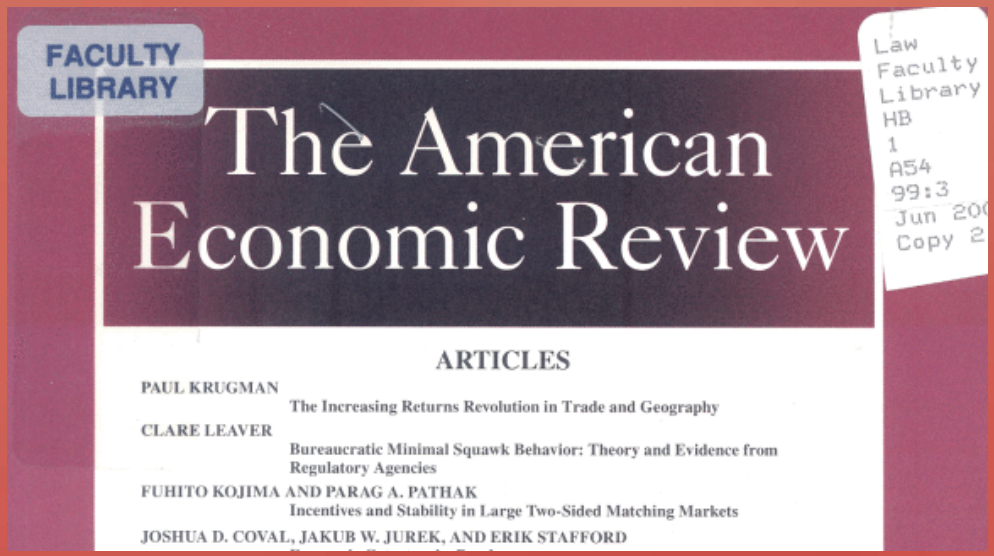
EDaWaX

European Data Watch Extended



„Linking Data and Publications" – Anforderungen und Vorgehensweise am Beispiel des Aufbaus eines publikationsbezogenen Datenarchivs.

Sven Vlaeminck / Dr. Hendrik Bunke | Leibniz Informationszentrum Wirtschaft (ZBW)
7. März 2013 | Ständiger Ausschuss Forschungsdaten-Infrastruktur | Berlin

Funded by
DFG

EDaWaX
European Data Watch Extended

RatSWD.
Rat für Sozial- und WirtschaftsDaten

ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Inhalte der Präsentation

- Projekthintergründe und Motivation
- Publikationsbezogene Forschungsdaten: Der Status Quo
  - Infrastrukturen von Fachzeitschriften zur Bereitstellung von Forschungsdaten
  - Charakteristika publikationsbezogener Forschungsdaten
- Anforderungen an ein publikationsbezogenes Forschungsdatenarchiv …
  - …auf Ebene der IT-Infrastruktur
  - …auf Ebene der Metadaten
- Ergebnisse der Befragung von Forschungsdatenzentren

Projekthintergründe und Motivation

Wirtschaftswissenschaftliche Forschung replizierbar machen.

Das Projekt EDaWaX | www.edawax.de

Funded by DFG

EDaWaX
European Data Watch Extended

RatSWD.

