Tagungsberichte

Zum qualifizierten Umgang mit Forschungsdaten

Ein Bericht über den Workshop „Wissenschaft im digitalen Wandel“ am 6. Juni 2017 in der Universität Mannheim

Johannes Fournier, Deutsche Forschungsgemeinschaft

Daten werden zusehends als wichtigster Rohstoff nicht nur für die Erzeugung neuen Wissens, sondern auch für die Entwicklung innovativer Dienstleistungen angesehen. Doch können nur gut qualifizierte, kompetente Fachleute die Potenziale gewaltiger Datenmengen erschließen – und dass solche Fachleute (noch) rar sind, ist eine Binsenweisheit. Allerdings gibt es eine Fülle von zumeist außercurricularen Fortbildungsangeboten, die einschlägige Kenntnisse zum qualifizierten Umgang mit Daten vermitteln. Die Veranstaltung „Wissenschaft im digitalen Wandel“[1] stellte die Frage, wie solche Angebote systematisch weiterentwickelt werden könnten. Der von rund 60 Teilnehmerinnen und Teilnehmern besuchte Workshop wurde von der Arbeitsgruppe „Intelligente und effiziente Nutzung von Open Data in Wissenschaft, Forschung und Wirtschaft“ organisiert, einer Untergruppe der vom Bundesministerium für Bildung und Forschung (BMBF) moderierten Plattform „Digitalisierung in Bildung und Wissenschaft“[2].

In seiner Eröffnung betonte *Wolfgang Marquardt* (Forschungszentrum Jülich), dass Daten immer schneller und mit immer größeren Volumina anfallen würden, die Interpretation dieser Daten mit dem enormen Wachstum jedoch nicht Schritt halte. Das erschwere nicht nur den Gewinn neuer Erkenntnisse in der Forschung, sondern wirke sich auch nachteilig auf die Entwicklung neuer Produkte und Dienstleistungen aus. Um auf diese Defizite zu reagieren, sei es erforderlich, Kenntnisse für den professionellen Umgang mit Forschungsdaten sowohl disziplin- als auch institutionenübergreifend zu vermitteln. Die Vermittlung solcher Kenntnisse müsse mittelfristig in die Curricula der Universitäten integriert werden. Dies sei jedoch ein weiter Weg und der Workshop solle daher gezielt aufzeigen, welche Zwischenschritte auf diesem Weg sinnvoll gestaltet werden könnten.

Keynote Sessions

In drei vormittäglichen Keynotes wurde zunächst eine breite Perspektive auf die Vermittlung datenbezogener Kompetenzen eingenommen. Vorgestellt wurden sowohl Aktivitäten auf europäischer Ebene als auch Formen der Kompetenzvermittlung, die stärker auf die Wirtschaft zielen. Zunächst wies *Thomas Ganslandt* (Universität Erlangen-Nürnberg) auf die zunehmende Verfügbarkeit von Versorgungsdaten sowie von molekularbiologischen Untersuchungsdaten, die Förderung der Medizininformatik und die Digitalisierung der Gesundheitswirtschaft als wesentliche Rahmenbedingungen für das zunehmende Interesse am Datenmanagement in der Medizin hin. In typischen Projektverläufen

1 Workshop „Wissenschaft im digitalen Wandel“, zuletzt geprüft am 02.07.2017, http://www.fz-juelich.de/ue/DE/Leistungen/Controlling_UE-C/WS_Wissenschaft/_node.html.

2 Plattformen, in denen Vertreterinnen und Vertreter aus Wirtschaft, Industrie, Wissenschaft und Zivilgesellschaft zusammenarbeiten, sind ein Instrument zur Umsetzung der „Digitalen Agenda“ der Bundesregierung, zuletzt geprüft am 02.07.2017, https://www.digitale-agenda.de/Webs/DA/DE/Home/home_node.html.

