Dossier
Open Data und Big Data: Neue Herausforderungen

Open by default – der Beitrag der SAGW zu Open Data

(ib) Der freie Zugang zu wissenschaftlich relevanten Datenbeständen ist heute ebenso ein Ziel guter wissenschaftlicher Praxis wie der Open Access zu Publikationen. Gemeinsam bilden sie Elemente der Open-Science-Initiativen, die eine offene, nachvollziehbare und transparente Forschung – sowohl was die Forschungsarbeit selbst als auch deren Ergebnisse betrifft – anstreben. Insbesondere die durch die öffentliche Hand finanzierten Förderorganisationen stehen deshalb in der Verantwortung, den genannten Initiativen in ihrem Zuständigkeitsbereich zum Durchbruch zu verhelfen.

«Open Data» im grösseren Kontext

«Open Data» ist zu einem Schlagwort der «Open-Familie» geworden. Es wird in diversen Zusammenhängen verwendet, gelegentlich unter Beifügung eines weiteren Begriffs, um das Einsatzgebiet näher zu umschreiben. So meint «Open Government Data» frei zugängliche Datenbestände der öffentlichen Verwaltung, während «Open Research Data» explizit die freie Verfügbarkeit von Daten, die im Zusammenhang mit Forschung entstanden sind, zum Ziel hat. Die SAGW setzt sich mit «Open Data» im Zusammenhang mit den durch sie betreuten Forschungsinfrastrukturen auseinander. Die Akademie ist hier als grösste Trägerinstitution für geisteswissenschaftliche Forschungsinfrastrukturen in der Schweiz ebenso in der Pflicht, wie sie es bei den geistes- und sozialwissenschaftlichen Zeitschriften beim «Open Access» ist[1].

In internationaler Perspektive wird der freie Zugang zu Forschungsdaten vielerorts propagiert, so etwa in der

Dossier
Open Data und Big Data: Neue Herausforderungen

[1] Zu den Aktivitäten der Akademie betreffend Open Access siehe Beat Immenhauser, Open Access on the road, in: Bulletin der Schweizerischen Akademie der Geistes- und Sozialwissenschaften 2017, Heft 3, S. 9 ff. (http://doi.org/10.5281/zenodo.839749).

Deklaration zur «European Open Science Cloud EOSC»[2], in den Bestimmungen für die Antragstellenden beim «European Research Council»[3], im OECD-Bericht über Forschungsinfrastrukturen[4] sowie in den «Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020»[5]. Auf nationaler Ebene setzt sich vor allem der Schweizerische Nationalfonds für «Open Research Data» ein, wozu er im Oktober 2017 eine eigene Policy eingesetzt hat[6].

Open Data = FAIR Data?

Oft wird Open Data im Zusammenhang mit den FAIR Guiding Principles genannt[7]. FAIR ist das Akronym für Daten, die «**f**indable, **a**ccessible, **i**nteroperable and **r**eusable» sein sollen. Entwickelt von FORCE11, einer «grass-root» Gemeinschaft von InformationsspezialistInnen anlässlich einer Tagung, haben die FAIR-Prinzipien grosse Beachtung und Resonanz gefunden, obwohl sie keinen Standard darstellen. Barend Mons beschreibt die Prinzipien folgendermassen: «FAIR simply describes the qualities or behaviours required of data resources to achieve – possibly incrementally – their optimal discovery and scholarly reuse»[8]. Eine optimale Auffindbarkeit und Wiederverwendung von Daten kann erzielt werden, wenn Metadaten über den Datenbestand nach allgemeinen Standards vorhanden sind, wenn Daten durch Identifikatoren referenziert werden können, wenn die Daten selbst über Schnittstellen maschinenlesbar vorgehalten werden und wenn der externe Datengebrauch transparent geregelt ist. FAIR Data ist aber nicht gleichbedeutend mit Open Data. Die FAIR Principles besagen lediglich, dass die Metadaten frei zugänglich sein müssen. Es braucht also beide Konzepte – Open Data und FAIR Data – in komplementärer Weise.

Big Data = Open Data?

Auch die Gleichsetzung Big Data gleich Open Data ist nicht korrekt. Big Data – wiederum ein Schlagwort – bezeichnet grosse Datenströme, die zu bestimmten Zwecken ausgewertet werden sollen. Diese Datenströme entstehen in der Regel kontinuierlich, so dass deren Umfang nicht überschaubar ist (z.B. Daten über das Einkaufsverhalten von Kundinnen und Kunden). Dieser Umstand bringt es mit sich, dass solche Massendaten mit anderen Verfahren analysiert werden müssen, als es bei herkömmlichen, abgrenzbaren Datenbeständen der Fall ist. Diese

---

[2] European Commission (2017), EOSC Declaration: European Open Science Cloud. New Research and Innovation Opportunities, Brüssel, https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none

[3] European Research Council (2018), Open Research Data and Data Management Plans, Version 2.0, https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf

[4] OECD (2017), «Strengthening the effectiveness and sustainability of international research infrastructures», OECD Science, Technology and Industry Policy Papers, No. 48, OECD Publishing, Paris, http://dx.doi.org/10.1787/fa11a0e0-en

[5] European Commission (2017), H2020 Programme. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[6] Schweizerischer Nationalfonds (2017), Open Research Data Policy, http://www.snf.ch/de/derSnf/forschungspolitische_positionen/open_research_data/Seiten/default.aspx

[7] The FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016, Scientific Data 3:160018), https://doi.org/10.1038/sdata.2016.18

[8] Mons, Barend et al. (2017), Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud, in: Information Services & Use 37 (2017), S. 49–56, hier S. 51, IOS Press, https://doi.org/10.3233/ISU-170824

Daten fallen in unterschiedlichen Kontexten an, die eher ausnahmsweise ihren Ursprung in einem Forschungsprojekt haben. Bei Daten aus einem privatwirtschaftlichen Umfeld (z.B. Telekommunikation, Social Media, Suchmaschinen) stellt sich die Frage der Zugänglichkeit in erhöhtem Masse, aber auch Registerdaten von staatlichen Behörden sind nicht ohne Weiteres für die Wissenschaft zugänglich, gilt es doch, Forschungsinteressen gegen Datenschutzbestimmungen abzuwägen.

Beitrag der SAGW

Mit den genannten Fragen des Datenzugangs, sei es zu herkömmlichen Forschungsdaten oder sei es zu Big Data, setzen sich die Autorinnen und Autoren des Dossiers in diesem Bulletin auseinander. Sie tun dies gleichsam in Vorbereitung auf die beiden «Data»-Veranstaltungen der SAGW am 2. («Geisteswissenschaftliche Forschungsplattformen in der Schweiz im Kontext von Open und FAIR Data», Bern) und am 9. November («Big Data in den Sozialwissenschaften – Herausforderungen und Chancen», Bern)[9].

Für die durch die SAGW verantworteten sieben Forschungsinfrastrukturen[10] sowie für die zur Akademie transferierten Editionen wird die SAGW eine Open Data Policy implementieren. Dabei gelten beispielsweise die Grundsätze, dass die Daten so offen wie möglich – open by default – angeboten werden, dass sie durch persistente Identifikatoren referenzierbar sind und dass die FAIR Prinzipien angewendet werden. Ziel ist es, diesen Prozess der Policy-Implementierung bis 2020 abgeschlossen zu haben.

Weitere Informationen

Open Access in der SAGW: http://www.sagw.ch/sagw/laufende-projekte/open-access.html

Tagung Open Data: http://www.sagw.ch/opendata

Tagung Big Data: http://www.sagw.ch/de/sagw/veranstaltungen/vst-2018-sagw/bigdata.html

[9] Siehe www.sagw.ch/veranstaltungen sowie Bulletin der SAGW 2 (2018), S. 34 f. (http://doi.org/10.5281/zenodo.1222084).

[10] Zu den Forschungsinfrastrukturen der SAGW siehe Immenhauser, Beat (2017): Forschungsinfrastrukturförderung der Schweizerischen Akademie der Geistes- und Sozialwissenschaften (Swiss Academies Factsheets 12, 1), http://doi.org/10.5281/zenodo.802093

Chancen und Hürden im Datenzugang für die Forschung

Georg Lutz, Direktor FORS

Daten sind der Rohstoff des 21. Jahrhunderts, auch für die Forschung. Weil von privaten Firmen gesammelte Daten kaum zugänglich sind, ist es umso wichtiger, dass Forschende existierende Daten, wie sie etwa das Bundesamt für Statistik (BFS) in grossem Umfang sammelt, einfach für Forschungszwecke nutzen können.

Personenbezogene Daten liefern wichtige Grundlagen, um gesellschaftliche, politische und administrative Prozesse zu verstehen und zu steuern. Die Menge gesammelter Daten wächst weiterhin rasant an und Personendaten werden schon lange nicht mehr in erster Linie von öffentlichen Einrichtungen gesammelt. Inzwischen legen private Firmen wie Facebook und Google umfangreiche Datensammlungen an und nutzen diese für kommerzielle Zwecke. Diese Unternehmen sehen Daten als privates Gut und machen Daten nicht für die Forschung verfügbar, zumindest solange keine Regulierung sie dazu zwingt.

Im Gegensatz dazu ist in Verwaltung und Forschung «Open Data» Standard geworden. Bund und viele Kantone machen Daten etwa über das Portal opendata.swiss zugänglich. Freier und offener Datenzugang zur Replikation und Verifizierung von Forschungsergebnissen ist seit jeher die Grundlage akademischer Forschung. In den Sozialwissenschaften haben Repositorien wie FORS (www.forscenter.ch) eine lange Tradition, Daten zu archivieren, zu dokumentieren und zur Verfügung zu stellen.