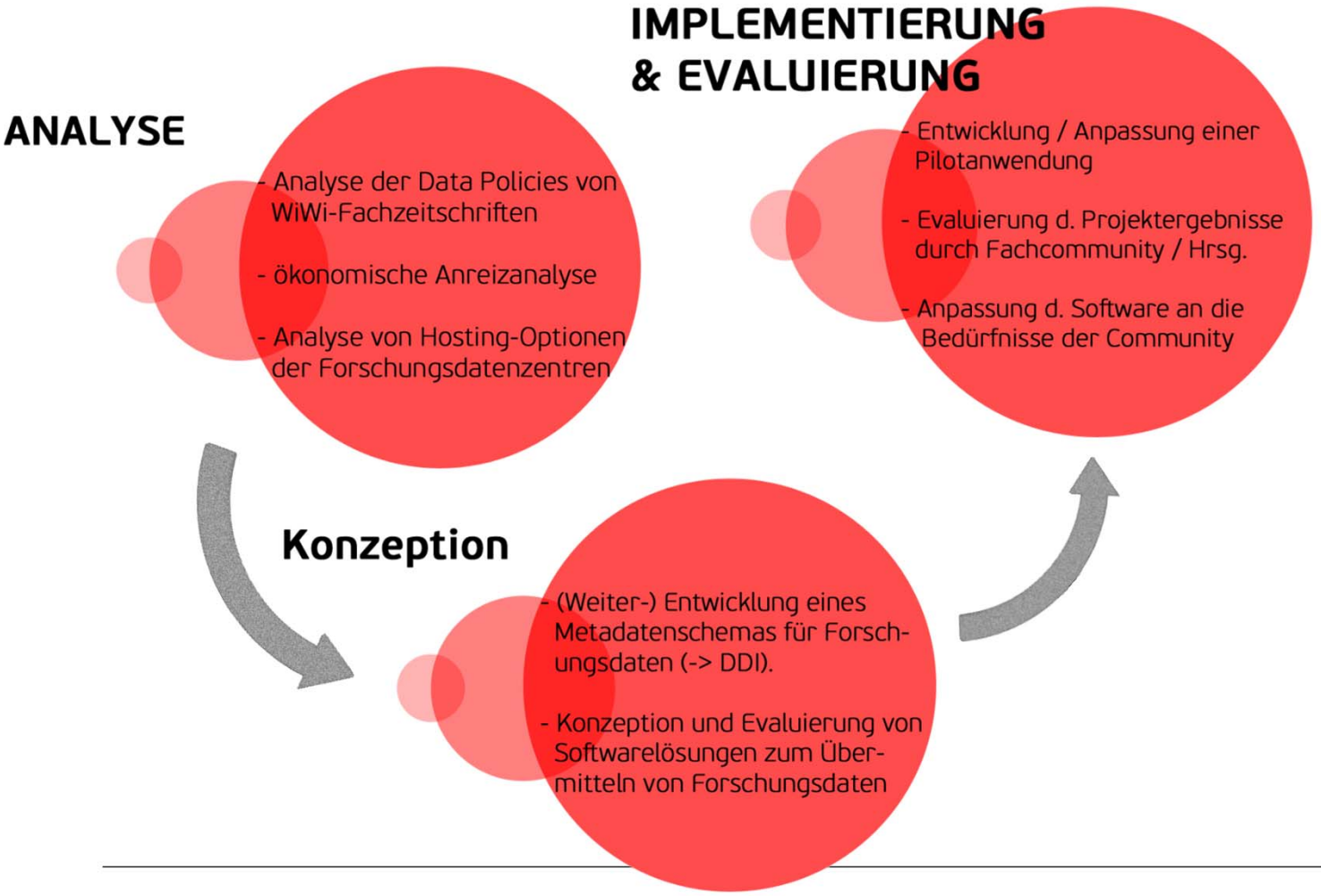
ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Projekthintergründe Wirtschaftswissenschaften

- Forschungsförderer (DFG/EC/Wissenschaftsrat) sehen Notwendigkeit der Verknüpfung von Publikationen & Daten.
- Zunehmend mehr empirische Publikationen (➔ Journals)
- Zumeist keine Möglichkeit deren Ergebnisse zu replizieren.
  - Gründe:
    - Anreize für Forschende „ihre“ Daten zu teilen, fehlen.
    - Fachzeitschriften verfügen selten über Richtlinien, die Forschungsdaten einfordern.
    - Infrastruktur zur Bereitstellung von publikationsbezogenen Forschungsdaten ist kaum vorhanden.

➔ Problematischer Befund, da Replizierbarkeit ein Eckpfeiler der wissenschaftlichen Methode ist!

Projektphasen und Ziele von EDaWaX



Projektziele:

> Implementierung eines Datenarchivs für eine Fachzeitschrift.