zeige sich eine durchgehende Datennutzung, beginnend mit dem datengetriebenen Formulieren von Hypothesen über die Kohortenrecherche, die Rekrutierung von Patientinnen und Patienten, die auf der prospektiven Arbeit mit Daten aufsetze und die Übernahme von Daten in die Case Reports bis hin zur Auswertung der Daten, ihrer Visualisierung und digitalen Langzeitarchivierung. Solches Arbeiten erfordere eine Vielfalt interdisziplinärer Kompetenzen wie die Erschließung von Rohdaten, deren semantische Interpretation, die systematische Betrachtung der Datenqualität, die Bild- und Signalverarbeitung, bioinformatische Analyse der molekularbiologischen Details, Statistik und Visualisierung, aber auch Kenntnisse in rechtlichen und ethischen Fragen. Da in der Medizininformatik nur wenige Ärztinnen und Ärzte arbeiten würden, da ein deutlicher Mangel qualifizierten Personals im Datenmanagement zu verzeichnen sei und da unter den bislang etablierten Data-Science-Studiengängen nur einer einen Schwerpunkt in der Medizin setze, biete die Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. (TMF) in jedem Jahr eine drei Tage dauernde TMF School an, die als permanente Nachwuchsförderung gelten könne. Neben einigen feststehenden Inhalten, etwa zu Fragen von Ethik und Datenschutz, würden regelmäßig wechselnde Kernthemen geschult. So hätten in vergangenen Jahren Kurse zur Systembiologie, zum Next Generation Sequencing oder zu klinischen Studien stattgefunden. Wichtig für die Schulung seien praktische Übungen, was das Vorhandensein der dazu nötigen Infrastruktur voraussetze, sowie die Möglichkeit zur Vernetzung der Kursteilnehmerinnen und -teilnehmer, die jeweils zu etwa drei Vierteln aus Universitäten bzw. Universitätskliniken kämen. Die Durchführung eines Kurses schlage heute mit ca. 17.000 € zu Buche, die für Miete, Vortragende und Reisekosten erforderlich seien; über Gebühren könnten von diesem Betrag ca. 12.000 € erwirtschaftet werden. Dem großen Bedarf nach Qualifizierung, der sich insbesondere mit Blick auf das 2018 anlaufende Förderkonzept Medizininformatik ergebe, könnten die TMF Schools mit ihren jeweils 25 Teilnehmerinnen und Teilnehmern allerdings nicht gerecht werden. Ein stärkerer Ausbau entsprechender Schulungen müsse auch im Blick haben, wie eine bessere „Datenkultur“ geschaffen werden könne, z.B. durch gezielte Anreize dafür als Datenexpertin oder Datenexperte zu arbeiten.

Yuri Demchenkov (Universität Amsterdam) bereicherte die Veranstaltung mit einem Bericht über ein EU-gefördertes Vorhaben. Das „EDISON Data Science Framework“[3] sei vor dem Hintergrund der Entwicklungen zu einer datengetriebenen europäischen Wirtschaft aufgesetzt worden. Nicht nur der in den Jahren von 2013 bis 2017 zu verzeichnende Anstieg von Stellen um 5,7 % belege die Notwendigkeit professioneller Ausbildung, sondern auch die im Bericht der High Level Expert Group zur „European Open Science Cloud“ (EOSC) formulierte Abschätzung, dass für jeweils 20 Forschende ein Data Steward tätig sein müsse. Daraus resultiere die Notwendigkeit, künftig über 80.000 Data Stewards zu beschäftigen. Mit dem im EDISON-Projekt entwickelten Framework biete sich eine Möglichkeit, die für unterschiedliche Berufsfelder einer Datenwissenschaft erforderlichen Anforderungen und Kompetenzen standardisiert zu beschreiben. Jede Datenwissenschaftlerin und jeder Datenwissenschaftler müsse Kernkompetenzen aus drei unterschiedlichen Feldern kombinieren, um methodische Fragestellungen entwickeln und bearbeiten zu können. Diese Felder seien „Data Analytics“ (wozu statistische Methoden oder maschinelles Lernen gehörten), „Data Engineering“ (z.B. Software, Tools, Infrastruktur) sowie „Domain Knowledge and Expertise“ (disziplin-spezifische

3 *EDISON. Building the data science profession*, zuletzt geprüft am 02.07.2017, http://edison-project.eu/.

