

RatSWD Working Paper Series

www.ratswd.de

RatSWD ■

Rat für Sozial- und
Wirtschaftsdaten

224

Replizierbare Forschung in den
Wirtschaftswissenschaften
erhöhen – eine Herausforderung
für wissenschaftliche Infrastruktur-
dienstleister

Sven Vlaeminck, Gert G. Wagner,
Joachim Wagner, Dietmar Harhoff,
Olaf Siegert

Oktober 2013



Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD)

Die *RatSWD Working Papers* Reihe startete Ende 2007. Seit 2009 werden in dieser Publikationsreihe nur noch konzeptionelle und historische Arbeiten, die sich mit der Gestaltung der statistischen Infrastruktur und der Forschungsinfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften beschäftigen, publiziert. Dies sind insbesondere Papiere zur Gestaltung der Amtlichen Statistik, der Ressortforschung und der akademisch getragenen Forschungsinfrastruktur sowie Beiträge, die Arbeit des RatSWD selbst betreffend. Auch Papiere, die sich auf die oben genannten Bereiche außerhalb Deutschlands und auf supranationale Aspekte beziehen, sind besonders willkommen.

RatSWD Working Papers sind nicht-exklusiv, d. h. einer Veröffentlichung an anderen Orten steht nichts im Wege. Alle Arbeiten können und sollen auch in fachlich, institutionell und örtlich spezialisierten Reihen erscheinen. Die *RatSWD Working Papers* können nicht über den Buchhandel, sondern nur online über den RatSWD bezogen werden.

Um nicht deutsch sprechenden Nutzer/innen die Arbeit mit der neuen Reihe zu erleichtern, sind auf den englischen Internetseiten der *RatSWD Working Papers* nur die englischsprachigen Papers zu finden, auf den deutschen Seiten werden alle Nummern der Reihe chronologisch geordnet aufgelistet.

Einige ursprünglich in der *RatSWD Working Papers* Reihe erschienenen empirischen Forschungsarbeiten sind ab 2009 in der RatSWD Research Notes Reihe zu finden.

Die Inhalte der *RatSWD Working Papers* stellen ausdrücklich die Meinung der jeweiligen Autor/innen dar und nicht die des RatSWD.

Herausgeber der RatSWD Working Paper Series:

Vorsitzender des RatSWD (2007/2008 Heike Solga; seit 2009 Gert G. Wagner)

Geschäftsführer des RatSWD (Denis Huschka)

Replizierbare Forschung in den Wirtschaftswissenschaften erhöhen – eine Herausforderung für wissenschaftliche Infrastrukturdienstleister

Sven Vlaeminck¹, Gert G. Wagner², Joachim Wagner³,
Dietmar Harhoff⁴ und Olaf Siegert¹

Die hier vorgestellten Projektergebnisse wurden im Rahmen des Projekts EDaWaX (European Data Watch Extended, www.edawax.de) erarbeitet. EDaWaX wird von der Deutschen Forschungsgemeinschaft (www.dfg.de) gefördert. Institutionell beteiligt sind an diesem Projekt der Rat für Sozial- und Wirtschaftsdaten (RatSWD), das Institut Inno-tec der LMU München in Kooperation mit dem Max Planck Institute for Intellectual Property and Competition Law (IMPRS-CI) sowie die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW). Neben den Autoren sind folgende Personen am EDaWaX Projekt beteiligt: Prof. Dr. Klaus Tochtermann (ZBW), Dr. Brigitte Preissl (ZBW), Patrick Andreoli-Versbach (IMPRS-CI), Dr. Frank Mueller-Langer (IMPRS-CI), Ralf Toepfer (ZBW) und Dr. Hendrik Bunke (ZBW).

Eine leicht modifizierte Version dieses Working Papers wurde in der Zeitschrift LIBREAS.Library Ideas veröffentlicht. Siehe: Sven Vlaeminck, Gert G. Wagner, Joachim Wagner, Dietmar Harhoff, Olaf Siegert, "Replizierbare Forschung in den Wirtschaftswissenschaften erhöhen". LIBREAS. Library Ideas, 23 (2013). <http://libreas.eu/ausgabe23/06vlaeminck/>

¹ ZBW – Deutsche Zentralbibliothek für Wirtschaftswissenschaften | Leibniz Informationszentrum
Wirtschaft

² RatSWD – Rat für Sozial- und Wirtschaftsdaten und DIW Berlin – Deutsches Institut für
Wirtschaftsforschung, Berlin

³ RatSWD – Rat für Sozial- und Wirtschaftsdaten und Leuphana Universität, Lüneburg

⁴ INNO-tec - Institut für Innovationsforschung, Technologiemanagement und Entrepreneurship (LMU
München) und Munich Center for Innovation and Entrepreneurship Research (MCIER)

Wissenschaftliche Untersuchungen auf Basis von Forschungsdaten nehmen in vielen Wissenschaftsdisziplinen zu – „Data Backed Science“ könnte in vielen Fachbereichen zu einem neuen Paradigma der Forschung werden.

Wenngleich die Wirtschaftswissenschaften sicher nicht mit datenintensiven Fächern wie der Hochenergiephysik oder aus dem Bereich der Erdbeobachtung (vgl. PARSE.Insight, 2010) verglichen werden können, finden sich dennoch zunehmend mehr Publikationen in Fachzeitschriften, in denen Forschende selbst erstellte oder extern verfügbare Datensätze nach eigenen Fragestellungen ausgewertet haben.

