Christoph Steinhof

Publikation von Forschungsdaten

Projektarbeit im Rahmen des Moduls M4.2 / WS15 im Studiengang
M.A. Informationswissenschaften der FH Potsdam:
Konzeptionelle Entwicklung eines Werkzeugs
für die Planung des Forschungsdatenmanagements
Prof. Dr. Heike Neuroth

in Kooperation mit dem DFG-Projekt

https://dmpwerkzeug.github.io/

08.02.2016

Inhaltsverzeichnis

1 Einleitung

> „No Publication without Data – no data without publication“
>
> – *Eefke Smit (Reilly et al., 2011, S. 84)*

Forschungsdaten sind schon immer integraler Teil wissenschaftlicher Publikationen. Die für die Schlussfolgerungen relevanten Forschungsdaten, werden in Tabellen, Graphen und Abbildungen zusammengefasst. Leser und Gutachter konsultieren diese Datenauswahl, um Forschungsergebnisse nachzuvollziehen und bewerten zu können.

Diese integralen Daten unterliegen einem hohen Aggregationsgrad. Sie sind dadurch nicht unabhängig vom Artikelkontext (nach)nutzbar und lassen sich auch nicht getrennt von der Publikation auffinden (Reilly et al., 2011, S. 37).

Die Daten innerhalb einer Publikation stellen nur einen kleinen Ausschnitt der während eines Forschungsvorhabens angefallen Daten dar. Es ist daher erstrebenswert, Forschungsdaten möglichst umfangreich und vollständig zu publizieren, um die Nachnutzung von Daten zu ermöglichen.

Die Publikation von Forschungsdaten, erleichtert den Zugang zu Daten, sorgt für ein besseres Verständnis der wissenschaftlichen Publikation und fördert die Anerkennung von Wissenschaftlern. Sie ermöglicht die Überprüfung von Forschungsergebnissen und ist Voraussetzung für die Nachnutzung der Daten.

Für eine bestmögliche Nachnutzung, müssen publizierte Forschungsdaten Zitierfähig sein. Dafür sind einheitliche Zitationskonventionen für Datenpublikationen notwendig. Die Auffindbarkeit von Datensätzen muss mithilfe eindeutiger Bezeichner (Persistent Identifier wie DOIs[1]) gewährleistet werden. Forschungsdaten müssen sicher gespeichert werden und über einen langen Zeitraum verfügbar bleiben. Dafür sind Erhaltungsstrategien notwendig, die auch die Pflege von Forschungsdaten sicherstellen (Reilly et al., 2011, S. 7 ff.).

Die in Abbildung 1 dargestellte Datenpublikationspyramide veranschaulicht die potentielle Menge an Forschungsdaten, die in der jeweiligen Publikationsform enthalten sein können. An der Spitze der Pyramide befinden sich die aggregierten Daten, die sich in Form von Tabellen und Abbildungen in einem traditionellen wissenschaftlichen Zeitschriftenartikel befinden. Darauf folgen zusätzliche Daten zu einem Artikel, die als Supplemental Material Dateien an der elektronischen Version eines Artikels hängen. An letzter Stellen stehen Daten in Repositorien und Datendokumentationen (Data Papers), auf die eine Publikation referenzieren kann.

In der vorliegenden Arbeit soll ein Überblick über die Möglichkeiten zur Forschungsdatenpublikation geben werden. Die Veröffentlichung von Forschungsdaten als Supplemental Material eines wissenschaftlichen Artikels wird in Abschnitt 2 diskutiert. In Abschnitt 3 wird auf die Publikation von Forschungsdaten

[1] https://www.doi.org/

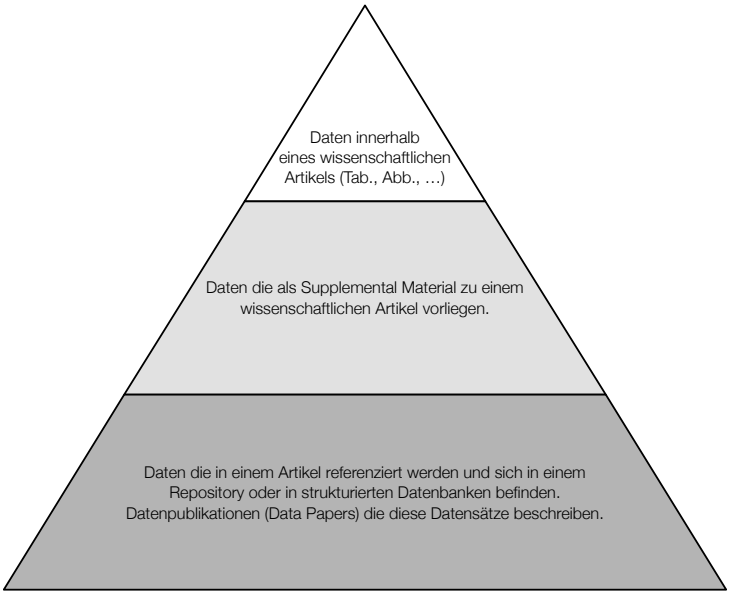

Abbildung 1: Datenpublikationspyramide (nach Reilly et al. (2011), S. 6)

durch eine neue Form des wissenschaftlichen Zeitschriftenartikels – dem Data Paper – eingegangen. Eine Einführung und Vergleich von Forschungsdaten-Repositorien die von Data Papers genutzt werden erfolgt im vierten Abschnitt.

