Forschungsdateninfrastruktur

Auf zu neuen Ufern

Stürmische Zeiten für datengestützte Natur- und Geisteswissenschaften: Das Projekt GeRDI hilft Forscherinnen und Forschern durch den Daten-Ozean.

Von Tobias Weber

WIR MÜSSEN UNS Alexis de Tocqueville als einen ausdauernden Menschen vorstellen. Als der französische Politiker und Publizist im Mai 1831 nach stürmischer Überfahrt in New York an Land ging, lagen noch neun Monate und 12.000 Kilometer durch Nordamerika vor ihm. Der Aufwand sollte sich lohnen: Auf dieser Reise recherchierte er unter anderem für sein Hauptwerk „De la Démocratie en Amérique". Heute sind viele Informationen für ein solches Vorhaben in Sekundenschnelle über den ganzen Globus zusammengetragen, ohne dass man dazu das Büro verlassen muss. Der Blick in die Vergangenheit ist wie so oft aufschlussreich: Nie waren die Möglichkeiten der Forschenden größer als heute, Daten für die Beantwortung wissenschaftlicher Fragestellungen heranzuziehen.

Seestück von Gerhard Richter.

Mit dem Verlassen des Schiffs hatte Tocqueville seinerzeit eine wesentliche Grenze überwunden – den Atlantik zwischen Europa und der Neuen Welt. Ähnlicher Seegang ist bei der Navigation durch die Meere datengestützter Wissenschaft zu erwarten. Im Folgenden wollen wir einen Blick auf diese raue See werfen, speziell auf ein Schiff, das vor gut einem Jahr in See gestochen ist: Das DFG-Projekt GeRDI (*Generic Research Data Infrastructure*) begleitet acht wissenschaftliche Communities bei ihrer datengestützten Forschung und hat sich zur Aufgabe gesetzt, mit Lösungen und Services zur Erschließung neuer Seerouten auf den Daten-Ozeanen beizutragen.

Mit Otto Neuraths Schiffsmetapher im Hinterkopf wollen wir zugleich vorausschicken, dass man die Wissenschaft nicht einfach einstellen kann, um Aufrüstungen im Trockendock vorzunehmen. Man muss also auf offener See die notwendigen Reparaturen tätigen, wenn nötig Planke für Planke. Zentral für GeRDI ist demnach eine Bestandsaufnahme: Wie arbeiten Forschende bereits mit Daten, und wie kann man diese Arbeit unterstützen?

Während einige Disziplinen wie die Physik oder die Lebenswissenschaften schon in der Businessklasse reisen und zum Teil auf jahrzehntelange Erfahrung mit großen Datenmengen zurückblicken, gibt es Disziplinen, die sich zum ersten Mal auf offene See wagen. Die digitalen Geisteswissenschaften wurden zwar nicht erst kürzlich aus der Taufe gehoben – Roberto

Alexis de Tocqueville (1805–1859), Porträt von Théodore Chassériau.

ABB.: GERHARD RICHTER / GERHARD-RICHTER-IMAGES; WIKIMEDIA CC

Der Alpenraum ist ausdrucksreich: Die Vielfalt landwirtschaftlicher Bezeichnungen der slawischen, romanischen und germanischen Sprachen sind im Projekt *VerbaAlpina* ein Forschungsgegenstand.

Busa bediente sich schon 1949 der elektronischen Datenverarbeitung, um Thomas von Aquin besser zu verstehen. Sie verlassen aber erst nach und nach den Bereich, in welchem sie durch digitale Techniken ihre Arbeit „nur" schneller oder effizienter erledigen können, und entwickeln darüber hinaus Ansätze, wie datengestützte Forschung auch wissenschaftliche Methoden verändern kann.

