Forschungsdaten in bester Lage: Nutzungsszenarien und Perspektiven digitaler Forschungsinfrastrukturen

Veranstalter: DFG-Projekt „Digitaler Wissensspeicher“, Berlin-Brandenburgische Akademie der Wissenschaften
Datum, Ort: 05.04.2016–06.04.2016, Berlin
Bericht von: Carmen Schwietzer, Berlin-Brandenburgische Akademie der Wissenschaften; Ulrike Wuttke, Union der deutschen Akademien der Wissenschaften

Vom 5.-6. April 2016 fand in Berlin der Workshop „Forschungsdaten in bester Lage: Nutzungsszenarien und Perspektiven digitaler Forschungsinfrastrukturen“ statt. Er wurde ausgerichtet vom DFG-Projekt „Digitaler Wissensspeicher“ an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW). In zwölf kurzen Vorträgen mit anschließenden Plenumsdiskussionen wurden die bisherigen Ergebnisse des Wissensspeicherprojekts sowie Zukunftsszenarien vorgestellt. Es waren Vertreterinnen und Vertreter thematisch relevanter Projekte und Infrastruktureinrichtungen als Sprecher geladen, um zentrale Fragen, die sich im Rahmen digitaler Forschungsinfrastrukturen stellen, im breiteren Kontext zu diskutieren: zur Bereitstellung, Integration und Vernetzung von großen heterogen und dezentral vorliegenden Datenbeständen sowie nach deren langfristiger Verfügbarkeit als zentrale Herausforderungen der Digitalisierung der Geisteswissenschaften und Digital Humanities.

Vier Themenbereiche prägten den Workshop: Erstens Forschungsinfrastrukturen, Forschungsdatenrepositorien und Metadatensuchmaschinen, zweitens die Nachhaltigkeit digitaler geisteswissenschaftlicher Forschung, insbesondere der Forschungsergebnisse und Infrastrukturen, bzw. die Aufgaben und Herausforderungen der Datenkuration, drittens Möglichkeiten der Präsentation und Visualisierung von Forschungsdaten in den Geisteswissenschaften und viertens rechtliche Fragen. ´

Den ersten Schwerpunkt des Workshops bildete die Vorstellung von Forschungsinfrastrukturen, Repositorien und Metadatensuchmaschinen sowie der Erfahrungsaustausch über deren Rolle und die zu lösenden Herausforderungen. MAXI KINDLING (Berlin) stellte das „Registry of Research Data Repositories“ (re3data.org[1]) vor. Bei diesem Service von DataCite[2] handelt es sich um das weltweit umfangreichste Verzeichnis von Forschungsdatenrepositorien aus allen Forschungsbereichen. Zum Zeitpunkt des Vortrags waren rund ein Zehntel der insgesamt rund 1.500 verzeichneten Repositorien den Geisteswissenschaften zuzuordnen (davon schätzte Kindling den Anteil wirklich disziplinspezifischer Forschungsdatenzentren auf rund 50 Prozent, da Mehrfachnennungen möglich sind). Kindling stellte definitorische Schwierigkeiten bezüglich des Begriffs „Forschungsdatenrepositorium“[3] sowie die trotz Standardisierungsbestrebungen auf nationaler und internationaler Ebene vorherrschende Vielfalt von Datentypen, Zugangsberechtigungen, verwendeten Datenlizenzen, Zertifikaten und Schnittstellen heraus. Besonders problematisch seien die unterschiedliche Metadatenqualität, der fehlende Zugang zu den in den Repositorien enthaltenen Daten und die unzulängliche Verwendung von PIDs.

SASCHA GRABSCH (Berlin), MARCO JÜRGENS (Berlin) und NIELS-OLIVER WALKOWSKI (Berlin) stellten in ihren Vorträgen den Digitalen Wissensspeicher[4] der BBAW unter unterschiedlichen Gesichtspunkten vor. Sascha Grabsch gab eine allgemeine Einführung zum Wissensspeicher, der neben einem zentralen Zugang zu allen digitalen Ressourcen der BBAW auch eine semantische Verknüpfung dieser heterogenen Bestände anbietet. Momentan sind über eine Million digitale Ressourcen, die im Kontext von 170 verschiedenen Projekten, wie Akademienvorhaben, Drittmittelprojekten und Initiativen entstanden sind, indexiert und durch Metadaten

[1] <www.re3data.org> (05.06.2016).

[2] <https://www.datacite.org/> (05.06.2016).

