Solche Auswahlkriterien für langfristig archivwürdige Daten seien ein Aspekt einer Archivierungspolicy, ein anderer sei die Festlegung fachspezifisch sinnvoller Aufbewahrungsfristen. Die DFG könne wichtige Anstöße für die Erarbeitung solcher fach- und/oder institutionsspezifischer Archivierungspolicies geben.

Das Rundgespräch endete mit den Schlussworten der Veranstalter, Sigrun Eckelmann (DFG) und Horst Thomsen (Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Vorsitzender des DFG-Unterausschusses für Informationsmanagement).

Thekla Kluttig