2 Supplemental Material

Online-Ausgaben von Zeitschriften boten erstmals die Möglichkeit, Forschungsdaten als Zusatzmaterial – meist als Supplemental Material oder Supporting Information bezeichnet – zusammen mit einem Artikel zu publizieren. Die Beschränkungen durch das traditionelle Format eines gedruckten Artikels wurde damit aufgehoben. Seitdem können zusätzliche Abbildungen, Tabellen, Multimediadateien, vollständige Datasets oder auch Programmcode mit der elektronischen Version eines Artikels bei den meisten Zeitschriften veröffentlicht werden.

Die Forschungsdaten bleiben auf diese Weise eng mit dem Artikel verknüpft und lassen sich, wenn auch nicht direkt, durch den Artikel zitieren. Durch die Publikation der Forschungsdaten als Zusatzmaterial, erhöht sich die Attraktivität des Artikels für andere Wissenschaftler. Für Autoren bedeutet das, dass sie mit steigender Nutzung ihres Artikels und mit höheren Zitationsraten rechnen können.

Kuipers et al. (2009) befragten Wissenschaftler und Verlage nach ihrem Umgang mit Forschungsdaten. Dabei gaben 15 % der befragten Wissenschaftler an,

Supplemental Material mit ihrem Artikel einzureichen. Rund 64 % der befragten Verlage akzeptieren Supplemental Material. Diese Verlage publizieren insgesamt über 90 % der in der Umfrage untersuchten Zeitschriften.

Laut Schriger et al. (2011) stieg der Anteil an Artikeln mit Supplemental Material in Medizinzeitschriften von 7 % im Jahr 2003 auf 25 % im Jahr 2009.

Aus den genannten Zahlen lässt sich schließen, dass der Großteil der Zeitschriften Supplemental Material akzeptiert und die Nutzung dieser Möglichkeit zur Datenpublikation in den letzten Jahren stark zugenommen hat. Dennoch wird der Großteil aller wissenschaftlichen Artikel ohne zusätzliche Forschungsdaten publiziert. Es ist zu vermuten, dass die Akzeptanz und die Nutzung von Supplemental Material in den STM-Disziplinen deutlich weiter verbreitet ist, als in den Geisteswissenschaften.

2.1 Vergleich von Supplemental Material Richtlinien

Verlage und Zeitschriften stellen Autoren Richtlinien für Supplemental Material zur Verfügung. Darin wird festgelegt, welche Art von Information als Supplemental Material akzeptiert werden und welche Anforderungen an Dateiformate und -größen gestellt werden.

Im Anhang 1 werden acht Supplemental Material Richtlinien nach typischen Anforderungen verglichen. Fünf dieser Richtlinien sind verlagsübergreifend und gelten für alle Zeitschriften die dort erscheinen (*PLoS*, *Springer*, *SAGE*, *Wiley*, *Cell*). Drei weitere stammen von einzelnen Zeitschriften (*PANS*, *Nature*, *The Lancet*). Es zeigt sich, dass die Richtlinien sehr heterogen sind:

- Für Dateiformate gibt es explizite Vorgaben bei *Nature*, bevorzugte Formate bei *Wiley* und im Prinzip völlige Wahlfreiheit bei *PLoS* und *Springer*. *Wiley*, *Cell* und *Nature* lassen nur Dateien mit max. 10 – 30 MB zu. *PLoS* und *Springer* haben hingegen keinerlei Dateigrößenbegrenzungen.
- *The Lancet*, *Cell* und *PANS* erwarten als Supplemental Material ein einziges PDF, welches die zusätzlichen Abbildungen, Tabellen und Dokumente enthält. Für große Datasets sind *Excel*-Dateien erlaubt.
- Alle Richtlinien erlauben Multimedia-Dateien. Doch die zugelassenen Videostandards bei *PANS*, *The Lancet* und *SAGE* sind als veraltet zu bezeichnen.

2.2 Herausforderungen

Supplemental Material bietet zwar die Möglichkeit zusätzliche Forschungsdaten mit einem Artikel zu publizieren, beschränkt sich in der Regel aber auf Daten, welche die Schlussfolgerung und Ergebnisse des Artikels weiter stützen (CellPress, 2015). Die vollständigen Daten eines Forschungsvorhabens lassen sich auf diese Weise nicht veröffentlichen. Zudem schränken die teilweise strengen und

heterogenen Richtlinien Art und Umfang der publizierbaren Daten ein. Keine der im vorherigen Abschnitt untersuchten Richtlinien macht Vorgaben zur Dokumentation oder Beschreibung der Zusatzmaterialien, was für eine Nachnutzung hilfreich wäre.

Supplemental Material wird von Abstracting & Indexing Services nicht gesondert indiziert. Die Auffindbarkeit von Zusatzmaterialien ist somit eingeschränkt. Auch auf den Seiten der Online-Artikel selbst gibt es in der Regel keine strukturierten Metainformationen, die das Supplemental Material beschreiben.

Nicht jede Zeitschrift macht Aussagen darüber, ob auch Zusatzmaterialen dem Peer-Review-Prozess unterliegen. Bei den im letzten Abschnitt untersuchten Verlagen und Zeitschriften konnte nur bei *Nature*, *The Lancet* und *Cell* festgestellt werden, dass Supplemental Material einem Peer-Review unterliegt (siehe Anhang 1).

Zwar haben 93 % der von Candela et al. (2015) untersuchten Zeitschriften Erhaltungsmaßnahmen (Preservation Policies) für ihre elektronischen Artikel getroffen. Aber nur 30 % verfügen über spezielle Policies für Supplemental Material.

Supplemental Material unterliegt den gleichen Zugangsberechtigung wie dem dazugehörigen Artikel. Es ist somit beispielsweise nicht möglich, Zusatzmaterialien unter einer Open Access Lizenz zu veröffentlichen, wenn der Artikel selbst zugangsbeschränkt ist.

