25. - 27. MÄRZ 2015

HERRENHÄUSER KONFERENZ
SCHLOSS HERRENHAUSEN, HANNOVER

TAGUNGSBERICHT

Big Data in a Transdisciplinary Perspective

Veranstalter: VolkswagenStiftung zusammen mit Dietmar Harhoff (München), Thomas Lippert (Jülich), Volker Markl (Berlin), Arnold Picot (München), Ralph Schroeder (Oxford) sowie Amir Zeldes (Georgetown University)

Janusköpfig ist das Gesicht von Big Data in der Öffentlichkeit: Während die Einen in Big Data euphorisch die Chancen einer Industrie 4.0 feiern, fürchten die Anderen die totale Überwachung des Einzelnen und letztlich das Ende der Demokratie. Auch in den Wissenschaften ist Big Data ein „Buzzword", das Menschen und Millionen Fördergelder mobilisiert – doch es fällt auf, dass bei den vielen Konferenzen zu diesem Thema jede Disziplin quasi unter sich bleibt. Mit ihrer Veranstaltung „Big Data in a Transdiciplinary Perspective" wollte die VolkswagenStiftung hier Abhilfe schaffen und lud nach Schloss Herrenhausen – und damit, wie der Generalsekretär der Stiftung **WILHELM KRULL** (Hannover) in seiner Begrüßung ausführte, genau an den Ort, an dem der große Universalgelehrte Gottfried Wilhelm Leibniz wirkte, die erste Rechenmaschine erfand und das Binärsystem mit 0 und 1 entwickelte. „Calculemus!", „Lasst uns rechnen!" soll Leibniz einmal ausgerufen haben.

In ihrem Eröffnungsvortrag **„Data, Scholarship and Disciplinary Practice"** spannte die US-amerikanische Soziologin und Informationswissenschaftlerin **CHRISTINE BORGMANN** (UCLA) [1] einen weiten Bogen von der notwendigen Klärung des Begriffs „Big Data" über die Nutzung von Daten in der Forschung heute bis zu den vielfältigen Problemen der Datennutzung in der Forschungspraxis. Ausgangspunkt bei Borgman – wie bei vielen anderen Sprecher(inn)en der Konferenz – war Douglas Laney's berühmte Bestimmung der Charakteristika von Big Data als die „Drei V": Volume, Velocity, and Variety [2]. Forschungsdaten definierte Borgman als "representations of observations, of objects, or other entities used as evidence of phenomena in research". Mit Hinweis auf Tony Heys Buch "The Fourth Paradigm" [3] rief Borgman das Versprechen einer mit Big Data verknüpften neuen

wissenschaftlichen Blütezeit auf, um gleichzeitig darauf hinzuweisen, dass auch in der Astronomie als kanonischer „Big Science“ die Datenmengen oft in kleine Einheiten zerschnitten werden, um dann analysiert zu werden: In Wirklichkeit bestehe zwischen Big Science mit großen Instrumenten, hohen Kosten, vielen Mitarbeitern und verteilter Arbeit auf der einen Seite und Little Science mit kleinen Geräten, geringen Kosten, kleinen Teams und lokaler Arbeit auf der anderen Seite kein grundsätzlicher Unterschied. Und dann führte Borgmann die lange Liste der Probleme beim Datenaufbau und der -nutzung in der Wissenschaft an: Von den fehlenden Anreizen, Daten weiterzugeben, über die Schwierigkeiten, bestehende Daten zu nutzen („Data is noise for another discipline“) und weiter aufzubereiten, bis hin zur Klärung der Rechtefragen. Doch das größte Problem sei zweifelsohne die fehlende Ordnung der Infrastruktur: Die Repositorien seien oft genug nicht institutionell abgesichert und damit nicht nachhaltig. Statt Big Data, so Borgman lapidar, drohe vielfach ein Zustand von No Data.

