Bausteine Forschungsdatenmanagement
Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

Die Entwicklung eines Forschungsdatenarchivs für Fachzeitschriften

Sven Vlaeminck[i]

2018

Zitiervorschlag

Vlaenimck, Sven. 2018. Die Entwicklung eines Forschungsdatenarchivs für Fachzeitschriften. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 1/2018: S. 57-63. DOI: 10.17192/bfdm.2018.12.7825.

[i]ZBW Leibniz-Informationszentrum Wirtschaft. ORCID: 0000-0002-7905-4209

1 Beschreibung

Das von der DFG zwischen 2011 und 2016 geförderte Projekt “European Data Watch Extended” (EDaWaX[1]) hatte zum Ziel, ein publikationsbezogenes Forschungsdatenarchiv für wirtschaftswissenschaftliche Fachzeitschriften als Webanwendung zu entwickeln. Die zentrale Funktionalität des Systems besteht in der Verlinkung von Fachartikeln mit den ihnen zugrunde liegenden Forschungsdaten, die in diesem Datenarchiv gespeichert sind. Diese Forschungsdaten sind mit einem Digital Object Identifier (DOI) versehen und somit zitierfähig. Die Metadaten der Forschungsdaten können zudem durch disziplinäre und interdisziplinäre Portale und Suchmaschinen über sog. “Harvesting” automatisch abgefragt werden. Zudem kommen bei der Applikation verschiedene Application Programming Interfaces (API)-Anbindungen zum Einsatz und erleichtern somit ‘nebenbei’ Forschenden ihre Arbeit bei der Erstellung von Metadaten. Der Anstoß für das Projekt kam aus der Wirtschaftsforschung selbst: Ausgelöst durch eine Reihe von Studien, die darlegten, wie schlecht es um die Reproduzierbarkeit publizierter Ergebnisse aus der empirischen Wirtschaftsforschung bestellt sei, sahen einige Zeitschriftenherausgeber Handlungsbedarf und starteten gemeinsam mit der ZBW – Leibniz Informationszentrum Wirtschaft die Projektinitiative.

2 Handlungsfelder

Mit dem Auftrag, ein publikationsbezogenes Forschungsdatenarchiv aufzubauen, war das Projekt in einem Bereich angesiedelt, der noch relativ nah am ‘klassischen’ Aufgabenfeld von Bibliotheken mit ihrer traditionellen Fokussierung auf Publikationen lag. Dennoch galt es zunächst im Zuge einer Anforderungs- und Umfeldanalyse Expertise aufzubauen. Die Entwicklung einer Infrastruktur für publikationsbezogene Forschungsdaten war somit eine gute Gelegenheit, um weitere Erkenntnisse zur Wissenschaftskultur in der Wirtschaftsforschung, zur Forschungsmethodik, zu den verschiedenen Arten von Daten und der von Forschenden häufig genutzten Software zu erlangen. Einige Fragen, die dezidiert mit der Anforderungs- und Umfeldanalyse abgedeckt wurden, waren:

- Wie adressieren Fachzeitschriften in der Domäne das Thema replizierbare Forschung bislang? Wie viele wirtschaftswissenschaftliche Zeitschriften verfügen über Richtlinien (“Data Policies”) zum Umgang mit Forschungsdaten? Wie sind diese Policies inhaltlich ausgestaltet?[2]

[1] http://www.edawax.de

[2] Vgl. Vlaeminck, Sven und Lisa-Kristin Herrmann. 2015. *Data Policies and Data Archives: A New Paradigm for Acadamic Publishing in Economic Sciences?* DOI: http://dx.doi.org/10.3233/978-1-61499-562-3-145.