Kenntnisse). Zusätzlich erforderlich seien Kenntnisse über das Datenmanagement ebenso wie über Geschäftsprozesse; gerade letztere seien unverzichtbar als Befähigung zu erkennen, welche Wertschöpfung auf der Grundlage von Daten erfolgen könne. Das „Competence Framework" biete die Möglichkeit, ein modernes Curriculum inklusive der nötigen Lernziele zu erstellen, wobei danach differenziert werden könne, ob Kurse sich eher an Anfänger oder bereits weit fortgeschrittene Experten wenden. Zugleich sei das Framework so beschaffen, dass die ganze Klasse unterschiedlicher und vielfältiger Berufe im Forschungsdatenumfeld adressiert werden könne. Auch könne das Framework genutzt werden um festzustellen, in welchen Feldern einzelne Lernende noch besonderen Entwicklungsbedarf haben. Kompetenz im jeweiligen Berufsfeld resultiere daraus, dass die im Training erlernten Fähigkeiten angewendet werden könnten und aus der routinierten Anwendung der jeweiligen Tätigkeiten hinreichende Erfahrung entstehe. Das Projekt werde zwar bald auslaufen, doch habe sich die Universität Amsterdam verpflichtet, die bisher erzielten Ergebnisse weiter zu pflegen.

Dirk Hecker (Fraunhofer-Institut für intelligente Analyse- und Informationssysteme IAIS) betonte, dass Daten zusehends als strategisches Asset zu begreifen seien; die unternehmerische Bedeutung von Daten zeige sich unter anderem daran, dass in Vorständen US-amerikanischer Firmen zunehmend Chief Data Officer berufen würden – eine in Deutschland bislang kaum besetzte Rolle. Die Entwicklung von Smart Home Devices und des Internets der Dinge werde weiter zu einem massiven Datenzuwachs beitragen; viele Entwicklungen seien durch die Anwendung von Open-Source-Tools getrieben, was auch der intensive Einsatz der Programmiersprachen R und Python belege. Das IAIS biete zwei- bis fünftägige Fortbildungen an, die sich dem Auftrag der Fraunhofer-Gesellschaft entsprechend in erster Linie an die Wirtschaft wenden. Ob die Kurse vom Führungspersonal oder bereits von Datenprofis besucht würden, die Rahmenbedingungen für die bis zu 12 Teilnehmer und Teilnehmerinnen pro Schulung seien stets gleich: ein knappes Zeitbudget, konkrete Fragen und Probleme aus dem Arbeitsalltag der zu Schulenden und der Einstieg erfolge häufig über die Werkzeuge, mit denen Daten bearbeitet würden. Am stärksten nachgefragt würden die Angebote zu Big Data Architecture, Big Data Analytics sowie Basic Data Analytics. Inzwischen habe die Fraunhofer-Gesellschaft eine Zertifizierungsstelle gegründet, so dass Kursteilnehmende nach erfolgreichem Bestehen anspruchsvoller Klausuren Zertifikate für ein Basic Level, Advanced Level oder Senior Level als Data Scientist erwerben könnten. Hecker wünschte sich eine Durchlässigkeit der IAIS-Zertifizierung zu universitären Abschlüssen und wies darauf hin, dass die Industrie ihre Fachleute für Daten aktuell vornehmlich in der eigenen Branche, in zweiter Linie über die Suche bei anderen Branchen finde und erst danach an den Universitäten suche.

Ergänzende Impulsvorträge

In sieben Impulsvorträgen wurden die breiten Ausführungen der Keynotes um weitere Facetten bereichert. Welche Kompetenzen zum Umgang mit Forschungsdaten vonnöten seien, wie diese adäquat vermittelt werden könnten und welche besonderen Rahmenbedingungen zu beachten seien, wurde dabei sowohl aus Sicht einzelner Forschungsvorhaben als auch aus institutioneller Perspektive beleuchtet.