CLIFFORD A. LYNCH von der Coalition for Networked Information CNI (Washington DC) legte den Fokus seines Vortrags auf **“The Challenges of Data Reuse: The Short and the Long Term”**. Dass einmal generierte Daten weiter vorgehalten werden müssen, steht für ihn außer Frage – auf kurze Sicht, um Untersuchungen reproduzieren zu können, und auf lange Sicht, zum Beispiel durch Rekombination und Reannotierung, um neue Fragen zu beantworten. Er sieht in der Wiederverwendung von Daten sogar einen zentralen Aspekt von Big Data. Doch bei der vorhandenen Geschwindigkeit der Datengenerierung sowie der parallel stattfindenden technischen Entwicklung sei es nicht möglich, tatsächlich alle Daten in die nächste Gerätegeneration zu migrieren. Hinzu komme, dass viele Repositorien nicht institutionell verstetigt sind. Also müsse man eigentlich überprüfen, welche Daten weiter benötigt würden und wie sie mit möglichst zeit- und kulturunabhängigen Metadaten vorgehalten werden sollten. Lynch schloss mit der Forderung nach einer multidisziplinären Diskussion über die Kriterien für die Datengenerierung, -aufbereitung und -vorhaltung, letztlich nach einer neuen Archivwissenschaft im digitalen Zeitalter.

Auch **PETER WITTENBURG** von dem europäisch-amerikanisch-australischen Netzwerk Research Data Alliance (Nimwegen) warb in seinem Vortrag „Data Science: Practices and Ambitions“ für gemeinsame Anstrengungen, die vielen Schwierigkeiten bei der gemeinsamen Datenarbeit zu überwinden. Denn – und hier zitierte Wittenburg die berühmte Metapher des Sheffielder

Mathematikers Clive Humby von 2006 – Daten sind das neue Öl, auch für die Wissenschaft. Aber der derzeitige Umgang mit Daten sei viel zu kostenintensiv und zu ineffizient. Arbeit an den Daten werde oft manuell statt automatisch durchgeführt mit entsprechenden steigenden Kosten. Es gebe immer noch Daten, die ohne Persistent Identifier generiert würden – und die damit schon bei ihrer Entstehung de facto Altdaten darstellten. Wittenburg entwarf demgegenüber das Modell einer „Data Fabric“, in der alle Momente des Korpusaufbaus aufeinander abgestimmt sind. Aber er gestand ein: Keiner weiß, was in nur zehn Jahren sein wird.

Mit dem Vortrag von **ANDREW PRESCOTT** (Glasgow) [4] beleuchtete die Konferenz die Situation von **„Big Data in the Arts and the Humanities“**. Prescott ist Mediävist und sozusagen ein Digital Humanist der ersten Stunde. Nach einer Tour d’Horizon der in Großbritannien geförderten großen Digital Humanities-Projekte stellte er sich die Frage, ob Big Data einfach nur ein Mehr an Daten oder eine substanzielle Veränderung der Wissensstruktur darstelle. Hier helfe die historische Erfahrung: Ähnlich wie die Verschriftlichung unserer Kultur ab dem 11. Jahrhundert im „Domesday Book“, dem ersten Grundbuch von England, zu einer Reorganisation der Wissensstrukturierung führte, werde Big Data unsere Wirklichkeit substanziell verändern. Im Gegensatz zur Wissenschaft, die auf Kausalitäten beruhe, seien es nun Korrelationen, die die Grundlage von Big Data darstellten und darüber hinaus Vorhersagen über die Zukunft ermöglichten. Durch Big Data würden die Geisteswissenschaften auf jeden Fall visueller, haptischer und explorativer. Die große wissenschaftstheoretische Herausforderung sah Prescott in der Entwicklung eines theoretischen Rahmenwerks, das er als "critical data studies“ bezeichnete: „Big Data needs Big Theory!“ Ziel müsse eine „humanization of Big Data“ sein. Denn Daten sind nicht Gegebenheiten, sondern werden erst aus der Beobachtung heraus gewonnen. Prescott zitierte den Glasgower Archäologen Jeremy Huggett: “Data are theory-laden, and relationships are constantly changing, depending on context” [5] und listete dann den Sieben-Punkte-Katalog von Craig Dalton und Jim Thatcher der Critical Data Studies [6] auf, darunter: Daten müssten in Zeit und Raum verortet werden; sie müssten als inhärent politisch und interessegeleitet verstanden werden, sie könnten nie für sich selber sprechen und „Rohdaten“ könne es in diesem Sinne nicht geben.