Die für diese Forschungsfragestellungen genutzten Daten stammen aus unterschiedlichsten Quellen. Im Unterschied zu stärker empirisch ausgerichteten Wissenschaftsdisziplinen, wie etwa der Psychologie, werden jedoch in der Regel in den Wirtschaftswissenschaften seltener eigene Forschungsdatensätze erstellt. Stattdessen greifen Wirtschaftsforscher/innen oftmals auf Daten der amtlichen Statistik zurück oder auf Studien, die durch spezialisierte Forschungseinheiten erhoben werden (z. B. der ALLBUS der GESIS⁵ oder das SOEP⁶ am DIW Berlin). Häufig werden auch relevante Datensätze von Firmen, wie etwa Bloomberg oder Thomson Reuters eingekauft. Eine Ausnahme bilden hier Untersuchungen im Bereich der experimentellen Wirtschaftsforschung, bei denen Wirtschaftswissenschaftler/innen die Ergebnisse eigener Versuche in selbst erstellten Datensätzen dokumentieren.

Doch obwohl zunehmend mehr Publikationen auf Basis der Auswertung von Daten erscheinen, gibt es in den Wirtschaftswissenschaften bislang kaum effektive Möglichkeiten, die in Fachzeitschriften publizierten Forschungsergebnisse im Kontext der zugehörigen Artikel zu replizieren, zu prüfen oder für eine Nachnutzung und zur Unterstützung des wissenschaftlichen Diskurses („Data Sharing“) bereit zu stellen. Selbst die

5 Der ALLBUS ist die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften und wird jährlich von der GESIS - Leibniz Institut für Sozialwissenschaften (Köln und Mannheim) durchgeführt. Siehe <http://www.gesis.org/de/allbus>.

6 Das Sozio-oekonomische Panel (SOEP) ist eine repräsentative Wiederholungsbefragung, die seit mehr als 25 Jahren durchgeführt wird. Im Auftrag des DIW Berlin werden jedes Jahr in Deutschland über 20.000 Personen aus rund 11.000 Haushalten befragt. Siehe http://www.diw.de/de/diw_02.c.221178.de/ueber_uns.html.

Forschungsdaten, die auf allgemein zugänglichen Datensätzen (z. B. ALLBUS oder SOEP) beruhen, werden in der Regel nicht in ihrer spezifischen Auswahl und Bereinigung archiviert. Damit sind Replikationen zwar nicht ausgeschlossen (da die Original-Daten ja frei zugänglich sind), aber eine Replikation fällt bei anspruchsvollen Analysen, die auf spezifischer Auswahl und Bereinigungen beruhen, schwer. Dies stellt sowohl die Fachdisziplin als auch wissenschaftliche Infrastrukturdienstleister wie Bibliotheken und Datenzentren vor zahlreiche Herausforderungen, die im Folgenden betrachtet werden.

Werden Wirtschaftsforschende von morgen „auf den Schultern von Riesen“ stehen?

Isaac Newtons Zitat *“if I have seen further it is by standing on the shoulders of giants”* beschreibt ein zentrales Prinzip von Fortschritt in der Wissenschaft, nach dem heutige Forschung und Erkenntnisse erst auf Basis des Wissens und der Erkenntnisse früherer wissenschaftlicher Arbeiten zustande kommen. Eine wichtige Methode frühere Erkenntnisse urbar zu machen, besteht darin Forschungsergebnisse replizieren zu können. Replizierbarkeit ist zudem das zentrale Prinzip empirisch-wissenschaftlichen Arbeitens. Bereits Studierende können durch Replizieren von Forschungsergebnissen ihr Wissen über wissenschaftliche Methoden erweitern.

Der US-Ökonom B.D. McCullough fasst die Bedeutung replizierbarer Forschung treffend zusammen: *“[...]replication ensures that the method used to produce the results is known. Whether the results are correct or not is another matter, but unless everyone knows how the results were produced, their correctness cannot be assessed. Replicable research is subject to the scientific principle of verification; non-replicable research cannot be verified. Second, and more importantly, replicable research speeds scientific progress. [...] Third, researchers will have an incentive to avoid sloppiness. [...] Fourth, the incidence of fraud will decrease.”* (McCullough, 2009, 118)

Doch wie ist es heute um diesen wichtigen Grundsatz in der Wirtschaftsforschung bestellt? Werden empirisch Forschende von morgen ohne großen Aufwand auch „auf den Schultern von Riesen“ stehen können?

Bisherige Untersuchungen, in denen systematisch Ergebnisse publizierter Artikel aus wirtschaftswissenschaftlichen Fachzeitschriften repliziert wurden, erzielten bedenklich geringe Replikationsquoten:

- Dewald, Thursby und Anderson versuchten 1986 die Ergebnisse von 54 Artikeln des *Journal of Money, Credit and Banking* replizieren. Sie waren in nur zwei Fällen erfolgreich (3,7%). (Dewald, Thursby und Anderson, 1986)
- 20 Jahre später versuchten die US-Wirtschaftsforscher/innen McCullough, McGeary und Harrison 62 Ergebnisse von Artikeln derselben Zeitschrift zu replizieren. In 14 Fällen konnten die Ergebnisse erfolgreich repliziert werden (22,6%). (McCullough, McGeary und Harrison, 2006)
- 2008 versuchten die drei Wissenschaftler/innen die Ergebnisse von 117 Artikeln der *Federal Reserve Bank of St. Louis Review* zu replizieren, waren aber nur 9 Mal (7,7%) erfolgreich. (McCullough, McGeary und Harrison, 2008)

Derart geringe Replikationsquoten sind nicht nur von einem wissenschaftlichen Standpunkt aus nicht zufriedenstellend, da zentrale Gütekriterien empirischer Forschung nicht eingehalten werden. Darüber hinaus können wirtschaftswissenschaftliche Forschungsergebnisse aber auch ökonomische und gesellschaftliche Auswirkungen haben, da politische und wirtschaftliche Maßnahmen häufig mit den Ergebnissen ökonomischer Forschungen begründet werden.