3 Data Paper und Data Journals

> „A data paper is a searchable metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal. Unlike a conventional research article, the primary purpose of a data paper is to describe data and the circumstances of their collection, rather than to report hypotheses and conclusions."
>
> – *Global Biodiversity Information Facility (GBIF)*

Ein Data Paper ist eine Form eines wissenschaftlichen Artikels, bei dem ein Datensatz und der Kontext seiner Entstehung beschrieben wird. Der beschriebene Datensatz selbst befindet sich in der Regel in einem Repositorium. Der Datensatz im Repositorium und das Data Paper sind dabei bidirektional verlinkt. Weitere Bezeichnungen für diesen Artikeltyp sind unter anderem: Data Note, Database Article oder Data Article (Candela et al., 2015). Die Qualität eines Data Papers soll durch Peer-Review-Prozesse gesichert werden.

Es existieren sowohl spezielle Data Journals, die ausschließlich Data Papers aus einem bestimmten Fachgebiet veröffentlichen (z. B. Earth System Science Data,

Scientific Data, Geoscience Data Journal), als auch Zeitschriften, die neben klassischen Artikeln auch Data Papers publizieren (z. B. Nature Conservation, ZooKeys).

Data Papers machen Forschungsdaten in der Regel unter Open Access Lizenzen öffentlich zugänglich und erleichtern durch die Beschreibung der Datensätze deren Nachnutzung. Durch die Verknüpfung eines Data Papers in einer Zeitschrift und dem Datensatz im Repositorium erhöht sich die Auffindbarkeit von Forschungsdaten.

Data Papers verfügen über einen Persistent Identifier und können wie traditionelle wissenschaftliche Artikel zitiert werden. Hohe Zitationsraten eines Data Papers können als Maß für die Relevanz des Datensatzes betrachtet werden und geben Autoren somit Anerkennung für die Publikation von Forschungsdaten. Dabei können auch Wissenschaftler Anerkennung bekommen, die sich nicht als Autoren für einen traditionellen Artikel qualifizieren würden, aber bei der Datenerhebung eine wesentliche Rolle gespielt haben.

Data Papers lassen sich unabhängig von einer wissenschaftlichen Erkenntnis publizieren. Komplette Datensätze können so bereits vor abschließenden Analysen oder der Beantwortung einer Forschungsfrage für die Wissenschaft zur Verfügung gestellt werden.

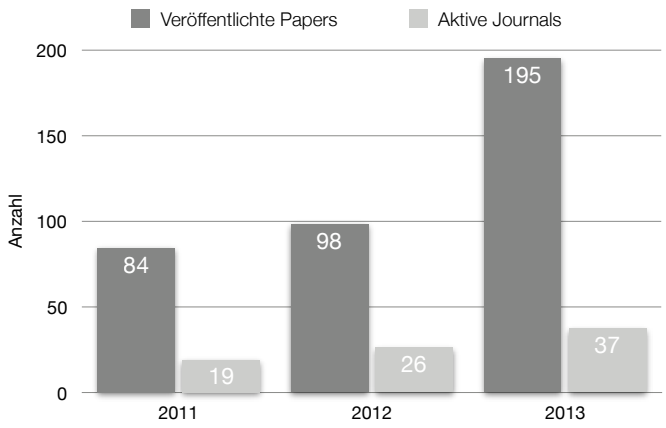

Abbildung 2: Publizierte Data Papers und Anzahl der aktiven Zeitschriften pro Jahr (nach Candela et al. (2015), Figure 3).

Candela et al. (2015) untersuchten 116 Zeitschriften, die Data Papers akzeptieren. Diese stammten hauptsächlich aus den Health Sciences, Life Sciences und Physical Sciences. Von den 116 untersuchten Zeitschriften sind nur sieben reine Data Journals, die ausschließlich Data Papers veröffentlichen. 113 dieser Zeitschriften sind Open Access Zeitschriften. Abbildung 2 zeigt die Anzahl der veröffentlichten Data Papers der untersuchten Zeitschriften in den Jahren 2011 – 2013. Im

Verhältnis zu der Anzahl der untersuchten Zeitschriften wurden pro Jahr nur relativ wenig Data Papers publiziert. Zwar hat sich die Anzahl der publizierten Artikel in diesem Zeitraum verdoppelt, dennoch haben im Jahr 2013 nur 32 % der in der Stichprobe enthaltenen Zeitschriften, Data Papers veröffentlicht.

3.1 Herausforderungen

Data Papers haben keine einheitlichen Standards zur Beschreibung der Datensätze. Zwar geben alle Data Papers an, wo die Datensätze zur Verfügung stehen, die Informationen darüber befinden sich aber an unterschiedlichen stellen im Artikel. Einige Data Papers haben eigene Abschnitte, die den Zugriff auf die Datensätze erläutern, andere zitieren den Datensatz nur in der Literatur. Dezidierte Informationen über die Qualität und den Anwendungsbereich der Datensätze fehlen häufig (Candela et al. (2015)). Es gibt eine großer Heterogenität im Aufbau von Data Papers und der verwendeten Metadaten.

Standardisierte Metadaten über den Datensatz selbst sind häufig nicht Teil des Data Papers und stehen nur im verwendeten Repositorium zur Verfügung. Eine bessere bidirektionale Integration von Metadaten im Repositorium und denen des Data Papers, könnte zu einer verbesserten Auffindbarkeit beitragen.

Zwar werden die meisten Data Papers einem Peer-Review unterzogen, aber es fehlen einheitliche Standards für die Begutachtung. Es ist häufig unklar, ob sich das Peer-Review nur auf den Artikel selbst bezieht oder auch die Datensätze selbst begutachtet. Die Bewertung der Qualität eines Datensatzes ist anspruchsvoll und hängt, neben der Art der geplanten Nachnutzung in einer bestimmten Domaine, auch von der Komplexität und Größe der Datensätze ab. Es fehlen zudem einheitliche Standards für die Bewertung der Datenqualität. Es ist davon auszugehen, dass sich das Peer-Review von Data Papers hauptsächlich auf den Artikel selbst fokussiert.