Zugang zu öffentlichen Daten

Weil private Datensammlungen für die Forschung kaum nutzbar sind, kommt den Erhebungen des Bundes für die Forschung eine grosse Bedeutung zu. Dies gilt insbesondere für die Daten des BFS, das unzählige Datensätze und Statistiken über die Website des BFS verfügbar macht. Erhältlich sind ebenfalls viele Mikrodaten, die das BFS erhebt.

Regelmässige Forschungsvorhaben von nationaler Bedeutung sowie internationale Forschungsvorhaben, die vom SNF mitfinanziert werden, haben auch Zugang zum Stichprobenrahmen für Personen- und Haushaltsbefragungen (SRPH) des BFS. Da der SRPH auf die viermal jährlich aufdatierte Registererhebung aus den Gemeinderegistern basiert, sichern die Daten dank der herausragenden Qualität eine wichtige Grundlage für wissenschaftliche Befragungen. Zudem ist es via BFS inzwischen auch möglich, verschiedene Datenverknüpfungsprojekte zu realisieren, um damit das Analysepotenzial mittels kombinierter Datensätze erheblich zu steigern und oft auch Kosten zu sparen.

Die Herausforderungen: FAIR und Datenschutz

«Open Data» orientiert sich heute an den FAIR-Standards («Findable, Accessible, Interoperable, Reusable»). Daten müssen dafür gut aufbereitet und dokumentiert werden. Beim Zugang zu personenbezogenen Daten gibt es jedoch einen erheblichen Zielkonflikt, der schwieriger zu hand-

haben ist. Konträr zum Primat von «Open Data» stehen der Schutz der Privatsphäre und Datenschutzgesetze, was insbesondere für die von Bund und Kantonen gesammelten Daten enorm wichtig ist, um das Vertrauen in diese Institutionen zu erhalten. Dieser Widerspruch ist immer wieder eine Hürde, um Daten möglichst rasch und unbürokratisch zugänglich zu machen. Erst recht, wenn es darum geht, Daten aus verschiedenen Quellen zu verknüpfen.

Damit dies trotzdem möglich ist, braucht es die Bereitschaft zu gegenseitigem Lernen und offenem Austausch. Forschende müssen die Abläufe innerhalb der Verwaltung verstehen und eine Sensibilität für hohe Datenschutzstandards entwickeln. Forschende haben ein Bedürfnis nach raschem und unkompliziertem Datenbezug und sie benötigen eine gewisse Flexibilität in der künftigen Datennutzung. Durch die Erfahrung mit konkreten Projekten hat sich in den letzten Jahren der Zugang zu BFS-Daten laufend verbessert.

Daten des BFS FAIR zugänglich zu machen, braucht allerdings auch zusätzliche Ressourcen, entweder vom Bund oder von Forschungsförderungseinrichtungen. Da es nicht Kernaufgabe des BFS ist, die Forschung mit Daten zu beliefern, führen knappe Ressourcen immer wieder zu Verzögerungen und erschweren damit den Datenzugang in der Praxis.

Zum Autor

Georg Lutz

Georg Lutz ist Direktor des Forschungszentrums Sozialwissenschaften FORS, in das auch das schweizerische Datenarchiv für sozialwissenschaftliche Daten eingegliedert ist. Er vertritt die Schweiz in CESSDA, dem «Consortium of European Social Science Data Archives», welches auf europäischer Ebene den Zugang von Forschungsdaten koordiniert und vernetzt. Zudem ist er Professor für Politikwissenschaft an der Universität Lausanne. Er forscht und lehrt dort zu politischen Institutionen und politischem Verhalten in vergleichender Perspektive sowie zu Schweizer Politik und Umfrageforschung.

Der Datenfundus beim Bundesamt für Statistik

Georges-Simon Ulrich, Direktor des Bundesamts für Statistik

Das Bundesamt für Statistik arbeitet sowohl mit regionalen, nationalen als auch internationalen Institutionen zusammen, um in fachlich unabhängiger Weise repräsentative Ergebnisse in allen Fachgebieten zu ermitteln. Im Folgenden wird aufgezeigt, wie und unter welchen Voraussetzungen Forschende die Daten des BFS benutzen können und welche rechtlichen Einschränkungen zu beachten sind.

Das Bundesamt für Statistik im Dienste der Forschung

Seit Jahren arbeitet das Bundesamt für Statistik (BFS) mit den Institutionen der Forschungsförderung sowie den Forschenden in universitären Hochschulen und Fachhochschulen bei der wissenschaftlichen Auswertung von statistischen Informationen zusammen. Dies betrifft vor allem die Sozial- und Wirtschaftswissenschaften, aber auch die epidemiologische und naturwissenschaftliche Forschung.

Im Zeitverlauf hat sich diese Zusammenarbeit aufgrund rechtlicher, politischer, institutioneller und technologischer Entwicklungen verändert. So wurden unter anderem mit der Integration der Schweiz in das Europäische Statistiksystem verschiedene Erhebungen wie zum Beispiel SILC ausgebaut. Zudem sind seit dem Inkrafttreten des bilateralen Vertrages auch Einzeldaten aus der Schweiz (im Verbund mit den Daten der EU-Mitgliedstaaten) bei EUROSTAT auf Gesuch hin für die schweizerische Forschung zugänglich.

Durchführung von Erhebungen im Interesse der Wissenschaft

Die Förderung von Forschungsprojekten von nationaler Bedeutung gehört gemäss Art. 3 BStatG zu den Grundaufgaben des BFS. Gleichzeitig sind die Bestimmungen des Bundesstatistikgesetzes zu beachten, wenn die Tätigkeit von Forschungsinstitutionen einen engen Bezug zur öffentlichen Statistik hat. Wichtige Anwendungsbereiche sind etwa die Ziehung von Stichproben, der Datenschutz und die Datenweitergabe sowie die Publikation von Resultaten und koordinative Tätigkeiten. Dabei ist zu beachten, dass Anzahl und Art der Direktbefragungen auf das erforderliche Minimum zu reduzieren ist (Art. 4 BStatG) und die nationale und internationale Koordination der gesamten Bundesstatistik durch das BFS zu erfolgen hat.

Die Bereitstellung von Einzeldaten für die Forschung entspricht einer immer stärkeren Nachfrage und wird in Zusammenarbeit mit verschiedenen Institutionen wie z.B. der Stiftung für die Forschung in den Sozialwissenschaften (FORS) mit Sitz in Lausanne sichergestellt.

Seit dem Jahr 2010 besteht die Möglichkeit, auch die Informationsanliegen und Datenbedürfnisse der Forschung durch spezifische Fragestellungen und Modulen im Rahmen von Omnibusbefragungen zu berücksichtigen. Falls durch die Auftraggeber eine entsprechende Finanzierung sichergestellt wird, können diese in Zusammenarbeit mit einer Bundesstelle die zu erhebenden Inhalte mitgestalten. Die Inhalte der Omnibusbefragung müssen ein breites nationales Interesse abdecken. Das BFS legt abschliessend fest, welche Omnibusbefragungen durchgeführt werden.

Für die Mehrfachnutzung braucht es eine koordinierte Datenhaltung, harmonisierte Metadaten, einheitliche Nomenklaturen und ein effizientes System der Pseudonymisierung bzw. Anonymisierung und Depseudonymisierung.

Regelung des Zugangs zu Daten

Der Zugang zu den Mikrodaten des BFS ist heute umfassend geregelt: Anonymisierte Einzeldaten können für Forschende mit einem Datenschutzvertrag für Zwecke der Statistik, der Forschung und der Planung auf maximal fünf Jahre befristet für ein einzelnes Projekt weitergegeben werden. Dabei muss der Datenempfänger das gleiche Datenschutzniveau gewährleisten wie der Datenlieferant. Konkret bedeutet dies, dass bei Einzeldaten der Schutzstufe 3 (besonders schützenswerte Personendaten) der Datenschutz schriftlich nachgewiesen werden muss.

Nach Abschluss des Projekts müssen die Daten vernichtet werden. Bei begründetem Verdacht auf Missbrauch kann das BFS geeignete Massnahmen wie die Einforderung einer Konventionalstrafe oder die Ablehnung der Abgabe weiterer Daten ergreifen.

Nutzung der Daten

Jährlich werden rund 650 Datenschutzverträge zwischen dem BFS und wissenschaftlichen Institutionen sowie Stellen von Bund, Kantonen oder Gemeinden geschlossen. Die Nutzung der Daten des BFS durch Forschende geht allerdings viel weiter. Allerdings lassen die umfassenden Auswertungen unseres Amtes über die Nutzung der einzelnen Angebote keine Rückschlüsse über die Identität der Nutzenden zu.

Zum Autor

Georges-Simon Ulrich

Prof. Dr. MBA Georges-Simon Ulrich ist seit dem 1. Oktober 2013 Direktor des BFS. Er hat in der Schweiz, in den USA und in Australien studiert. Er ist zudem Professor für strategisches Management und Forschungsmethoden an der HWZ Zürich. Herr Ulrich war zwischen 1992 bis 2013 als Unternehmer und in verschiedenen leitenden Positionen in der Markt- und Meinungsforschung und als Direktor von LUSTAT tätig. Herr Ulrich ist Mitglied des Büros der UNO-Statistik-Kommission und war Mitglied der Partnership Group des Europäischen Statistik-Systems.

Les répercussions des Big Data sur les procédés de recherche en psychologie

Emilie Joly-Burra et Paolo Ghisletta, Faculté de psychologie et des sciences de l'éducation, Université de Genève

S'il est aujourd'hui indéniable que les Big Data ont créé une véritable révolution dans le domaine des computer-sciences et du marketing, c'est au tour de la psychologie de modifier ses pratiques en fonction de cette nouvelle réalité de génération, récolte, stockage et analyse des données.