Welche weitreichenden Implikationen nicht replizierte Wirtschaftsforschung haben kann, verdeutlicht der kürzlich bekannt gewordene Fall der US-Topökonomin Kenneth Rogoff und Carmen Reinhart, deren (fehlerhafte) Forschungsergebnisse sowohl von Oli Rehm, EU- Kommissar für Wirtschaft und Währung, als auch vom US-Vizepräsidentenskandidat Paul Ryan als Begründung für wirtschaftliche Austeritätspolitik verwendet wurden.⁷

⁷ Der Fall Rogoff/Reinhart wurde in vielen Medien und wissenschaftlichen Blogs diskutiert. Einen Eindruck davon vermittelt z.B. <http://www.zeit.de/2013/27/staatsverschuldung-rechenfehler-thomas-herndon> oder <http://www.nextnewdeal.net/rortybomb/researchers-finally-replicated-reinhart-rogoff-and-there-are-serious-problems>.

Warum ist wirtschaftswissenschaftliche Forschung oft nicht replizierbar?

Die Gründe für die fehlende Replizierbarkeit wirtschaftswissenschaftlicher Forschung lassen sich auf verschiedenen Ebenen verorten:

- Zum ersten sind fehlende Anreizmechanismen für Forschende zu nennen, die von ihnen bearbeiteten Forschungsdaten mit anderen Wissenschaftler/innen zu teilen (data sharing). Das Wissenschaftssystem belohnt die Arbeit des Data Sharing, die oftmals sehr zeitintensiv ist, nicht. Dies steht im krassen Gegensatz zur Wertigkeit von klassischen Publikationen, die, wie Anderson et al. feststellen, nur “[...] the advertising for the data and code that produced the published results” sind (Anderson et al, 2008, 101).
- Forschende fürchten zudem, dass Data Sharing für sie mit Nachteilen verbunden sein könnte: Denn die bereits unter großem Arbeitsaufwand aufbereiteten Daten werden auch einer wissenschaftlichen Community zur Verfügung gestellt, die sie nutzen kann, obwohl von deren Seite kein Beitrag zur Datenaufbereitung geleistet wurde. Dies kann zu einer Schieflage in der Forschung führen, da Wissenschaftler/innen, die neue Daten generieren und für andere aufbereiten, dafür keine Reputation erlangen und es somit schwerer haben, Universitätskarrieren zu verfolgen. Zudem befürchten viele Forscher/innen einen Missbrauch der Daten durch Dritte, zum Beispiel durch falsche Interpretation oder durch Nutzung der Daten ohne korrekte Zitation der Urheberin bzw. des Urhebers. Schließlich ist drittens die Rechtslage bei der Weitergabe von Datensätzen in vielen Fällen nicht ausreichend geklärt, was ebenfalls zu einer großen Zurückhaltung im Bereich „Data Sharing“ führt. (Siegert, Toepfer und Vlaeminck, 2012, 219)
- Derzeit verfügen nur wenige wirtschaftswissenschaftliche Fachzeitschriften über Richtlinien, die den Umgang mit den verwendeten Forschungsdaten regeln: So genannte Data Availability Policies beispielsweise verpflichten Autor/inn/en empirischer Artikel dazu, die zur Erlangung der Ergebnisse verwendeten Forschungsdaten sowie die erstellte Syntax der statistischen Auswertung gemeinsam mit

dem publizierten Artikel bereit zu stellen. Solche Richtlinien folgen somit oftmals dem vom Gary King 1995 formulierten „Replication Standard“ (King, 1995). Zudem sind verpflichtende Policies ein wirksamer Weg die Verfügbarkeit von Forschungsdaten zu erhöhen, solange keine anderen wirksamen Anreize für das „Teilen“ von verwendeten Forschungsdaten bestehen.

- Es werden nur selten geeignete Infrastrukturkomponenten für das Management von publikationsbezogenen Forschungsdaten eingesetzt, so dass oftmals nicht einmal uniforme Zitationen der hinterlegten Daten möglich sind. Professionelle Forschungsdatenzentren bieten bislang nur selten Services für solche Daten an. Andere Lösungen, wie beispielsweise Dataverse, ein leistungsfähiges Instrument für das Management und die Dokumentation von publikationsbezogenen Forschungsdaten, werden nur von wenigen Fachzeitschriften eingesetzt.