> Entwicklung v. Anreizschemata zum "Data Sharing".

> Empfehlungen zur Gestaltung von Data Policies.

> Empfehlungen zur Speicherung/Hosting von publikationsbezogenen Forschungsdaten.

> Metadatenschema zur Beschreibung von Forschungsdaten.

Publikationsbezogene Forschungsdaten: Der Status Quo bei Fachzeitschriften

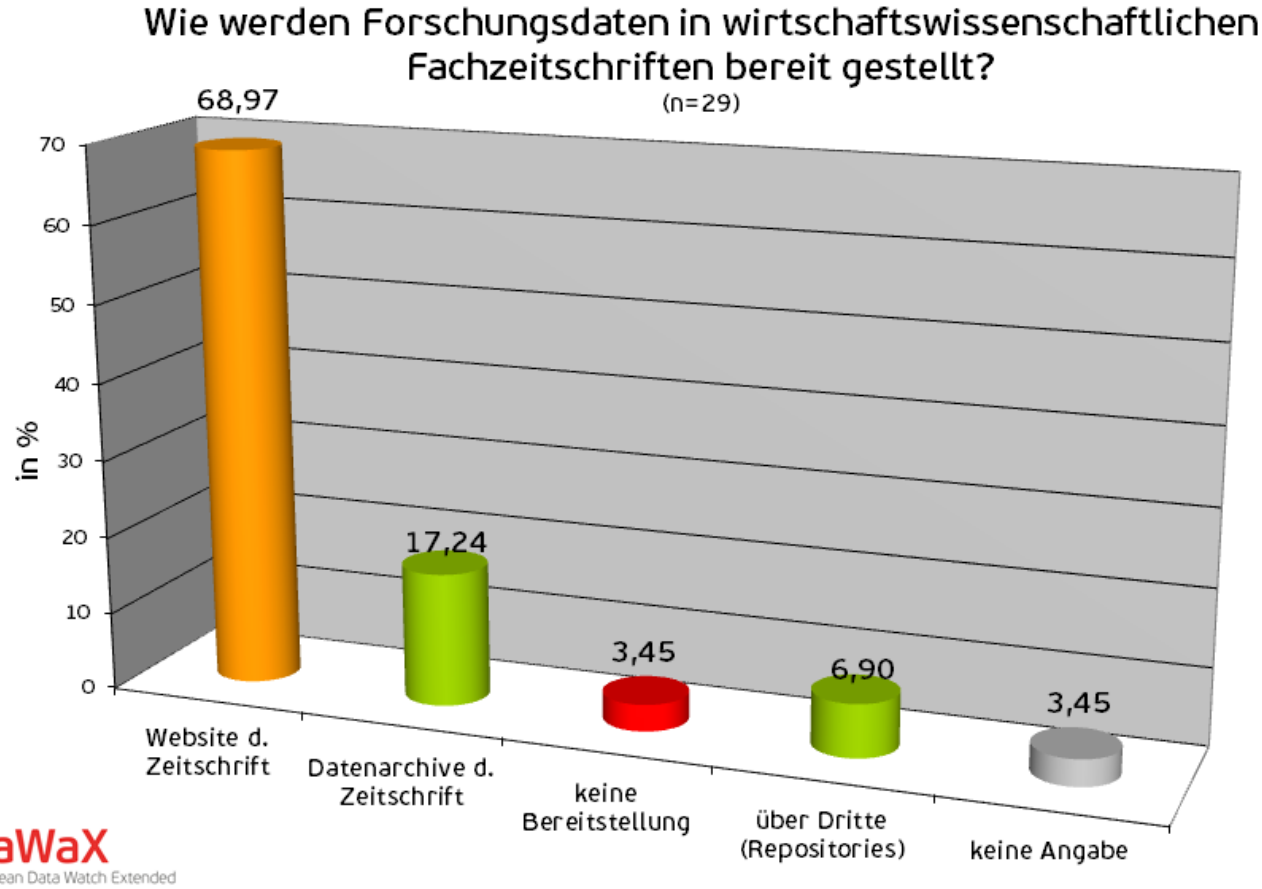
Infrastrukturen von Fachzeitschriften zur Bereitstellung von Forschungsdaten

EDaWaX
European Data Watch Extended

RatSWD.