VerbaAlpina: Kultur- und Sprachgeschichte des Alpenraums

Ein Beispiel hierfür steht auch auf der Passagierliste von GeRDI: Das Projekt *VerbaAlpina* der LMU München. Sein Ziel es ist, den einzelsprachlich und dialektal stark fragmentierten Alpenraum kultur- und sprachgeschichtlich zu erschließen. Die Sammlung wird mit multimedialen Methoden ausgewertet und aufbereitet. Die Daten werden unter anderem in einer interaktiven Karte dargestellt, wodurch sich neue Analyseansätze ergeben. Darüber hinaus werden auch Citizen Science-Methoden angewandt – also die Möglichkeit, dass sich Laien aktiv an Forschung beteiligen: Ortsübliche Ausdrücke können über ein Webinterface eingegeben werden. Dieser wertvolle Datenbestand soll unter anderem durch das GeRDI-Projekt eine größere Reichweite erlangen und mit anderen Daten verknüpft werden.

HiOS: Sturzfluten und wild abfließendes Wasser in Bayern

Die Mitarbeiter und Mitarbeiterinnen des Projekts HiOS (*Hinweiskarte Oberflächenabfluss und Sturzflut*) sind Teil einer weiteren Community, die wir auf unserem Schiff mitnehmen wollen. Nicht erst seit den katastrophalen Überflutungen im niederbayerischen Simbach besteht das valide Interesse, diese Ereignisse besser zu verstehen, um schnell auf sie zu reagieren oder sie sogar voraussagen zu können. Zentral für die Fragestellung der Binnenwasserforschung sind hydrologische und hydrodynamische Modelle, welche die Hochwasserwellen und die daraus resultierenden Kräfte quantifizierbar machen. Diese benötigen eine Vielzahl an unterschiedlichen Daten: von der Niederschlagsmenge über Bodenbeschaffenheit und Fauna bis zur wirtschaftlichen Nutzung des Einzugsgebiets. Aus solchen Modellen können Verfahren entwickelt werden, die es Kommunen erlauben, sich auf Hochwassergefahren vorzubereiten. Das Management sowohl der Eingabe- wie auch Ausgabedaten spielt dabei eine zentrale Rolle. Hier unterstützt das Leibniz-Rechenzentrum als Projektpartner sowohl in HiOS als auch in GeRDI das Vorhaben.

Wissenschaftliche Daten interdisziplinär teilen

Man könnte nun den Seeleuten zurufen: „Wozu braucht Ihr denn eine generische Dateninfrastruktur? Haben wir nicht genügend Daten vor Ort? Die Daten sind außerhalb einer Disziplin doch ganz und gar unverständlich!“ Zunächst kann man feststellen, dass Daten an sich weder Disziplin- noch Institutionengrenzen kennen. Sie werden aber oftmals zur Beantwortung einer bestimmten Fragestellung für eine spezifische Disziplin gesammelt oder digitalisiert. Daraus folgt jedoch nicht, dass sie für andere Disziplinen uninteressant sind.

Beispiele hierfür finden sich leicht: Wenn der Einfluss von Wetter auf die Sprachgewohnheit in einem bestimmten Alpental Untersuchungsgegenstand wird, gewinnen meteorologische Daten auch für Linguisten an Bedeutung. Historisches Kartenmaterial wiederum kann Aufschluss über die Genese von Flussverläufen geben und entsprechende Modelle befüttern.

Beide Beispiele überspringen die Grenze zwischen Natur- und Geisteswissenschaften mit Leichtigkeit, da die geographische und temporale Indexierung gemeinsame Dimensionen aufspannt. Beim multi- und interdisziplinären Teilen von Daten sind wir gerade erst dabei, das Potential zu erschließen, geschweige denn auszuschöpfen.

Hier liegt allerdings schon eine der großen Herausforderungen: Während Meteorologen ihre Niederschlagsdaten mit Gauß-Krüger-Koordinaten versehen, nutzen Linguisten unter Umständen Ortsnamen oder ein anderes Koordinatensystem. Historiker verschlagworten ihre Karten möglicherweise mit der Technik der Herstellung (Kupferstich, Lithographie etc.), die ein Hydrologe eher nicht als Suchbegriff in Erwägung zieht. Die unterschiedliche Sicht auf die Daten macht eine semantische Unschärfe deutlich: Es ist unklar, wann verschiedene Disziplinen von derselben Sache sprechen. Dies zeigt sich auch in den zahlreichen Standards für die Metadaten,

ABB.: FREILICHTMUSEUM GLENTLEITEN (2); SIGRID MITTERER

2016 versank Simbach in den Fluten. Das Projekt HiOS untersucht Starkregen-Ereignisse, die auch der Grund für diese Katastrophe waren.