Jens Dierkes (Staats- und Universitätsbibliothek Göttingen) betonte die Notwendigkeit der Koordination vieler Beteiligter, um datenbezogene Kompetenzen an einem universitären Standort zu

vermitteln. Die Universität Göttingen befasse sich schon seit gut 15 Jahren mit der Digitalisierung und habe einen Schwerpunkt im Bereich der Digital Humanities. Zur Kompetenzvermittlung würden insbesondere Studierende und der wissenschaftliche Nachwuchs angesprochen – mit vielen und unterschiedlichen Formaten wie z.B. Beratungen, Coffee Lectures, E-Learning-Modulen, Workshops und Sommerschulen. Oft stehe die Sensibilisierung für das Datenmanagement im Vordergrund. Die inhaltlichen Bedarfe seien über Befragungen von Lehrenden und Studierenden erhoben worden, die vielen Aktivitäten im Rahmen eines vom Präsidium der Universität geförderten Projekts umgesetzt, das nach vier Jahren nun mit Blick auf eine mögliche Fortsetzung evaluiert werde. Als Entscheidungsträger seien auch die Dekaninnen und Dekane wichtig für den Erfolg der Göttinger Aktivitäten, weil sie Personen für die Mitwirkung in der AG E-Research benennen und zudem als Multiplikatoren in die Fakultäten wirken könnten.

Der Frage, wie eine mittelgroße Universität Studierende und Forschende digital qualifizieren könne, widmete sich *Torsten Eymann* (Universität Bayreuth). Als Prorektor wolle er einzelne Fachgebiete dazu motivieren, sich dem Thema Big Data weiter zu öffnen. Fehlende Anreize für die Beschäftigung mit diesem Gebiet seien dabei ein Haupthindernis: Da der Umgang mit Forschungsdaten in den Curricula nicht vorgesehen sei, spiele das Thema in der Lehre keine Rolle und sei nicht klausurrelevant. Um Interesse zu wecken, setze die Universität Bayreuth gezielte Anreize, die auf die unterschiedlichen Gruppen zugeschnitten seien: für Professorinnen und Professoren seien Preise für die digitale Lehre ausgelobt, der wissenschaftliche Nachwuchs könne eigene Zertifikate erwerben und zur Unterstützung der digitalen Lehre könne auf E-Tutorinnen und E-Tutoren rekurriert werden. Angebote zum Thema Data Science würden von der Biologie angenommen, in der theoretischen Physik sei das Interesse am Höchstleistungsrechnen sehr stark. Die Erfahrungen bisher würden zeigen, dass Lehrende zur Kompetenzvermittlung oft Ideen und Anregungen „von außen“ aufgreifen – was erneut belege, dass heute übliche Arten der Kompetenzvermittlung noch allzu oft als Workaround aufgesetzt seien.

Jörn Ungermann (Forschungszentrum Jülich) führte aus, welche Anforderungen der Umgang mit Daten aus erdbeobachtenden Satelliten mit sich bringt. Da in der Atmosphärenforschung kontinuierlich sehr hohe Datenmengen anfallen, stellten sich beträchtliche Herausforderungen an die mathematische und physikalische Modellierung, um Rohdaten in einer Prozesskette qualitätsgesichert und fehlerfrei zu Spektren und ggf. auch anderen Produkten zu verarbeiten. Ziel dieser Arbeiten sei, Modelle für Vorhersagen zu Wetter- und Klimaereignissen nutzen zu können. Deshalb müssten die Daten zugleich langfristig aufbewahrt werden, was angesichts der exponenziell anwachsenden Datenmengen eine besondere Herausforderung auch für die Archivierung darstelle. Das Kompetenzprofil für in der Atmosphärenforschung aktive Datenwissenschaftlerinnen und Datenwissenschaftler umfasse vor allem Statistik, numerische Analysis und maschinelles Lernen, daneben seien Kenntnisse über Datenstrukturen, Datenkuratierung und professionelle Software-Entwicklung erforderlich. Da es nicht immer möglich sei, all diese Kompetenzen in einer Person zu finden, sei es wichtig, ein Team so zu besetzen, dass die unterschiedlichen, jeweils erforderlichen Expertisen vertreten seien.

