Hans Pfeiffenberger

Data Publishing und Open Access

In den letzten Jahren hat sich der Begriff des Data Publishing – in Verbindung mit dem Offenen Zugang auch zu Forschungsdaten - etabliert. Das Publishing wird hier „mit großem P“ verstanden, d.h.: es soll um einen formalen Publikationsvorgang gehen, nicht nur um das publik machen, etwa auf einem beliebigen Server im Web. Die Formalien und deren Prioritäten, Methoden und Anforderungen wurden dabei zuweilen verschieden gesehen; erst in letzter Zeit zeichnet sich ein Konsens hin zu den FAIR Data Principles[1] ab: Die Daten sollen findbar, zugänglich, interoperabel und nachnutzbar sein. Bei genauem Hinsehen gehen die Prinzipien weit über das hinaus, was die vier Begriffe suggerieren – insbesondere die Maschinen-Tauglichkeit („machine-actionable“) legt einen hohen Maßstab sowohl an die (Meta-)Daten als auch an die Infrastrukturen, mittels derer sie erfasst und verbreitet werden.

Über die eher technischen Standards für Daten, Formate und Protokolle hinaus – deren Bedeutung noch zu diskutieren sein wird – besetzt der Begriff des Publizierens zwei für die Wissenschaft und deren Akteure äußerst wichtige Funktionen: Zum einen wird mit dem formalen Publizieren stets auch ein Maß an Qualitäts(zu)sicherung verbunden – sei es durch Editorial oder Peer Review oder eine andersartige Kritik oder Diskussion durch Dritte. Zum anderen erwirbt der Schöpfer eines so veröffentlichten Elementes des wissenschaftlichen Wissensbestandes – gemäß Guter wissenschaftlicher Praxis - das Anrecht zitiert zu werden und damit die Anerkennung der wissenschaftlichen Leistung.

Noch im Jahr 2010 war allerdings keineswegs unumstritten, dass die Erzeugung eines qualitätsgesicherten Datensatzes eine *eigenständige* wissenschaftliche Leistung sein kann. Dies ist etwa an den „Grundsätzen zum Umgang mit Forschungsdaten“[2] der Allianz der deutschen Wissenschaftsorganisationen unter dem Zwischentitel „Wissenschaftliche Anerkennung“ zu erkennen. Entsprechend wurde die Aufforderung einschlägiger Daten-Repositories, etwa PANGAEA, Daten zu zitieren, trotz des Angebots einfach kopierbarer Referenzangaben nur selten befolgt. Die Zitate waren auch kaum auffindbar. Einerseits verlagerten Autoren von

---

Alle URLs überprüft zwischen 21.10.2016 und 25.10.2016

[1] Force11, „FAIR Data Guiding Principles“, https://www.force11.org/fairprinciples

[2] „Grundsätze zum Umgang mit Forschungsdaten“, Allianz der deutschen Wissenschaftsorganisationen, 2010;
http://www.allianzinitiative.de/handlungsfelder/forschungsdaten/grundsaetze/

Artikeln den Bezug auf die Daten weiterhin gern in die Danksagungen (Acknowledgements), andererseits wurden auch „echte“ Zitate nicht in die Indexdatenbanken wie Web of Science oder Scopus aufgenommen. Nicht zuletzt dieses Problem hatte bereits zuvor zur Idee des Datenjournals geführt, das Artikel *über* Daten veröffentlicht, gewissenmaßen als leichter formal zitierbaren Stellvertreter für die Daten. Dabei wird bei einigen dieser Journale, etwa „Earth System Science Data“, gegründet 2008, das Prinzip des Peer Review auf die Daten selbst ausgedehnt, und so neben der Zitier“würdigkeit“ des Artikels auch der Qualitätssicherung Rechnung getragen[3].