Einen überraschend tiefen Einblick in die Industrie gab **STEPHAN FISCHER** (Ditzingen) in seinem Vortrag über die Trumpf Werkzeugmaschinenbau GmbH mit dem Titel **„Data-Value Services as a**

Differentiator for Machine Tools“. Als ehemaliger Abteilungsleiter bei SAP wechselte er Anfang 2014 als neuer IT-Direktor in das sich aktuell auf Lasertechnik stützende Industrieunternehmen, um es in das vernetzte digitale Zeitalter zu führen. Sei es in einem ersten Schritt darum gegangen, die physische mit der virtuellen Welt zu verknüpfen und beispielsweise mit Sensoren die Qualität der Lasernadel zu prüfen („smart data“), gehe es derzeit um die Optimierung des gesamten Produktionssystems („smart factory“) anhand der massenhaft erzeugten Daten und des Maschinellen Lernens – die Zukunft jedoch werde darin liegen, das Internet of Services als Business Modell zu entwickeln. Bei dem Prozess der Digitalisierung müssten essentielle Fragen geklärt werden, z. B. wie die analogen Daten in digitale Form umgewandelt, wie Daten verwaltet werden und wie Daten sicher vom Kunden zu Trumpf gelangen – oder von Trumpf zu wissenschaftlichen Institutionen. Im „Smart Data Innovation Lab“ arbeiten Trumpf und andere Partner aus der Wirtschaft mit der Wissenschaft zusammen, zum Beispiel um zu berechnen, für wann mit einer Wartung zu rechnen ist. Von dem Datenaustausch mit der Wissenschaft, so Fischer, erhoffe man sich strategische Vorteile.

Auch der Leiter des Instituts für Arbeitsmarkt- und Berufsforschung **STEFAN BENDER** (Nürnberg) sieht in der Datenfreigabe für die Wissenschaft einen Vorteil für die Planung von zukünftigen politischen Maßnahmen wie auch mittelbar für das Branding von Deutschland. In seinem Vortrag **„Researcher Access, Economic Value and the Public Good“** forderte er die Entwicklung von Dokumentations-standards, die Definition von Datenreproduzierbarkeit und vor allem einen geeigneten Umgang mit Fehlern bei Big Data. Weiterhin führte Bender die Unterscheidung von “made data”/“designed data“ und “found data”/“organic data“ ein, die aber nicht miteinander konkurrierten, sondern zusammengeführt werden könnten und sich daher ergänzten. Denn Big Data sei zwar billiger in der Generierung, nicht aber in der Bereinigung. Bender interpretierte die bekannte Öl-Metapher noch einmal neu: Daten könnten auch großen Schaden wie eine Ölpest verursachen.

Für den Physiker mit Soziologielehrstuhl **DIRK HELBING** (Zürich) gibt es zurzeit ein Ungleichgewicht zwischen den Erkenntnissen, die wir über die Natur und die wir über unsere Gesellschaft haben. Er stellte sich die Frage: **„How we can build a smart resilient digital society?”** [7]. Big Data könne uns dabei helfen, dieses bestehende Ungleichgewicht zu beseitigen. Helbing stellt sich hierfür eine Welt vor, mit vielen verteilten und selbstorganisierten Systemen und einer dezentralisierten Kontrolle bzw. Intelligenz, die auf Grundlage der Daten entscheidet. Ein solches „Planetarisches Nervensystem“

zusammen mit einem „Living Earth Simulator“, der verschiedene Änderungen und Einflüsse auf der Welt simulieren könnte, wäre seiner Meinung nach im Stande, grundlegende Einsichten in unsere Gesellschaft zu enthüllen. Gleichzeitig machte Helbing auf das Verfallsdatum von Daten aufmerksam, da bestimmte Datensätze nach kurzer Zeit jeglichen Wert verlieren. Hierzu gehörten bestimmt auch einige der Twitter-Nachrichten, die parallel zur Konferenz unter dem Hashtag **#HKBigData** verschickt wurden.