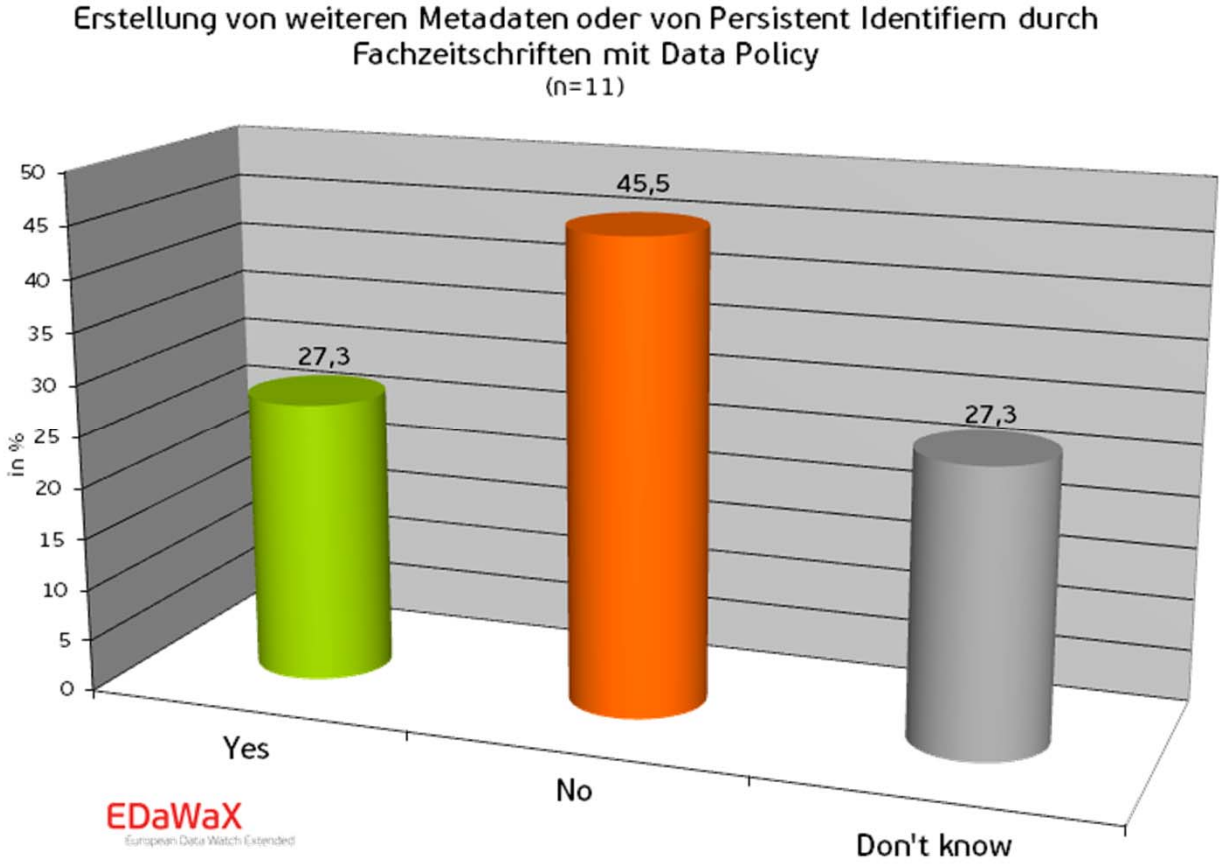
ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Bereitstellung von Forschungsdaten