Depuis plusieurs décennies déjà, les chercheurs en psychologie ont compris l'intérêt de récolter de très larges volumes de données pour mieux comprendre les processus sous-tendant le fonctionnement psychologique tel qu'en témoignent, par exemple, les nombreuses études longitudinales sur de larges échantillons ou cohortes de population (ex. Berlin Aging Study, LIVES). Cependant, ces dix dernières années ont vu une nette augmentation de la quantité d'indicateurs récoltés pour comprendre le fonctionnement humain (ex. données socio-économiques, cognitives, affectives, de personnalité, génétiques, neuro-fonctionnelles). Les processus mêmes de création et de récolte des données ont profondément évolué avec l'avènement de l'ère des médias sociaux et du «quantified self». Des terabites de posts Facebook, de tweets ou de données physiologiques issues de trackers d'activité physique ou de smartphones sont créés chaque jour, et il ne suffit que de quelques lignes de codes pour les extraire.

La nécessité de développer des infrastructures de stockage et d'analyse de cette masse de données en constante augmentation apparait donc plus évidente que jamais. Pour répondre à ce besoin, la discipline vit actuellement un véritable essor dans le développement de techniques et d'outils statistiques permettant de tester les théories psychologiques sur de gros volumes de données[1]. On voit ainsi apparaitre de nouveaux outils statistiques mêlant, par exemple, apprentissage automatique («machine learning») et modèles à équations structurales (ex. Structural Equation Model Forests[2]).

Applications actuelles et défis futurs

Les modèles cités ci-dessus ont ainsi permis d'identifier les prédicteurs les plus influents de différentes dimensions du bien-être ainsi que leurs effets d'interactions non linéaires[3]; de montrer que deux des 65 prédicteurs les plus puissants de survie dans un échantillon de 6203 personnes sont de nature psychologique et pas uniquement médicale ou sociodémographique[4]; ou encore, dans le domaine de la psychologie de la personnalité, les patterns de «likes» Facebook ont fourni une meilleure prédiction des traits de personnalité que l'évaluation classique par questionnaires[5].

Si les Big Data semblent ouvrir un vaste champ des possibles pour la recherche en psychologie, cela n'est pas sans soulever un ensemble de défis. D'une part, plusieurs psychologues se retrouvent confrontés à des défis techniques tels qu'un manque de connaissances approfondies en programmation, souvent nécessaires au traitement des Big Data. D'autre part, la facilité d'accès à des données personnelles potentiellement sensibles pose des questions éthiques majeures, comme l'a rappelé le récent scandale du Facebook - Cambridge analytica. Sans même parler d'utilisation commerciale des données, les chercheurs se doivent de garder le respect de l'humain et de la vie privée au centre de leur démarche, et d'adopter des pratiques empêchant les retombées négatives du traitement de données personnelles. Finalement, se pose la question de la véracité des données et du risque de surinterpréta-

tion des modèles obtenus sur de très larges échantillons, ce qui peut conduire à une surestimation de l'importance des effets mis en lumière (d'où l'importance de considérer non seulement la significativité statistique d'un résultat, mais également la taille de son effet, un aspect mis en avant en psychologie).

De par leurs connaissances des théories psychologiques de l'humain, de la psychométrie et des statistiques, les psychologues se trouvent dans une position privilégiée pour donner du sens à ces masses colossales de données[6]. Dans une discipline où la recherche est résolument guidée par le test strict d'hypothèses théoriques, l'ère des Big Data permet d'entrevoir un enrichissement mutuel entre recherche guidée par la théorie (Theory driven) et recherche guidée par les données (Data driven) afin d'aboutir à une compréhension plus fine du fonctionnement humain dans toute sa complexité.

Références

[1] Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 738.

[2] Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21*(4), 566.

[3] Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods, 21*(4), 583.

[4] Aichele, S., Rabbitt, P., & Ghisletta, P. (2016). Think fast, feel fine, live long. A 29-year study of cognition, health, and survival in middle-aged and older adults. *Psychological Science, 27*(4), 518–529. https://doi.org/10.1177/0956797615626906

[5] Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036–1040.

[6] Cheung, M. W. L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 738.

Les auteurs

Emilie Joly-Burra

Emilie Joly-Burra a obtenu son Master en Psychologie à l'Université de Genève en 2013 et poursuit actuellement son travail de doctorat sur l'utilisation des méthodes mixtes (quantitatives et qualitatives) pour étudier la relation entre les buts et le bien-être chez les personnes âgées. Elle enseigne également des travaux dirigés en statistiques appliquées à la psychologie en Bachelor et Master.

Paolo Ghisletta

Paolo Ghisletta est professeur de méthodologie et d'analyse de données à la Section de psychologie de la Faculté de psychologie et des sciences de l'éducation de l'Université de Genève, et doyen de la filière Bachelor of Science in psychology francophone à la Formation universitaire à distance, Suisse (Brigue). Il s'intéresse aux changements psychologiques au cours de la vie et aux méthodologies pour les étudier. (Pôle de recherche national LIVES: Surmonter la vulnérabilité: Perspective du parcours de vie. Universités de Lausanne et de Genève)

Methodische Herausforderungen bei der Nutzung von Big Data

Sophie Mützel, Soziologisches Seminar der Universität Luzern

Durch die Digitalisierung entstehen riesige, unstrukturierte Datenmengen. Die Big Data fordern die Sozialwissenschaften heraus. Es braucht neue Fähigkeiten und Zusatzwissen zur Datenaufbereitung, von Computerlinguistik bis zu algorithmischen Verfahren. Aber es können auch neue Forschungsfragen beantwortet werden.

Der Begriff *Big Data* bezeichnet Daten, die aus vielen Datenpunkten bestehen, unterschiedliche Datentypen umfassen, die oft nicht numerisch, relational und unstrukturiert sind und die ganz unterschiedliche Phänomene betreffen können. Als *neue* grosse Datenmengen entstehen diese Datenpunkte z.B. bei der Nutzung von sozialen Netzwerkplattformen, Suchmaschinenanfragen, Kaufportalen und modernen Kommunikationsmedien beim Einsatz von Kredit- und anderen Geldkarten. So produziert jedes Anklicken, jedes Aufrufen einer Website, jede Eingabe und auch jedes Löschen eines Kommentars in Echtzeit weitere Datenpunkte, sogenannte digitale Datenspuren. Aber auch in Archiven lagern in Datenbanken Daten, die über lange Zeiträume reichen, z.B. Geburts- und Handelsregister, Statistiken, Reden, Zeitungsartikel oder Fotos als digitalisierte *alte* grosse Daten.

Diese Arten von Daten unterscheiden sich von den Daten, die in repräsentativen Umfragen, Interviews oder Experimenten kostenintensiv gezielt für die Forschung erhoben werden und zum klassischen Grundstock sozialwissenschaftlicher Untersuchungen gehören. Diese alten und neuen grossen Daten sind aus Beobachtungen oder Einträgen entstanden und wurden aus kommerziellen oder administrativen Interessen erhoben. Für Unternehmen wie Netzwerkplattformen sind die digital generierten Beobachtungsdaten Grundlage ihres Geschäftsmodells mit dem Ziel der Gewinnoptimierung. Forschende nutzen nun diese existierenden Daten aufgrund ihrer Forschungsinteressen.

Umnutzung von Daten

Diese Umnutzung der Daten stellt besondere Herausforderungen an methodische Fähigkeiten und auch an die methodologische Reflexion. Soziologen, die mit neuen grossen Daten forschen, müssen reflektieren, wie diese Daten konstruiert und strukturiert sind, um in der Folge zu klären, welche Fragen sich mit den jeweiligen Datensätzen überhaupt beantworten lassen. So sind Daten von Netzwerkplattformen nicht repräsentativ, sondern unvollständig: Aussagen über eine Grundgesamtheit sind nicht möglich. Ferner enthalten die Daten sensible, persönliche Informationen von einzelnen Nutzenden, die auch in anonymisierter Form schnell de-anonymisiert werden können und somit Forschende vor zentrale ethische Fragen stellen. Zudem sind solche Daten algorithmisch vorstrukturiert und alles andere als sauber, sondern häufig z.B. durch Spam oder andere maschinelle Prozesse ‹verschmutzt›. Im Vergleich dazu sind Archivdaten eher repräsentativ nutzbar, doch sind auch hier textbasierte Daten häufig verschmutzt: So können sich bspw. in einem ursprünglichen Textkorpus unterschiedliche Schreibweisen einzelner Begriffe finden. Auch wenn die Reflektion über Daten zur soziologischen Kernkompetenz gehört, erfordert die Notwendigkeit der Aufbereitung der Rohdaten zu soziologisch auswertbaren Daten neue Fähigkeiten.

Und: Die Analyse von alten und neuen grossen Datenmengen erfordert von Soziologinnen und Soziologen nicht nur neue methodische Fähigkeiten in der Konstruktion soziologischer Datensätze. Auch sind neue methodische Fähigkeiten für die Analyse solcher Daten vonnöten.

Computerlinguistik, Netzwerkanalyse und Algorithmen

Die Bearbeitung grosser alter oder neuer Textdatenmengen erfordert Kenntnisse aus der Computerlinguistik, insbesondere zur maschinellen Verarbeitung der natürlichen Sprache (*Natural Language Processing*). Zudem ermöglichen die Kenntnisse der Netzwerkanalyse die Analyse relativer Daten, wie sie aufgrund der technischen Anlage von Netzwerkplattformen durch ihre Nutzenden z.B. im Austausch mit Freunden auf Facebook produziert werden. Des Weiteren sind neue Fähigkeiten im Einsatz mit algorithmischen Verfahren nötig, die induktiv nach Mustern in Daten suchen. Anders als kausale Erklärungen lassen sich nun deskriptive Erklärungen treffen.