Dass Bibliotheken eine besondere Rolle beim Forschungsdatenmanagement zukomme, betonte *Jan Brase* (Staats-und Universitätsbibliothek Göttingen). Es sei nämlich ureigene Aufgaben von Bibliotheken, Informationen für die wissenschaftliche Arbeit bereitzustellen. Auch Daten seien Information,

woraus Brase folgerte, dass Bibliotheken in diesem Bereich aktiv werden müssten. Als Beispiele für einschlägiges Engagement führte Brase die Verschlagwortung von Videofilmen, die Suche nach visualisierten chemischen Strukturen, die Unterstützung bei der Netzwerkanalyse textueller Information und den Nachweis von Forschungsdaten in bibliotheksseitig gepflegten Publikationslisten von Forschenden an. Der Einwand eines Zuhörers, dass die Pflege gerade komplexer Datensätze die Entwicklung und Kuratierung von Forschungssoftware erfordere, was eher in die Zuständigkeit von Rechenzentren falle, führte zur Feststellung, dass „Datenwissenschaft" als Querschnittsaufgabe anzusehen und es somit essenziell sei, die unterschiedlichen Akteurinnen und Akteure mit ihren je spezifischen Kompetenzen dialogfähig zu machen.

Auf die besonderen Möglichkeiten, die das Agieren in einem Verbund biete, wies *Joachim Schachtner* (Universität Marburg) hin. So ermögliche ein Verbund den raschen Austausch von Erfahrungen ebenso wie das Setzen fachspezifischer Schwerpunkte bei den im Verbund engagierten Partnern. Mit dem Aufbau einer hessischen Forschungsdateninfrastruktur solle dieser Gedanke umgesetzt werden. Zudem biete das BMBF-geförderte Projekt „Forschungsdatenkurse für Studierende und Graduierte" (FOKUS) den hessischen Hochschulen die Chance, die künftige Generation von Forscherinnen und Forschern früh zu adressieren. Um einschlägiges Wissen und Kompetenzen zu vermitteln, solle in diesem Projekt insbesondere untersucht werden, welche Fertigkeiten fachbezogen oder generisch vermittelt werden müssten, was didaktisch gut funktioniere und welche institutionellen und curricularen Rahmenbedingungen erforderlich seien. Schließlich solle auch analysiert werden, welchen Beitrag die geplante Kompetenzvermittlung dazu leiste, an der Hochschule selbst ein qualitativ gutes Forschungsdatenmanagement aufzusetzen. Das Projekt, für das gerade Personal gesucht werde, werde im Sommer 2018 starten. In Orientierung an existierenden Konzepten seien schon Überlegungen dazu angestellt worden, welche Module für die Schulungen entwickelt werden müssten.

Timo Dickscheid (Forschungszentrum Jülich) führte aus, dass für die Kartierung von Hirnarealen, die auf der Grundlage von Gewebeschnitten menschlicher Gehirne erfolge, vor allem Anwendungen aus dem Bereich des maschinellen Lernens zum Einsatz kommen. In der Praxis habe es sich bewährt, für die Definition des methodischen Vorgehens sowie für die Erhebung und Modellierung der Daten Fachwissenschaftlerinnen und Fachwissenschaftler mit ihrem spezifischen Domänenwissen eng mit Datenanalysten zusammenarbeiten zu lassen. Die konkrete Entwicklung der auf Höchstleistungsrechnen basierenden Auswertungen werde eher getrennt davon von einschlägig ausgewiesenen IT-Fachkräften vorgenommen. Um dem großen Bedarf an Spezialisten für das Deep Learning gerecht zu werden, habe Jülich gemeinsam mit der Fachhochschule Aachen einen Bachelor of Science aufgesetzt. Nötig seien auch erfahrene Software-Entwicklerinnen und -Entwickler, die Wissen über den Aufbau skalierbarer Systeme hätten, sowie Fachleute für das Metadaten-Management.