- Welche Forschungsdaten und welche zusätzlichen Informationen werden überhaupt benötigt, um Replikationen in den Wirtschaftswissenschaften im Zusammenhang mit publizierten Ergebnissen zu ermöglichen?[3]
- Welche Perspektive haben Wirtschaftsforschende zum 'data sharing'? 'Teilen' sie ihre Daten?[4] Wie können Anreize zum Data Sharing gestaltet werden?[5]
- Existieren ggf. schon passende Services an deutschen oder europäischen Forschungsdatenzentren, auf die aufgebaut werden könnte?[6]
- Welche rechtlichen Aspekte sind beim Data Sharing zu beachten?[7]

Stark zusammengefasst brachten die Analysen des Projekts in Bezug auf die technische Entwicklung folgende Ergebnisse:

- Volkswirtschaftliche Fachzeitschriften entdecken zunehmend das Thema "Replizierbarkeit von Forschungsergebnissen" für sich. 2015 verfügte bereits ein Viertel von 130 untersuchten Fachzeitschriften über eine Richtlinie zum Umgang mit Forschungsdaten. Als Quasi-Standard scheint sich die Data Availability Policy der American Economic Association (AEA) durchzusetzen.[8] Diese benennt exemplarisch auch die Dateien und Informationen, die für Replikationen benötigt werden – und die das geplante Datenarchiv somit speichern und mit Metadaten versehen muss. Zu wirtschaftswissenschaftlichen Forschungsdaten zählen vielfach (aber nicht ausschließlich) quantitative Datensätze, Syntaxfiles/ Programmcodes, Beschreibungen der Datensätze oder auch sog. Readme-files, welche die einzelnen Daten und deren Einbindung in den Forschungs- und Analyseprozess darlegen.
- Da in den Wirtschaftswissenschaften auch viel auf Basis proprietärer (z.B. Finanztransaktionsdaten) oder geschützter Daten (z.B. personenbezogene Daten und andere Mikrodaten, wie etwa das Sozio-oekonomische Panel[9]) geforscht und publiziert wird, muss die Applikation in der Lage sein, auch solche Use Cases zu berücksichtigen. Dies kann etwa darüber geschehen, dass ein Link (besser: DOI) zur Landingpage der genutzten Daten in der Applikation hinterlegt wird (alternativ eine ausführliche Beschreibung und Kontaktangaben zur Erlangung

[3] Vgl. Vlaeminck, Sven. 2013. Data management in scholarly journals and possible roles for libraries– Some insights from EDaWaX. *LIBER Quarterly* 23 (1). DOI: http://doi.org/10.18352/lq.8082. Der Artikel stellt beispielhaft dar, welche Anforderungen vor allem in Hinsicht auf quantitative Daten (z.B. ökonometrische Analysen), Simulationen und Experimente bestehen.

[4] Vgl. Andreoli-Versbach, Patrick und Frank Mueller-Langer. 2014. Open access to data: An ideal professed but not practised. *Research Policy* 43 (9). DOI: https://doi.org/10.1016/j.respol.2014.04.008.

[5] Vgl. Vlaeminck, Sven und Gert Wagner. 2013. On the role of research data centres in the management of publication-related research data. *LIBER Quarterly* 23 (4). DOI: http://doi.org/10.18352/lq.9356.

[6] Vgl. Mueller-Langer, Frank und Patrick Andreoli-Versbach. 2017. Open access to research data: Strategic delay and the ambiguous welfare effects of mandatory data disclosure. *Information Economics and Policy* (40). DOI: https://doi.org/10.1016/j.infoecopol.2017.05.004.

[7] Vgl. u.a. http://auffinden-zitieren-dokumentieren.de/auffinden/zugangswege/.

[8] https://www.aeaweb.org/journals/policies/data-availability-policy.

[9] http://www.diw.de/de/diw_02.c.221178.de/ueber_uns.html.

der verwendeten Daten) sowie die zur Auswertung dieser Daten erstellte Syntax bzw. der Programmcode einer Anwendung.