Prinzipiell änderte sich die wissenschaftliche Anerkennung des Beitrags von Daten etwa ab den Jahren 2012/13. Unter den einschlägigen Dokumenten seien drei mit sehr unterschiedlicher Stoßrichtung hervorgehoben:

- Der Report „Science as an open enterprise“(SAOE)[4] der Royal Society (of London), begründete die eigenständige Bedeutung von Daten, Software und anderen digitalen Artefakten im wissenschaftlichen Ökosystem– und zwar vor allem dann, wenn diese so offen wie möglich verfügbar sind. Der buchstäblich erste Satz lautet: „Open inquiry is at the heart of the scientific enterprise“. Bemerkenswerterweise wird dort eine Definition von „Intelligent Openness“ gegeben, die den FAIR Prinzipien sehr ähnlich ist – mit einer großen Ausnahme: FAIR nimmt die Forderung „assessable“, gemeint ist die Möglichkeit der Qualitätseinschätzung, nicht als eine der vier grundlegenden Forderungen auf.
- Die DFG-Empfehlungen zur „Sicherung guter wissenschaftlicher Praxis“[5] enthalten erstmals in der Version von 2013 in den Regelungen zur Autorschaft (Empfehlung 11) eine Anerkennung des Beitrags von Daten: “Als Autoren einer wissenschaftlichen Originalveröffentlichung sollen alle diejenigen ... firmieren, die ... zur Erarbeitung, Analyse und Interpretation der Daten ... selbst wesentlich beigetragen ... haben“.
- Die Antragsrichtlinien der amerikanischen National Science Foundation vom Januar 2013[6] ziehen wissenschaftliche Artikel, Daten, Software, sogar Patente, als (potentiell) gleichwertige „Produkte“ wissenschaftlicher Arbeit der Antragsteller in Betracht. Anstelle einer (möglichst langen)

---

[3] Hans Pfeiffenberger und David Carlson, "Earth System Science Data" (ESSD) - A Peer Reviewed Journal for Publication of Data“, 2011 , D-Lib Magazine, 17 (1/2) doi:10.1045/january2011-pfeiffenberger

[4] „Science as an open enterprise“, The Royal Society Science Policy Centre report 02/2012, https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf

[5] „Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission ‚Selbstkontrolle in der Wissenschaft’ “ DFG/Wiley 2013; DOI:10.1002/9783527679188.oth1

[6] „Grant Proposal Guide“ Chapter II.C.2.f(i)(c) , NSF 2012 (effektiv Januar 2013); https://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpgprint.pdf

> Liste von Artikeln werden qualifizierte „Produkte“ erwartet „Products - A list of: (i) up to five products most closely related to the proposed project; and (ii) up to five other significant products, whether or not related to the proposed project. Acceptable products must be citable and accessible including but not limited to publications, data sets, software, patents, and copyrights. ... Only the list of 10 will be used in the review of the proposal.“

Die Erläuterung des Terminus „intelligent openness“ im SAOE-Report greift sehr deutlich das Problem auf, dass zwar der Grundsatz der Offenheit gelten muss, aber die Daten etwa aus einem schon finanziell sehr wesentlichen Teil der Forschung, der Medizin, nicht völlig offen zugänglich sein können (translationale medizinische Forschung; „Nationale Kohorte“[7]). Auch diese Daten dennoch so offen wie möglich zugänglich zu machen, stellt eine besondere mathematische, technische und organisatorische Herausforderung dar. (Stichworte dazu sind etwa Differential Privacy, gehärtete IT-Systeme, Zertifizierung autorisierter Nutzer). Die betroffene Community muss dann einen Konsens finden, was als hinreichend offen und was als publiziert gelten kann.

Nachdem so die grundsätzliche Frage der heutigen Bedeutung eigenständiger Datenprodukte in der Wissenschaft geklärt ist, lohnt die Frage: Warum kam dies erst jetzt auf die Tagesordnung? Schließlich waren Daten zumindest für die Naturwissenschaften schon seit dem Beginn solcher Forschung von entscheidender Bedeutung. Schon seit der Zeit der Sumerer und Assyrer (ab 2000 v.Chr.) wurden über Jahrhunderte die Daten der Mond- und Sonnenfinsternisse erfasst, um erst viel später (700 bzw. 585 v.Chr) zur Vorhersage „nachgenutzt“ zu werden. Erst Tycho Brahe’s exakte Beobachtung der Planeten ermöglichte es dem Analysten, deren Bewegungen in kompakten Formeln (Kepler’sche Gesetze) zu beschreiben und vorherzusagen (wenn auch nicht in einer Theorie zu erklären – dies blieb Newton vorbehalten).