Für den uneingeschränkten Zugriff auf die Datensätze sind Data Papers auf Dritte angewiesen; den Forschungsdaten-Repositorien und Archiven. Sie müssen sich auf die Erhaltungsstrategien dieser Dienste verlassen können.

4 Forschungsdaten-Repositorien

> „Digitale Forschungsdaten-Repositorien sind Informationsinfrastrukturen, die digitale Forschungsdaten möglichst dauerhaft – anhand der Anforderungen der jeweiligen Nutzergruppe – speichern und organisieren, um die Auffindbarkeit und Zugänglichkeit der Daten zu sichern. Forschungsdaten-Repositorien werden durch disziplinäre Anforderungen geprägt (z. B. Form und Format der Daten)."
>
> – *Heinz Pampel (Pampel, 2014)*

Forschungsdaten können auch publikationsunabhängig in Repositorien bereitgestellt werden. In den meisten Repositorien besteht die Möglichkeit, von Datensätzen auf dazugehörige Veröffentlichungen zu verweisen. Dies erfolgt idealerweise über die Verlinkung von Persistent Identifiers. In Abbildung 3 wird schematisch der Publikationsverlauf eines Data Papers und eines traditionellen Forschungsartikel mit Zusatzmaterialien dargestellt. Datensätze, die in einem Data Paper beschrieben werden, liegen immer in einem Repositorium vor. Zusatzmaterialien bei einem Forschungsartikel können entweder zusammen mit dem Artikel auf der Plattform der elektronischen Zeitschrift als Supplemental Material, oder in einem Repositorium bereitgestellt werden. Befindet sie das Zusatzmaterial in einem Repositorium wird meist innerhalb des Artikels auf den Datensatz im Repositorium verwiesen.

4.1 Anforderungen an Forschungsdaten-Repositorien

Bei der Untersuchung unterschiedlicher Data Papers haben Candela et al. (2015) bereits Anforderungen an Forschungsdaten-Repositorien, die Data Journals stellen, zusammengetragen. Diese lassen sich wie folgt zusammenfassen:

- Repositorien und deren betreibende Organisationen müssen international oder institutional anerkannt und vertrauenswürdig sein
- Repositorien müssen die Langzeitverfügbarkeit und den permanenten Zugriff auf die Daten gewährleisten (Policies)
- Repositorien müssen Datensätzen Persistente Identifier zuweisen (z. B. DOI)
- Repositorien müssen den Zugriff auf Datensätze kostenfrei ermöglichen

4.2 Arten von Forschungsdaten-Repositorien

Es lassen sich zwei Arten von Repositorien unterscheiden. Es gibt eine Vielzahl von spezialisierten, fachgebietsspezifischen Repositorien mit spezifischen Anforderungen an Datentypen und -formaten, die in diesen Repositorien verwaltet werden. Daneben gibt es allgemein ausgerichtete, multidisziplinäre Repositorien, die eine Vielzahl von unterschiedlichen Datentypen akzeptieren.

Mit re3data.org[2] existiert ein globales Verzeichnis für Forschungsdaten-Repositorien. Damit lassen sich fachspezifische als auch interdisziplinäre Repositorien für die Bereitstellung von Forschungsdaten recherchieren.

[2] http://www.re3data.org/

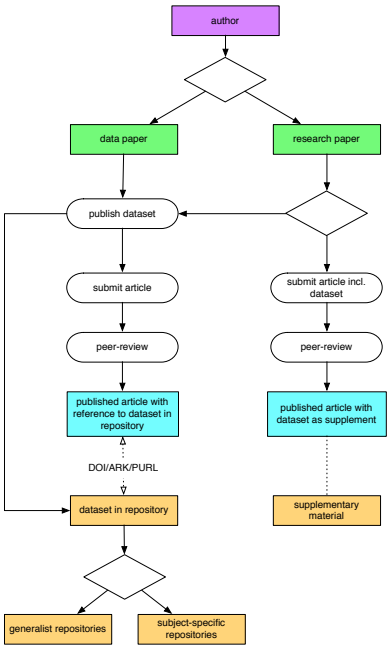

Abbildung 3: Publikationsverlauf im Vergleich

4.3 Vergleich verschiedener Forschungsdaten-Repositorien

Im Anhang 2 werden sechs Forschungsdaten-Repositorien verglichen. *Zenodo*[3], *figshare*[4], *Dryad*[5], *DANS-EASY*[6] und *Harvard Dataverse*[7] sind Beispiele für interdisziplinäre Repositorien. Mit *PANGAEA*[8] wurde zusätzlich ein fachgebietsspezifisches Repositorium untersucht.

Die Kategorien nach denen die Repositorien verglichen werden, wurden durch die Aggregation der Funktionsbeschreibung der jeweiligen Repositorien und den Anforderungen an Forschungsdaten-Repositorien aus Abschnitt 4.1 zusammengestellt.

Alle untersuchten Repositorien erfüllen die Anforderungen an Forschungsdaten-Repositorien. Hinter den Repositorien stehen anerkannte Institutionen als Betreiber. Alle Repositorien verfügen über Preservation and Archiving Policies. Jeder Datensatz wird mit einem DOI versehen und ist frei zugänglich.

