Offener Zugang zu wissenschaftlichen Primärdaten

H.Pfeiffenberger, Alfred Wegener Institut, Helmholtz-Gemeinschaft

Wert und Nutzen von Primärdaten sind in vielen Disziplinen schon länger unbestritten - aus unterschiedlichen Gründen: "Gute wissenschaftliche Praxis", Nachnutzung, bis hin zu "Data Driven Science". Neben einer Diskussion des offenen Zugangs wird auch versucht, erste Hinweise auf zukünftige Akteure bei Kuratierung, Archivierung und Publikation von Primärdaten zu geben.

In den letzten Jahren hat bei Konferenzen im Umfeld des „offenen Zugangs zu wissenschaftlichem Wissen“ die Frage des Umgangs mit Primärdaten prominenten Raum gewonnen. Der Wert und Nutzen von Primärdaten - und damit auch deren sorgfältige Archivierung – ist in vielen Disziplinen seit langem unbestritten. Das Neue in der Diskussion scheint daher zunächst „nur“ die Frage der Kosten bzw. Rechte zu sein – bis man bemerkt, dass Geistes- und Sozialwissenschaften (HSS) sich diesem Gedanken nicht minder interessiert nähern, als bestimmte(!) Disziplinen aus dem Feld der „üblichen Verdächtigen“ („Science, Technology, Medicine“, STM), die dies schon länger zu besetzen scheinen. [1]

Darüber hinaus aber gibt es für die „neue“ Aufmerksamkeit sowohl zunehmend gewichtige Gründe im Wissenschaftssystem selbst („gute wissenschaftliche Praxis“) als auch – wichtiger noch – in Form radikaler (informations-)technische Änderungen in der Praxis vieler Disziplinen, sowohl aus dem STM- als auch HSS-Bereich. Beide Triebfedern, sich mit wissenschaftlichen Primärdaten zu befassen, führen dabei in „natürlicher“ Weise dazu, dass ein offener Zugang allen anderen Regelungen und „Geschäftmodellen“ aus der Sicht des Fortschritts der Wissenschaft vorzuziehen ist.

Es soll gleich zu Beginn betont werden, dass die systematische Behandlung einer ganzen Reihe von Ausnahmen ebenso wichtig sein wird, wie die präzise Formulierung dessen, was man unter offenem Zugang zu Primärdaten verstehen will. So ist etwa nach kurzem Nachdenken klar, dass eine Vielzahl „sensitiver“ Daten eben nicht in „voller Auflösung“ frei zugänglich sein kann - ob es sich um sensitive medizinische oder soziale Daten („im Dorf X leben Y% Alkoholiker“) handelt oder die genaue Positionen schützenswerter Habitate. Auf der anderen Seite stellt sich die Frage, welcher Art Umgang mit offen zugänglichem Material der Wissenschaft adäquat und wie diese zu realisieren ist – wie also Daten zu zitieren sind und welcher Umfang und welche Qualität zitierenswert ist. Schließlich kann das Gewähren offenen Zugangs mit der Inanspruchnahme kostspieliger Infrastrukturen verbunden sein und bei anonymem Zugriff auch die Gefahr von „DoS“ (Denial of Service) Angriffen eröffnen.

[1] z.B. während der „Berlin4“-Konferenz, Golm 2006 http://berlin4.aei.mpg.de/program.html und der EU-Konferenz über „Scientific Publishing in the European Research Area -Access, Dissemination and Preservation in the Digital Age“, Brüssel, 2007 sowie der vorausgegangenen „Konsultation“; http://ec.europa.eu/research/science-society/page_en.cfm?id=3459

Ein kurzer Artikel kann das ganze Spektrum der Möglichkeiten und Herausforderungen des offenen Zugangs zu wissenschaftlichen Daten nicht erschöpfend behandeln. Daher kann nur anhand einer Reihe von Beispielen verdeutlicht werden, dass viele Probleme noch zu lösen sind und Wissenschaftskulturen verändert werden müssen, um der Wissenschaft ein neues, zugleich notwendiges und viel versprechendes Instrumentarium zu erschließen.