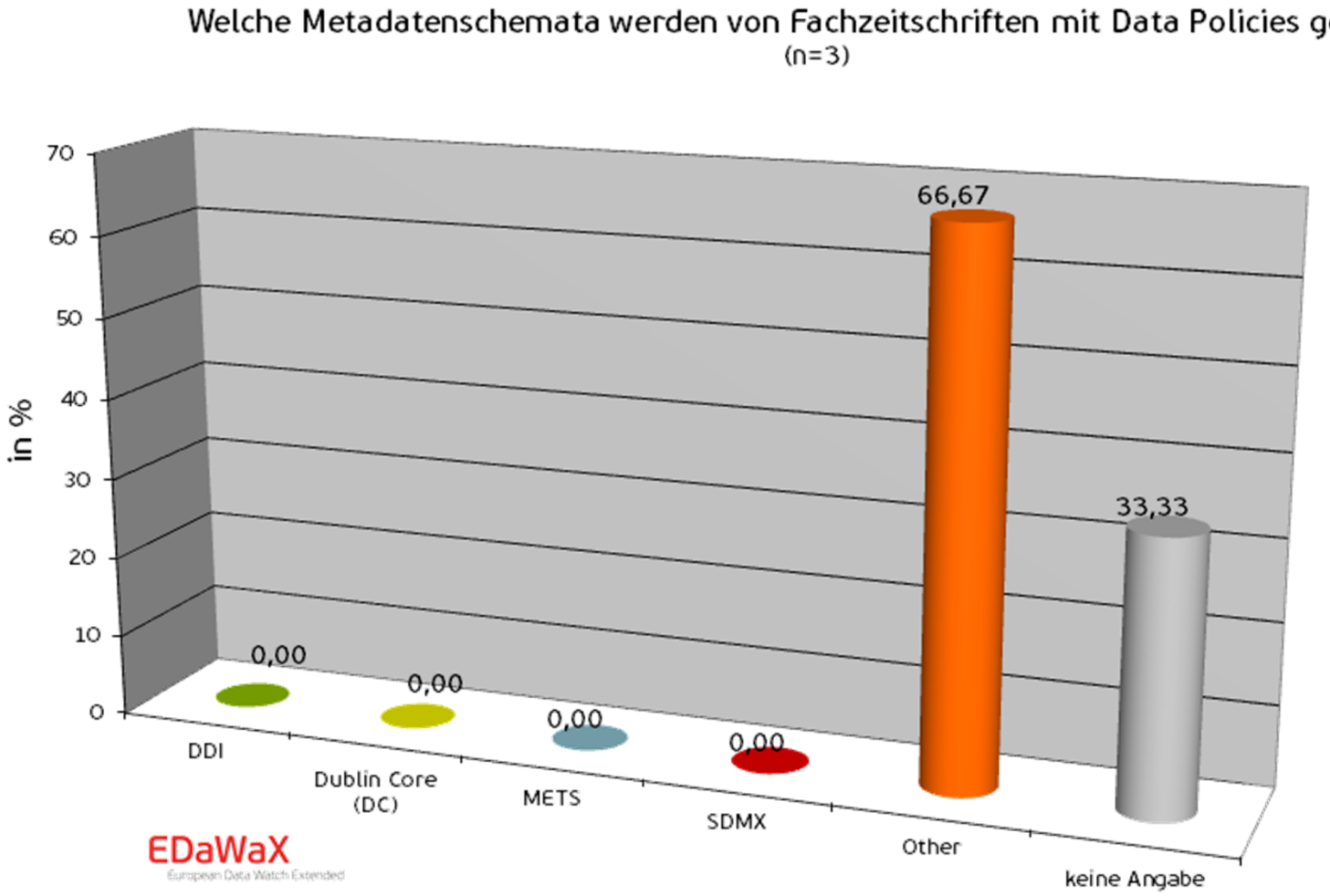
> Forschungsdaten werden zumeist als ‚Supplementary Material' auf den Journalwebseiten bereit gestellt.

Erstellen von Metadaten & PIs durch Journals:



> Weniger als ⅓ der Befragten erstellen weitere Metadaten / PI

Welche Metadatenschemata werden genutzt?