Bezüglich der Auffindbarkeit, werden mit Ausnahme von *DANS-EASY* alle

[3] http://www.zenodo.org/
[4] http://figshare.com/
[5] http://datadryad.org/
[6] https://easy.dans.knaw.nl/ui/home/
[7] https://dataverse.harvard.edu/
[8] http://www.pangaea.de/

Repositorien von Abstracting & Indexing Services wie *SCOPUS* oder *Thomson Reuters* indiziert. Darüber hinaus bieten *Zenodo*, *figshare*, *Dryad*, *Harvad Dataverse* und *PANGAEA* Standard-APIs für Suchanfragen und Daten-Harvesting.

DANS-EASY und *PANGAE* geben Dateigrößenbeschränkungen von 100 MB pro Datei an, bei den restlichen Repositorien liegt diese bei 2 – 5 GB pro Datei. Pro Datensatz können mehrere Dateien gleichzeitig hochgeladen werden.

Die Bereitstellung von Daten in einem Repositorium ist in der Regel kostenfrei. Erst bei dem Überschreiten einer bestimmten Datenmenge können Kosten entstehen.

figshare, *Zenodo* und *Harvard Datavers* verfügen über Kollaborationsfunktionen, wie zum Beispiel die Erstellung eigener Sub-Repositorien. Außer *DANS-EASY* bieten alle Repositorien Exportfunktionen für Zitationen an, bei *Zenodo* und *figshare* ist diese Funktion besonders umfangreich.

Der Verweis auf eine wissenschaftliche Publikation ist bei *Dryad* und *PANGAEA* zwingend erforderlich. *Zenodo* akzeptiert eine Vielzahl von Persistent Identifiers zur Verknüpfung von Publikationen, *figshare* bitte die Möglichkeit URLs anzugeben. Bei *DANS-EASY* und *Harvard Dataverse* können Zitationen einer Publikation in einem Freitextfeld eingetragen werden.

Zenodo, *figshare* und *Dryad* haben ORCID[9] integriert. Damit lassen sich Datensätze eindeutig mit Datenproduzenten assoziieren. Darüber hinaus können Zenodo und figshare die Bereitstellung von Forschungsdaten automatisch an bestimmte Förderinstitutionen melden.

4.4 Herausforderungen

Besonders bei den interdisziplinären Repositories ist häufig unklar, wie die Authentizität der Datensätze sichergestellt wird. Das fachgebietsspezifische Repositorium *PANGAEA* überprüft jeden eingereichten Datensatz vor der Veröffentlichung. Bei interdisziplinären Repositorien, die eine Vielzahl von Daten ohne Überprüfung akzeptieren, muss dem Datenproduzenten selbst vertraut werden.

Zitationen von Datensätzen in Repositorien sind nur auf der Ebene des Datensatzes möglich. Es existieren keine etablierten Möglichkeiten, um nur Teile von Datensätzen zu zitieren. Auch die Auffindbarkeit beschränkt sich auf die angegebenen Metadaten zu einem Datensatz. Ein Retrieval auf die enthaltenen Datensätze selbst, bieten die untersuchten Repositories nicht an.

Viele Funktionalitäten von Repositories ähneln sich, es fehlt aber an übergreifenden Standardisierungen. Besonders die Metadaten, die zu einem Datensatz angegeben werden können, unterscheiden sich bei interdisziplinären Repositorien deutlich.

[9] http://orcid.org/

Zwar verfügen alle untersuchten Repositorien über Preservation und Archiving Policies, doch der Wert und die Einhaltung dieser Policies sollte Zertifiziert werden, um als vertrauenswürdiges Repositorium gelten zu können. Von den untersuchten Repositorien hat allein *DANS-EASY* ein solches Zertifikat, das Data Seal of Approval[10]. Weitere Zertifizierungsmöglichkeiten für Repositorien wären die Trustworthy Repositories Audit & Certification (TRAC)[11] oder das nestor-Siegel für vertrauenswürdige digitale Langzeitarchive[12].

5 Zusammenfassung

Supplemental Material bot erstmals die Möglichkeit, Forschungsdaten zusammen mit einem Wissenschaftlichen Artikel zu publizieren. Mittlerweile wird von dieser Möglichkeit umfangreich gebrauch gemacht. Richtlinien von Zeitschriften und Verlagen zu Supplemental Material unterschieden sich stark und schränken Umfang und Art der Daten, die Veröffentlicht werden können, teilweise erheblich ein. Meist werden nur reine Datensätze veröffentlicht, strukturierte Metadaten und eine standardisierte Dokumentation dieser Datensätze fehlt in der Regel.

Seit 2010 existiert eine gemeinsame Arbeitsgruppe von NISO (National Information Standards Organization) und NFAIS (National Federation of Abstracting and Information Services) zur Standardisierung von Supplemental Material. Das *Supplemental Journal Article Materials Project* hat das Ziel einheitliche Empfehlungen für den Umgang mit Supplemental Material zu entwickeln. Dabei wird auf Auswahl, Bearbeitung, Bereitstellung, Auffindbarkeit, Referenzierung und Erhaltung von Supplemental Material eingegangen und ein Metadatenformat für die technische Umsetzung vorgeschlagen (NISO/NFAIS, 2013).

Erste Zeitschriften sind dazu übergegangen, kein Supplemental Material mehr zu akzeptieren, da das Volumen so stark zugenommen hat, dass kein Peer-Review mehr möglich ist (z. B. Journal of Neuroscience). Stattdessen wird darauf gedrängt, zusätzliche Daten in entsprechenden Forschungsdaten-Repositorien abzulegen (Reilly et al., 2011, S. 45). Datensatz und Artikel können dabei bidirektional miteinander verlinkt sein.