[3] Als kleinsten gemeinsamen Nenner habe man sich laut Kindling innerhalb von re3data.org auf folgende Definition geeinigt: „A research data repository is a subtype of a sustainable information infrastructure which provides long-term storage and access to research data [...]“. Die Begriffe Forschungsdaten und Langfristigkeit wurden im Plenum kontrovers diskutiert.

[4] <http://wissensspeicher.bbaw.de/> (05.06.2016).

beschrieben. Neben technischen Herausforderungen, wie der nachträglichen Entwicklung von Schnittstellen, stellte Grabsch insbesondere organisatorische und administrative Aspekte heraus. So waren mangels Dokumentation für die Erschließung der laufenden und beendeten digitalen Projekte der Akademie umfangreiche Gespräche mit (ehemaligen) Projektmitarbeiterinnen und -mitarbeitern erforderlich. Dazu kamen ressourcenintensive Umbaumaßnahmen an (laufenden) Projekten im Rahmen der Erschließung.

Auf die grundlegenden Technologien des Wissensspeichers als Infrastruktur zur Integration von äußerst heterogenen digitalen Ressourcen ging Marco Jürgens ein.[5] Besonders hervorzuheben sind die eine Volltextsuche weit übersteigenden Such- und Präsentationsfunktionen, die u.a. auf Apache Lucene, Donatus[6] und DBPedia Spotlight[7] beruhen. Weitere Entwicklungsarbeit ist hinsichtlich der Automatisierung und der Workflows notwendig. Niels-Oliver Walkowski hinterfragte in seinem Vortrag den Wissensbegriff im Kontext des Wissensspeichers („Ist Wissen die Repräsentation eines Diskurses durch die Verknüpfung digitaler Ressourcen?“) und gab einen Ausblick auf die weitere Entwicklung des Projekts im Rahmen der Modellierung geisteswissenschaftlicher Forschungsprozesse und der Dokumentation von Wissenschaftsprozessen (SDM[8] und NeMO[9]). In den Diskussionen spielten die Rolle des Wissensspeichers als Endbaustein einer Veröffentlichungskette, d.h. als Teil der wissenschaftlichen Kommunikation, und für die Vermittlung von Best Practices eine wichtige Rolle, sowie die komplexen Frage nach dem wissenschaftlichen Mehrwert des Wissensspeichers und seiner Zielgruppe.

Die technische Umsetzung sowie der wissenschaftliche Mehrwert einer Metasuchmaschine standen auch im Mittelpunkt der Vorstellung der Judaica-Suchmaschine durch HARALD LORDICK (Essen).[10] Lordick argumentierte, dass konfigurierbare nutzerzentrierte Suchmöglichkeiten aus den vorhandenen Suchinfrastrukturen heraus relativ einfach zu entwickeln seien. Der wissenschaftliche Mehrwert bestehe im Vergleich zu kommerziellen Anbietern wie Google in der Transparenz über den Datenraum und der Kuration durch Fachwissenschaftlerinnen und -wissenschaftler, die zu einer weitaus höheren Relevanz der Treffer führe. Das Konzept wird momentan im Rahmen von DARIAH-DE als Branded Search[11] weiterentwickelt. JÖRG LEHMANN (Berlin, Bern) stellte die *lessons learned* bezüglich der Datenakquise und der technischen Umsetzung beim Aufbau von CENDARI[12] vor, einer virtuellen Forschungsinfrastruktur für die europaweite Recherche in Archivbeständen zum Mittelalter und zum Ersten Weltkrieg.[13] LISA DIECKMANN (Köln) widmete ihren Beitrag den Herausforderungen bei der Zusammenführung heterogener kunst- und kulturhistorischer Daten am Beispiel des digitalen Bildarchivs prometheus.[14] Im Mittelpunkt der Diskussion stand vor allem die heterogene Metadatenqualität bzw. deren Verbesserung durch die verstärkte Einbindung von Wörterbüchern und Normdaten, z.B. der GND.

Der zweite Themenbereich widmete sich der Nachhaltigkeit von Forschungsinfrastrukturen und der Forderung nach Institutionen oder Dienstleistern, die sich

[5] Zu technischen Details siehe: <http://wissensspeicher.bbaw.de/> (05.06.2016).

[6] <http://archimedes.fas.harvard.edu/cgi-bin/donatus/> (05.06.2016).

[7] <https://github.com/dbpedia-spotlight/dbpedia-spotlight> (05.06.2016).

[8] Siehe u.a. Steffen Hennicke (et al.), D3.4: Research Report on DH Scholarly Primitives, 2015 <http://dm2e.eu/files/D3.4_2.0_Research_Report_on_DH_Scholarly_Primitives_150402.pdf> (05.06.2016).