Diese kurzskizzierten, methodischen Herausforderungen im Umgang mit grossen Daten sollten jedoch Soziologinnen und Soziologen nicht abhalten, sie zu nutzen, denn Analysen von neuen und alten grossen Datenmengen ermöglichen es, alte soziologische Fragen neu zu beantworten. So können wir bspw. alte Aussagen über soziale Konflikte überprüfen wie auch neue Einsichten in soziale Dynamiken gewinnen.

Literatur

— Bail, Christopher A. 2014. «The cultural environment: measuring culture with big data», in: *Theory and Society* 43: 465-482.

— Mützel, Sophie. 2015. «Facing Big Data: Making sociology relevant», in: *Big Data & Society* 2, http://journals.sagepub.com/doi/full/10.1177/2053951715599179

— Salganik, Matthew. 2018. *Bit by Bit*. Princeton: Princeton University Press.

Zur Autorin

Sophie Mützel

Prof. Sophie Mützel, Ph.D., ist Ordinaria am Soziologischen Seminar der Universität Luzern. Ihre Arbeitsschwerpunkte liegen im Bereich der Soziologie von Algorithmen, der Digitalisierung des Alltags sowie der Entstehung neuer Märkte aus wirtschafts- und kultursoziologischen Perspektiven. Sie leitet das Forschungsprojekt «Facing Big Data: Methods and skills needed for a 21st century sociology» im NFP 75 «Big Data».

Datenstandards in den Geistes- und Sozialwissenschaften

Lukas Rosenthaler, Leiter des «Data and Service Centers for the Humanities»

Digitale Daten sind in der Forschung heute selbstverständlich. Damit die Datensätze langfristig les- und verfügbar bleiben, müssen gewisse Standards eingehalten werden. Insbesondere auf der Ebene der Internet-Services besteht noch Handlungsbedarf. Gelingt es, die Daten zu standardisieren, eröffnet das ganz neue Möglichkeiten für die Forschung.

Moderne Forschung in den Geistes- und Sozialwissenschaften ist heute ohne den Einsatz von digitalen Hilfsmitteln kaum mehr denkbar. Quellen werden digitalisiert, um den einfachen Zugang zu ermöglichen und die Forschungsarbeit zu erleichtern. Viele Quellen sind nur noch in digitaler Form zugänglich. Andererseits werden auch im Forschungsprozess selbst digitale Daten erzeugt. Schon das Transkribieren eines Textes mit einem Textsystem erzeugt digitale Daten. Werden Hilfsmittel wie Tabellenkalkulationsprogramme, Desktop-Datenbanken etc. verwendet, so entstehen weitere digitale Daten, welche eine recht komplexe Struktur aufweisen können. Um den wissenschaftlichen Austausch pflegen zu können, sollten sowohl die digitalen Quellen als auch die Forschungsdaten allgemein anerkannten Datenstandards genügen. Diese sollten möglichst offen, breit akzeptiert und so einfach wie möglich sein, um eine lange Lebensdauer – das heisst, die langfristige Les- und Interpretierbarkeit – zu gewährleisten.

Datenstandards

Grundsätzlich gibt es zwei Ebenen, auf denen Datenstandards eine wichtige Rolle spielen. Die erste Ebene bilden die Datenformate, welche festlegen, in welcher Form die digitalen Daten festgehalten werden. Im Digitalen besteht das Alphabet nur aus 2 Zeichen, «0» und «1», deren Interpretation und Bedeutung durch das *Datenformat* festgelegt wird. Die zweite Ebene bilden die Internet-Services von Repositorien, welche die Forschungsdaten zur Verfügung stellen. Es ist heute noch üblich, dass jedes Repositorium oder jedes Forschungsprojekt, das seine Daten auf dem Internet zur Verfügung stellt, eigene Standards für diesen Zugang einsetzt. Dies führt dazu, dass die Wiederverwendung solcher Daten sehr aufwändig und schwierig ist. Während es im Bereich der Datenformate unterdessen – je nach Datentyp unterschiedliche – gut etablierte Standards gibt, besteht auf der zweiten Ebene noch grosser Handlungsbedarf.

Dateiformate

Ein seit langer Zeit etablierter Standard für Daten beruht auf XML («eXtended Markup Language»), welche es erlaubt, viele Arten von Daten in einem einfachen und auch für den Menschen interpretierbaren Form festzuhalten. XML eignet sich für viele unterschiedliche Formen von Information und ist durchaus auch im Bereich von quantitativen Daten gut einzusetzen. Für Text-basierte Daten (Transkriptionen etc.) wurde schon vor vielen Jahren mit TEI/XML ein Standard geschaffen, der eine sehr grosse Verbreitung und Akzeptanz gefunden hat.

Im Bereich von digitalen Quellen sind mindestens für Bilder (digitale Reproduktionen, Faksimile etc.) weit akzeptierte Formate vorhanden. Das TIFF-Format wird seit Jahren erfolgreich für digitale Bilder verwendet. Allerdings ist der Umgang mit dem TIFF-Format nicht ganz einfach, da es einige ‹exotische› Varianten zulässt, welche die Interoperabilität behindern können. Zudem sind TIFF-Dateien in der Regel sehr gross. Seit einiger Zeit wird deshalb das JPEG2000-Format propagiert, welches vor allem in Bezug auf die Dateigrösse Vorteile aufweist. Für andere Medientypen (Ton, Bewegtbild) sind die Standards weniger offensichtlich. Während bei Tondateien das PCM oder mp3-Format weite Verbreitung gefunden hat, gibt es für Video- und Filmdateien keine generellen Standards.

Datenbanken, welche nicht auf XML basieren (z.B. weit verbreitete Desktop-Datenbanken), verwenden proprietäre Formate. Es zeichnet sich in diesem Bereich jedoch ab, dass auf dem semantischen Web (Linked Open Data, LOD) basierende Datenrepräsentationen sehr attraktiv sind. Das

World Wide Web Consortium (W3C) hat in diesem Bereich offene Standards definiert (z.B. das Resource Description Framework RDF und die Terse RDF Triple Language turtle), mit denen komplexe und hochgradig verknüpfte Daten einfach repräsentiert werden können. Es zeichnet sich ab, dass in Zukunft viele geistes- und sozialwissenschaftliche Daten in dieser Form festgehalten werden.

Internet-Services

Im Bereich der Internet-Services hat sich das International Image Interoperability Framework (IIIF, siehe http://iiif.io) als Standard durchgesetzt, um Bilder in flexibler Art und Weise austauschen zu können. Im Bereich der vernetzten Daten (Datenbanken) stehen mit JSON-LD und RESTful APIs Werkzeuge zur Verfügung, um wissenschaftliche Daten in einer standardisierten Form auszutauschen. JSON-LD ist ein Text-basiertes, strukturiertes Format, mit dem vernetzte Daten flexibel und Internet-kompatibel präsentiert werden können. Eine RESTful API ist eine standardisierte Schnittstelle, um digitale Informationen einfach über das Internet maschinenlesbar zur Verfügung zu stellen.

Fazit

Die zunehmende Standardisierung von Datenformaten und vom Zugriff auf Repositorien eröffnet neben dem vereinfachten Austausch von Information weitere, vollkommen neue Möglichkeiten der Interoperabilität: Forschungswerkzeuge können unabhängig vom individuellen Repositorium entwickelt und eingesetzt werden. Damit können Daten aus verschiedenen Repositorien aggregiert, verbunden und verknüpft werden.

Zum Autor

Lukas Rosenthaler

Prof. Dr. Lukas Rosenthaler arbeitet am Digital Humanities Lab der Universität Basel und ist der Leiter des «Data and Service Centers for the Humanities» (DaSCH), einer nationalen Forschungsinfrastruktur für die Geisteswissenschaften. Er beschäftigt sich seit vielen Jahren mit der Problematik der langfristigen Verfügbarkeit von Forschungsdaten und der Entwicklung von digitalen Forschungsumgebungen.

Voraussetzung für die nachhaltige Sicherung von digitalen Services und Daten

Alice Keller, Zentralbibliothek Zürich

Zur nachhaltigen Sicherung von Daten laufen in den Bibliotheken verschiedene Projekte. Verschiedene Faktoren wie Qualität, gute Verankerung und langfristiger Nutzen tragen zum Erfolg bei. Aber für wirkliche Innovationen fehlt oft das Geld.

Der Schrecken sitzt der britischen BCC heute noch in den Knochen: Im Jahre 1986 publizierte sie mit grossem Aufwand das Domesday Project. Es galt als Neuauflage des 900 Jahre alten Grundbuchs (Domesday Book) und sollte einen Einblick in das moderne Leben Grossbritanniens bieten. Über eine Million Erwachsene und Schulkinder erfassten digitale Texte und Fotos zu ihrem Land und Leben. Als Trägermedium wurde eine Laserdisc gewählt, damals State-of-the-Art. Dann, nur 25 Jahre später, waren die Daten unlesbar, während das pergamentene Dokument aus dem Jahr 1086 weiterhin problemlos und ohne Hilfsmittel lesbar war. Die Rettung der Daten auf der Laserdisc und ihre Überführung in eine moderne Web-Umgebung war eine höchst komplexe Aufgabe und glückte nur dank eines Grosseinsatzes zahlreicher IT-Spezialisten.[1]

Um solchen Situationen, nämlich dem Informations- und Serviceverlust, durch digitale Obsoleszenz vorzubeugen, ist es wichtig, dass sich Wissenschaftler bereits während der Projektphase mit Fragen der Nachhaltigkeit befassen.