Data Papers fördern die Nachnutzung von Forschungsdaten durch die genaue Dokumentation der Entstehung und Verwendung der Datensätze. Die Forschungsdaten selbst befinden sich in Repositorien, auf die vom Data Paper verwiesen wird. Da es sich bei Data Paper um einen relativ neuen Artikeltyp handelt, gibt es noch keine einheitlichen Auffassung darüber, was ein Data Paper leisten muss. Zwar werden die meisten Data Paper einem Peer-Review unterzogen, häufig ist aber nicht klar, ob sich diese Begutachtung nur auf den Artikel selbst bezieht oder auch die Qualität der beschriebenen Datensätze bewertet. Dabei

[10] http://datasealofapproval.org/
[11] http://www.crl.edu/archiving-preservation/digital-archives/metrics/
[12] http://www.langzeitarchivierung.de/Subsites/nestor/DE/nestor-Siegel/siegel.html

ist auch offen, mit welchen Methoden sich die Qualität von Forschungsdaten eigentlich bestimmen lässt. Es fehlt des Weiteren an einheitlichen Standards zur Beschreibung der Datensätze und ihrer Metadaten. Eine bessere Integration der Metadaten zwischen Data Journals und Repositorien ist wünschenswert.

Besonders interdisziplinäre Repositorien haben niedrige Hürden für die Veröffentlichung von Datensätzen. Sie stehen jedem offen und es gibt beinahe keine Einschränkungen bezüglich der Datenformate und deren Volumen. Jeder Datensatz bekommt ein DOI zugweisen und ist somit eindeutig zitierbar. Publikationen, die mit den Datensätzen in Verbindungen stehen, lassen sich assoziieren. Preservation Policies sichern die unbegrenzte Bereitstellung dieser Daten für die Öffentlichkeit zu.

Fachgebietsspezifische Repositorien erleichtern durch Zentralisierung die Auffindbarkeit bestimmter Datenformate in den entsprechenden Domänen. Sie sind in der Lage, Metadatenstandards für ihre Datensätze festzulegen und helfen, durch Vorgaben an Datenproduzenten und interne Qualitätssicherungsmaßnahmen, die Datenheterogenität im Repositorium zu minimieren. Es gibt praktisch keine Beschränkung der Menge an Daten die veröffentlicht werden können. Besonders multidisziplinäre Repositorien haben geringe Eintrittsbarrieren und stehen jederzeit für die Ablage von Forschungsdaten zur Verfügung – unabhängig von einer wissenschaftlichen Publikation.

Erstrebenswert ist die Verbindung von Forschungsergebnissen in wissenschaftlichen Publikationen mit den dazugehörigen Forschungsdaten in entsprechenden fachspezifischen Repositorien und der Dokumentation dieser Datensätze in Data Papers. Die Verknüpfung dieser drei Teile ist dabei durch die Verwendung von einem Netzwerk persistenter Identifier, wie DOIs, zu gewährleisten.

Die Publikation von Forschungsdaten ist nicht mehr nur ein Nebenprodukt der Veröffentlichung wissenschaftlicher Arbeiten. Die Datensätze selbst entwickeln sich zu eigenständigen zitierfähigen wissenschaftlichen Einheiten. Die Häufigkeit ihrer Zitierung ist ein Maß für die Qualität und Relevanz als Forschungsprodukt selbst und resultiert in Anerkennung für Autoren.

Literatur

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P. et al. (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, *7* (1), 107–113. doi:10.2218/ijdc.v7i1.218

Candela, L., Castelli, D., Manghi, P. & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66* (9), 1747–1762. doi:10.1002/asi.23358

CellPress. (2015). Supplemental Information Guidelines. *Cell Press*. http://www.cell.com/supplemental-information.

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, *12* (0), CIDCR1–CIDCR75. doi:10.2481/dsj.OSOM13-043

Kuipers, T. & Van der Hoeven, J. (2009). *Insight into digital preservation of research output in Europe. Survey Report*. PARSE.Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe.

Lawrence, B., Jones, C., Matthews, B., Pepler, S. & Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation*, *6* (2), 4–37. doi:10.2218/ijdc.v6i2.205

Lawrence, R. (2012). Data publishing: Peer review, shared standards and collaboration. Southampton.

NISO/NFAIS (Hrsg.). (2013). *Recommended Practices for Online Supplemental Journal Article Materials (NISO RP-15)*. NISO/NFAIS.

Pampel, H. (2014, Dezember). Ausgewählte Aspekte digitaler Informationsversorgung (SoSe 14). Vortrag, Humboldt-Universität zu Berlin, Institut für Bibliotheks- und Informationswissenschaft (IBI).

Reilly, S., Schallier, W., Schrimpf, S., Smit, E. & Wilkinson, M. (2011). Report on integration of data and publications. *Opportunities for Data Exchange (ODE)*.

Schriger, D.L., Chehrazi, A.C., Merchant, R.M. & Altman, D.G. (2011). Use of the Internet by Print Medical Journals in 2003 to 2009: A Longitudinal Observational Study. *Annals of Emergency Medicine*, *57* (2), 153–160.e3. doi:10.1016/j.annemergmed.2010.10.008

Tempest, D. (2012). Journals And Data Publishing: Enhancing, Linking And Mining. Southampton.

Vision, T. (2015). Data and the scientific literature: new directions in what data gets published, how it happens & why it matters. In *National Data Integrity Conference-2015*. Colorado State University. Libraries.