[9] Siehe <http://nemo.dcu.gr/> (05.06.2016) und Luise Borek (et al.),TaDiRAH: a Case Study in Pragmatic Classification, DHQ 10 (2016):1 <http://www.digitalhumanities.org/dhq/vol/10/1/000235/000235.html, 05.06.2016>

[10] <http://steinheim-institut.de/vf> (05.06.2016).

[11] Siehe u.a. Tobias Gradl, Andreas Henrich, Christoph Plutte, ‚Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Förderation von Kollektionen. ' In: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum / Thomas Stäcker. 2015 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). DOI: <10.17175/sb001_020> (05.06.2016).

[12] <http://www.cendari.eu/> (05.06.2016).

[13] Siehe Jakub Beneš (at al.), The Cendari White Book of Archives, 2016 http://www.cendari.eu/thematic-research-guides/white-book-archives (05.06.2016).

[14] <http://www.prometheus-bildarchiv.de/index> (05.06.2016).

nach Ablauf der Entwicklungsphase um die Pflege und den Betrieb der Anwendungen kümmern, bzw. die Sicherstellung der langfristigen Verfügbarkeit der digitalen Forschungsdaten über die Projektlaufzeit hinaus. JOHANNES STIGLER (Graz) stellte hierzu vier Thesen auf: Erstens muss die Datenkuration von Anfang an mitbedacht werden; zweiten sollte die Repräsentationsschicht nicht programmiert, sondern modelliert werden; drittens muss eine technische Standardisierung erfolgen, z. B. durch die Etablierung von Workflows und viertens müssen Lösungen für die Langzeitarchivierung gefunden werden. Auch wenn einige dieser Thesen wie Zukunftsmusik klingen mögen und im Plenum kritisch diskutiert wurden – vor allem die „on the fly" Modellierung der Ansichten in GAMS[15] stieß bei den Fachwissenschaftlerinnen und –wissenschaftlern auf Widerstand – legen sie dennoch den Finger in die richtige Wunde. Die in den digitalen Geisteswissenschaften noch immer vorherrschenden proprietären Lösungen können nur durch institutionelle Strategien und einen möglichst breiten Erfahrungsaustausch aller relevanter Stakeholder, einschließlich der Fachcommunity, in nachhaltigere Lösungen und Strukturen überführt werden. Über Strategien zur Langzeitarchivierung und Datenkuration referierten auch FELIX SCHÄFER (Berlin), der die Arbeitsschritte für die Übernahme von Forschungsdaten in IANUS[16] vorstellte, und PATRICK SAHLE (Köln) für das Kölner Data Center for the Humanities (DCH)[17]. Da IANUS davon ausgeht, dass es als disziplinspezifisches, zentrales Forschungsdatenzentrum im Gegensatz zu institutionellen Forschungsdatenzentren wahrscheinlich erst sehr spät kontaktiert wird, wird großer Wert auf die Bereitstellungen von Informationen zur Sicherung der Nachhaltigkeit der Forschungsdaten gelegt, die weit über die engere IANUS-Community als repräsentativ gelten dürften.[18] Neben einer allgemeinen Vorstellung des Schichtenmodells des DCH stellte Sahle das „Resource Description Schema" des DCH vor, wobei es im Plenum zu einer angeregten Diskussion zur Definition des Terminus „Ressource" kam.[19] Sahle argumentierte auch, dass die oftmals geforderten zehn Jahre Langzeitarchivierungsfrist für Forschungsdaten eher ein naturwissenschaftliches Paradigma seien. In den Geisteswissenschaften, in denen Forschungsdaten eine andere Rolle spielen, gelten andere zeitliche Paradigmen (wenn auch das einmal von Klaus Graf „until judgment day" formulierte Paradigma etwas weit geht). Gerade die Ergebnisse geisteswissenschaftlicher Grundlagenforschung (wie Editionen, Wörterbücher, Personen-, Orte-, und Sachdatenbanken) haben eine sehr lange Halbwertzeit.

