Strategien zur verbesserten „Nachnutzung“ von Datens(ch)ätzen

Das Grund-Problem: Es werden fortlaufend im Rahmen von mit öffentlichen Mitteln geförderten Forschungsvorhaben, aber auch von Ämtern und Behörden, in erheblichem Umfang physikalische, chemische und biologische Messdaten bei Feldmessungen und in Experimenten erhoben. Diese werden in der Regel unter einer bestimmten Fragestellung ausgewertet, zusammengefasst und in stark aggregierter Form publiziert. Häufig könnten die Primär-Daten, z. B. kombiniert mit anderen Informationen, auch zur Beantwortung weiterer Fragestellungen beitragen. Jedoch ist es derzeit unüblich, dass die für eine zweckdienliche Aufbereitung, Dokumentation und langfristige Sicherung der Daten erforderliche Mehrarbeit im Wissenschaftsbetrieb anerkannt wird. Dies hat zur Folge, dass sie unterbleibt und es folglich insbesondere für Forscher aus anderen Arbeitsgruppen sehr schwierig oder unmöglich ist, die nicht in einer Zeitschrift publizierten, weniger zusammengefassten Primär-Daten zu erhalten. In der Regel geraten diese früher oder später in Vergessenheit, wodurch der Fortschritt der Wissenschaft und eine effiziente Nutzung der Fördermittel behindert wird.

Ein Lösungsvorschlag: Daten sollen nicht mehr ausschließlich in stark aggregierter, verrechneter Form in wissenschaftlichen Publikationen veröffentlicht werden, sondern außerdem in weniger aufbereiteter Form als Primärdaten eine eigenständige, zitierfähige Identität erhalten. Hierbei ist im Einzelfall zu entscheiden, ob eine Publikation der primären Rohdaten oder von in gewissem Umfang aufbereiteten Daten sinnvoll ist. Mit solch einer im Wissenschaftsbetrieb anerkannten Datenpublikationen (s. z. B. http://www.earth-system-science-data.net/) erhält ein/e Autor/in eine zitierfähige Veröffentlichung und damit einen im Wissenschaftsbetrieb relevanten Anreiz, die Daten zügig zu publizieren und damit der Wissenschaftsgemeinschaft zur Verfügung zu stellen. Publikationen, die die Daten verwenden, verweisen auf die Datenpublikation.
So ist beispielsweise der Datensatz: Storz, D et al. (2009): Planktic foraminiferal flux and faunal composition of sediment trap L1_K276 in the northeastern Atlantic. *PANGAEA* doi:10.1594/PANGAEA.724325 http://dx.doi.org/10.1594/PANGAEA.724325
Grundlage für Storz et al. (2009) *Deep-Sea Research I*, **56(1)**, 107-124, doi:10.1016/j.dsr.2008.08.009 . Die Möglichkeit einer anerkannten Zitierung von Forschungsdaten erhöht nicht nur die Sichtbarkeit von diesen, sondern stellt auch eine starke Motivation für Wissenschaftler dar, ihre Daten zu publizieren, um wissenschaftliche Anerkennung durch Zitate zu erlangen.

Der Digital Object Identifier (DOI) wurde eingeführt, um Einheiten geistigen Eigentums eindeutig identifizieren und verwalten zu können. Mit Unterstützung der DFG wurde die Technische Informationsbibliothek Hannover (TIB) die weltweit erste DOI-Registrierungs-Agentur für wissenschaftliche Daten, die z. B. geowissenschaftliche Primärdaten registriert. Die Daten selbst werden in einem sogenannten Datenrepositorium dauerhaft gesichert. Die TIB vergibt für jeden Datensatz einen DOI Namen als dauerhaften Identifizierer, wodurch der Datensatz weltweit über jeden Webbrowser sehr einfach zugänglich ist.

Die praktische Umsetzung: Die folgenden Details dieses Artikels fokussieren bewusst auf der Bereitschaft Daten zur Verfügung zu stellen. Eine zweite Frage ist die technische Machbarkeit (z. B. welches Datenformat, welche Software, wo archivieren) und das Klären wichtiger, teils fachspezifischer Fragen (z. B. bzgl. der Metadaten; was muss alles festgehalten werden, damit man in 10 Jahren noch weiß, wie gemessen wurde und wie kann man dieses Wissen nach bestimmten Standards festhalten). Grob gesagt, stehen uns heute die dafür nötigen Werkzeuge zur Verfügung auch wenn noch nicht alle Detailfragen abschließend geklärt sind. Hiermit beschäftigen sich dafür geeignete, den Wissenschaftsbetrieb stützende

Institutionen, wobei in den vergangenen Jahren verschiedene Pilotprojekte durchgeführt worden. Ein Beispiel für eine gelungene Dateninfrastruktur stellt das *Publishing Network for Geoscientific & Environmental Data* (PANGAEA) (http://www.pangaea.de) dar, die jedoch einen nicht zu unterschätzenden Personalaufwand zur Aufbereitung der Daten für die Langzeitarchivierung hat(te).

Wie kann ich für meinen Datensatz einen DOI erhalten, sodass er offiziell zitiert werden kann?

Zunächst muss ein Datenrepositorium gefunden oder, bei Bedarf, aufgebaut werden, dass bereit und in der Lage ist, die Primärdaten langfristig verfügbar zu halten und den Empfehlungen des UA-INF entspricht (www.dfg/lis). Repositorien sind Dokumentenserver, auf denen wissenschaftliche Materialien archiviert und weltweit entgeltfrei zugänglich gemacht werden (http://open-access.net/de/allgemeines/was_bedeutet_open_access/repositorien/). Dann müssen die Datenerheber (in Kooperation mit bzw nach den Vorgaben des Datenrepositorium) überlegen, welche Daten für andere potentiell von Interesse sein können (z. B. welche raum-zeitliche Auflösung, nach welcher Kalibrierung und Umrechnung in gängige Einheiten) und welche Datendokumentation für eine effiziente und fehlerfreie Nachnutzung erforderlich ist.