Einen technischen Blick auf Big Data lieferte **SHIVAKUMAR VAITHYANATHAN** von IBM Big Data Analytics (San José), der zunächst drei unterschiedliche Big Data-Problemstellungen vorstellte: 1) Fragestellungen mit einer schier riesigen Datenmenge, 2) Fragestellungen, die mit einer großen Anzahl an Modellen, die verschiedene Aspekte abdeckten, behandelt würden, und 3) Fragestellungen, bei denen nur geringe Mengen an Daten vorhanden seien, aber bei der anhand von Simulationen eine riesige Datenmenge erzeugt werde. Diesen Herausforderungen begegneten aktuell Datenwissenschaftler (Data Scientists), die aus der Menge von Daten Erkenntnisse extrahierten. Hierzu müsse der Datenwissenschaftler beide Welten kennen (die noch „normale“ IT Welt und die „Big Data“ Welt) und zwischen beiden Welten vermitteln und übersetzen. Das große Ziel von Big Data Analytics sei daher, eine solche Übersetzung automatisch durchzuführen und somit die Ideen des Datenwissenschaftlers automatisiert in die Welt der Softwareumgebung Hadoop und Co. umzuwandeln.

In mehreren Zeitfenstern stellten bei der Herrenhäuser Konferenz insgesamt 29 Nachwuchswissenschaftler(innen) aus 16 Ländern, für die die Stiftung Reisestipendien zur Verfügung gestellt hatte, in dreiminütigen **Lightning Talks** ihre Forschungsprojekte aus verschiedenen wissenschaftlichen Disziplinen vor. Ihre Vorträge und die Poster wurden auf Basis von Voten der Konferenzteilnehmer am Ende der Herrenhäuser Konferenz prämiert. Für die beste Präsentation wurde der Historiker **IAN MILLIGAN** (University of Waterloo) für seine Darstellung des Projekts **„Finding Community in the Ruins of GeoCities“** ausgezeichnet, für das beste Poster der Sozialwissenschaftler **JOSH COWLS** (Oxford) für **„Using Big Data for Valid Research: Three Challenges“**.

Eine sehr lebhafte Sektion der Konferenz war juristischen Fragen gewidmet. Big Data sind letztlich Daten, für die seitens der Wissenschaft keine informierte Einwilligung des Einzelnen („informed

consent“) eingeholt worden ist oder eingeholt werden kann. Hier setzte die Wirtschaftswissenschaftlerin **JULIA LANE** (Straßburg/Melbourne) [8] in ihrem Vortrag **„Big Data, Science Policy, and Privacy“** an. Man müsse sich erst einmal bewusst machen, dass man mit Big Data auch zu völlig falschen Ergebnissen kommen könnte – eine These, die Julia Lane mit den Ereignissen rund um den Bombenanschlag von Boston verdeutlichte, bei dem ein durch Big Data-Analysen unschuldig verdächtigter Mann, weil öffentlich der Tat bezichtigt, Selbstmord beging. Dies hat zudem ein Rechtsproblem aufgeworfen: „What is the legal framework for found data on human beings?“ Die informierte Einwilligung, die in den USA in der sogenannten Common Rule zum Schutz von menschlichen Forschungssubjekten festgelegt ist, sei heute eine Fiktion, da in Zeiten von Big Data keine Anonymisierung von Daten mehr möglich sei. Der Einzelne habe oft keine Ahnung, welche Daten alle von ihm gespeichert seien und dass er über deren Zusammenführung jederzeit identifizierbar sei. Doch wie dann weiter sozialwissenschaftliche Forschung durchführen? Julia Lane forderte einen Runden Tisch, an dem Wissenschaft, Förderorganisationen und öffentliche Hand eine Roadmap entwerfen, um gemeinsam dieses Problem anzugehen.