Begriffsklärung

Der Begriff Nachhaltigkeit verdankt seiner Verwendung im Brundtland-Bericht aus dem Jahre 1987 seine breite Bedeutung. «Sustainable development» wurde definiert als Entwicklung, «die den Bedürfnissen der heutigen Generationen entspricht, ohne die Möglichkeiten künftiger Generationen zu gefährden, ihre eigenen Bedürfnisse zu befriedigen und ihren Lebensstil zu wählen»[2]. Ökologische, ökonomische und soziale Ziele sollen also nicht zueinander im Wettbewerb stehen, sondern gleichrangig angestrebt werden. Obwohl die Definition von Brundtland nicht eins zu eins auf digitale Services und Daten übertragen werden kann, so ist doch die Erkenntnis wichtig, dass die Nachhaltigkeit mehrere Dimensionen umfasst, die zueinander in Balance stehen müssen. Die Autorin hat versucht, diese Dimensionen im Kontext wissenschaftsnaher Projekte zu analysieren, und hierbei *qualitative*, *zeitliche*, *organisatorische* und *räumliche* Kriterien identifiziert.[3]

Verfolgt man die Diskussionen von Fachkollegen im Bereich datenintensiver Projekte im Allgemeinen oder digitaler Bibliotheken im Besonderen, so ist auffallend, dass finanzielle Überlegungen oft im Vordergrund stehen. Erwähnt wird hier die Widersprüchlichkeit zwischen der befristeten Finanzierung und dem wissenschaftlichen oder kulturhistorischen Anspruch, dass diese Daten oder Services langfristig erhalten und zugänglich bleiben.

[1] http://www.bbc.co.uk/history/domesday/story

[2] Hauff, V. (Hrsg.). Unsere gemeinsame Zukunft. Der Brundtland-Bericht der Weltkommission für Umwelt und Entwicklung. Greven: Eggenkamp, 1987, hier S. XV.

[3] Alice Keller. Nationale Förderprogramme: eine Analyse der Nachhaltigkeit von Bibliotheksprojekten. Bern 2017. Online verfügbar auf http://www.kpm.unibe.ch/weiterbildung/weiterbildung/zertifikatsarbeiten/index_ger.html

Evaluation der Nachhaltigkeit von nationalen Bibliotheksprojekten

Das Thema der Nachhaltigkeit von nationalen Bibliotheksprojekten war auch das Thema einer umfassenden Analyse aus dem Jahr 2017.[4] In Bibliotheksprojekten sind typischerweise Daten von Services nicht zu trennen. Zu den Daten, die von Bibliotheken professionell erfasst, gesammelt und zur Verfügung gestellt werden, gehören bibliografische Daten, «digital born»-Volltexte, Digitalisate von Text- oder Bilddokumenten, Transkriptionen, Geodaten, Forschungsdaten u.a.m.

In ihrer Studie hat die Autorin folgende sechs Projekte auf ihre Nachhaltigkeit hin analysiert: «E-Depot» (Service zur zentralen Speicherung lizenzierter Zeitschriften), «ElibEval» (Usability-Evaluation von Online-Angeboten), «e-rara.ch» (Online-Plattform für digitalisierte alte Drucke), «Kartenportal.CH» (virtuelle Fachbibliothek «Geodaten und Karten»), «nationales Konsortium der Hochschulbibliotheken» (Lizenzierung von Zeitschrifteninhalten), «swissbib» (Metakatalog der Schweizer Hochschulbibliotheken). Hierbei entwickelte und nutzte sie folgende Definition der Nachhaltigkeit:

Ein Projekt gilt als nachhaltig, wenn es gute Qualität und Wirkung aufweist und über das Projektende hinaus einen dauerhaften Nutzen bietet. Zur Sicherung der Dauerhaftigkeit ist die Intervention oder Innovation organisatorisch, strategisch und finanziell in der Trägerorganisation verankert und verfügt ausserdem über das Potenzial, erfolgreich auf andere Kontexte transferiert zu werden.[5]

Fünf der sechs untersuchten Projekte wurden von den entsprechenden Projektleitenden grundsätzlich als nachhaltig bezeichnet, nur eines wurde als Service nicht weitergeführt. Allerdings bedeutet dies nicht, dass alle Kriterien der Nachhaltigkeit bei allen fünf Projekten erfüllt sind. So werden beispielsweise bei «e-rara.ch» noch nicht alle Objekte vollständig digital langzeitarchiviert. Beim Metakatalog «swissbib» fehlt ein nachhaltiges Finanzierungsmodell. Das Projekt «ElibEval» konnte zwar erfolgreich in ein Kompetenzzentrum integriert werden, aber es fehlen Mittel zur Weiterentwicklung. Das «nationale Konsortium» bietet ausgezeichnete Dienstleistungen an, ist aber organisatorisch und rechtlich nicht nachhaltig aufgestellt. Einzig das «Kartenportal.CH» erfüllt alle Kriterien der Nachhaltigkeit. Das «E-Depot» schliesslich musste nach Projektende eingestellt werden, da das System seinerzeit technisch nicht skalierbar war.

Erfolgsfaktoren

Es haben sich folgende Erfolgsfaktoren herauskristallisiert, die positiv zur Nachhaltigkeit von Projekten beitragen.

— Gute Zusammenarbeit zwischen den beteiligten Partnern; sie identifizieren sich mit dem Service und tragen ihn ideell mit.
— Gemeinsame Weiterentwicklungen, die stark verbindend wirken.
— Gute Qualität der Services (Funktionalitäten und Daten) mit Referenzcharakter über die Schweizer Grenze hinaus.
— Personelle Kontinuität durch langjährige, engagierte Projektleiter und Fachspezialisten.
— Open Access: Die Angebote sind für den Endkunden kostenlos.

[4] Dito.

[5] Dito, S. 8.

Fast alle Projekte bzw. Services beruhen in hohem Masse auf Basisleistungen, die von Bibliotheken im Rahmen ihrer Kernaufgaben erbracht werden (z.B. Metadatenerfassung, Digitalisierung, E-Medienangebot). Zusätzlich profitieren die Services von Overhead-Leistungen, die von beteiligten Bibliotheken i.d.R. kostenlos zur Verfügung gestellt werden. Insgesamt betrachtet, profitieren die beteiligten Bibliotheken davon, dass die Resultate ihrer Kernaufgaben durch die neuen Services optimal der Öffentlichkeit präsentiert und breit genutzt werden. Ausserdem erzielen sie dank Bündelung von Kräften Synergien und eine stärkere Marktposition. Es herrscht also eine Win-win-Situation für das Projekt und die beteiligten Bibliotheken.

Kein Geld für Innovationen

Beachtet man die Auflage, dass bei vielen Förderprogrammen Eigenleistungen in der Höhe von mindestens 50% erbracht werden müssen und es sich ausschliesslich um Anschubfinanzierungen handelt, so überrascht es nicht, dass die Bibliotheken Projekte wählen, die auf ihren Kernaufgaben aufbauen. Schliesslich ist allen klar, dass man nach Ablauf der Projektphase auf eigenen Beinen stehen muss. Dies dürfte einerseits die Nachhaltigkeit positiv beeinflussen, andererseits werden hierdurch Projektvorschläge für radikale Innovationen erschwert oder gar verhindert. Projekte in Bibliotheken bleiben typischerweise Zusatzdienste – oft sehr wichtige, manchmal auch nur «Nice-to-have»-Angebote.

Wichtig ist auch die Anerkennung der Grenzen der Projekte hinsichtlich ihrer Nachhaltigkeit. Obwohl wie eingangs erwähnt viele der neu entstandenen Services und Infrastrukturen nicht mehr wegzudenken sind, ist es bisher keinem der Services gelungen, Projektstellen zu verstetigen bzw. neue, unbefristete Personalstellen zu schaffen. Wie das Beispiel des nationalen Konsortiums zeigt, bräuchte es hierfür tragfähigere Strukturen und eine gesicherte dauerhafte Finanzierung.[6]

[6] Gegenwärtig handelt es sich beim Konsortium um eine *einfache Gesellschaft*. Vgl. hierzu: R. Ball und P. Boutsiouci: Literaturversorgung, Collection Management und das Konsortium der Schweizer Hochschulbibliotheken. In: Bibliotheken der Schweiz: Innovation durch Kooperation. Hrsg. A. Keller und S. Uhl. Berlin: De Gruyter, 2018. S. 145–159. Online verfügbar: https://doi.org/10.1515/9783110553796-007

Zur Autorin

Alice Keller

Alice Keller hat nach ihrem Studium in den Naturwissenschaften eine Stelle an der ETH-Bibliothek angenommen. Sie fand die Bibliotheksarbeit interessant und hat sich zur wissenschaftlichen Bibliothekarin ausbilden lassen und schliesslich auch auf diesem Gebiet promoviert (Humboldt-Univ. Berlin). Auslandserfahrung hat sie während mehrerer Jahre an der Bodleian Library Oxford gesammelt. Bevor sie in die Schweiz zurückkehrte, hat sie beim De Gruyter Verlag, Berlin/München, das Lektorat Library & Information Science geleitet. Seit 2014 arbeitet sie als Chefbibliothekarin Bestandsentwicklung an der Zentralbibliothek Zürich.

Voraussetzungen für die Nutzung von Forschungsdaten

René Schneider, Haute école de gestion, Genf (HES//SO)

Forschungsdaten und ihre Nachnutzung werden die akademische Landschaft in den kommenden Jahren umgreifend verändern. Dies jedoch nur, sofern an den richtigen Stellen die richtigen Voraussetzungen geschaffen werden. Dabei haben zwei Elemente sowohl eine Hebel- als auch eine Scharnierfunktion: Metadaten und persistente Identifikatoren.

Der Begriff der «Nutzung» von Forschungsdaten lässt auf den ersten Blick vermuten, dass es sich allein um die Nutzung aktueller, häufig spricht man auch von aktiven Forschungsdaten handelt, also all jener Daten, die die Forscher in ihrer täglichen wissenschaftlichen Arbeit durch Messung, Beobachtung, Modellierung, Anreicherung, Ableitung oder ganz einfach Digitalisierung erzeugen. Diese Daten werden dann, wenn man von einem Idealbild ausgeht, mit anderen interessierten Forschern oder der Allgemeinheit, die in der Regel die finanziellen Mittel zur Verfügung stellt, geteilt oder ausgetauscht; ganz im Sine des Open Access.