Der offensichtliche Unterschied im Zeitalter der digitalen (Online-)Wissenschaft ist das Wuchern der Datenmengen, ebenso wie das der Anzahl der Quellen und die Vielfalt der Methoden, mit diesen Daten umzugehen. Dies geht so weit, dass diese nicht mehr adäquat und im Detail als Tabellen oder im Methodenteil als Text wiederzugeben sind. Die womöglich größte unmittelbare Herausforderung, die diese Entwicklung für die Wissenschaft und Wissenschaftler und Wissenschaftlerinnen hervorruft, ist die Frage der Nachvollziehbarkeit ihrer Ergebnisse

---

[7] „Was ist die NAKO Gesundheitsstudie?“, http://nako.de/allgemeines/was-ist-die-nako- gesund heitsstudie/

und – sogar und gerade im beginnenden Zeitalter der Open Science – die Frage der nachhaltigen Nutzbarkeit. Das Problem besteht hier zunächst in der für Menschen nicht mehr ohne digitale Hilfsmittel überschaubaren und nicht bearbeitbaren Menge. In der zweiten und dritten Dekade dieses Jahrhunderts werden die digitalen Methoden aber über die simple, 1-zu-1 Abbildung oder Nachahmung analoger Methoden mittels analytischer Software und „digitalisierter" Informationsformate weit hinausgehen:

Maschine learning – im wohl einfachsten Fall ein neuronales Netz, das mit einem Datensatz trainiert wird - kann wissenschaftliche Aussagen und Vorhersagen produzieren. Zum Beispiel könnten Lücken in Beobachtungsdaten auf besonders „intelligente" Weise gefüllt oder extrapoliert werden, aber durchaus auch Hypothesen vorgeschlagen[8] werden. Aber worum handelt es sich bei den Outputs eines Machine Learning Systems? Extrapolierte Daten sind jedenfalls keine Beobachtungen. Und selbst wenn ein solches System theoriegestützt trainiert wurde (und sei es implizit durch die Auswahl des Lernmaterials), so sind dessen Ergebnisse doch nicht mehr in gleicher Weise wie bei einem klassischen Model zu verstehen und nachzuvollziehen. Welcher wissenschaftliche Wert, jenseits der pragmatischen Nutzung der Vorhersagen, ist also diesen Ergebnissen beizumessen und wie sind diese Methoden ggf. zu dokumentieren und zu publizieren?

Diese Entwicklungen und ihre Faktoren werden in den nächsten Jahren einen Umbau - oder gar eine Umwälzung? - des wissenschaftlichen Publizierens, insbesondere des Publizierens von Daten, Software, ihres Kontextes und ihrer Provenienz, prägen. Wir haben in den letzten Jahren zumindest in den Naturwissenschaften erlebt, wie für alle praktischen Zwecke das Publizieren von Artikeln auf Papier durch die Darstellung und Online-Verbreitung derselben Information im PDF-Format ersetzt wurde. Dabei geht weder die Granularität und die Strukturierung des Inhalts noch dessen Präsentation wesentlich über eine 1-zu-1 Abbildung hinaus. Die eigentliche Innovation bestand in der Einführung eines persistenten und eindeutigen Identifiers, als dessen technische Realisierung sich hier eindeutig der DOI durchgesetzt hat. Die heute anerkannten Methoden, Datensätze zu publizieren, nämlich über Repositories, bei denen (spätestens) bibliographische Metadaten erfasst und zugeordnet werden, deren Darstellung (!) üblicherweise einen DOI erhält, sind wiederum im Grunde eine Abbildung der Online-Methoden des Publizierens von Artikeln auf Datensätze.