Mit den Möglichkeiten der Präsentation und Visualisierung von Forschungsdaten in den Geisteswissenschaften wurde im Verlauf des Workshops ein zentrales Thema der diesjährigen Tagung des DHd-Verbandes in Leipzig[20] aufgegriffen. Insbesondere der Beitrag von MARIAN DÖRK (Potsdam) war Visualisierungen als Mittel zur Analyse und Interpretation wachsender Informationsräume in den Geisteswissenschaften und ihrem Potential während des Forschungsprozesses gewidmet. Ziel von Dörks Projekt „Vizualising cultural collections"[21] ist – wie nicht zuletzt des Digitalen Wissensspeichers auch – die Erforschung der Möglichkeiten der explorativen Sichtung digitaler Sammlungen bei großer Transparenz des Datenraums. Im Mittelpunkt der Diskussion standen Fragen der Nachhaltigkeit von Visualisierungstools und der vorgestellten Visualisierungen sowie die Entwicklung von Komponenten des Nutzertrackings, d.h. der Nachvollziehbarkeit beliebter Wege durch digitale Sammlungen.

Wie in allen Diskussionen zur Verfügbarmachung von Forschungsdaten in den digitalen Geisteswissenschaften spielten auch die rechtlichen Rahmenbedingungen eine wich-

[15] <http://gams.uni-graz.at/> (05.06.2016). GAMS steht als österreichischer Beitrag zu DARIAH als Open Source Lösung zur Nachnutzung zur Verfügung <https://github.com/acdh/cirilo>, 05.06.2016).

[16] <http://www.ianus-fdz.de/> (05.06.2016).

[17] <http://dch.phil-fak.uni-koeln.de/> (05.06.2016).

[18] <http://www.ianus-fdz.de/it-empfehlungen/> (05.06.2016).

[19] Die Folien zu diesem Vortrag finden sich unter: <http://dch.phil-fak.uni-koeln.de/sites/dch/Materialien_Aktivitaeten/2016/ForschungsdatenInBesterLage.pdf> (05.06.2016).

[20] <http://dhd2016.de/> (05.06.2016).

[21] <https://uclab.fh-potsdam.de/projects/vikus/> (05.06.2016).

tige Rolle. Das fachliche „Know-how“ lieferte PAUL KLIMPEL (Berlin) mit einem kurzen historischen Abriss über die Entwicklung des Urheberrechts und einem Plädoyer für die großzügige Verwendung von CC-0 Lizenzen zumindest für Metadaten und Beschreibungstexte, um die Nachnutzung von Forschungsdaten zu erleichtern.[22] Des Weiteren wies Klimpel darauf hin, dass die Regeln des Urheberrechts und die Verwendung offener Lizenzen in keinem intrinsischen Zusammenhang mit dem Schutz des geistigen Eigentums im wissenschaftlichen Sinne stehen. Diese werden durch die gute wissenschaftliche Praxis, das heißt der Pflicht zur Kennzeichnung einer Quelle zur Genüge abgedeckt.

Zum Abschluss wurden durch die Teilnehmerinnen und Teilnehmer des Workshops an fünf Thementischen Fragen zur Datenkuration, Visualisierungen, Technik/Infrastruktur/Frameworks, Langzeitarchivierung sowie Infrastruktureinheiten und Forschungsdaten in kleineren Runden diskutiert. Die im Plenum vorgestellten Ergebnisse der Diskussionen überlappten sich teilweise oder ergänzten sich. So wurden beispielsweise bei der Diskussion über die Aufgaben der Datenkuration und des Berufsbilds des „Data Curators“ Parallelen zu den Ergebnissen der Thementische „Langzeitarchivierung“ und „Infrastruktureinheiten“ gezogen. Die Diskussion am Thementisch „Visualisierung“ setzte sich kritisch mit den Potenzialen und Risiken der Visualisierung von Forschungsdaten auseinander. Der Thementisch „Technik/Infrastruktur/Frameworks“ stellte technische Schwierigkeiten sowie Erfolgskriterien für die Technikentwicklung vor, wobei sich Überschneidungen mit den Diskussionen an den Thementischen „Langzeitarchivierung“ und „Infrastruktureinheiten“ ergaben. Ein wichtiger roter Faden, war die Notwendigkeit der Sensibilisierung der Wissenschaftlerinnen und Wissenschaftler für die Nachhaltigkeit ihrer digitalen Forschung und das Bewusstsein Infrastrukturprojekte, aber auch Bibliotheken und Rechenzentren, als kompetente Partner in diesem Bereich wahrzunehmen.

Weitere wichtiges Themen des Workshops waren die bisher ungelöste Frage der Kassation (Was wollen wir aufheben und wie lange?) und die ungewisse Zukunft vieler (auf Projektbasis geförderter) geisteswissenschaftlicher Infrastrukturen und Datenzentren, die nicht dazu beiträgt, das Vertrauen der Fachwissenschaftlerinnen und –wissenschaftler in die Zukunftsträchtigkeit der digitalen Forschung zu bestärken. Eine wichtige Aufgabe für die Zukunft ist somit die Gewährleistung der Nachhaltigkeit digitaler geisteswissenschaftlicher Infrastrukturen und Datenzentren. Wenn dieses Problem ungelöst bleibt, werden die Repositorien, Metasuchmaschinen oder avancierten Forschungsumgebungen wie der Digitale Wissensspeicher nur kurz in die Vergangenheit reichen oder komplett verschwinden.