Gute wissenschaftliche Praxis

Die Empfehlung 7 der „Vorschläge zur Sicherung guter wissenschaftlicher Praxis" der DFG vom Januar 1998[2] gibt den Institutionen von Autoren (!) vor, Primärdaten, die Grundlage von Veröffentlichungen waren, für 10 Jahre „auf haltbaren und gesicherten Datenträgern" aufzubewahren. Man darf vermuten, dass eine systematische Überprüfung der Realität einen wenig erfreulichen Grad der Umsetzung enthüllen würde. Zudem darf die Wirksamkeit dieser Empfehlung – selbst wenn sie eingehalten würde – bezweifelt werden, da eine gelegentliche Nachfrage wohl in vielen Fällen wohl nicht als im Wortsinne selbstverständlich gesehen würde.

Das der Empfehlung von 1998 zugrunde liegende Problem ist auch in den darauf folgenden Jahren in einer Reihe von spektakulären Fällen zu Tage getreten – und ruft bei renommierten Journalen eher hilflos wirkende Betrachtungen hervor[3], die auch auf die Schwierigkeit des *Zugriffs* auf die zugrunde liegenden Daten abheben (die Daten werden auf CDs in Pappkartons vermutet). Zu einer systematischen Bekämpfung unakzeptabler Praktiken gerade im medizinischen Bereich werden daher sogar radikale Änderungen im Ablauf der Forschung vorgeschlagen, deren Kern die Begutachtung *der Planung* von Studien und deren statistischer Methoden sowie die vollständige Offenlegung *aller* gewonnen Daten enthält.[4]

Andererseits stellt sich die Frage, ob es z.B. bei Forschung mit gesellschaftlicher oder wirtschaftlicher Bedeutung ausreichend ist, „auxiliary material" wie Daten und Programme auf Informationssystemen unter der Kontrolle der Universität oder gar des Autors selbst vorzuhalten – die politisch geladene Kontroverse über die „hockey stick curve" bezog als verunsicherndes Element gerade eine solche Verteilung von Informationskomponenten ein.[5]

Nachnutzung wertvoller Daten

Einen weitaus konstruktiveren Nutzen bietet der Zugang zu Daten aus Experimenten und Beobachtungen. Zum Beispiel umfassen jene in den „Earth System Sciences" nicht nur die physikalischen und biologischen sondern auch die sozialen Aspekte der globalen Umwelt und deren Wandel[6]. Einerseits handelt es dabei um nicht wiederholbare Beobachtungen des jeweiligen Zustands des Erdsystems, zum Anderen

[2] http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf
[3] Marris, E. „Should journals police scientific fraud?". Nature. 439 (2006): 520-521. doi:10.1038/439520a.
[4] Smith R, Roberts I (2006) Patient safety requires a new way to publish clinical trials. PLoS Clin Trials 1(1): e6. DOI: 10.1371/journal.pctr.0010006
[5] „Climate change: is the US Congress bullying experts?" Nature 436, 7 (2005) doi:10.1038/436007a
[6] Earth System Science Partnership (ESSP), http://www.essp.org/

sind viele dieser Beobachtungen mit extrem hohen Kosten verbunden – im gerade angelaufene Internationale Polarjahr 2007/2008 werden etwa 50.000 Beteiligte aus 60 Nationen unter Einsatz von ca. einer Milliarde Euro einen Datenschnappschuss der für den globalen Wandel besonders sensitiven Polargebiete erfassen. Dieser umfassende Datensatz wird als das wesentliche Ergebnis des Polarjahrs gesehen, da er auf Jahrzehnte eine Referenz für (fast) alle weiteren Untersuchungen dienen wird.

Schon der Umfang der eingesetzten Mittel und die Bedeutung des Themas verlangen natürlich die möglichst vollständige Nutzung der Daten – die zumeist von ihren unmittelbaren Erzeugern nur auf bestimmte Fragestellungen hin genutzt werden. Angesichts der Komplexität des Erdsystems ist es für einzelne Wissenschaftler oder sogar ganze Institute auch gar nicht möglich, alle potentiell bedeutsamen Korrelationen ihrer Daten mit gleichartigen auf globaler Skala oder mit Daten aus anderen Disziplinen zu untersuchen. Gerade letzteres aber birgt vermutlich die größten oder wichtigsten Erkenntnisgewinne – etwa: wie wirken sich physikalische Klimaänderungen auf die Biosphäre oder auf die Lebensbedingungen der Menschen in den Polargebieten aus?