---

[8] Scott Spangler et al., „Automated hypothesis generation based on mining scientific literature", in Proceedings, KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 1877-1886; DOI:10.1145/2623330.2623667

Es steht zu vermuten, dass dem Publizieren – ob von Artikeln oder Daten – und dessen Akteuren und ihren Infrastrukturen ebensolche disruptive Innovationen bevorstehen, wie dies dem Einzelhandel, Telefonherstellern, dem Automobilbau etc. widerfahren ist oder gerade widerfährt.

So wird dem Bioinformatiker Barends Mons ein im Kontext einer Keynote[9] geäußertes Bonmot zugeschrieben, in Zukunft würden wohl nicht mehr die Daten Supplement zu den Artikeltexten seien, sondern die Artikel (Narrative) Supplement zu Daten. Der Autor dieses Kapitels hat selbst schon vor längerer Zeit bei einer Innovations-Konferenz der STM-Verleger bemerkt[10], dass deren Produkte (gemeint waren die Volltexte von Zeitschriften-Artikeln) die besten verfügbaren Metadaten zu Forschungsdaten seien – wenn nur Artikel und Forschungsdaten untereinander verlässlich zitiert wären.

Etwas allgemeiner gefasst, ist abzusehen, dass in Zukunft wissenschaftliche Information aus einem Gewebe von digitalen Objekten, nämlich Texten, Daten, Software und insbesondere (den Identifiern für) Personen bestehen bzw. daraus abgeleitet werden wird. Eine erste, „einfache" kommerzielle Realisierung ist in dem neuen Publikationskonzept des Verlages Elsevier unter dem Namen „Research Elements"[11] zu bestaunen. Der wirkliche Wert dieser Elemente oder „research objects" lässt sich dabei nur durch die Fäden oder Kanten des Gewebes, nämlich die Links bzw. Zitate, erschließen und dies angesichts einer explodierenden Anzahl auch nur dann, wenn dieses Gewebe – und natürlich vorzugsweise auch die verbundenen Objekte selbst - in maschinenlesbarer Form vorliegen.

Zudem entsteht ein Kontinuum der Größen, Qualitäten und Formate der publizierten Objekte, dessen Ausdehnung heute bereits absehbar ist: Einzelne Datenobjekte können eine Terabyte umfassende Datei sein oder ein Datenbankeintrag entsprechend einer Tabellenzeile. Daten und Software bzw. dazugehörige Artikel können ebenso wie Narrative einem Peer Review unterzogen[12] sein, oder ihre Qualitätssicherung ist durch sorgfältigste Dokumentation ihrer Entstehung[13] reali-

---

[9] Barends Mons, „Bringing Data to Broadway", RDA Plenary Amsterdam, 22.09.2014; https://www.rd-alliance.org/plenary-meetings/fourth-plenary/plenary4-programme.html, https://collegerama.tudelft.nl/Mediasite/Play/0844aefac5bb49ca9032069c6edc668f1d?catalog=3984a02f-bf33-4c70-a080-94a04d3e8112, ab Minute 33:50

[10] Hans Pfeiffenberger, „Data, Big Data and Publications", STM Innovations Seminar 2012 – Innovation Impacts in STM, London, 07.12.2012, http://hdl.handle.net/10013/epic.48775

[11] https://www.elsevier.com/books-and-journals/research-elements

[12] etwa: Le Quéré, C., et al.: Global Carbon Budget 2016, Earth Syst. Sci. Data Discuss., doi:10.5194/essd-2016-51, in review, 2016.

[13] etwa: Reference documentation http://www.argodatamgt.org/Documentation des globalen ARGO Projekts, http://www.argo.ucsd.edu/About_Argo.html

siert, etwa durch Links zu Standard Operating Procedures bzw. dem verwendeten Protokoll oder durch dedizierte Web-Präsenzen[14], welche die Entstehung, beteiligte Personen, Ereignisse und genutzte Werkzeuge in freier Form zusammenführen. Es ist festzuhalten, dass die Koordinaten eines konkreten Forschungsergebnisses in diesem Kontinuum in hohem Masse disziplinabhängig sind.