> ...und wenn doch, werden keine Standards verwendet.

Ergebnisse: Infrastruktur v. Fachzeitschriften

- Forschungsdaten zu Artikeln sind meist als zip-Files den Artikeln angehängt.
- Wenige Zeitschriften nutzen eine weitergehende Infrastruktur.

- Forschungsdaten werden kaum durch Redaktionen mit Metadaten angereichert…
- …und wenn doch, werden keine Standards genutzt.
  - Problematisch für Auffinden und Nachnutzen der Daten.
  - Problematisch für Zitieren der Datensätze (-> Anreize!)

Charakteristika publikationsbezogener Forschungsdaten

EDaWaX
European Data Watch Extended

RatSWD.

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Publikationsbezogene Forschungsdaten: Rechtliche Aspekte

- Unter rechtlichen Aspekten sind grob drei Typen von publikationsbezogenen Forschungsdaten zu unterscheiden:
  a. Weitgehend „unproblematische“ Daten (Autor hält Rechte)
  b. Proprietäre Daten (Daten gehören Firma)
  c. Vertrauliche/Personenbezogene Daten (Mikrodaten (➔ §40 BDSG); Firmendaten;...)
- Nicht alle diese Daten können / dürfen bei Zeitschriften-Einreichungen übermittelt werden.

➔ Statt der Daten selbst, können nur Metadaten, weitere Supplements und Zugangswege dargestellt werden

Publikationsbezogene Forschungsdaten…

- …stammen aus verschiedensten Quellen (z.B. FDZs; Autoren; kommerzielle Provider)
- …sind unterschiedlich, was die rechtlichen Anforderungen bzgl. der Weitergabe angeht („offen"/proprietär/vertraulich)
- …sind teilweise nicht mit Metadaten beschrieben
- …werden in fachwissenschaftlichen Portalen / Katalogen oft nicht nachgewiesen
- …sind (wenn überhaupt) Bestandteil der Supplements eines Artikels, keine eigene „Entität"

➔ Umfangreiche Anforderungen an IT-Infrastruktur…

Anforderungen an ein publikations-bezogenes Forschungsdatenarchiv

Eine Zusammenfassung - (Work in Progress)

EDaWaX
European Data Watch Extended

RatSWD.

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

1. Anforderungen an die IT-Infrastruktur

- Modularer Aufbau der IT-Infrastruktur wg. verteilten Datenressourcen notwendig.
- IT-Infrastruktur benötigt folgende Komponenten:
  - Einfach bedienbares User-Frontend (Upload Daten/Metadaten)
  - Instanz (Katalog) zur Speicherung/Anzeige von Metadaten
  - Instanz zur Speicherung, für Hosting und LZA von Forschungsdaten
- Weitgehend barrierefreier Zugang zu Datensätzen

1. Anforderungen an die IT-Infrastruktur (2)

- Schnittstellen (APIs) mit Lese- / Schreibzugriff:
  - Zur Speicherung von Forschungsdaten (extern)
  - Zur Abfrage/Austausch von Katalogdaten (Metadaten) zu verschiedenen Systemen und Portalen, wie
    - Forschungsdatenkatalogen (z.B. CESSDA; DataCite)
    - Bibliothekskataloge (z.B. GVK)

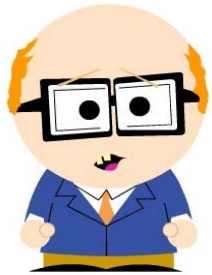
    ➔ Zugriff über standardisierte Schnittstellen (z.B. RDF, OIA-PMH)
- Für das Hosting von Forschungsdaten gibt es folgende Anforderungen:
  - Unterstützung/Implementierung von APIs
  - Referenzierbarkeit der einzelnen Datensätze nötig (z.B. DOI)

2. Anforderungen an Metadaten

- Zur Verlinkung von Daten und Publikationen werden verschiedene Metadaten benötigt:
  - Metadaten zur Publikation (Eindeutige Identifier -> PPN)
  - Metadaten zu den Daten (Eindeutige Identifier -> DOI) da|ra Registrierungsagentur für Sozial- und Wirtschaftsdaten
  - Metadaten zum Autor (Eindeutige Identifier -> GND)
- Metadaten müssen standardisiert sein, damit sie
  - einen Austausch zwischen verschiedenen Portalen und Systemen („mapping“) ermöglichen.