DER AUTOR

Tobias Weber ist als wissenschaftlicher Mitarbeiter am Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften zuständig für das Projekt GeRDI.

die eine digitale Information beschreiben sollen. Die Integration von Daten über gewachsene Fächergrenzen hinweg ist keine triviale Aufgabe.

Diese Aufgabe wird in Zukunft für die meisten Wissenschaften obligatorisch werden, die sich um Fördergelder für datengestützte Wissenschaft bewerben. In europäischen Projektanträgen für das Horizon 2020-Programm der Europäischen Union sind Datenmanagementpläne bereits vorgeschrieben. Datenintegration ist hierbei nur ein Bestandteil: Sollen wissenschaftliche Ergebnisse über Fächergrenzen hinweg nutzbar sein, müssen sie auffindbar, zugreifbar, interoperabel und nachnutzbar sein. Die Richtlinien schreiben zudem vor, Daten ohne Zugriffsbeschränkung der Öffentlichkeit zur Verfügung zu stellen, wenn nicht valide Gründe wie Datenschutz oder der Schutz geistigen Eigentums dagegensprechen. Diese Anforderungen können Forschende kaum alleine erfüllen. Es braucht die Zuarbeit von wissenschaftlichen Dienstleistern, wie sie auch im GeRDI-Projekt vertreten sind: Bibliotheken (das Leibniz-Informationszentrum Wirtschaft aus Hamburg/Kiel), Rechenzentren (das Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften in Garching und das Rechenzentrum der TU Dresden), Infrastrukturverbünde (das Deutsche Forschungsnetz mit Sitz in Berlin) und Programmierer (der Software-Engineering-Lehrstuhl der Universität zu Kiel).

Forschung betreiben mit GeRDI

Die FAIR-Prinzipien (aus dem Englischen: findable, accessible, inter-operable and re-useable) sind auch die Grundkoordinaten für das GeRDI-Projekt. Wie kann man nun mithilfe dieser Prinzipien navigieren, in See stechen, neue Ufer erreichen – also Forschung betreiben? Nur wenn die Daten auffindbar sind, können sie nachgenutzt werden. Deshalb ist eine zentrale Komponente in GeRDI das Abfragen der Metadatenbestände diverser Datenrepositorien und daraus resultierend der Aufbau eines Suchindex, auch über die Grenzen der Wissenschaft hinweg (nicht alle Daten, die in der Wissenschaft genutzt werden, kommen originär aus einem akademischen Kontext). An dieser Stelle wird auch die semantische Herausforderung aufgegriffen: Eine Anfrage verschiedener Koordinaten oder gar Ortsnamen auf dem Index sollte gleiche oder zumindest ausreichend ähnliche Ergebnisse liefern. Um sich auf dem Daten-Ozean zurechtzufinden, braucht es Kartenmaterial – der Index erfüllt diese Funktion.

Da die unterschiedlichen Disziplinen teils sehr spezifische Perspektiven auf das Thema „Suche“ haben, kann der Index neben einer generischen Weboberfläche auch in gewohnte Umgebungen eingebunden werden: Die Suche der Universitätsbibliothek, die virtuelle Forschungsumge-

Um Simulationen mit Daten zu füttern, müssen Forschende ins Feld. Hier eine Messstation zur Berechnung des Abflussvolumens.

bung oder der Kommandozeilenzugriff können unterschiedliche Sichten auf den Index bereitstellen. Ist das Suchergebnis zufriedenstellend, kann eine Auswahl getroffen und gespeichert werden. Daraus resultiert ein „Daten-Rezept", das beim Daten-Bibliothekar eingelöst und dann nachgekocht werden kann: Angefangen beim Statistik-Skript auf dem eigenen Laptop bis hin zur Simulation, die einen Supercomputer mit 240.000 Rechenkernen benötigt.