Solche Daten sehr langfristig zu erhalten ist schon seit langem als bedeutend erkannt und hat anlässlich des Internationalen Geophysikalischen Jahres 1957/1958 zur Gründung des World Data Center Systems[7], unter dem Dach des International Council for Science, d.h. unter der Ägide von Wissenschaftsorganisationen, geführt. Neuerdings gibt es eine seit 2003 von den G8, d.h. direkt regierungsseitig, getragene internationale Bemühung[8] ein Global Earth Observation System of Systems (GEOSS), d.h. eine Föderation verteilter Datenzentren und Datensätze, einzurichten. Im Gegensatz zum WDC-System existiert für GEOSS allerdings (noch?) keine Politik des offenen Zugangs, da dadurch eine Vielzahl von Datenquellen, die unter nationalen Restriktionen stehen, ausgeschlossen würden.

Die vierte Dimension wissenschaftlicher Arbeit

In den vorangegangenen Beispielen ergibt der offene Zugang schon dann Nutzen, wenn er weiterhin nur durch einzelne, menschliche Nutzer wahrgenommen wird. Gänzlich neue Dimensionen können wohl erschlossen werden, wenn Daten – und dazu gehören dann auch die Texte wissenschaftlicher Veröffentlichungen – in großem Maßstab maschinell organisiert und analysiert werden können. Manche glauben, dass sich dadurch - nach Theorie, Experiment und Modellierung („Simulation“) - das reine Arbeiten anhand vorhandener Daten als vierte Methode oder Paradigma wissenschaftlichen Arbeitens („Data Driven Science“) etablieren wird[9].

Allein schon besonders effektive und „übersichtliche“ Zugriffs- und Darstellungsmöglichkeiten für eine Vielzahl von Informationen, die sich auf dasselbe Objekt oder Phänomen beziehen, ergeben Erkenntnismöglichkeiten, die bisher außerhalb des Menschenmöglichen lagen. Ein Beispiel hierfür ist das entstehende

[7] World Data Center System / International Council for Science, http://www.icsu.org/5_abouticsu/STRUCT_InterBod_2.php?query=WDC
[8] The intergovernmental Group on Earth Observations (GEO), http://www.earthobservations.org
[9] NSF/JISC Repositories Workshop 2007, http://www.sis.pitt.edu/~repwkshop/

internationale „Virtual Observatory“[10] der Astrophysik und Astronomie, das den integrierten und interoperablen Zugriff auf astronomische Archive in der Gestalt eines virtuellen Observatoriums ermöglichen soll.

Aber erst Data- und Textmining versprechen durch Informationserkennung, –extraktion und –integration – etwa in der biomedizinischen Forschung [11] - eine noch viel weiter gehende Verstärkung des menschlichen Erkenntnisprozesses, die weit über verbesserte Such- und Darstellungsmöglichkeiten hinausgeht. Diese Methoden setzen einen ungehinderten Zugang gleichermaßen zu Text, Bildern und Daten voraus, insbesondere, um jederzeit innovativen Neulingen den „Marktzugang“ zu ermöglichen.

Wie offen können/dürfen Daten zugänglich sein?

Die letzten der aufgezeigten Beispiele verdeutlichen, dass die Forderung nach offenem Zugang sich auf alle Arten von Information beziehen und sehr weitgehende Rechte umfassen muss, um den größtmöglichen Gewinn für die Wissenschaft zu erzielen. Schon die Definition der Budapest Open Access Initiative ist so umfassend und es sollte nicht ohne Not dahinter zurückgegangen werden: "[The] free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself."[12]

Bekannte Problematiken sowohl aus den Sozialwissenschaften (beim Schutz personenbezogener Daten), der Ökologie (vom Schutz von Nestern bis zu ganzen Landschaften), der Erdbeobachtung (nationale Sicherheit, - vermeintliche - Sicherung von Wettbewerbsvorteilen) oder ganz praktische betriebliche Bedenken (Überlastung durch digitalen „Vandalismus“) zeigen aber, dass wohldefinierte Begrenzungen zumindest im Bereich von Daten unvermeidlich sind. Es wird also darauf ankommen, solche Grenzen möglichst schadlos zu setzen.