Vorhandene kommerzielle Suchmaschinen können (und wollen) wissenschaftliche Forschungsdaten nicht den wissenschaftlichen Bedürfnissen entsprechend abbilden und zugänglich machen. Eine Zusammenführung und Vernetzung von Forschungsdaten durch die nachhaltige Einrichtung von institutionell gestützten Forschungsinfrastrukturen ist daher unabdingbar. Dabei gilt es ein besonderes Augenmerk auf die Datenkuration zu legen, um den Zugriff, die Auffindbarkeit und damit die Sichtbarkeit der digitalen Ressourcen langfristig zu gewährleisten. Hier hilft der Einsatz von technischen Standards und standardisierten Workflows, aber auch die attraktive Präsentation mittels innovativer Visualisierungsmethoden. Die kuratierten Daten und die entwickelten Softwarekomponenten der Forschungsinfrastrukturen müssen frei zur Verfügung stehen und ohne Hürden nachnutzbar sein, um eine möglichst breite Akzeptanz innerhalb der wissenschaftlichen Community zu erlangen.

Konferenzübersicht:

Maxi Kindling (Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin), Überblick über die Landschaft der Forschungsdaten-Repositorien auf Basis von re3data.org

Sascha Grabsch (Digitaler Wissensspeicher, BBAW), Der Digitale Wissensspeicher als

[22] Siehe z.B. John H. Weitzmann und Paul Klimpel, Handreichung Recht (1.2), 2016 <http://dx.doi.org/10.12752/2.0.002.2> (05.06.2016).

Baustein in der Publikationskette digitaler Forschungsdaten

Johannes Hubert Stigler (Zentrum für Informationsmodellierung, Austrian Centre for Digital Humanities, Universität Graz), Die Sicherstellung der Verfügbarkeit von Forschungsdaten als Aufgabe von Langzeitarchivierung am Beispiel eines FEDORA-basierten Repositoriums

Felix Schäfer (IANUS, Deutsches Archäologisches Institut Berlin), Vom Produzenten ins Archiv. Arbeitsschritte zur Datenkuratierung bei IANUS

Patrick Sahle (Cologne Center for eHumanities, Universität Köln), Datenübernahme und ‚Resource Description Schema' im Kölner Data Center for the Humanities (DCH)

Marco Jürgens (Digitaler Wissensspeicher, BBAW), Lessons Learned: Erfahrungen aus der Entwicklung des Digitalen Wissensspeichers

Paul Klimpel (iRights-law), Wissenschaftliche Freiheit und rechtliche Fußangeln

Harald Lordick (Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte, Universität Duisburg-Essen), Möglichkeiten und Grenzen einer Judaica-Suchmaschine im digitalen Forschungsumfeld

Jörg Lehmann (FU Berlin, Universität Bern), Das Rezept der CENDARI Data Soup (CENDARI White Book of Archives)

Niels-Oliver Walkowski (TELOTA, BBAW), Integration, Interaktion, Evaluation: Neue Möglichkeiten für die Analyse digital unterstützter Forschungsaktivitäten

Marian Dörk (Institut für Angewandte Forschung Urbane Zukunft, FH Potsdam), Von Repräsentation zu Interpretation: Visualisierung in den Geisteswissenschaften

Lisa Dieckmann (Kunsthistorisches Institut, Universität Köln), Von Malerei bis Performancekunst – Über Möglichkeiten und Herausforderungen bei der Zusammenführung heterogener kunst- und kulturhistorischer Daten

Diskussion an Thementischen:
Data Curation (Alexander Czmiel)
Vernetzung Kunstgeschichte/Visualisierung (Lisa Dieckmann)
Technik/Infrastruktur/Frameworks (Josef Willenborg)
Langzeitarchivierung (Hubert Stigler)
Infrastruktureinheiten/Archive/Bibliotheken und Forschungsdaten (Markus Schnöpf)

Tagungsbericht *Forschungsdaten in bester Lage: Nutzungsszenarien und Perspektiven digitaler Forschungsinfrastrukturen*. 05.04.2016–06.04.2016, Berlin, in: H-Soz-Kult 19.07.2016.