Anhang

A1 Vergleich von Supplemental Material Richtlinien

	PLoS	Springer	Wiley	SAGE	CELL	PANS	Nature	The Lancet
Dateiformate	keine Einschränkung	keine Einschränkung	keine Einschränkung (Empfehlungen s. u.)	PDF und alle MS Office Format (Word, Excel, Powerpoint, Project, Access, etc.). Abbildungen, Video und Audi siehe unten.	Ein PDF mit allen Abblidungen, Tabellen, Referencen, etc.. Video und Audio möglich.	Es wird ein PDF aus den geleiferten Dokumenten erstellt. Ausnahmen für Video und 3D Abbildungen. Nur angegebne Dateitypen erlaubt.	PDF (preferred), .txt, .rtf, .wpd, .ps, .eps, .htm, .xls, .xlsx, .mov (preffered), .wav, .mpg, .mp4, .mp3, Systems Biology Markup Language (.sbml, .xml, .owl)	Ein PDF mit allen Abblidungen, Tabellen, Referencen, etc.. Video und Audio möglich.
Dateigröße	keine Einschränkung	keine Einschränkung	Bitte um max. 10 MB pro Datei.	Bitte um max. 10 MB pro Datei.	PDF max. 10 MB Video max. 150 MB	Video max. 10 MB	Datei max. 30 MB (150 MB insg.)	Video/Audio max. 50 MB
Tabellen	keine speziellen Richtlinien	"Spreadsheets should be converted to PDF if no interaction with the data is intended."	keine speziellen Richtlinien	keine speziellen Richtlinien	"PDF that contains all supplemental tables. If a supplemental table cannot fit onto three 8.5" x 11" pages, please instead supply the table separately as an Excel file."	"Supply Word, RTF, or LaTeX files (LaTeX files must be accompanied by a PDF with the same file name for visual reference); include only one table per file. Do not use tabs or spaces to separate columns in Word tables."	"Note that Tables may be included in Supplementary Information, but only if they are unsuitable for formatting as Extended Data tables (for example, tables containing large data sets or raw data that are best suited to Excel files)."	Teil des PDFs
Abbildungen	keine speziellen Richtlinien	"A collection of figures may also be combined in a PDF file."	keine spezielle Richtlinien (Empfehlungen: GIF, TIF (or TIFF), EPS, PNG, JPG (or JPEG), BMP, PS (Postscript))	GIF, TIF, EPS, PNG, JPG, BMP, PS "Embedded graphics (i.e. a GIF pasted into a Word file) are also acceptable."	"PDF that contains all supplemental figures and legends"	TIFF (LZW), EPS, PDF, JPEG, GIF "Provide a brief legend for each supporting figure after the supporting text … figures may not be embedded in manuscript text. … Do not save figure numbers, legends, or author names as part of the image. Composite figures must be preassembled."	"Images should be a maximum size of 640 x 480 pixels (9 x 6.8 inches at 72 pixels per inch)."	im PDF "minimum resolution of 300 dpi, width 107 mm"
Audio	keine speziellen Richtlinien	keine speziellen Richtlinien	Empfehlungen: MP3, AAC, WMA	MP3, AAC, WMA, WAV, SPHERE, TIMIT	keine speziellen Richtlinien	keine Angaben	WAV, MP3	MP3 max. 50 MB
Video	"We expect reasonable video quality and prefer 128 kbit/s AAC audio ZD and 480p H.264 video in an MPEG-4 (mp4) container." "Preferred size limit of videos is 10 MB. If making the dimensions smaller or recompressing the video compromises the image quality or usefulness of the video, we can accept the video file as is."	"Resolution: 16:9 or 4:3. Minimum video duration: 1 sec. Video files do not contain anything that flashes more than three times per second (so that users prone to seizures caused by such effects are not put at risk)"	"All video clips must be created with commonly-used codecs, and the codec used should be noted in the supporting information legend"	MOV, MPEG, AVI "All video clips must be created with commonly-used codecs, and the codec used should be noted in the supporting information legend"	MP4, MOV, AVI, MPG Video max. 150 MB "Frame rate: 15 frames per second minimum Recommended frame size: 492x276 (16:9) Field order: none (progressive, not interlaced) Aspect ratio: widescreen 16:9 Video codec: H.264 (+AAC preferred) Video bitrate: at least 260 kbps (750 kbps preferred)Audio codec: AAC Audio bitrate: 128 kbps"	AVI, MOV, WMV, GIF, MPEG Video max. 150 MB "All movies should be submitted at the desired reproduction size and length."	"For optimal-quality videos please use a H.264 encoding, the standard aspect ratio of 16:9 (4:3 is second best) and do not compress the video."	MPG, MOV, AVI, GIF Video max. 150 MB "aspect ratio of 16:9"
Datasets	keine speziellen Richtlinien	"Spreadsheets should be converted to PDF if no interaction with the data is intended. If the readers should be encouraged to make their own calculations, spreadsheets should be submitted as .xls files (MS Excel)."	"If a native dataset is supplied, the program and/or equipment used should be given. For specialist software (e.g. LaTeX), the software and version number used should be given."	keine speziellen Richtlinien	keine speziellen Richtlinien	"Supply Excel (.xls), RTF, or PDF files. This file type will be published in raw format and will not be edited or composed."	"tables containing large data sets or raw data that are best suited to Excel files"	keine Angaben
Peer-Review	keine Angaben	keine Angaben	"It will not be available for review prior to publication"	"Supplementary files will be subjected to peer-review alongside the article."	keine Angaben	keine Angaben	"Supplementary Information (SI) is peer-reviewed material "	"All material … will be peer reviewed"
URL	http://journals.plos.org/plosone/s/supporting-information	http://www.springer.com/authors/manuscript+guidelines?SGWID=0-40162-12-339499-0	https://authorservices.wiley.com/bauthor/suppinfo.asp	https://uk.sagepub.com/en-gb/eur/supplementary-files-on-sage-journals-sj-guidelines-for-authors	http://www.cell.com/supplemental-information	http://www.pnas.org/site/authors/preparation.xhtml	http://www.nature.com/nature/authors/submissions/final/suppinfo.html	http://www.thelancet.com/lancet/information-for-authors/web-extra-guidelines