Diese Ergebnisse flossen als Anforderungen in das Lastenheft für die softwaretechnische Entwicklung der Applikation ein. Ergänzend zu den Umfeldanalysen wurde anschließend eine Stakeholderanalyse durchgeführt, da bei der Entwicklung einer derartigen Applikation grundsätzlich verschiedene Anspruchsgruppen und Workflows denkbar sind: So könnten beispielsweise die Interaktion ausschließlich zwischen Forschenden und dem Datenarchiv stattfindet, aber auch zwischen Fachzeitschriften bzw. deren HerausgeberInnen und der Autorenschaft, oder auch zwischen allen drei Akteuren. 'Klassische' Forschungsdatenrepositorien setzen auf eine Interaktion zwischen Forschungsdatenzentrum und Forschenden bzw. Forschergruppen. Für die Entwicklung des publikationsbezogenen Datenarchivs wurde hingegen ein Modell favorisiert, in dem vor allem Fachzeitschriften bzw. deren HerausgeberInnen mit Ihren Autorinnen und Autoren interagieren. Dies ist vorrangig darin begründet, dass Fachzeitschriften in ihren Data Policies unterschiedliche Anforderungen an Autorinnen und Autoren stellen.

Somit wurden für die Konzeption des Datenarchivs zwei primäre Nutzergruppen identifiziert: Autorinnen und Autoren, die Ihre Daten einstellen, sowie Zeitschriftenredaktionen welche die eingestellten Daten auf Basis ihrer Anforderungen qualitativ evaluieren.

Der Workflow der Anwendung wurde letztlich so gestaltet, dass teilnehmende Redaktionen zunächst die AutorInnen empirisch-orientierter Fachartikel im System registrieren (bzw. dies bei der Nutzung von OJS[10] als Publikationssystem dies automatisiert erfolgt). Die AutorInnen stellen dann die in der Data Policy der Zeitschrift spezifizierten Forschungsdaten in die Applikation ein. Je nach Typ der Daten kommt eine leicht modifizierte Eingabemaske zur Eingabe der zugehörigen Metadaten (siehe unten) zur Anwendung.

Ist die Einreichung abgeschlossen, werden die HerausgeberInnen vom System über die vorliegenden Replikationsfiles informiert und aufgefordert, diese anhand der Kriterien der Data Policy ihrer Fachzeitschrift zu evaluieren und auch die Metadaten auf Korrektheit zu prüfen. Entsprechen Daten und Metadaten den Anforderungen, wird anschließend im System ein DOI für die Replikationsfiles generiert. Die publizierenden Verlage müssen somit nur diese DOI verlinken. Ebenso ergänzen die Redaktionen den DOI zum Fachartikel, wodurch die Replikationsdaten auf den Fachartikel verlinken.

Um die Replikationsdaten angemessen beschreiben zu können, war fraglich, welches Metadatenschema für den Aufbau eines solchen Datenrepositoriums geeignet ist. Beachtet wurden Fragen wie:

[10] https://pkp.sfu.ca/ojs/.

- Welche Metadaten werden benötigt? Zu welchem Zweck werden diese Metadaten erhoben? Was für Nachnutzungsszenarien sollten für die Auswahl eines Metadatenschemas und für die Entwicklung der Infrastruktur beachtet werden?

Grundsätzlich standen verschiedene Metadatenschemata zur Debatte. Insbesondere wurden die Schemata von DublinCore[11], MODS[12], DataCite[13], DDI[14] und da|ra[15] geprüft. Verwendet und implementiert wurde schließlich das Schema der Forschungsdatenregistratur da|ra. Dieses Schema ist (etwa im Vergleich zu DDI) eher schlank gehalten, verfügt aber (im Gegensatz zu DublinCore oder dem DataCite-Schema) über fachspezifische Felder, die wichtige Metainformationen zum Datensatz enthalten (z.B. Informationen zur Grundgesamtheit oder zur Zahl der Variablen). Darüber hinaus können die Metadaten zwecks Bezugs eines persistenten Identifiers ohne weiteres Mapping direkt an die DOI-Registrierungsagentur übermittelt werden.