Parallel dazu gibt es eine – je nach Perspektive – zumindest gleichberechtigte Betrachtungsweise, die eher auf die «Nachnutzung» der Forschungsdaten ausgerichtet ist und vom Gedanken der Langzeitarchivierung geleitet wird.

Nutzung und Nachnutzung

Geht es in beiden Fällen um die Bereitstellung von Forschungsdaten, unterscheiden sich diese zwei Perspektiven jedoch um einen ganz wichtigen Aspekt, jenen der Zeit nämlich. Der Einfachheit halber könnte man sagen, dass es im einen Fall um ein «So schnell wie möglich», im anderen Fall um ein «So lange wie möglich» geht; einmal darum, die Forschungsdaten so schnell wie möglich zur Verfügung zu stellen und zu teilen, ein anderes Mal darum, die Forschungsdaten so zu bearbeiten, dass sie möglichst lange aufbewahrbar und wiederverwendbar sind. Im einen Fall sprechen wir von einem Zeitpunkt, der möglichst schnell nach Ablauf des Forschungsprojekts beginnt, wenn nicht sogar schon – je nach Laufzeit des Projekts – während des Projekts. Im anderen Fall um Zeiträume, die mindestens Jahrzehnte umfassen.

Trotz dieser Gegensätzlichkeit gibt es zugleich ein oder mehrere verbindende Elemente, unabhängig davon, zu welchem Zeitpunkt Forschungsdaten zur Verfügung gestellt oder benötigt werden. Dieser Zusammenhang wird häufig als das FAIRness-Prinzip bezeichnet. Dies besagt, dass Forschungsdaten auffindbar (Findable), zugänglich (Accessible), interoperabel (Interoperable) und wiederverwendbar (Re-usable) sein müssen. Die Daten müssen also ab dem Zeitpunkt der Publikation oder der Übergabe an ein Langzeitarchiv dauerhaft identifizierbar, referenzierbar und zitierbar sein … und alles Weitere folgt darauf.

Metadaten

Konkret liegen dafür zwei Instrumente bereit: die Metadaten sowie die – weniger geläufigen – sogenannten persistenten Identifikatoren (PID). Der Bereich der Metadaten, d.h. der Daten, die die eigentlichen Forschungsdaten beschreiben, lässt sich grob in technische und deskriptive Metadaten unterteilen, wobei technische Metadaten häufig automatisch erstellt werden. Problematischer sind die deskriptiven Metadaten, die Auskunft darüber geben sollen, was sich hinter den eigentlichen Daten verbirgt. Häufig können die Forscher sie ansatzweise selbst erstellen, sie bedürfen in der Regel aber der Nachbearbeitung seitens eines Datenkurators.

Diesem Datenkurator, dessen Berufsbild sich erst allmählich eigenständig zu manifestieren beginnt, obliegt es, die für eine (Nach-)Nutzung relevanten Daten so aufzubereiten, dass sie entweder möglichst schnell in entsprechenden Repositorien als FAIRe Daten zur Verfügung stehen oder dass sie so transformiert werden, dass sie auch noch während oder nach zehn oder mehr Jahren trotz einer bis dahin veränderten Hard- und Softwarelandschaft in diese nahtlos bzw. nach entsprechenden Migrationen eingefügt bzw. wieder ausgewertet werden können.

Persistente Identifikatoren

Ähnlich komplex verhält es sich mit den persistenten Identifikatoren, die – wie eine ISBN für Bücher – eine eindeutige (oder mathematisch korrekt, eine eineindeutige) Zuordnung zwischen einem Objekt und einer Kennziffer erlauben. Diese PID werden an Daten, Personen, Organisationen u.v.m. vergeben. Was die Daten der Geisteswissenschaften betrifft, sind dies vor allen Dingen die Digitalisate und hier teils sehr feingranulare Bestandteile davon sowie die darauf verweisenden Ketten von Referenzen und Zitationen. Eine besondere Rolle bei den Personen spielen dabei neben den Forschern auch die reellen bzw. in der Realität sehr wirkmächtigen virtuellen Personen, von denen in den einzelnen Werken die Rede ist. Hierfür stehen erste Lösungen zur Verfügung, etwa DOIs, ARK/N2T, ORCID, ISNI usw. Die wissenschaftliche Landschaft aber ist noch weit entfernt von einer Vernetzung aller zu vergebenden PID. Zudem werden Mechanismen benötigt, die in der Lage sind, die Vielfalt der Metadaten und der Identifikatoren dynamisch zu verwalten, so dass es gelingen kann, den Forschern neben dem «Mehr an Arbeit», das das Forschungsdatenmanagement von ihnen verlangt, auch Anreize zu bieten, bspw. über die Berechnung eines h-Indexes für die Zitation ihrer Daten oder andere geeignete Altmetrics, etwa um die Ansichten oder Downloads bereitgestellter Datensätze nachzuweisen.

Handlungsbedarf

Es zeigt sich auch, dass hierbei insbesondere Institutionen, die sich selbst über den Begriff der Perennität definieren oder deren Daseinszweck mit der Perennität der zu verwaltenden Objekte verbunden ist, also genauer Bibliotheken und Archive, bei der Vergabe und Verwaltung der PID eine entscheidende Rolle spielen können, sofern sie es denn möchten oder aber dazu in der Lage sind. Aber selbst darüber hinaus bedarf es in der Schweiz und hier insb. im Bereich der Geisteswissenschaften noch einer Reihe gemeinsamer Anstrengungen seitens aller mit der Forschung befassten Partner, damit Forschungsdaten schnellstmöglich und dauerhaft geteilt und nachgenutzt werden können.

Zum Autor

René Schneider

Prof. ord. Dr. phil. René Schneider, M.A., ist Professor für Informationswissenschaft an der Haute école de gestion – HES//SO in Genf und Leiter des Masterstudiengangs Information Science. Seine Hauptinteressen gelten dem Forschungsdatenmanagement und der Nützlichkeit von Information.

Geschäftsmodelle für Forschungsdatenarchive – die Empfehlungen der OECD

André Golliez, Präsident Swiss Data Alliance

Unter dem Titel «Business Models for Sustainable Research Data Repositories» hat die OECD im Dezember 2017 einen Bericht zur langfristigen Finanzierung von Forschungsdatenarchiven publiziert.[1] Experten aus 18 Ländern haben sich an mehreren Workshops mit diesem Thema auseinandergesetzt und gleichzeitig knapp 50 Forschungsdatenarchive zu ihren Geschäftsmodellen befragt. Aus dem Bericht resultierten fünf Empfehlungen an Entscheidungsträger, um die langfristige Existenz von Forschungsdatenarchiven finanziell zu gewährleisten.

Problemstellung: Die zunehmenden Anforderungen an die Forschungsdatenarchive sind finanziell nicht abgesichert

Es gibt weltweit eine Vielzahl von Datenarchiven, welche für die langfristige Verfügbarkeit von Forschungsdaten verantwortlich sind. Da das Datenvolumen und die Anforderungen für einen offeneren Zugang zu diesen Daten stetig steigen, geraten diese Datenarchive zunehmend unter finanziellen Druck und sehen sich in ihrer langfristigen Existenz bedroht. Der Bericht der OECD untersucht die Einkommensströme, Kosten, Angebote und Geschäftsmodelle von 48 Forschungsdatenarchiven. Er führt auf dieser Basis zu einer Reihe von Empfehlungen, die einen Rahmen für die Entwicklung nachhaltiger Geschäftsmodelle abstecken und politischen Entscheidungsträgern sowie Geldgebern bei der ausgewogenen Festlegung von Regulierungen und Anreizen für die Datenarchive helfen.

Die wissenschaftlichen, wirtschaftlichen und sozialen Vorteile einer offenen Wissenschaft sind breit anerkannt und führen dazu, dass offene Daten als Bedingung für die Forschungsfinanzierung verlangt werden. Es gibt eine Reihe von Forschungsgebieten, die fast ausschliesslich von der Verfügbarkeit globaler Datenquellen abhängen und auf die Forschungsdatenarchive angewiesen sind. Forschungsdatenarchive werden daher zu einem zentralen Bestandteil der Forschungsinfrastruktur, und ihre angemessene Finanzierung ist langfristig sicherzustellen.

Spezifische Geschäfts- und Finanzierungsmodelle sind möglich

Die Gestaltung und Nachhaltigkeit von Geschäfts- und Finanzierungsmodellen für Forschungsdatenarchive hängt von vielen Faktoren ab. Dazu zählen die Rolle, welche ein Datenarchiv im nationalen Rahmen sowie im Kontext der betreffenden Wissenschaftsdomäne spielt, der Entwicklungsstand und Reifegrad des Archivs, die Merkmale der Benutzergemeinschaft oder die Art der Datenprodukte, welche die Höhe der Investitionen in die Pflege der Daten bestimmen. Alle diese Punkte müssen bei der Auswahl und Entwicklung geeigneter Geschäfts- und Finanzierungsmodelle berücksichtigt werden – es gibt sicherlich keine «one size fits all»-Lösung.

Von den befragten Datenarchiven wurden insgesamt 95 unterschiedliche Ertragsquellen gemeldet. In der Regel kombinieren die Archive strukturelle und institutionelle Finanzierungen mit Erträgen aus verschiedenen Formen von forschungs- und vertragsbezogenen Dienstleistungen. Eine weitere Finanzierungsquelle ergibt sich aus Gebühren für die Aufbewahrung der Daten sowie für datenbasierte Mehrwertdienste des Archivs.

