

Der gläserne Forscher

Unermüdlich häuft die Wissenschaft Daten an. Jeden Tag werden es mehr. Zum Schatz der Erkenntnis werden die Informationen aber erst, wenn sie aufbereitet und für andere zugänglich sind. Doch genau das ist nicht einfach. Es gibt Widerstände, nicht nur finanzieller Art.

von Bernd Kramer

Wissen ist Macht. Und Daten sind es erst recht. Experimente, Umfragen, Exkursionsberichte – unermüdlich türmen Forscher Messwerte auf. Das Problem: Ihre Zahlenkolonnen betrachten viele Wissenschaftler als Privateigentum. „80 bis 90 Prozent aller Forschungsdaten verschwinden in der Schublade“, schätzt Dr. Stefan Winkler-Nees, der sich bei der Deutschen Forschungsgemeinschaft (DFG) mit dem Thema befasst: Es sind Ergebnisse von Vorstudien, die nicht weiter genutzt werden; Datensätze, von denen niemand erfährt, weil sie nie zu Publikationen führen; Expeditionsprotokolle, die nach den Jahren einfach vernichtet werden. „Die meisten wissenschaftlichen Daten werden außer von denen, die sie erheben, von niemandem mehr angefasst.“ Doch langsam wächst der Druck zum Datenteilen. Wissenschaftler sollen ihre Forschungsdaten offenlegen. Forscher werden gläsern, das ist die Hoffnung – und die Befürchtung.

Zehn Jahre Pflicht zur Datensicherung – für Hochschulen ein ehrgeiziges Ziel

Der Preis ist hoch

Es gibt gute Argumente, weshalb Forscher Rohdaten offenlegen sollten. Die Deutsche Forschungsgemeinschaft (DFG) sieht darin einen Teil guter wissenschaftlicher Praxis. In ihren im Juli vergangenen Jahres überarbeiteten und aktualisierten Richtlinien dazu heißt es: „Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden.“ Dadurch könnten sich Hochschulen Datenmanipulationen und -fälschungen wappnen. Denn wie Dr. Daniele Fanelli von der Universität Edinburgh 2009 in einer Studie herausgefunden hat, haben immerhin fast zwei Prozent aller Wissenschaftler schon einmal an ihren Daten gedreht, um das Ergebnis zu frisieren. Aufdecken lässt sich solch ein Betrug eben nur, wenn die Daten gesichert und zentral in den Hochschulen aufbewahrt werden.

Das allerdings kostet. Als ein „ambitioniertes Ziel“ bezeichnet die Hochschul-

rektorenkonferenz in ihren Mitte Mai veröffentlichten Empfehlungen zur guten wissenschaftlichen Praxis denn auch den DFG-Vorschlag einer zehnjährigen Datenspeicherung. Die Hochschulen müssten elektronische Datenspeicher anschaffen.

Was eine sichere Aufbewahrung kostet, ist schwer zu beziffern. Als Daumenregel gilt: Etwa zehn Prozent der Forschungsausgaben fallen für die Datensicherung an. Dr. Stefan Winkler-Nees von der DFG geht davon aus, dass es in der Praxis aber oft deutlich günstiger sein kann, Faustregel hin oder her. Und ohnehin: Winkler-Nees empfiehlt, die Pflege all der Datenmengen nicht als Kostenfaktor zu sehen – sondern als ein Pfund, mit dem Forscher und Hochschulen wuchern können.

Wie viele neue Erkenntnisse man gewinnen könnte, wenn alle Forscher auf die vielen versteckten Datenmengen der Wissenschaftsgemeinde zugreifen könnten – diese Idee treibt die Anhänger von Open Data, der freien Verfüg- und Nutzbarkeit von Forschungsdaten (s. *duz* MAGAZIN 03/2011, S. 8ff.). Open Data ist das Pendant zum Open Access, dem kostenfreien Zugang zu wissenschaftlichen Veröffentlichungen – und die logische Fortsetzung: Auch die Allianz der Wissenschaftsorganisationen, der Zusammenschluss der wichtigsten Forschungsorganisationen, empfahl bereits vor drei Jahren, neben Publikationen auch die zugrundeliegenden Rohdaten zugänglich zu machen.

Der offene Datenzugriff braucht Regeln, genau wie Open Access (s. *duz* MAGAZIN 08/13, S. 39ff.) – „sonst ist es nicht praktikabel“, sagte Prof. Dr. Jörg Hacker, Präsident der Nationalakademie Leopoldina, in einem Interview. Doch Open Data kostet nicht nur Geld, sondern kann auch helfen zu sparen. Künftig würden zum Beispiel nicht mehr so viele Messungen aus purer Unkenntnis doppelt gemacht.

Zuerst müssen aber die Wissenschaftler überzeugt werden. Der Konkurrenzdruck um Reputation und Forschungsgelder ist groß. Da will sich niemand gerne in die Karten schauen lassen und womöglich noch anderen mit hart erarbeitetem Material zu Ruhm verhelfen. Zum gläsernen Forscher zu werden, das ist auf den ersten Blick keine verlockende Aussicht.