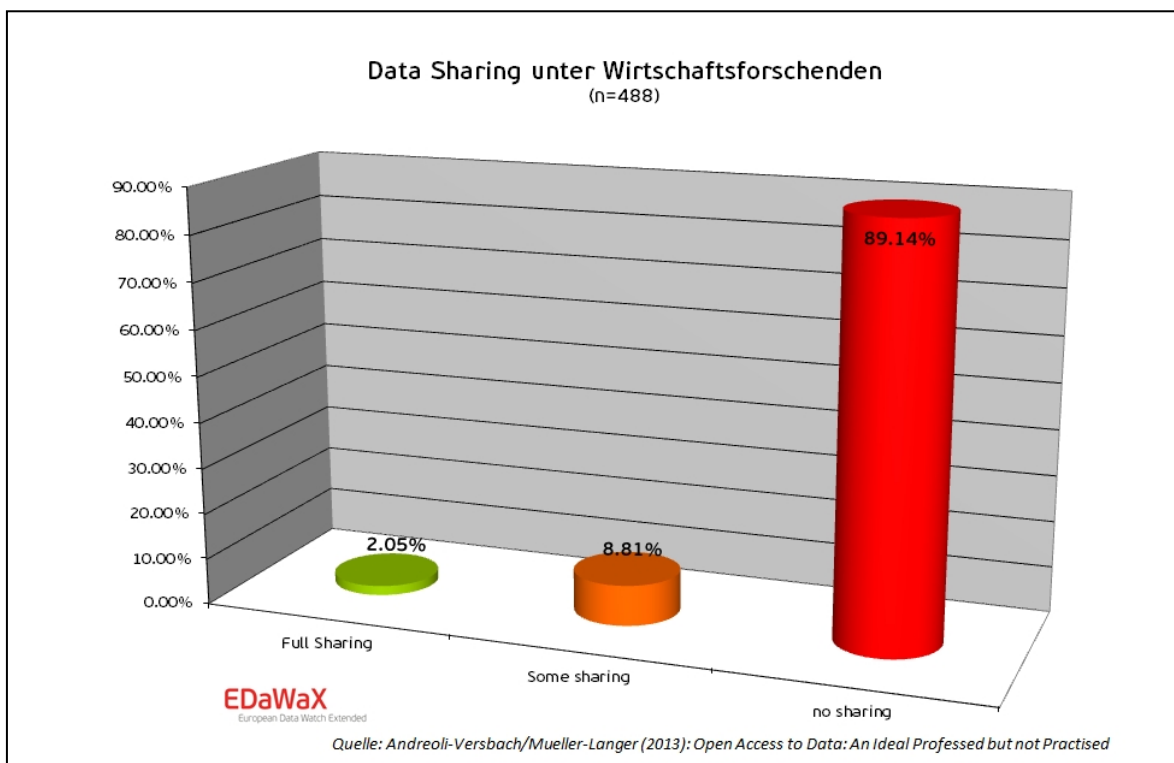
Das DFG-geförderte Forschungsprojekt EDaWaX (European Data Watch Extended – www.edawax.de) untersucht seit Herbst 2011 unter anderem die oben genannten Aspekte. In den folgenden Absätzen werden die zentralen Ergebnisse der bisherigen Analysearbeitspakete von EDaWaX vorgestellt. Im Anschluss daran werden die daraus resultierenden Anforderungen an eine Software für das Management von publikationsbezogenen Forschungsdaten skizziert. Die Entwicklung einer solchen Pilotapplikation ist ein Kernziel des Projekts.

I Data Sharing und Anreize für das Teilen von Forschungsdaten

Während in einigen Wissenschaftsdisziplinen kollaboratives Arbeiten mit Forschungsdaten bereits die Regel ist, zeigen Ergebnisse des EDaWaX-Projekts, dass die Wirtschaftswissenschaften hiervon noch weit entfernt sind. Im Rahmen von EDaWaX analysierten Andreoli-Versbach und Mueller-Langer in einer Studie das Data-Sharing-Verhalten von fast 500 empirisch arbeitenden Wirtschaftsforscher/innen (Andreoli-Versbach und Mueller-Langer, 2013). Die beiden Münchener Wirtschaftswissenschaftler untersuchten

zu diesem Zweck die Webseiten der Wissenschaftler/innen auf die Verfügbarkeit von in Artikeln verwendeten Forschungsdaten. Die erzielten Ergebnisse zeichnen ein deutliches Bild: Nur rund 2 % der Untersuchten teilten die von Ihnen verwendeten Forschungsdaten häufig, weitere 9% tun dies zumindest gelegentlich. Im Umkehrschluss bedeutet dies, dass mehr als 89% der Untersuchten keine Forschungsdaten mit anderen „teilen“. Dieses Ergebnis macht deutlich, dass in den Wirtschaftswissenschaften noch einige Überzeugungsarbeit geleistet werden muss, um die Verfügbarkeit von Forschungsdaten merklich zu steigern.

Das Thema „Anreize für Data Sharing“ ist somit von hoher empirisch-praktischer Relevanz. Daher widmet sich das EDaWaX Projekt auch diesen Fragestellungen und entwickelte verschiedene Vorschläge, die das Teilen von Daten fördern können und unterschiedliche Gruppen innerhalb des Wissenschaftsbetriebs adressieren:



- Forschungsförderer, Wissenschaftsorganisationen und Universitäten können Data Sharing dadurch fördern, dass sie die Aufbereitung und das Teilen von Forschungsdaten mit anderen Wissenschaftlern als wissenschaftliche Leistung honorieren. Eine solche Anerkennung

könnten beispielsweise bei karriererelevanten Entscheidungen (Stellenbesetzungen, Professuren), bei der Vergabe von Stipendien sowie Projekt- und Haushaltsmitteln, oder auch als ein Kriterium im Rahmen von Evaluierungen von Forschungseinrichtungen eingesetzt werden.

- Infrastrukturanbieter können dazu beitragen, indem sie leicht nutzbare und barrierefreie Zugänge zu frei verfügbaren Forschungsdaten anbieten sowie eine Zitierbarkeit gespeicherter Forschungsdaten sicherstellen. Auch die Entwicklung von unterstützenden Services zur Metadatengenerierung sowie zur Auffindbarkeit in Fachportalen und Katalogen ist eine Maßnahme, um die Sichtbarkeit dieser wichtigen Ressourcen für die Forschung zu erhöhen.
- Lehrende können im Rahmen der Ausbildung von Studierenden und Graduierten die Thematik aufgreifen, auf die Vorteile und wissenschaftliche Notwendigkeit des Data Sharings hinweisen und geeignete Infrastrukturen dafür vorstellen.
- Letztlich können Wirtschaftsforschende selbst dazu beitragen, die Akzeptanz von Data Sharing zu erhöhen, indem sie im Rahmen von „Good Practises“ aufzeigen, welche positiven Erfahrungen mit dem Teilen von Daten erzielt werden konnten.