Im letzten Impulsvortrag wies *Stefan Liebig* (Universität Bielefeld) auf den besonderen Charakter sozialwissenschaftlicher Daten hin, der sich vor allem mit Blick auf deren Schutzwürdigkeit ergebe. Nicht zuletzt die Akkreditierung von Forschungsdatenzentren habe eine Entwicklung begünstigt, in deren Folge die Sekundäranalyse in den Sozialwissenschaften immer stärker neben die Primärerhebung von Daten trete. Deshalb seien nicht nur Kompetenzen für die Erstellung und Analyse eigener Daten, sondern zunehmend auch Methodenwissen für die sekundäre Auswertung von Daten erforderlich,

um einschlägige Qualitätskriterien adäquat bedienen zu können. Methodenwissen beziehe sich dabei auf die Erhebung, Aufbereitung und Verknüpfung von Daten, aber auch auf rechtliche Rahmenbedingungen und Techniken wie Statistik oder Textanalyse. Zur Vermittlung der jeweiligen Expertise seien Hands-on Sessions am erfolgversprechendsten: eine Bearbeitung konkreter, datenbezogener Fragen am Computer, wie sie etwa in kooperativ von Infrastruktureinrichtungen wie dem Rat für Sozial- und Wirtschaftsdaten oder dem Leibniz-Institut für Sozialwissenschaften GESIS gemeinsam mit Universitäten durchgeführten Formaten erfolge, sei der richtige Ansatz.

Diskussion und Ausblick

In der Abschlussdiskussion wurde deutlich, dass die Schnittstellen zwischen akademischen Fachgebieten und IT-Kenntnissen ausgelotet und gestaltet werden müssen. Die sogenannten Bindestrich-Informatiken hätten durchaus das Potenzial, Kenntnisse in beiden Bereichen gut zu vermitteln. Vielfach sei es aufgrund von Größe und Komplexität der unterschiedlichen Anforderungen jedoch notwendig, die je erforderlichen Kompetenzen auf einzelne Mitglieder eines Teams oder eines Verbunds aufzuteilen. In diesem Fall sei es unerlässlich, die Kommunikationsfähigkeit zwischen den unterschiedlichen Beteiligten zu verbessern.

Da profunde Kenntnisse zur Analyse und Bearbeitung von Daten in Projekten benötigt werden, wurde zudem angeregt, dass die z.B. für die Bearbeitung eines Forschungsprojekts gewährten Fördermittel zu Beginn eben dieses Projekts gezielt verwendet werden dürften, um einschlägige Fähigkeiten zu vermitteln. Sollte ein solcher Ansatz etwa aus zuwendungsrechtlichen Gründen nicht möglich sein, müsste eine anderweitige Finanzierung gefunden werden, um notwendige Schulungen unmittelbar vor Anlaufen eines Projekts durchführen zu können.

Alle, die am Workshop teilnahmen, waren sich einig darin, dass eine Sensibilisierung für die Relevanz gründlicher Kenntnisse über Forschungsdaten möglichst früh beginnen müsse. Damit komme vor allem den Universitäten eine besondere Rolle zu. Da Universitäten nach Disziplinen strukturiert seien, sollte auch die Vermittlung einschlägiger Kompetenzen zunächst von den jeweiligen Fachgebieten ausgehen und bereits in Tutorien, in Promotionsseminaren oder in regulären Treffen von Arbeitsgruppen erfolgen. Allerdings müsse durchdacht werden, ob die jeweils durchgeführten Schulungsmaßnahmen auch im Interesse der Universität lägen, da eine dezidiert über-lokale Ausrichtung solcher Maßnahmen aufgrund beihilferechtlicher Bestimmungen problematisch sein könne.

Die Erträge des Workshops wurden inzwischen für die Ausschreibung eines Ideenwettbewerbs genutzt.[4] Hier sollen Konzepte für datenbezogene Schulungsmaßnahmen eingereicht werden, die perspektivisch als Module in ein sich entwickelndes Curriculum für Berufsfelder in einer Datenwissenschaft integriert werden können. Die fünf besten Konzepte sollen mit bis zu 20.000 € prämiert werden und so in die Umsetzung gelangen.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2017H3S88-93

4 „Ideenwettbewerb zur ‚Wissenschaft im digitalen Wandel'", zuletzt geprüft am 14.09.2017, http://www.wissenschaft-im-digitalen-wandel.de/wissdw/DE/Home/ideenwettbewerb.pdf?__blob=publicationFile.