Derzeit sind viele Forschungsdatenarchive weitgehend von öffentlichen Mitteln abhängig. Es stellt sich daher die Frage, wie diese öffentlichen Mittel am effektivsten bereitgestellt werden – nach welchem Mechanismus und von welcher Agentur, welcher Behörde oder welcher Institution. In diesem Zusammenhang ist anzuerkennen, dass das Angebot eines bestimmten Datenarchivs für die verschiedenen Akteure des öffentlichen Sektors einen unterschiedlichen Wert hat.

[1] https://www.oecd-ilibrary.org/science-and-technology/business-models-for-sustainable-research-data-repositories_302b12bb-en

Empfehlungen für die Entscheidungsträger in der Wissenschaftspolitik

Die OECD gelangt zu den folgenden fünf Empfehlungen für die wissenschaftspolitischen Entscheidungsträger in ihren Mitgliedsländern:

— Forschungsdatenarchive sollten von allen Beteiligten als wesentlicher Bestandteil der Infrastruktur für eine offene Wissenschaft anerkannt werden.
— Alle Forschungsdatenarchive sollten ein klar definiertes Geschäftsmodell haben.
— Politische Entscheidungsträger und Verantwortliche der Forschungsförderung sollen die unterschiedliche Art und Weise, wie Datenarchive finanziert werden können, und die Vor- und Nachteile verschiedener Geschäftsmodelle bei ihren Entscheidungen berücksichtigen.
— Die Geschäftsmodelle der Forschungsdatenarchive sind mit ihrem Auftrag (erteilte Mandate) sowie den gesetzten Anreizen (Eigenfinanzierung) abzustimmen.
— Die Datenarchive sollten Möglichkeiten zur Kostenoptimierung in Betracht ziehen, um die ihnen anvertrauten Daten langfristig effektiv verwalten zu können.

Für die Schweiz bedeutet die Umsetzung dieser Empfehlungen in erster Linie, sich einen Überblick über die bestehende Landschaft an Forschungsdatenarchiven zu verschaffen und deren unterschiedliche Finanzierungs- und Geschäftsmodelle zu verstehen («Swiss Research Data Repositories Landscape»). Auf dieser Grundlage kann eine mittel- bis langfristige Finanzierungspolitik für die Forschungsdatenarchive als zentraler Bestandteil der Forschungsinfrastruktur der Schweiz formuliert werden.

Zum Autor

André Golliez

André Golliez hat Anfang der 80er-Jahre an der ETH Zürich Informatik studiert und anschliessend über zehn Jahre im IT-Management der UBS gearbeitet. Seit 1999 ist er als selbständiger IT-Berater für Banken, öffentliche Verwaltungen und Forschungsinstitutionen tätig. Seit 2010 widmet André Golliez sich der Datenpolitik in der Schweiz – zunächst als Präsident des Vereins Opendata.ch und seit März 2017 auch als Präsident der neu gegründeten Swiss Data Alliance.

Fiabilité des Big Data du point de vue de la statistique publique

Bertrand Loison, vice-directeur de l'Office fédéral de la statistique
Diego Kuonen, fondateur et directeur général de Statoo Consulting

L'avènement des Big Data modifie le contexte dans lequel les organisations produisant des statistiques officielles opèrent. Bien que les Big Data offrent des opportunités, il n'en demeure pas moins vrai qu'un certain nombre de défis importants doivent encore être relevés afin d'en optimiser leur utilisation dans le contexte de la statistique publique.

L'ère des Big Data devrait avoir un impact important sur les organisations pour lesquelles la production et l'analyse de données et d'informations constitue le cœur de métier. Les instituts nationaux de statistique (INS) n'y font pas exception. Ils sont responsables de la production de la statistique publique qui est largement utilisée par les décideurs politiques et d'autres acteurs importants de la société. On peut raisonnablement poser comme postulat que la façon dont les INS adopteront ou pas les Big Data aura des implications pour l'ensemble de la société.

Les statistiques officielles sont souvent considérées comme allant de soi. Cependant, là où la confiance fait défaut, la société manque d'un pilier important pour une discussion pragmatique et l'élaboration de politiques publiques fondées sur des données probantes. Les normes et standards professionnels jouent un rôle vital pour assurer la confiance envers les statistiques officielles. La statistique publique dispose de ses propres codes de déontologie[1], [2], [3]. La prise en compte des Big Data dans la production de la statistique publique devra se faire dans le respect de la déontologie scientifique.

La confiance engendrée par le respect de ces codes de déontologie offre une position privilégiée aux INS en matière d'acquisition de données. De nombreux INS à travers le monde, dont l'Office fédéral de la statistique (OFS) en Suisse, ont déjà accès, conformément à la loi, aux sources de données gouvernementales. Certains pays ont légiféré ou sont en train de le faire pour permettre aux producteurs de statistiques officielles de pouvoir accéder gratuitement aux données de tierces parties (entreprises, ...). De plus, à des fins statistiques, de nombreux INS sont autorisés à appareiller des données provenant des différentes sources. La Suisse n'y fait pas exception[4].

Un nouvel écosystème

L'émergence de nouvelles sources de données crée pour les INS un bénéfice potentiel, mais cela rend aussi leurs produits moins uniques, puisque d'autres acteurs du marché de l'information ont commencé à produire des statistiques.

Le potentiel pour de nouvelles statistiques officielles est cependant bien réel. Par exemple, les données de localisation des téléphones mobiles pourraient être utilisées pour des statistiques quasi instantanées sur la population diurne et le tourisme. Les messages issus des médias sociaux pourraient être utilisés pour plusieurs types d'indicateurs, comme par exemple un indicateur précoce de la consommation. L'inflation pourrait être estimée à partir de l'information sur les prix disponible sur le web, et ainsi de suite. Toutefois, pour saisir ces opportunités, un certain nombre de défis doivent être surmontés.

Défis

Le principal défi auxquels les statisticiens officiels sont confrontés dans leur utilisation des Big Data est celui de la véracité des données qui représente le fondement de la confiance dans les données. Elle comprend la fiabilité, la solidité et la validité des données, leur qualité, ainsi que la transparence des processus de production des données. Un autre défi de taille concerne la méthodologie. De nombreuses sources de type Big Data, comme par exemple les messages issus des médias sociaux, sont composés de données d'observation et ne sont pas délibérément conçus pour l'analyse des données, et n'ont donc pas de population cible, ni de structure et ni de qualité. C'est pourquoi il est difficile d'appliquer les méthodes statistiques traditionnelles basées sur la théorie de l'échantillonnage.

Pour les INS, la question est donc de savoir comment la qualité des statistiques officielles peut être garantie si elles sont tout ou partiellement produites à partir de Big Data. L'utilisation des Big Data va induire un changement de paradigme et une utilisation accrue des méthodes d'analyse complémentaires (p. ex. l'analyse prédictive par des techniques statistiques avancées, la science des données et/ou l'apprentissage automatique).

La protection de la vie privée et les questions juridiques constituent d'autres défis, de même que les droits d'auteur et de propriété des données.

Pour les INS, il est essentiel de répondre à ces préoccupations par le biais de pratiques telles que la transparence quant à l'utilisation des sources de données et à la manière dont elles sont utilisées. Il en va de la crédibilité de la statistique publique.

L'avenir de la statistique publique

En cette période d'abondance croissante de données, la production d'informations statistiques potentiellement pertinentes pour la société n'est plus une activité intrinsèquement limitée aux INS.

Etant donné la concurrence croissante que les données générées par d'autres sources représentent vis-à-vis des INS en tant que porteurs des statistiques officielles, une réévaluation du positionnement stratégique de ceux-ci est nécessaire.

L'avenir des statistiques officielles à l'ère des Big Data fait encore l'objet de discussions. Le fait que la communauté internationale de la statistique publique doive s'adapter à une nouvelle réalité et répondre aux opportunités et aux défis auxquels elle est confrontée ne fait, lui, aucun doute.

L'OFS qui est membre du Global Working Group on Big Data for Official Statistics depuis 2017 a identifié ces défis et y a apporté une première réponse en publiant, en novembre 2017, sa stratégie sur l'innovation des données[5].

Les auteurs

Bertrand Loison

Le Prof. Dr Bertrand Loison, MPA IDHEAP, est vice-directeur de l'Office fédéral de la statistique (OFS) et chef de la division des registres, membre nommé du Comité de planification de la Cyberadministration Suisse et représentant de la Suisse au sein du «UN Global Working Group on Big Data for Official Statistics (ONU)». Il est également responsable du groupe de travail «New Data Sources» en charge de l'implémentation au sein de l'OFS de la stratégie d'innovation sur les données. Ses travaux se focalisent sur les changements induits par les nouvelles sources de données sur les offices nationaux de statistiques. Il est également Professeur en systèmes d'information au sein de la Haute école de gestion Arc (HES-SO).

Diego Kuonen

Le Prof. Dr Diego Kuonen, PhD en statistiques, statisticien accrédité (CStat et PStat) et scientifique accrédité (CSci), est fondateur et directeur général de Statoo Consulting (www.statoo.ch). Le Prof. Dr Diego Kuonen, CStat PStat CSci, intervient depuis de nombreuses années auprès de grands groupes industriels et de services en Europe. Depuis 2016, il est également Professeur en Data Science au sein de la Faculté d'économie et de management (GSEM) de l'Université de Genève, et fondateur et directeur de son nouveau programme de Master en Business Analytics. Actuellement, il est également le principal conseiller stratégique et scientifique externe de la Direction et du conseil de Direction de l'OFS dans le domaine d'expertise Big Data Analytics.

Notes

[1] United Nations (A/RES/68/261 from 29 January 2014) Fundamental principles of official statistics. Disponible à l'adresse: https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-F.pdf (consulté le 25 mai 2018).

[2] La Suisse est membre du SSE depuis la signature le 26.10.2004 de l'accord bilatéral Suisse – Union européenne sur la statistique. Disponible à l'adresse: https://www.eda.admin.ch/dea/en/home/bilaterale-abkommen/ueberblick/bilaterale-abkommen-2/statistik.html (consulté le 25 mai 2018).