Abschließend wurden im Projekt verschiedene auf dem Markt präsente Softwarelösungen und Dienste gesichtet und analysiert. In die Evaluierung wurden kommerzielle Lösungen (u.a. figshare) ebenso einbezogen, wie die von uns favorisierten Open Source-Pakete (u.a. Dataverse, CKAN, Zenodo,...). Neben den oben umrissenen Anforderungen wurden weitere Kriterien in die Evaluierung einbezogen:

- Die Applikation muss eine separate Speicherung von Metadaten und Datensätzen an physikalisch unterschiedlichen Orten ermöglichen (Anforderung bzgl. des Organisationsmodells).
- Die Applikation muss über leistungsfähige APIs verfügen, um zentrale Prozesse der Applikation auch über diese Schnittstellen steuern zu können (Speicherung, DOI-Vergabe, Dateneingabe,...)
- Die Applikation soll auch Use Cases unterstützen, in denen nur die Metadaten von Datensätzen bereitgestellt werden können, nicht aber die Datensätze selbst. Ggf. verfügbare Landingpages dieser Datensätze sollen jedoch als Ressource verlinkbar sein.
- Die Applikation muss in der Lage sein, unterschiedliche Workflows von Fachzeitschriften zu unterstützen.
- Das ausgewählte Metadatenschema muss in die Applikation implementiert werden können.
- Die Übermittlung von Forschungsdaten und Metadaten soll über ein komfortables Webfrontend gewährleistet werden und Anbindungsmöglichkeiten an Linked Data (RDF) bieten.

Im Ergebnis fiel die Wahl auf CKAN[16], ein Open Source-Paket, das von der Open Knowledge Foundation (OKFN) gepflegt wird und von zahlreichen Open Data Portalen bzw.

[11] http://dublincore.org/specifications/.
[12] https://www.loc.gov/standards/mods/.
[13] https://schema.datacite.org/.
[14] http://www.da-ra.de/home/.
[15] https://ckan.org/.
[16] https://ckan.org/.

-Applikationen genutzt wird. Die auf Python basierende Software war an unsere Anforderungen anpassbar, wenngleich auch mit nicht unerheblichem Programmieraufwand für die Implementierung des Metadatenschemas und eines journalspezifischen Workflows. Alle weiteren der oben genannten Anforderungen erfüllte die Software ebenfalls.

3 Beteiligte Akteure

In Bezug auf die Entwicklung einer Pilotapplikation waren neben den technischen und inhaltlichen Anforderungen auch Aspekte wie die Usability von Bedeutung. Alle Bereiche, die mit Funktionalitäten und Usability zusammenhingen, wurden in enger Abstimmung mit den intendierten Stakeholdern entwickelt. Unser Ansatz war hierbei, dass eine Software nur von den intendierten Nutzerinnen und Nutzern verwendet wird, wenn diese ihre Bedürfnisse und Anforderungen erfüllt sehen, der Aufwand zur Nutzung der Applikation nicht zu hoch ist und nicht mit ihrer Forschungsarbeit sowie sonstigen Aufgaben & Interessen kollidiert.[17] Daher wurde die intendierte Zielgruppe an der Entwicklung der Anwendung regelmäßig beteiligt. So waren die Herausgeberinnen und Herausgeber einer Fachzeitschrift im Steering Commitee des Projektkonsortiums vertreten. In späteren Planungsstadien wurden auch Autorinnen und Autoren einbezogen. Herzstücke der Arbeit mit der Wissenschaftscommunity waren Workshops und Veranstaltungen, in denen wir den jeweiligen Entwicklungsstand der Applikation vorstellten und zur Kommentierung einluden.