II Forschungsdatenmanagement in wirtschaftswissenschaftlichen Fachzeitschriften

In einer zweiten Studie befasste sich das EDaWaX-Projekt mit der Verfügbarkeit und Qualität von Forschungsdatenrichtlinien in wirtschaftswissenschaftlichen Fachzeitschriften. Dazu wurde ein Sample von 141 vornehmlich volkswirtschaftlich ausgerichteten Fachzeitschriften (darunter vielen Top-Journals) dahingehend evaluiert, wie viele Fachzeitschriften über entsprechende Policies verfügen, wie diese im Detail ausgestaltet sind und ob diese die Replizierbarkeit von Forschung fördern.⁸ Unsere Untersuchung erzielte folgende Ergebnisse:



- 40 (28,4%) von 141 untersuchten Fachzeitschriften verfügten über eine irgendwie geartete Richtlinie in Bezug auf verwendete Forschungsdaten.

⁸ Eine ausführliche Auswertung der durchgeführten Untersuchung findet sich in Vlaeminck, S. ; Siegert, O. (2012). "Welche Rolle spielen Forschungsdaten eigentlich für Fachzeitschriften? Eine Analyse mit Fokus auf die Wirtschaftswissenschaften." RatSWD Working Papers, Nr. 210, Berlin. Verfügbar unter: http://www.ratswd.de/download/RatSWD_WP_2012/RatSWD_WP_210.pdf. Siehe zudem: Vlaeminck, S. (2013). "Data Management in Scholarly Journals and possible Roles for Libraries – Some Insights from EDaWaX." LIBER Quarterly, 23 (1). Verfügbar unter: <http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-114595>.

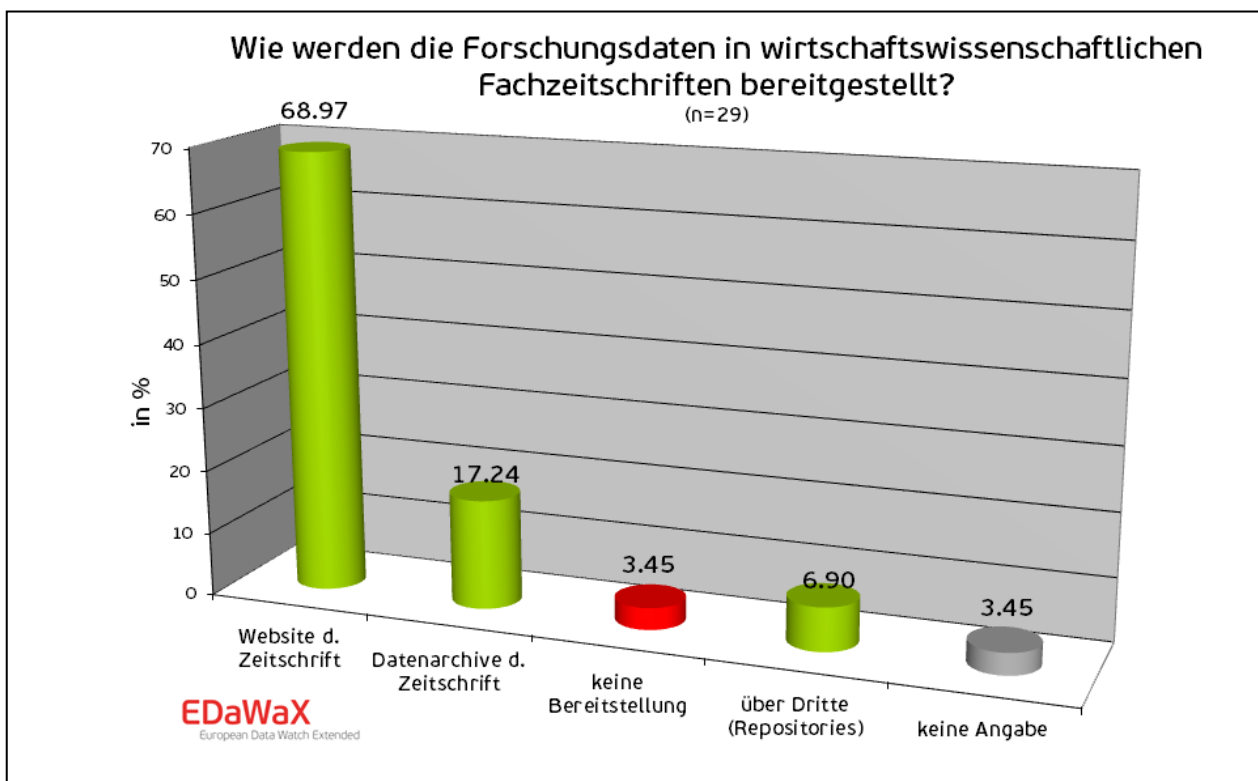
- 29 von diesen 40 Zeitschriften (rund 21% des Gesamtsamples) nutzten dabei eine so genannte Data Availability Policy, die darauf abzielt, die genutzten Forschungsdaten (wo rechtlich möglich) mitsamt des Artikels bereit zu stellen. 11 Fachzeitschriften nutzen so genannte „Replication Policies“, die Autor/inn/en dazu verpflichten im Fall aufkommender Anfragen, genutzte Forschungsdaten und (teilweise auch) den statistischen Berechnungscode bereitzustellen. Empirische Untersuchungen zeigten jedoch deutlich, dass solche Richtlinien von Autor/inn/en oftmals ignoriert werden und somit als sehr schwache Form solcher Richtlinien gelten müssen (Dewald, Thursby und Anderson, 1986).
- 24 von 29 Zeitschriften mit einer Data Availability Policy waren für Autor/inn/en verpflichtend. Dies ist von daher bedeutend, da aufgrund der schwach ausgeprägten Anreize zur Bereitstellung von Forschungsdaten vor allem verpflichtende Richtlinien befolgt werden.
- 12 der untersuchten Policies (knapp 10% des Gesamtsamples) fordern zudem von ihren Autor/inn/en die Übermittlung der verwendeten Syntax, mit deren Hilfe die Berechnungsschritte der durchgeführten statistischen Auswertung nachvollzogen werden können. Dies ist unter Replikationsaspekten von besonderer Wichtigkeit.