[3] Le code de bonnes pratiques des statistiques européennes est également valable en Suisse. Disponible à l'adresse: https://www.bfs.admin.ch/bfs/fr/home/ofs/engagement-qualite.html (consulté le 12 mai 2018).

[4] L'appariement de données à des fins statistiques est réglé à l'art. 14a de la loi sur la statistique fédérale (LSF; RS 431.01). Disponible à l'adresse: https://www.bfs.admin.ch/bfs/fr/home/services/appariement-donnees/generalites.html (consulté le 26 mai 2018).

[5] Stratégie d'innovation sur les données. Disponible à l'adresse: https://www.bfs.admin.ch/bfs/fr/home/actualites/quoi-de-neuf.gnpdetail.2017-0673.html

re3data – ein internationales Verzeichnis von Forschungsdateninfrastrukturen

Frank Scholze, KIT-Bibliothek, Karlsruher Institut für Technologie

Forschungsdaten sind wertvoll und allgegenwärtig. Die nachhaltige Verfügbarkeit von Forschungsdaten ist eine Herausforderung für alle Beteiligten in den wissenschaftlichen Communities. Dennoch bieten Langzeitarchivierung und permanenter Zugang zu Forschungsdaten grosse Chancen für die Wissenschaft. Es kann jedoch sehr schwierig sein, geeignete Infrastrukturen (meist Repositorien) zu finden, um Forschungsdaten zu speichern, zu bearbeiten und nachzunutzen.

re3data (https://www.re3data.org/) ist ein internationales Verzeichnis von Forschungsdaten-Repositorien, das nicht auf bestimmte akademische Disziplinen beschränkt ist. re3data fördert eine Kultur des Teilens, einen besseren Zugang und eine bessere Sichtbarkeit von Forschungsdaten und unterstützt die FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable). Es vermittelt Repositorien für die dauerhafte Speicherung und den Zugriff auf Datensätze für Forschende, Förderer, Verleger und Journalisten. Bei der Konzeption und Implementierung des Dienstes wurde deutlich, dass das Auffinden und Speichern von Forschungsdaten, das Analysieren, Aufbauen und Integrieren von Infrastruktur und Diensten auf der Grundlage der bereitgestellten Informationen im Zentrum stehen sollte.

Informationsquelle für ein breites Publikum

Als Einstieg in wissenschaftliche Ressourcen ist re3data nicht auf Forschende beschränkt. Es dient bereits als Informationsquelle für Förderorganisationen, politische Entscheidungsträger und die öffentliche Verwaltung. re3data ist z.B. ein Datenanbieter für den EU Open Science Monitor. Rund 2100 Repositorien sind Stand Juni 2018 indexiert und decken ein breites Spektrum an Forschungsdaten ab. Diese Daten stehen ausser Forschenden auch für Datenjournalisten oder die interessierte Öffentlichkeit zur Verfügung. Da re3data nicht domänenspezifisch ist, deckt es eine grosse Anzahl von Fachgebieten aus den Ingenieur-, Natur-, Geistes- und Sozialwissenschaften sowie den Lebenswissenschaften ab.

Das Metadatenschema

re3data stellt strukturierte Metadaten für jedes Repositorium bereit und benötigt daher ein Metadatenschema, das von vielen anderen als De-facto-Standard übernommen wurde. Das Schema von re3data deckt verschiedene Aspekte von Repositorien ab (Allgemeine Informationen, Richtlinien, Rechtliche Aspekte, Technische Standards, Qualitätsstandards, Art des Zugriffs). Zur schnellen visuellen Orientierung spiegeln sich diese Kategorien auch im Icon-System von re3data wider, in dem komplexe Eigenschaften in einfachen grafischen Symbolen ausgedrückt werden.

Die Beschreibung von Forschungsdateninfrastrukturen ist aufgrund der fortgesetzten und vielseitigen Entwicklung des Forschungsdatenmanagements ein dynamischer Prozess. re3data begann 2012 mit einigen grundlegenden Eigenschaften und entwickelte sich zu einem umfassenden Metadatenschema. Derzeit ist das Metadatenschema in Version 3.0 mit insgesamt über 141 Eigenschaften verfügbar. Die anstehenden Anforderungen aus der Forschungsgemeinschaft wurden jedoch bereits identifiziert und weitere Versionen sollen veröffentlicht werden.

Vielfältige Suchmöglichkeiten

re3data zeigt Repository-Ergebnisse als eine lange einfache Liste, aber es bietet auch verschiedene Funktionen für Forscher, um Repositorien ohne viel Aufwand zu explorieren und zu finden. Neben der Suche nach Einträgen nach Stichworten enthält das Tool auch eine facettierte Suchfunktion und ein visuelles Suchwerkzeug. Dadurch können Benutzer die Entitäten für jede Eigenschaft mithilfe des Metadatenschemas filtern, sodass sie die Ergebnisliste ihren spezifischen Anforderungen anpassen können. Beispielsweise kann ein Journalist den gewünschten Themenbereich schnell eingrenzen oder Repositorien für ein bestimmtes Land suchen.

Mit seinem internationalen Bestand an Repositorien bietet re3data auch grundlegende statistische Analysen. Für weitere Untersuchungen innerhalb der Forschungsinfrastrukturlandschaft werden alle Daten über eine sogenannte API angeboten. Damit die Daten von re3data so einfach wie möglich nachgenutzt werden können, sind sie unter einer Creative Commons CCo-Lizenz kostenlos zugänglich.

Zum Autor

Frank Scholze

Frank Scholze ist Direktor der Bibliothek des Karlsruher Instituts für Technologie (KIT) und Mitglied des Bundesvorstandes des Deutschen Bibliotheksverbandes (dbv) sowie einer Reihe von wissenschaftlichen Ausschüssen und Beiräten; unter anderem bei DARIAH-DE (Digital Research Infrastructures for the Arts and Humanities). Frank Scholze ist Sprecher der AG Elektronisches Publizieren der Deutschen Initiative für Netzwerkinformation (DINI) und ist an zahlreichen Projekten im Bereich Digitaler Bibliotheken und Forschungsinformation beteiligt.

Literatur

— Pampel, H., Bertelmann, R., Scholze, F., Kindling, M., Vierkant, P. (2015). Stand und Perspektive des globalen Verzeichnisses von Forschungsdaten-Repositorien re3data.org. In: Müller, P. (Eds.), *8. DFN-Forum Kommunikationstechnologien: Beiträge der Fachtagung 08.06-09.06.2015 in Lübeck*, (GI-Edition: lecture notes in informatics; 243), Bonn: Gesellschaft für Informatik, pp. 13–22. http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1169893:3

Forschungsplattformen im Kontext von Open und FAIR Data

(bk) Der November 2018 gehört den geisteswissenschaftlichen Forschungsplattformen. Die Schweizerische Akademie der Geistes- und Sozialwissenschaften präsentiert an zwei Tagungen die neusten Entwicklungen im Bereich Open Data, FAIR Data und Big Data und zeigt die Auswirkungen auf die Geistes- und Sozialwissenschaften.

Open und FAIR Data

Am Freitag, 2. November 2018, findet im Kursaal Bern die Tagung «Geisteswissenschaftliche Forschungsplattformen in der Schweiz im Kontext von Open und FAIR Data» statt. Die Schweizer Forschungslandschaft umfasst eine Fülle von geistes- und kulturwissenschaftlich relevanten Datenbeständen. Wie alle wissenschaftlichen Infrastrukturen sind auch die geisteswissenschaftlichen Forschungsplattformen mit Forderungen nach Open und FAIR Data konfrontiert.

Open Data fügt sich in die Forderung nach Open Science insgesamt ein, die offene Wissenschaftsprozesse und frei zugängliche Forschungsergebnisse verlangt. Die FAIR-Data-Principles wiederum sind ein Set von Richtlinien, welche die Auffindbarkeit, Zugänglichkeit, Interoperabilität sowie Wiederverwendbarkeit von Daten gewährleisten sollen. Die geisteswissenschaftlichen Forschungsinfrastrukturen und -plattformen erfahren aus Sicht der SAGW nicht überall eine adäquate materielle und finanzielle Berücksichtigung, obwohl das Thema in der Schweiz generell an Relevanz gewinnt.

An der Tagung werden inhaltliche, technische, finanzielle und strategische Voraussetzungen identifiziert, damit geisteswissenschaftliche Forschungsplattformen die Anforderungen nach Open und FAIR Data umsetzen können.

Big Data in den Sozialwissenschaften

Eine Woche später, am 9. November 2018, geht es um Big Data. Im Berner Hotel Kreuz findet die Tagung «Big Data in den Sozialwissenschaften – Herausforderungen und Chancen» statt. In der Schweiz wird die Diskussion

zu Big Data vorwiegend unter informatikbezogenen und juristischen Aspekten geführt. Es bleibt unklar, was mit Big Data genau gemeint ist. Oft werden unrealistische Erwartungen an die Nutzung solcher Daten für die sozialwissenschaftliche Forschung geknüpft.

An der Tagung diskutieren wir die zu erwartenden Ergebnisse und Erkenntnisse für die gesellschaftliche und politische Praxis. Mit sozialwissenschaftlichen Anwendungsbeispielen zeigen wir, dass die Beschäftigung mit Big Data Eingang in die Forschung gefunden hat und dass relevante wissenschaftliche sowie anwendungsorientierte Resultate vorliegen.

Veranstaltungen

Geisteswissenschaftliche Forschungsplattformen in der Schweiz im Kontext von Open und FAIR Data
Freitag, 2. November 2018, Kursaal Bern

Big Data in den Sozialwissenschaften – Herausforderungen und Chancen
Freitag, 9. November 2018, Hotel Kreuz, Bern