Außerhalb der GeRDI-Core-Services (Suche und Selektion) versteht sich das Projekt als Vermittler der Datenservices, die für Forschende von Institutionen vor Ort angeboten werden. Letztlich muss jede wissenschaftliche Infrastruktur, die erfolgreich sein will, nach diesem Prinzip der Subsidiarität gebaut sein. Das Katalogisieren, Auffinden und Selektieren von Daten geschieht zentral, der Bezug, die Verarbeitung und die Speicherung lokal an den ortsansässigen Rechenzentren und Bibliotheken. Einige Services zur Analyse und Archivierung von Forschungsdaten werden im Rahmen von GeRDI entwickelt, allerdings nur um zu zeigen, wie Dienstleister ihre Services an die Forschungsdateninfrastruktur einbinden können. Hier wird auch deutlich, welche Probleme GeRDI nicht lösen wird: Daten zu speichern, zu kuratieren und zur Verfügung zu stellen – dies alles wird weiterhin dezentral geschehen, sei es bei der Universitätsbibliothek oder mit den Services eines fachspezifischen Datenrepositoriums. GeRDI will die Routen zwischen diesen Daten-Häfen aufzeigen, nicht die Kontore leerplündern.

Welche Risiken birgt die datengestützte Wissenschaft?

Man darf nicht ausschließlich von den Chancen der erschlossenen Datenreichtümer sprechen, dabei aber die Risiken verschweigen. Die Einführung der datengestützten Wissenschaft hat durchaus disruptiven Charakter: Während ernstzunehmende Publikationen durch einen Review-Prozess gehen müssen, gibt es für Daten noch keine vergleichbare flächendeckende Qualitätssicherung. Und die Probleme, die selbst der Review-Prozess bisher nicht zu lösen vermag, bestehen bei Daten natürlich ebenso fort: Die mangelhafte Reproduzierbarkeit von Datenerhebungen und -auswertungen ist genau das Grundproblem der Publikationen, die Schlüsse aus den Daten ziehen. Offene Methoden und transparente Verfahren sind die besten Mittel gegen diese *replication crisis*.

Messstationen für Temperatur, Feuchtigkeit und Luftdruck stehen oft an entlegenen Orten. Ein Grund mehr, die Daten so zu teilen, dass viele Forschende profitieren.

Die Möglichkeit, öffentlich kritisiert und widerlegt zu werden, ist gerade das wesentliche Instrument zur Qualitätssicherung in der Wissenschaft. GeRDI will hierzu einen Beitrag leisten und für Open Science-Methoden werben. Der Bezug der (Meta-)Daten via GeRDI-Infrastruktur soll die manuelle wie automatische Reproduzierbarkeit wissenschaftlicher Erkenntnisse unterstützen.

Letztlich hat die kritische Betrachtung selbst vor Tocqueville nicht haltgemacht. Heute würden die meisten Forschenden seiner Einschätzung widersprechen, die Ureinwohner hätten keine Besitzansprüche auf den Kontinent gehabt, da sie Nomaden waren. Wer zu neuen Ufern aufbricht und Neuland betritt, mag eine offene Denkweise haben, ist aber nicht gegen alle Fehlurteile gefeit. ■

WWW

www.gerdi-project.de
(GeRDI, ein bundesweites Projekt zum Aufbau einer vernetzten Forschungsdateninfrastruktur)

www.verba-alpina.gwi.uni-muenchen.de
(Projekt VerbaAlpina der LMU München)

www.hios-projekt.de
(Projekt HiOS der TU München)

BEIDE ABB.: LEHRSTUHL HYDROLOGIE UND FLUSSGEBIETSMANAGEMENT / TECHNISCHE UNIVERSITÄT MÜNCHEN