Das EDaWaX-Projekt untersuchte neben diesen Richtlinien grob stichprobenartig auch die Datenarchive der 29 Fachzeitschriften mit Data Availability Policy. Ziel war zu prüfen, wie stark die Einhaltung der jeweiligen Richtlinien auch tatsächlich eingefordert wird.

Für die untersuchten Ausgaben 1/2010 und 1/2011 konnte festgestellt werden, dass bei nur 4 Fachzeitschriften mehr als jeder zweite Artikel im Archiv mit Forschungsdaten versehen war, was nur rund 3% des initialen Samples entspricht. Dieser Befund zeigt, dass auch bei Fachzeitschriften, die ihre Aufgabe unter anderem in der wissenschaftlichen Qualitätssicherung sehen, die Verfügbarkeit von Forschungsdaten und deren Einbindung in Reviewprozesse noch eine untergeordnete Rolle spielen.

III Infrastruktur für publikationsbezogene Forschungsdaten in den Wirtschaftswissenschaften

Im Zuge der Untersuchung der Forschungsdatenrichtlinien der Fachzeitschriften, analysierte unser Projekt zudem die Infrastruktur, mit der publikationsbezogene Forschungsdaten bei Fachzeitschriften mit Data Availability Policy bereitgestellt werden. Dabei konnten wir feststellen, dass fast 70% der Fachzeitschriften Forschungsdaten über die Verlags- oder Herausgeberwebseite anbieten, wo diese Daten vorzugsweise als zip-File heruntergeladen werden können und dem Artikel angegliedert sind. Diese Praxis bietet jedoch kaum Funktionalitäten bzw. Mehrwerte und ist aus verschiedenen Gründen nicht empfehlenswert:



- Die publikationsbezogenen Forschungsdaten sind häufig nicht mit eigenen persistenten Identifiern ausgestattet. So hinterlegte Forschungsdaten können somit nicht standardisiert zitiert werden und Autor/inn/en folglich nur schwer für die Datenaufbereitung und -bereitstellung honoriert werden.

- Im Regelfall existieren keine weiteren Metadaten zu so bereit gestellten Forschungsdaten. Dies bedeutet im Umkehrschluss, dass diese Daten nicht als einzelne Entität nachgewiesen und aufgefunden werden können, sondern nur und ausschließlich über den jeweiligen Artikel. Auch können solche Forschungsdaten daher nicht mit anderen Publikationen, Autor/inn/en oder deren Einrichtungen verlinkt werden.

Weitere 17 % der untersuchten Fachzeitschriften nutzen ein spezielles Datenarchiv. Hierbei handelt es sich oftmals um eine spezielle Webseite, in denen alle Artikel aufgelistet sind, zu denen Forschungsdaten vorliegen (siehe z.B. JMCB). Jenseits dieser Übersicht verfügen diese Webseiten jedoch über keine weiteren Funktionalitäten; auch hier werden die Forschungsdaten zumeist nur in Form eines zip-Files bereitgestellt.

Anders ist dies bei Zeitschriften, die Dataverse für ihr Datenarchiv nutzen (z.B. Economics E-Journal / Review of Economics and Statistics). Dataverse kann als gute und sinnvolle Praxis für das Management von publikationsbezogenen Forschungsdaten gelten: Neben der obligatorischen Downloadmöglichkeit finden sich sowohl standardisierte Zitationsvorschläge der Forschungsdaten sowie einige Metadaten zu den Forschungsdaten und der zugehörigen Publikation. Auch sind begrenzte Datenanalysen und Visualisierungen möglich.

Darüber hinaus existieren auch funktionale Eigenentwicklungen, wie das Datenarchiv der Jahrbücher für Nationalökonomie und Statistik, auch wenn diese vom Design und von den bereit gestellten Funktionalitäten nicht mit Dataverse mithalten können.

Zwei interdisziplinäre Zeitschriften aus unserem Sample (Science und Nature) gehen einen anderen Weg: Sie verweisen die Autor/inn/en auf externe Forschungsdatenrepositorien, bei denen die verwendeten Forschungsdaten hinterlegen werden sollen. Zu diesem Zweck findet sich für eine viele Wissenschaftsdisziplinen eine Liste mit geeigneten Repositorien. Für den Bereich Social Sciences, zu dem auch die Wirtschaftswissenschaften zählen, fehlen jedoch entsprechende Angaben.

Management von publikationsbezogenen Forschungsdaten – eine Aufgabe für Forschungsdatenzentren?

Da nur wenige Fachzeitschriften über funktionale Forschungsdatenarchive verfügen, evaluierte unser Projekt, ob das Management publikationsbezogener Forschungsdaten möglicherweise auch ein Arbeitsfeld für Forschungsdatenzentren (FDZ), Bibliotheksverbände und einzelne Bibliotheken oder für Archive darstellt. Speziell die FDZ verfügen im Umgang mit Forschungsdaten über teilweise jahrzehntelange Erfahrungen, weshalb diese Institutionen auch im Bereich des Managements von publikationsbezogenen Forschungsdaten eine wichtige Rolle spielen könnten.