Schließlich, und dies unterscheidet Forschungsdaten und –software mehr als alles andere vom klassischen Artikel, sind dies häufig „lebende" Objekte: Sie werden – wie in den zuvor gegebenen Daten-Beispielen - kontinuierlich oder schubweise erweitert, verfeinert und zuweilen auch korrigiert. Diese wissenschaftliche Arbeit an den Daten wird im Allgemeinen nicht so stattfinden, dass per DOI auf ein Daten- oder Software-Repository zugegriffen, an einer Kopie gearbeitet und eine neue Version (mit neuer DOI) hochgeladen wird. Vielmehr findet die Arbeit in Arbeitsdatenbanken oder an einzelnen Modulen in Softwareversionsverwaltungssystemen, etwa Github, statt. Zum Zweck der langfristigen und vertrauenswürdigen wissenschaftlichen Nachvollziehbarkeit bleibt dann nur, anlässlich von Veröffentlichungen „Schnappschüsse" zu ziehen – für Software etwa mittels der mit Github verbundenen Funktion von Zenodo[15], bei Daten etwa in Jahresausgaben.

Auch um den damit verbundenen Gefahren der Inkonsistenz zwischen einer tatsächlich genutzten Version in der Arbeitsdatenbank und der im Repository sowie der Aufblähung der Datenmengen entgegenzuwirken, werden zur Zeit im Rahmen der Research Data Alliance (RDA) Konzepte und Methoden erarbeitet[16] und erprobt, wie Identifier für Extrakte und zeitliche Stände dynamischer Datensätze bereit zu stellen wären.

Unter anderem auch bei der RDA finden Arbeiten zur Konzeption der Dateninfrastruktur von morgen[17] statt, die das beschriebene Gewebe von research objects in einer Anzahl, welche die heutige um viele Größenordnungen übersteigt, speichern, erhalten und zugänglich machen kann. Es muss festgehalten werden, dass in diesem Kontext zum Beispiel diskutiert wird, dass aus Skalierungsgründen nicht jedem aus der Vielzahl von Objekten ein DOI als Identifier mitgegeben

---

[14] The Global Carbon Project, http://www.globalcarbonproject.org

[15] Martin Fenner, „Software Citation Workflows", 2015 https://blog.datacite.org/software-citation-workflows/

[16] Andreas Rauber et al., „Data Citation of Evolving Data Recommendations of the Working Group on Data Citation (WGDC)", RDA 2015; https://www.rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf

[17] Larry Lannom und Peter Wittenburg, „Global Digital Object Cloud (DOC) - A Guiding Vision", 2016; http://hdl.handle.net/11304/a8877a1a-9010-428f-b2ce-5863cec4aff3

werden kann, oder dass auch in großem Umfang Daten nur befristet erhalten werden können.

Konkrete technische Vorhersagen für diesen Bereich können kaum gewagt werden – etwa dazu, welche Art von Identifier sich letztendlich durchzusetzen vermag, ob das Zitieren „evolvierender" Datensätze nach der RDA-Methodik flächendeckend gelingen wird oder ob und wie es gelingen mag, die Abläufe zur gegenseitigen Verlinkung und korrekten Zitierung zwischen Artikeln und Daten zu automatisieren.