Ein Zeichen der Qualität

Einige Initiativen versuchen daher die Wissenschaftler bei ihrer Ehre zu packen. Wie das geht, demonstriert Pangaea (Data Publisher for Earth and Environmental Science), ein Vorreiterprojekt im Open-Data-Bereich. Vor rund 20 Jahren haben Dr. Michael Diepenbroek, Meeresgeologe von der Universität Bremen, und Dr. Hannes Grobe vom Alfred-Wegener-Institut, dem Helmholtz-Zentrum für Polar- und Meeresforschung in Bremerhaven, angefangen, die Datenbibliothek für Geowissenschaftler und Umweltforscher aufzubauen. Mittlerweile birgt Pangaea einen Schatz von 350 000 Forschungsdatensätzen, die auf dem Server hinterlegt sind. Das sind rund sechs Milliarden Einzelmessungen.

Der Weg zum Datenschatz führt über eine Website, die aussieht wie die Suchmaschine Google: Spartanisch weiß die Oberfläche, der Cursor blinkt in einem Eingabeschlitz in der Seitenmitte. Wer zum Beispiel CO₂ eintippt, bekommt mehr als 10 000 Treffer aus allen Ecken

„80 bis 90 Prozent aller Forschungsdaten verschwinden in der Schublade“

der Welt: Aktuelle Luftuntersuchungen aus der Antarktis oder Wassermessungen eines Forschungsschiffes rund um die Färöer-Inseln, die in das Jahr 1968 zurückreichen. Mit einem Klick können die Daten heruntergeladen und nach eigenen Fragestellungen ausgewertet werden.

„In der Community sind wir eine wichtige Anlaufstelle“, sagt Diepenbroek. Pangaea hat sich etabliert unter den Geowissenschaftlern und Umweltforschern. Wissenschaftler, deren Messungen in die Pangaea-Bibliothek aufgenommen werden,

können das nämlich mittlerweile als Qualitätsausweis für ihre Arbeit werten: Diepenbroek und seine Kollegen unterziehen alle eingereichten Tabellen und Zahlenkolonnen einem strengen Check. Und das bei rund 10000 neuen Datensätzen jedes Jahr, die auf der Pangaea-Homepage hochgeladen werden. Sind die Ergebnisse gut dokumentiert? Sind die Werte plausibel? Gibt es extreme Ausreißer?

„Viele Wissenschaftler denken, dass ihre Daten für andere Nutzer bereits allgemein verständlich wären“, sagt Diepenbroek, „das sind sie in der Regel nicht.“ Ein Beispiel: Ein Doktorand hatte Wassermesswerte auf dem Server hinterlegen wollen – allerdings ohne die dazugehörigen Geo-Koordinaten anzugeben. Für andere Forscher wären die Daten dadurch wertlos. „300 Meter rechts hinter der Boje ist keine Angabe, mit der ein Wissenschaftler etwas anfangen kann“, sagt Diepenbroek. Der Doktorand musste die fehlenden Angaben nachliefern. Das Vorgehen des Pangaea-Teams ähnelt damit der Begutachtung von Zeitschriftenaufsätzen. Die aufgenommenen Daten gelten damit als peer-reviewed – von der Fachgemeinschaft für gut befunden.

Auch Rohdaten zitieren

Das allein reicht allerdings noch nicht, um Forschern ihre Messschätze zu entlocken. Deshalb gilt bei Pangaea: Wer sich die Zahlenwerke eines Forschers herunterlädt, muss den Originaldatensatz in der eigenen Arbeit als Quelle aufführen. „Daten müssen zitiert werden wie die übrige Literatur auch“, fordert Diepenbroek. Das Kalkül: Die Bereitschaft zum Teil steigt, wenn nicht nur Aufsätze und Buchbeiträge eines Forschers in den neusten wissenschaftlichen Veröffentlichungen seiner Kollegen genannt werden, sondern auch die Datensätze. Zitate sind schließlich die Währung der Wissenschaft, ihr Gradmesser für Einfluss und Reputation.

Der Kampf um den Rohstoff des Wissens ist das Eine, was die Wissenschaftler bei Open Data fürchten. Datenbanken innerhalb einer Hochschule oder Forschungseinrichtung können aber auch ein wichtiges Steuerungsinstrument sein. „Universitäten werden von der Politik dazu angehalten, sich auf ihre Stärken zu besinnen und ihr Profil zu schärfen“, sagt Dr. Sebastian Herwig, der die Abteilung Forschungsinformation der Universität Münster leitet, „dafür brauchen sie aber erst einmal Informationen über sich selbst.“ Mit einem Überblick über alle an ihr gesammelten Daten könnte eine

Hochschule den Ministerien und Förderern dann wichtige Anhaltspunkte liefern. Seit gut zwei Jahren baut die Universität Münster daher eine Datenbank auf, in der ihr gesamtes Forschungsgeschehen zusammengetragen wird.