Seit 2005 hat die TIB bereits über 600.000 DOI Namen für Primärdaten vergeben (siehe oben). Die Vergabe erfolgt immer in Kooperation mit bestehenden Datenzentren, die für die Pflege und Persistenz der Daten verantwortlich sind. Eine in Frage kommende Daten haltende Einrichtung schließt mit der TIB einen Vertrag ab, in der sie sich verpflichtet, die Primärdaten langfristig verfügbar zu halten. Mit der Registrierung speichert die TIB die Metadaten (die Beschreibung der Primärdaten) und veröffentlicht sie ggf. in ihren Katalogen, die mit den Primärdaten verlinkt werden.

Einige mögliche Komplikationen:

1. Wie verhindere ich, dass jemand aus meinen bereits publizierten Daten ein Paper macht, dass wesentliche Dinge vorwegnimmt, die ich selbst noch in einer Zeitschrift publizieren will? Wenn dies nicht möglich ist, bedeutet es, dass man die Daten erst publizieren will, nachdem man sie selbst vollständig ausgewertet hat, denn es kann sein, dass eine Zitierung als Datenautor weniger zum wissenschaftlichen Ansehen beiträgt und/oder weniger befriedigend ist als die eigene weitere Analyse der selbst erhobenen Daten, die jedoch Zeit und ggf. das Abwarten weiterer Messungen erfordert. Dies kann auch zu unterschiedlichen Interessen bei Ko-Erhebern eines Datensatzes führen, z. B. zwischen Nachwuchs-WissenschaftlerInnen und DauerstelleninhaberInnen („besser jetzt eine Datenpublikation als vielleicht später eine Koautorenschaft bei einem Zeitschriftenartikel“ und *vice versa*).
2. Häufig sind noch keine für das Fach adäquate Strukturen zur Langzeitarchivierung etabliert, wobei hieran jedoch gearbeitet wird und sich diese Situation jedoch mit zunehmender Bereitschaft zur Datenpublikationen verbessern wird.
3. Die Metadaten zur Erläuterung des Datensatzes müssen sorgfältig und überlegt zusammengestellt werden, damit vorgehaltenen Primärdaten langfristig für unterschiedlichste, z.T. fächerübergreifende Fragestellungen verwendet werden können. Bei komplexen Daten(strukturen) kann die Rücksprache mit den Datenerhebern sehr sinnvoll sein um Fehlinterpretationen zu vermeiden.
4. Datenerheber können aus Sorge wegen bisher nicht erkannter Unstimmigkeiten zögern, dass die Daten ohne weitere Rücksprache und damit potentielle Korrekturmöglichkeit verwendet werden.

5. DOIs verursachen Kosten zu Lasten des Datenproviders. Alternativ kann die Publikation durch Netzpublikation, URN, PURL, OpenURL, Ex Libris SFX und SRef geprüft werden, wobei jedoch das Kosten-Nutzen Verhältnis abgewogen werden muss.

Abschließend sei noch einmal darauf hinweisen, dass für die angewandte und Grundlagen-Forschung im Umweltbereich relevante Daten nicht nur im Rahmen von Forschungsprojekten erhoben werden, sondern auch durch viele verschiedene Ämter und Behörden auf Landes- und Bundesebene (z. B. BfG, LUAs, Bundesschifffahrtsamt) und teilweise, z. B. im Zuge von staatlich finanzierten Umweltverträglichkeitsprüfungen, auch auf kommunaler Ebene. Dies bedeutet, dass darauf hingewirkt werden muss, dass auch diese Daten einfacher und zügiger für weitere Forschungszwecke verfügbar gemacht werden. Hier sind im Bereich des Deutschen Wetterdienstes in den vergangenen Jahren große Fortschritte erzielt worden, die hoffentlich beispielhaft für andere Ämter und Behörden sein können.

Weitere Informationen

Grundsätzliche Gedanken, technische Aspekte, praktische Durchführung & Beispiele: http://edoc.hu-berlin.de/series/dini-schriften/2009-10/PDF/10.pdf

Informationen und Aktivitäten der DFG: http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/veroeffentlichungen/index.html#5

Aus internationaler Sicht: "To share or not to share": http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

Informationsmanagementexperten: Brase & Klump "Zitierfähige Datensätze": http://edoc.gfz-potsdam.de/gfz/10493

Diese Informationen wurden aufgrund von Diskussionen in der DFG-Senatskommission für Wasserforschung von Prof. Dr. Ursula Gaedke (Uni Potsdam) zusammengestellt mit Unterstützung von u.a. Dr. Jan Brase (DOI Registrierung, Technische Informationsbibliothek (TIB), Tel.: +49 511 762 19869; http://www.datacite.org; http://www.tib-hannover.de/de/die-tib/doi-registrierungsagentur/; http://www.janbrase.de, Prof. Dr. Susanne Crewell (Uni Köln), Dr. Birgit Gemeinholzer (Botanischer Garten, FU Berlin) und Dr. Stefan Winkler-Nees, (DFG, Wissenschaftliche Literaturversorgungs- und Informationssysteme, Tel. +49 (228) 885-2212, Stefan.Winkler-nees@dfg.de). Mit Fragen wenden Sie sich bitte an Dr. Winkler-Nees oder Dr. Brase.