Eines aber ist bei einer ganzheitlichen Betrachtung mehr als deutlich: Alle „Geschäfte" und Transaktionen, die über das Internet stattfinden, sind einer massiven Konzentration unterworfen („the winner takes it all"). Dies ist trotz der Existenz exklusiver Nutzungsrechte sogar im Publikationsbereich in Form eines Oligopols im STM-Bereich zu erkennen. Sofern eine Leistung keiner persönlichen Interaktion mit dem Endabnehmer bedarf - und dies kann man im Online-Bereich mit etwas einmaligem Aufwand sehr häufig erreichen – werden Zwischenhändler oder andere Intermediäre leicht ausgeschaltet: Die Lieferung erfolgt von einem Global Player direkt an den Endkunden. Auch in der Informationstechnik ist von einer Industrialisierung die Rede (gemeint ist die Verlagerung von Unternehmens-IT „in die Cloud", wo in der Tat Serverräume die Größe von Fabrikhallen annehmen.)

Dieser Effekt ist überall dort besonders wirksam, wo die Grenzkosten pro Kunde minimal gegenüber den Fixkosten sind (z.B. Verlage bei reiner Online-Lieferung) und / oder der Wert der Leistung für den Kunden mit der Anzahl der anderen Kunden desselben Unternehmens steigt (Netzwerkeffekt; Beispiel ResearchGate). Diesen Effekten werden sich lokale Infrastrukturen wie Bibliotheken und Rechenzentren an Universitäten und Forschungseinrichtungen nicht dauerhaft entgegenstemmen können.

Der wichtigste Ratschlag, der zur Zeit im IT-Bereich kursiert - welcher einer schnelleren Änderung unterworfen scheint als das Bibliothekswesen - ist der, seine Leistungen gegenüber dem Unternehmen ganz neu zu denken – „Was braucht der Kunde", nicht: „Was ist unser Tagesgeschäft; was wollen wir dem Kunden gern anbieten". Tatsächlich wird der Bereich Information und Data Science (zurzeit unter dem Schlagwort „Digitalisierung") in ganz normalen Unternehmen, aber auch in Forschungseinrichtungen[18] dieser Tage als ein strategisch

[18] „Die Ressource Information besser nutzbar machen!", Positionspapier der Helmholtz-Gemeinschaft, 2016, https://www.helmholtz.de/os-positionspapier/, Pressemitteilung dazu,

wichtiger (neu) entdeckt. Ob Industrie 4.0 oder Science 4.0: Eine Institution, die rechtzeitig entdeckt, wie ihre Informationsinfrastruktur unter den neuen Bedingungen aufgestellt sein muss, wird in der hoch dynamischen Umgebung einen womöglich unbezahlbaren Wettbewerbsvorsprung erreichen.

Diese Herausforderung wird mittlerweile auch auf höherer Verantwortungsebene aufgenommen. So hat der 2014 gegründete Rat für Informationsinfrastrukturen im Jahr 2016 erste „Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland“[19] herausgegeben. Deren erste konkrete lautet: „Förderpolitisch empfiehlt der RfII, Projektförderungen von Forschungsdaten- Infrastrukturen nachhaltig auszurichten, da diese Art der Finanzierung für langfristig benötigte Dienste ein Risiko darstellt.“. Es ist daher zu erwarten, dass die dort ebenfalls geforderte „Etablierung einer Nationalen Forschungsdateninfrastruktur (NFDI)“ kurzfristig an Fahrt gewinnen wird.

Bestehende Informationsinfrastrukturen – ob klassische oder als Projekt erst in den letzten Jahren entstandene – müssen sich entscheiden, ob sie die Disruption mitgestalten oder sich eine Nische im neu entstehenden Ökosystem von Diensten suchen wollen. So oder so werden sie sich neu erfinden müssen, wie bereits im DFG-geförderten Projekt RADIESCHEN[20] festgestellt, oder auf ein Abstellgleis geraten.

12.10.2016, https://www.helmholtz.de/aktuell/presseinformationen/artikel/artikeldetail/digitale_forschungsdaten_offen_zugaenglich_machen/

[19] Rat für Informations-Infrastrukturen, „Leistung aus Vielfalt“, 2016; http://www.rfii.de/?wpdmdl=1998

[20] DFG Projekt „Rahmenbedingungen einer disziplin-übergreifenden Forschungsdaten-Infrastruktur“, http://www.forschungsdaten.org/index.php/Radieschen