Die dazu durchgeführte Internetrecherche förderte ein explizites Angebote für solche Daten zu Tage: Das am ICPSR (Inter-university Consortium for Political and Social Research, Ann Arbor, USA) angesiedelte Publication Related Archive (PRA) wird jedoch bislang nur von wenigen Autor/inn/en für die Hinterlegung von Forschungsdaten genutzt. Nach dem Ausfüllen einer uniformen Data Deposit Form werden die „Replication Datasets“ per Webupload an das ICPSR übermittelt. Das PRA verfügt jedoch nicht über Schnittstellen (APIs - Application Programming Interfaces), sondern lässt sich ausschließlich über den Webupload nutzen – ein unkomfortabler Weg für Forschende und Redaktionen.

Um mehr über etwaige Angebote und Services von wissenschaftlichen Infrastrukturdienstleistern in Bezug auf das Management von publikationsbezogenen Forschungsdaten zu erfahren, wurde durch das EDaWaX-Projekt eine Onlinebefragung unter potenziell interessanten Institutionen durchgeführt. Insgesamt 46 Organisationen wurden für die Befragung angeschrieben, darunter waren 36 nationale und internationale Forschungsdatenzentren (FDZs) und Datenservicezentren (DSZs), 1 Archiv, 7 Bibliotheksverbände und Bibliotheken sowie drei weitere Institutionen. 22 Organisationen beteiligten sich an der Befragung (rund 48%).⁹

⁹ Eine detailliertere Auswertung dieser Befragung ist auf dem Projektblog verfügbar, siehe <http://www.edawax.de/2013/01/results-of-the-edawax-online-survey-on-hosting-options-for-publication-related-research-data/>.

Die Befragungsergebnisse zeigen, dass vor allem Datenzentren ein relevanter Speicherort für publikationsbezogene Forschungsdaten sein können, da sie verschiedene Voraussetzungen dafür bereits erfüllen.¹⁰ Dennoch gibt es bislang unter den befragten Organisationen keine Services für solche Daten. Zudem erfüllt gegenwärtig keine Institution in Gänze alle Anforderungen für Speicherung und Hosting von publikationsbezogenen Forschungsdaten in den Wirtschaftswissenschaften.

Im Einzelnen ergaben sich folgende Ergebnisse:

- Etwa Dreiviertel aller befragten Einrichtungen akzeptieren grundsätzlich externe Forschungsdaten, inklusive publikationsbezogener Forschungsdaten. Allerdings gibt es z.T. Einschränkungen, etwa aufgrund der fachlichen oder regionalen Zuständigkeit oder hinsichtlich der qualitativen Anforderungen an solche Datensätze.
- Fast ebenso hoch (annähernd 75%) ist die Anzahl der Datenzentren, die den zugehörigen Berechnungscode (Syntax) prinzipiell speichern und hosten. Falls zur Berechnung empirischer Ergebnisse spezielle (selbstgeschriebene) Software verwendet wurde, wird diese allerdings nur von etwa 40% der Befragten für Speicherung und Hosting akzeptiert.
- An eingesetzten Metadatenschemata dominiert klar mit 70% das Schema der Data Documentation Initiative (DDI), einem Standard für die Beschreibung von sozialwissenschaftlichen Daten vor Dublin Core (30% - Mehrfachnennungen möglich). Knapp zwei Drittel zeichnen die Datensätze zudem mit Persistenten Identifikatoren aus und machen sie so eindeutig zitierbar. Etwa Dreiviertel aller Befragten leisten zudem Unterstützung bei der Eingabe der Metadaten durch Forschende.
- Schnittstellen für die externe Suche oder den Upload von Datensätzen werden bislang nicht durch die befragten Einrichtungen angeboten. Kaum verbreitet ist auch der Einsatz von semantischen Technologien wie z.B. RDF.

¹⁰ Die angeschriebenen Bibliotheksverbünde und Archive beteiligten sich nicht an der Befragung. Ob das Thema von diesen Institutionen als nicht relevant angesehen wird, kann folglich nur vermutet werden.

Schlussfolgerungen für die Entwicklung eines publikationsbezogenes Forschungsdatenarchivs für wirtschaftswissenschaftliche Fachzeitschriften

Auf Basis der Ergebnisse der oben dargestellten Analysen konnte im Rahmen von EDaWaX verschiedene Anforderungen für die Entwicklung einer Pilotapplikation entwickelt werden, die das Management von publikationsbezogenen Forschungsdaten in Fachzeitschriften sinnvoll unterstützt. Im Unterschied zu bereits vorhandenen Lösungen wie Dataverse und das ICPSR-PRA setzt die zu entwickelnde Applikation dabei auf eine dezentral einsetzbare und individuell anpassbare Softwarelösung, die zudem Raum für eigene Anpassungen durch Fachzeitschriften lässt.