Als die Hochschule vor einigen Jahren einen Antrag für die Exzellenzinitiative vorbereitete, fiel plötzlich auf, wie wenig sie über das Forschungsgeschehen im eigenen Haus wusste. So wurde eine Expertin in ihrem Fach gar nicht berücksichtigt. Nicht weil sie zu wenig geforscht hatte – sondern weil ihre Publikationen für die Uni-Oberen plötzlich nicht mehr auffindbar waren. „Dabei hatte sie nur geheiratet und unter einem anderen Namen weiterveröffentlicht“, sagt Sebastian Herwig. Änderungen des Familienstandes verfolgt die Personalabteilung. Das Publikationsregister der Bibliothek aber war blind für den Namenswechsel. Der Datenwarr ist perfekt.

Für die Universität Münster ein Schlüsselerlebnis. Welche Wissenschaftler kommen und gehen – das wusste bisher nur die Personalabteilung. Wer welche Projektmittel eingeworben hat – diese Informationen liegen in der Haushaltsstelle. Separat davon verwaltet die Bibliothek Publikationen. Kurz: „Jede Stelle machte ihre eigene Excel-Tabelle“, sagt Sebastian

Herwig. In der Datenbank ist jetzt für jeden der rund 6500 Wissenschaftler aus über 100 Fächern an der Universität ein eigener Eintrag vorgesehen, eine Art festes Personenprofil, zu dem laufend sämtliche Forschungsleistungen verzeichnet werden sollen. Auch bei einem Namenswechsel.

„Das einheitliche Forschungsinformationssystem ist eine weitestgehend vorgefertigte, webbasierte Anwendung, das erleichtert die Verteilung innerhalb der Uni“, sagt Herwig. Zwar könnten die Hochschulen ein solches System auch selbst entwickeln. Herwig aber rät zum Einkauf einer Standardanwendung mit der Möglichkeit zur flexiblen individuellen Anpassung, „da an Hochschulen häufig die nachhaltige Betreuung und Entwicklung von Eigenlösungen nicht gegeben“ sei. Die Kosten für den Aufbau liegen für eine mittlere bis große Uni im unteren sechsstelligen Bereich für die Softwareeinführung, schätzt Herwig. Dazu kommen Personalkosten. In Münster beschäftigen sich zwei Vollzeitkräfte mit Betrieb und Weiterentwicklung der Daten-

bank. Im Moment entscheidet noch jede Hochschule für sich, welche Angaben sie mit welcher Software über ihre Forscher erhebt. Doch das soll sich ändern.

Vor ziemlich genau einem Jahr empfahl der Wissenschaftsrat, in Deutschland einen „Kerndatensatz Forschung“ zu entwickeln; ein einheitliches Sammelregister, das für alle Hochschulen greifen soll – national, und idealerweise auch über die Landesgrenzen hinweg. „Mit dem Kerndatensatz Forschung werden wir in der Lage sein, unsere Leistungsfähigkeit zum Beispiel in einem Forschungsfeld klar anzugeben und uns damit international zu vergleichen“, sagte Wissenschaftsrat-Chef Prof. Dr. Wolfgang Marquardt. Die Details zur totalen Transparenz entwickelt derzeit das Institut für Forschungsinformation (IFQ) (s. S. 36f). Dass es nicht ohne die einzelnen Forscher geht, ist auch den Vordenkern des Kerndatensatzes klar.

Wer sich drückt, geht leer aus

In Münster müssen sich die Wissenschaftler selbst um die Pflege ihren Angaben kümmern. Weil das Zeit kostet und sich Forscher ungern in die Karten sehen lassen, geschieht das nicht ohne sanften Druck: Wer seine Daten nicht regelmäßig über ein Webformular einträgt, hat aus Sicht der Uni-Leitung nicht geforscht – und könnte bei der nächsten internen Mittelvergabe leer ausgehen. 70 Prozent der Wissenschaftler beteiligen sich mittlerweile an der zentralen Forschungsdatenerfassung der Universität. Die Datenbank listet 75000 Publikationseinträge sowie 5500 Promotionen und 500 Habilitationen auf.

Der nächste Schritt im Datenmanagement wäre, die Informationen wiederum mit den Messwerten, Umfrageangaben und Laborergebnissen zu verknüpfen. „Dokumentationsdaten und Forschungsprimärdaten sind wie Bruder und Schwester“, sagt Sebastian Herwig.

Ähnlich denkt Stefan Winkler-Nees von der DFG. Er kann sich ein soziales Netzwerk vorstellen, in dem Wissenschaftler Profile anlegen und dort für jedermann ihre Forschungsprojekte und Publikationen auflisten und Datensätze zur Verfügung stellen. Eine Art Forschungs-Facebook, in dem man alles auf einen Blick hat. Modelle dafür gibt es bereits. „In Deutschland“, sagt Stefan Winkler-Nees, „sind wir aber noch weit davon entfernt.“ ■

Bernd Kramer
ist duz-Redakteur.

„In der Community
sind wir eine
wichtige Anlaufstelle“

„Dokumentationsdaten und
Forschungsprimärdaten sind
wie Bruder und Schwester“