Dass heute keine informierte Einwilligung mehr möglich ist, diese Ansicht teilte auch der deutsche Jurist **THOMAS HOEREN** (Münster) in seinem Vortrag **„From Alibaba to Abida. Legal Issues concerning Big Data“**. In Zeiten von Big Data gebe es fast keine nicht-persönlichen Daten mehr. Er bezeichnete die deutsche Rechtsprechung zur Schufa als das erste richtige Big Data-Gesetz, da es erstens wissenschaftliche Standards bei dem Datenumgang und zweitens Transparenz festschreibe: Jeder Bürger hat jederzeit das Recht, Auskunft über die dort über ihn gespeicherten Daten zu bekommen. Ansonsten warf Hoeren viele Fragen auf: Wer haftet für falsch erhobene Daten? Gibt es ein Eigentumsrecht an Daten und wenn ja, wem gehört was? Wie sieht es mit den Persönlichkeitsrechten aus? Welche Rolle spielen die beiden großen Rechtstraditionen, das angelsächsische Common Law und das Römische Recht, beim Umgang mit den Daten? Big Data, so Hoerens Fazit, wird das gesamte Gesetzeswerk verändern. Im vom BMBF geförderten Begleitprojekt „Assessing Big Data“ (ABIDA – daher auch der Titel seines Vortrags) an dem Hoeren beteiligt ist, werden die vielschichtigen Entwicklungen von Big Data-Anwendungen, Datenströmen und Geschäftsmodellen kontinuierlich beobachtet und erfasst.

Wie nüchtern Juristen die derzeitige Situation in Zeiten von Big Data einschätzen, machte auch **NIKOLAUS FORGÓ** (Hannover) in seinem Vortrag deutlich. Dieser trug den pointierten Titel: „**Ignore the Facts, Forget the Rights: European Principles in an Era of Big Data"**. Forgo setzte am sogenannten „Volkszählungsurteil" vom 15. Dezember 1983 an, als das Bundesverfassungsgericht in einer Grundsatzentscheidung das Grundrecht auf informationelle Selbstbestimmung als Ausfluss des allgemeinen Persönlichkeitsrechts und der Menschenwürde etablierte. Das Urteil galt als Meilenstein des Datenschutzes und ging in die Charta der Grundrechte der Europäischen Union in Artikel 7 und insbesondere in Artikel 8 (2) ein: Personenbezogene Daten dürfen „nur nach Treu und Glauben für festgelegte Zwecke und mit Einwilligung der betroffenen Person oder auf einer sonstigen gesetzlich geregelten legitimen Grundlage verarbeitet werden. Jede Person hat das Recht, Auskunft über die sie betreffenden erhobenen Daten zu erhalten und die Berichtigung der Daten zu erwirken." Doch wie ist die Wirklichkeit heute? Sie sei von Kontrollverlust des Einzelnen über „seine" Daten gekennzeichnet und damit mit Selbstverlust: „If the product is for free, you are the product". Drei Problemfelder müssten gleichzeitig und weltweit geklärt werden: Fragen des Eigentumsrechts, der Achtung der Privatheit wie auch des Urheberrechts.

Von juristischen Fragen zurück zu technischen Aspekten und zu den großen Herausforderungen für die Wissenschaft im Allgemeinen führte die letzte Session der Konferenz. In seiner Einführung sah **VOLKER MARKL** (Berlin) den zentralen Aspekt von Big Data im Aufeinandertreffen der zwei Welten, nämlich der Welt des Datenmanagements und der Welt der Datenanalyse. Darüber hinaus brachte er noch zwei weitere Eigenschaften von Daten in die Diskussion. Zum einen könnten Daten – hier setzte Markl einen anderen Akzent als beispielsweise Bender und Wittenburg – an Wert verlieren, wenn sie geteilt werden, und zum anderen werde es immer schwieriger, die riesigen Datenmengen von einem Server zum anderen zu verschieben, denn letztlich seien Daten „as elastic as a brick of stone". Daraus folgernd fügte er der mehrfach angesprochenen Öl-Metapher noch einen weiteren Aspekt hinzu, den des Kriegs um diese Ressource.

Markls anschließende ausführliche Darstellung der verschiedenen Herausforderungen für die Forschung und die Forschungsförderung stand am Beginn der Podiumsdiskussion, an der neben ihm auch **DAVID CARR** vom Wellcome Trust (London), der KI-Experte **OSCAR CORCHO** (Madrid), **JOSHUA M. GREENBERG** von der Alfred P. Sloan Foundation (New York) und **STEFAN WINKLER-NEES** von der

Deutschen Forschungsgemeinschaft (Bonn) teilnahmen. Dabei war die fehlende Infrastruktur, die von fast allen Teilnehmern bemängelt wurde, der rote Faden der Diskussion. Weitere wichtige Gesprächspunkte waren der Bedarf nach einer gut durchgeführte Datenpflege, die Stärkung der Umgangsfähigkeit mit Daten in der Gesellschaft sowie die Frage nach der Maximierung des Potenzials von bereits existierenden Daten – speziell wenn man sich die schiere Daten- und Informationsflut bei Wikipedia vor Augen hält – und die geeigneten wissenschaftlichen Fragen, welche anhand der vorhandenen Daten untersucht werden können.