Rollenverteilung in einer zukünftigen Informations-Umgebung

Im Bereich der wissenschaftlichen Textpublikation hat sich über Jahrhunderte eine nach Disziplinen differenzierte Arbeitsteilung zwischen publizierenden und begutachtenden Forschern, kommerziellen, Gesellschafts- und Universitätsverlagen, Universitäts-, Spezial- und Nationalbibliotheken eingestellt. Es wäre vermessen, eine endgültige, zukünftige Rollenverteilung bei der Publikation, Kuratierung und Archivierung von Primärdaten schon heute konkret zu benennen. Allerdings sind doch aus den Beispielen einige wünschenswerte, notwendige oder wahrscheinliche *Eigenschaften* eines zukünftigen Systems und seiner „Player“ abzulesen, die sinnvollerweise auch mit

[10] International Virtual Observatory Alliance (IVOA), http://www.ivoa.net
[11] Hofmann-Apitius, M. „Paradigm Changes Affecting the Practice of Scientific Communication in the Life Sciences”. Scientific Publishing in the European Research Area. Brüssel. 15. Februar 2007. http://ec.europa.eu/research/science-society/document_library/pdf_06/hofmann-022007_en.pdf.
[12] Budapest Open Access Initiative, http://www.soros.org/openaccess/read.shtml

Beobachtungen der Dynamik von Standards und Angeboten im Internet abzugleichen wären:

Den maximalen Erkenntnisgewinn versprechen global und interdisziplinär korrelierte oder sogar integrierte Datenzugänge. Dies setzt neben dem offenen Zugang zu den einzelnen Datensätzen auch deren interoperable Beschreibung und Formatierung voraus – eine Aufgabe, die zumindest derzeit noch deutlich schwieriger und umstrittener ist als das vergleichbare MetadatenProblem für Texte – und darüber hinaus für jede Disziplin, womöglich sogar für jeden Datentyp, neu zu lösen! Keinesfalls wird sie vom Daten beitragenden Wissenschaftler gelöst werden, sondern im Zweifel vom Personal des Repository, das also eine gewisse Affinität zu Disziplin und Datentyp besitzen muss.

Es zeichnet sich also ab, dass die meisten Hochschulen nicht in der Lage sein werden, ein qualitativ und quantitativ angemessenes „Institutional Repository" für Daten *jeder* der an ihr vertretenen Disziplinen zu betreiben. Selbst wenn die gleiche Technologie wie die für Text-Repositories eingesetzt würde[13]: Der Flaschenhals wird das disziplinspezifisch kompetente Personal sein. Aus verwandtem Grund sind aber ebenso wenig zentrale nationale Repositories für Daten aller Disziplinen denkbar – sie würden zu völlig unbeweglichen Großbürokratien degenerieren.
In einigen Disziplinen reicht für bestimmte Datentypen heute sogar weltweit eine oder einige wenige Datenbanken (z.B. für Sequenzdaten aus der Genomforschung) – mit entsprechend qualifiziertem und - unter anderem wegen der Bewegungsfreiheit - motiviertem Personal.

Welche Rolle werden also in Zukunft Universitäts- oder Institutsbibliotheken, Daten- oder Rechenzentren im Bezug auf die wissenschaftlichen Primärdaten spielen? Sicher wird jede von diesen Einrichtungen vor allem neu überlegen müssen, welche Dienstleistungen sie den Nutzern an ihrer Einrichtung bei der Produktion, dem (lokalen, befristeten) Umgang mit und der Publikation von Daten anbieten kann und muss. Das erste und mindeste Ziel muss sein, den Studenten und Forschern technische Mittel und Best-Practise-Handreichungen oder Schulungen anzubieten, damit diese der Aufforderung nachkommen können „gute wissenschaftliche Praxis", lege artis, auszuüben und ihnen zu vermitteln, wie und wo die Daten langfristig „gut aufgehoben" sind und wie sie ihre Rechte (und die des Steuerzahlers?) angemessener zur Wirkung bringen, als dies derzeit bei Textpublikationen der Fall ist.

Neben dieser (zukünftigen) „Pflicht"-Rolle werden zumindest einige der Einrichtungen als „Kür" auch Betreiber spezialisierter Dienste für ganze nationale oder gar globale Communities werden können. Dies setzt jedoch neben einer ausreichenden technischen Infrastruktur vor allem den Aufbau von Kompetenzen und intensive, im Zweifel globale Kooperation und Positionierung in diesem Spezialgebiet voraus.

[13] z.B. IVOA Roadmap 2006 http://www.ivoa.net/Documents/Notes/RoadMap/IVOARoadMap-2006.pdf und Plankton-Net Präsentation bei der ECDL 2006, http://www.lib.uoa.gr/dorsdl/dorsdl2006-macario.ppt