A2 Vergleich von Forschungsdaten-Repositorien

	Zenodo	figshare	Dryad	DANS-EASY	Harvard Dataverse	PANGAEA
Persistent identifier system(s)	DOI	DOI	DOI	DOI / URN	DOI	DOI
Fachgebiete	interdisziplinär	interdisziplinär	interdisziplinär	interdisziplinär	interdisziplinär	Earth and Environmental Science
Auffindbarkeit	Suche im Repository und über API, Thomson Reuters Data Citation Index	Suche im Repository und über API, Google Scholar Indizierung, Thomson Reuters Data Citation Index	Suche im Repository und über API, Tabelle1, SCOPUS	Suche im Repository	Suche im Repository und über API, Thomson Reuters Data Citation Index, SCOPUS	Suche im Repository und über API, Thomson Reuters Data Citation Index, SCOPUS,
Lizenzierungsmodell	Creative Commons Zero (CC0) vorausgewählt, 100 weiter Lizensen zur Auswahl	CC-BY vorausgewählt, weitere freie Lizensen stehen zur auswahl (CC0, GPL 1-3, Apache 2.0, MIT). Instituionelle Nutzer können eigenen Lizensen hinterlegen.	Creative Commons Zero (CC0)	Creative Commons Zero (CC0)	Creative Commons Zero (CC0) . Es besteht die Möglichekiten ein *custom data usage license agreement* zu verwenden.	CC-BY vorausgewählt, alle weitern CC-Lizensen auswählbar
Embargofunktion	Ja	Ja	Ja	Ja	Nein	Ja
Collaboration	Community Collections (eigene sub-Repositorys mit Nutzerverwaltung)	Private Ordner können geteilt werden	Nein	Nein	Dataverse (eigene sub-Repositorys mit Nutzerverwaltung)	Nein
Dateibeschränkungen	2 GB pro Datei	5 GB pro Datei	keine Angaben	100 MB pro Datei (listet bevorzugte Formate)	2 GB pro Datei	100 MB pro Datei (listet bevorzugte Formate)
Versionierung	Ja	Ja	Ja	Ja	Ja	Ja
Dateibetrachter	Vorschau von PDFs und Archivinhalten.	Vorschau von Abbildungen, tabellarischen Daten, Videos, Office Dokumeten, PDFs, Geodaten und Archivinhalten.	Nein	Nein	Betrachtung von Geodaten durch WorldMap integration. Betrachtung und Analyse von tabellarischen Daten durch TwoRavens Integration.	Vorschau von tabellarischen Daten.
Export von Zitation	Ja (8 Optionen)	Ja (8 Optionen)	Ja (2 Optionen)	Nein	Ja (2 Optionen)	Ja (2 Optionen)
API	Ja (OAI-PMH und weitere)	Ja	Ja (OAI-PMH und weitere)	Nein	Ja (SWORD und weitere)	Ja (OAI-PMH)
Reportingmöglichkeit an Förderinstitution	Angabe der *grant number* für EU-Projekte möglich (OpenAIRE)	Angabe der *grant number* möglich.	Nein	Nein	Nein	Nein
Verweis auf Publikationen	Ja, über einer Vielzahl von Identifiern möglich (DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, URLs ...).	Ja, ausschließlich über URLs.	Datensatz muss zwingend mit einem Zeitschriftenartikel assoziiert werden (DOI oder PubMedID)	Volltextfeld	Volltextfeld	Ja, DOI
Weiter Features	DropBox und GitHub Integration	Desktop Uploader. Widget zum Einbetten von Datensätzen.	-	-	DropBox Integration. Widget zum Einbetten von Datensätzen.	-
Kosten	Keine	bis 20 GB frei, Business Model für Institutionen und Verlage	120 $ pro Datenpublikation (unless there is a sponsor or fee waiver in place. Additional charges apply to data packages in excess of 20 GB.)	Keine	Keine	Keine (PANGAEA would appreciate a financial contribution of 300.- € per data supplement of a publication)
Preservation and Archiving Policies	Ja (https://zenodo.org/policies)	Ja (https://figshare.zendesk.com/hc/en-us/articles/207056827-Preservation-Policies)	Ja (https://datadryad.org/pages/policies#preservation and http://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreservationpolicyUK.pdf)	Ja. Data Seal of Approval (https://assessment.datasealofapproval.org/assessment_101/seal/html/)	Ja (http://dataverse.org/best-practices/harvard-dataverse-preservation-policy)	Ja (http://www.pangaea.de/curator/files/pangaea-data-policy.pdf)
Quality management	Ja	Ja	Ja	Ja	Nein	Ja
Author identifier system(s)	ORCID	ORCID	ORCID	Nein	Nein	Nein
Institution	European Organization for Nuclear Research (CERN); OpenAIRE	Digital Science, Holtzbrinck Publishing Group.	University of North Carolina; Metadata Research Center; National Evolutionary Synthesis Center; Dryad	DANS; Netherlands Organisation for Scientific Research; Royal Netherlands Academy of Arts and Sciences	Harvard University; Institute for Quantitative Social Sciences	Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung; Center for Marine Environmental Sciences (MARUM)
URL	http://www.zenodo.org/	http://figshare.com/	http://datadryad.org/	https://easy.dans.knaw.nl/ui/home/	https://dataverse.harvard.edu/	http://www.pangaea.de/
re3data.org record	http://doi.org/10.17616/R3QP53	http://doi.org/10.17616/R3PK5R	http://doi.org/10.17616/R34S33	http://doi.org/10.17616/R3401D	http://doi.org/10.17616/R3C880	http://doi.org/10.17616/R3XS37