Folgende Anforderungen wurden daher an die zu entwickelnde Applikation gestellt:

- Die Applikation muss dazu in der Lage sein die verwendeten Forschungsdaten und weitere für die einfache Replikation von Ergebnissen notwendige Daten zu speichern. Hierzu zählen insbesondere Datensätze, Beschreibungen der Daten und des Vorgehens bei Datenanalyse und Replikationen, Syntax sowie weitere Materialien wie verwendete Fragebögen, Codebücher und Versuchsskizzen, aber auch Computerprogramme / Applikationen und deren Sourcecode.¹¹
- Sie soll darüber hinaus umfangreiche Metadaten (inkl. Persistenten Identifiern) zu diesen Objekten zu speichern können und dabei unterschiedlich granulare „Level“ von Metadaten berücksichtigen können. Diese unterschiedlich granularen „Level“ bieten bestimmte Funktionalitäten und erfüllen verschiedene Aufgaben (diese „Level“ haben wir mit den Funktionen „Zitierbarkeit herstellen“, „Auffindbarkeit unterstützen“ und „Verlinkbarkeit sicherstellen“ bezeichnet). Hintergrund dieser Anforderung ist, dass Forschende weitgehend selbst auswählen sollen, wie viel Aufwand sie bei der Erstellung dieser wichtigen Beschreibungen aufwenden möchten und

¹¹ Diese Anforderungen ermöglichen eine Speicherung der genutzten Forschungsdaten sowohl für ökonomische Fragestellungen, als auch für Simulationen sowie für Ergebnisse experimenteller Wirtschaftsforschung.

mit welchen „Funktionalitäten“ sie die Metadaten zu ihren Forschungsdaten ausstatten möchten.¹²

- Die Applikation muss eine separate Speicherung von Metadaten und Datensätzen an physisch unterschiedlichen Orten gewährleisten. Dadurch soll sichergestellt werden, dass Metainformationen auch für proprietäre oder vertrauliche Datensätze vorliegen können, die an anderen Orten vorliegen (z.B. bei Forschungsdatenzentren oder bei kommerziellen Datenprovidern) und auf die aus datenschutzrechtlichen oder anderen Gründen kein direkter Zugriff erfolgen kann. Somit können Metainformation zu solchen Datensätzen in die Applikation eingebunden werden, ohne dass die Daten selbst dafür vorliegen müssen. Hierfür ist das Vorhandensein und die Implementierung leistungsfähiger APIs eine Vorbedingung.

Auf Basis der erarbeiteten Anforderungen wurden prinzipiell für EDaWaX in Frage kommende Softwareapplikationen ermittelt und analysiert. Dabei wurden die Applikationen CKAN, Dataverse, Nesstar und webbasierte Dienste wie figshare untersucht. Für die Auswahl der geeigneten Software wurden u.a. folgende allgemein technische Anforderungen entwickelt:

- Die Applikation muss über leistungsfähige APIs verfügen, die Lese- und Schreibzugriffe ermöglichen.
- Sie soll über ein übersichtliches, intuitiv bedienbares Userinterface verfügen.
- Es soll eine Open Source verfügbare Software sein, die überdies von zahlreichen Fachzeitschriften ohne größeren Implementierungsaufwand nachgenutzt werden kann.
- Die Applikation soll Anbindungsmöglichkeiten an Linked Data (RDF) bieten.

¹² Detailliert werden unsere Überlegungen zu den verwendeten Metadatenschemata und den unterschiedlichen „Granularitäts-Leveln“ auf unserem Projektblog vorgestellt: <http://www.edawax.de/2013/08/metadata-for-publication-related-data-archives-as-much-as-necessary-and-as-little-as-possible/>.

Nach eingehender Analyse und Auswertung wurden die Software CKAN für die weitere Entwicklung ausgewählt. CKAN wird von der britischen Open Knowledge Foundation (OKFN) entwickelt und mit einigem Personaleinsatz betreut.

Die Entwicklung der Pilotapplikation ist inzwischen weit fortgeschritten und wird in den kommenden Monaten technisch weiterentwickelt und dann prototypisch bei der renommierten Fachzeitschrift Schmollers Jahrbuch – Journal of Applied Social Science Studies implementiert.

Ein Implementierungs- und Evaluierungsworkshop, bei der Herausgeber/innen von Fachzeitschriften die Applikation begutachten, wird im Herbst 2013 wertvolles Feedback für weitere Verbesserungen und die zukünftige Entwicklung der Applikation liefern.

Literaturverzeichnis

- Anderson, R. G.; Greene, W. H.; McCullough, B. D.; Vinod, H. D. (2008). The Role of Data/Code Archives in the Future of Economic Research. In: *Journal of Economic Methodology*, 15(1), S. 99-119.
- Andreoli-Versbach P.; Mueller-Langer, F. (2013). Open Access to Data: An Ideal Professed but not Practised, RatSWD Working Paper Series, Nr. 215, Berlin. Verfügbar unter: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2224146.
- Dewald, W. G.; Thursby, J.G.; Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76 (4), S. 587-603.
- King, G. (1995). Replication, replication. In: *PS: Political Science and Politics*, 28, 443–499. Verfügbar unter: <http://gking.harvard.edu/gking/files/replication.pdf>.
- McCullough, B.D. (2009): Open Access Economics Journals and the Market for Reproducible Economic Research. In: *Economic Analysis and Policy* 39 (1), S. 117-126.
- McCullough, B.D.; McGeary, K. A.; Harrison, T. D. (2006). Lessons from the JMCB Archive. In: *Journal of Money, Credit, and Banking*, 38 (4), S. 1093-1107.
- McCullough, B.D.; McGeary, K. A.; Harrison, T. D. (2008). Do economics journal archives promote replicable research? In: *Canadian Journal of Economics*, 41(4), S. 1406-1420.
- PARSE.Insight Projekt (2010): Insight into digital preservation of research output in Europe. Case studies report. Deliverable D3.3. Verfügbar unter: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf.
- Siegert, O.; Toepfer R.; Vlaeminck, S. (2012). Forschungsdatenmanagement in den Wirtschaftswissenschaften – Ausgewählte Dienste und Projekte der Deutschen Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft (ZBW). In: Altenhöner, R. / Oellers, C. (Hrsg.): *Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen*, Berlin, Scivero Verlag.