Insgesamt brachte die Herrenhäuser Konferenz herausragende internationale Vertreterinnen und Vertreter der unterschiedlichen Disziplinen zusammen und bot damit eine transdisziplinäre Diskussion auf hohem intellektuellem Niveau. Darin lag der besondere Mehrwert der Veranstaltung, nicht nur, weil sie Gelegenheit zum Kennenlernen bot, sondern weil auch eine ganze Reihe von Problemen und Herausforderungen, die von allen Disziplinen nur gemeinsam gelöst werden können, identifiziert wurden – und das, obwohl Big Data ein Containerbegriff mit unscharfen Konturen ist. Auf der wissenschaftstheoretischen und -soziologischen Ebene erscheint die Forderung nach "Critical Data Studies" mit der notwendigen historisch-kritischen Einbettung von Daten wichtig. Auf der technischen Ebene scheint die Frage der Datenaufbereitung, -vorhaltung und -reproduzierbarkeit von zentraler Bedeutung. Auf statistisch-methodischer Ebene ist es der Umgang mit Fehlern bei Big Data-Analysen, der sicher die Diskussion der Zukunft beherrschen wird. In der juristischen Dimension ist überdeutlich, dass das Recht, wie es aktuell verfasst ist, nicht haltbar und das Rechtsverständnis an die neue digitale Epoche angepasst werden muss. Auf der gesellschaftlichen Ebene steht die Forderung nach der Stärkung der Umgangsfähigkeit mit Daten, der Data Literacy, im Raum. Auf der übergeordneten Ebene schließlich stellt sich die Frage, welchen Anspruch die Gesellschaft hat, dass die Daten Common Goods sind, mit denen die Wissenschaft arbeiten kann – und die Chancen von Big Data nicht nur der Internetwirtschaft überlassen werden.

Vera Szöllösi-Brenig und Christoph Kolodziejski, VolkswagenStiftung

E-Mail: Szoelloesi-Brenig@volkswagenstiftung.de; Kolodziejski@VolkswagenStiftung.de

WEITERE INFORMATIONEN

FOTOS UND AUDIOS: http://www.volkswagenstiftung.de/bigdata

VolkswagenStiftung
Kastanienallee 35
30519 Hannover
bigdata@volkswagenstiftung.de

Anmerkungen

[1] Christine L. Borgman: “Big Data, Little Data, No data. Scholarship in the Networked World“, MIT Press 2015; Christine L. Borgman and Marianne Krasny: „Scholarship in the Digital Age. Information, Infrastructure, and the Internet“, MIT Press 2007

(2] Laney, Douglas: "3D Data Management: Controlling Data Volume, Velocity and Variety" (PDF). Gartner. Retrieved 6 February 2001.

[3] “The Fourth Paradigm: Data-Intensive Scientific Discovery”. edited by Tony Hey, Stewart Tansley & Kristin Tolle, Microsoft 2009

[4] http://de.slideshare.net/burgess1822/prescottherrenhausen [7.5.2015]

[5] Jeremy Huggett: “Promise and Paradox: Accessing Open Data in Archaeology”, Proceedings of the Digital Humanities Congress 2012

[6] Craig Dalton and Jim Thatcher: “What does a critical data studies look like, and why do we care? Seven points for a critical approach to ‘big data’”, Society and Space 2014 http://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data/ [7.5.2015]

[7] https://www.youtube.com/watch?v=mO-3yVKuDXs (Helbings Vortrag auf Youtube)

[8] Julia Lane, Victoria Stodden, Stefan Bender and Helen Nissenbaum (Hg.): “Privacy, Big Data, and the Public Good: Frameworks for Engagement”, Cambridge University Press 2014