Im ersten Workshop wurde die Pilotapplikation den Herausgeberinnen und Herausgebern von 15 Fachzeitschriften vorgestellt. Während das Feedback grundsätzlich positiv ausfiel, wurden auch verschiedene Anregungen geäußert, die Auswirkungen auf die weitere technologische Entwicklung der Software hatten. Neben kleineren Rückmeldungen zum Look and Feel der grafischen Benutzeroberfläche (GUI) wurden auch Punkte genannt, die umfangreichere Programmierarbeiten nach sich zogen. Hierzu zählte etwa die komplette Überarbeitung der Darstellung und Abfrage der Metadaten sowie Veränderungen im Rechtemanagement. Das Vorhandensein von nur zwei verschiedenen Gruppen von Accounts (einer für Autorinnen / Autoren und einer für Herausgeberinnen / Herausgeber wurde als zu grob angesehen, so dass eine weitere Accountgruppe im System implementiert wurde (etwa für Mitarbeiterinnen / Mitarbeiter der Redaktionen oder Reviewer mit abgestuften Rechten).

Ein interessanter Diskussionspunkt war insbesondere in Bezug auf das intendierte Metadatenschema der Applikation auszumachen: Während das Projektmanagement, Informationswissenschaftler und Metadatenexpertinnen eher die Möglichkeiten eines umfangreicheren Metadatenschemas betonten, drangen die Wirtschaftsforschenden im Gegensatz dazu auf eine Reduktion der Metadatenfelder und eine Verschlankung

[17] Vgl. dazu Feijen, Martin. 2011. *What Researchers Want*. SURFfoundation. https://www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/What_researchers_want.pdf.

des Schemas. Insbesondere sollte die Anzahl der Pflichtfelder auf ein Minimum reduziert werden. Diese Vorschläge wurden im Zuge der Weiterentwicklung der Pilotapplikation hin zu einem produktiven Dienst aufgegriffen und umgesetzt.

4 Handlungsempfehlungen

Die Einbeziehung der intendierten Nutzungsgruppen bei der Entwicklung des Datenarchivs ist ein wichtiger Pfeiler der Softwareentwicklung im Bereich des Forschungsdatenmanagements. Um die Nutzenden nicht unnötig mit einer Vielzahl an Metadatenfeldern zu konfrontieren, wurden nach Rückmeldungen aus der Community darauf geachtet, die Befüllung der Felder, wo möglich, zu (semi-)automatisieren. Derartige Erleichterungen bestanden bei der Metadateneingabe u.a. bei Feldern, die verschiedentlich wiederholt werden (Name, Affiliation, integrierte Normdaten) - hier haben wir einmal gemachte Eingaben übernommen. Zudem setzten wir auf Unterstützung bei der Eingabe von Nutzerdaten in Form einer Suggest-Funktion (ähnlich der Eingabe bei Google), die wir aus einem Subset der Gemeinsamen Normdaten (GND) gezogen und per JavaScript integriert haben. Hiermit konnten wir zwei Fliegen mit einer Klappe schlagen: Einerseits werden Autorinnen und Autoren sowie Institutionen eindeutig identifiziert, andererseits werden Aufwand sowie Fehleranfälligkeit bei der Metadatenerstellung für die Nutzenden deutlich reduziert.

Mit Auslaufen der Projektförderung wurde im Sommer 2016 der Produktivbetrieb unter dem Namen ZBW Journal Data Archive aufgenommen[18]. Der Dienst wird im Herbst 2018 auf seine Relevanz evaluiert. Somit liegt der Fokus einerseits auf der Akquise weiterer Fachzeitschriften, andererseits auf technischen Weiterentwicklungen, um den Dienst noch komfortabler nutzen zu können. Bisherige Rückmeldungen aus den Wirtschaftswissenschaften zeigen, dass es ein großes Interesse an diesem Service gibt. Insbesondere Fachzeitschriften aus dem deutschsprachigen Raum, die von mittelständischen Verlagen publiziert werden, zeigen Interesse an der Applikation, zumal diese Verlage oftmals nicht die Mittel und die technische Expertise haben, eigene Services aufzubauen und zu betreiben.

[18] http://journaldata.zbw.eu/.