Für EDaWaX:

- Implementierung von drei unterschiedlich granularen „Leveln“ von Metadaten geplant (Wahl der Forschenden)
- Anbindung an Linked Data/Semantic Web geplant.



Fachzeitschrift

Publikation

Physikalische Speicherung von Forschungsdaten

EDaWaX
European Data Watch Extended

Erstellen v. Metadaten

Upload der Forschungsdaten

J/N

User-Frontend

(Metadaten-)Katalog:
Verzeichnis der „Adressen“ und Metadaten von Publikationen und Forschungsdaten

Abfragen / Metadaten-Harvesting

FDZ

GVK Gemeinsamer Verbundkatalog

cessda

ECONBIZ Find Economic Literature

sowiport

DataCite

Externe Ressourcen (Metadatenkataloge)

AGE-SPECIFIC RISE OF INCOME AND CONSUMPTION INEQUALITY [DATASET]

< View Previous Study Listing

hdl:1902.1/20442
Version: 1– Released: Mon Mar 04 03:51:47 EST 2013

If you use these data, please add the following citation to your scholarly references. Why cite?

Data Citation	Zhu, Guozhong, 2013, "Age-specific Rise of Income and Consumption Inequality [Dataset]", http://hdl.handle.net/1902.1/20442 Economics: The Open-Access, Open-Assessment E-Journal [Distributor] V1 [Version]
	Citation Format Print
Publications	Guozhong Zhu (2013). Age-specific Rise of Income and Consumption Inequality. Economics Discussion Papers, No 2013-21, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2013-21/

Data Citation Details

Title	Age-specific Rise of Income and Consumption Inequality [Dataset]
Study Global ID	hdl:1902.1/20442
Authors	Zhu, Guozhong (Guanghua School of Management, Peking University, China)
Production Date	2013
Distributor	Economics: The Open-Access, Open-Assessment E-Journal
Contact	Korinna Werner-Schwarz (IfW), korinna.werner-schwarz@economics-ejournal.org
Distribution Date	2013
Deposit Date	March 04, 2013
Provenance	Economics: The Open-Access, Open-Assessment E-Journal Dataverse

Description and Scope

Description	Based on Panel Study of Income Dynamics (PSID) and Consumer Expenditure Survey (CEX), the author presents evidence that the rise of income/consumption inequality over the past decades is more significant among younger households. This is consistent with the theory that the secular rise of inequality is due to increasing heterogeneity in earning ability. The author further shows that such age-specificity implies significant changes to the previously documented life-cycle profiles of inequality which are the basis of many important economic inferences.

TAKING A DSGE MODEL TO THE DATA MEANINGFULLY [DATASET]

< View Previous Study Listing

hdl:1902.1/13661
Version: 1– Released: Thu Nov 26 05:21:18 EST 2009

Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

Select all files | Download All Selected Files

Total Downloads: 282

2. Data		
irelanddata.xls MS Excel - 175 KB - 108 downloads	Download	data in xls file: quarterly
irelandoutput.txt Plain Text - 22 KB - 90 downloads	Download	output file in txt format showing the computer output
ireland.prg Plain Text - 1 KB - 84 downloads	Download	program file used in CATS in RATS

„Wo lässt sich ein publikationsbezogenes Datenarchiv sinnvoll speichern & hosten?“

(Einige) Ergebnisse unserer Online-Befragung von 22 Forschungsdatenzentren, Bibliotheken und Archiven.

Physikalische Speicherung von Forschungsdaten

EDaWaX
European Data Watch Extended

RatSWD.

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

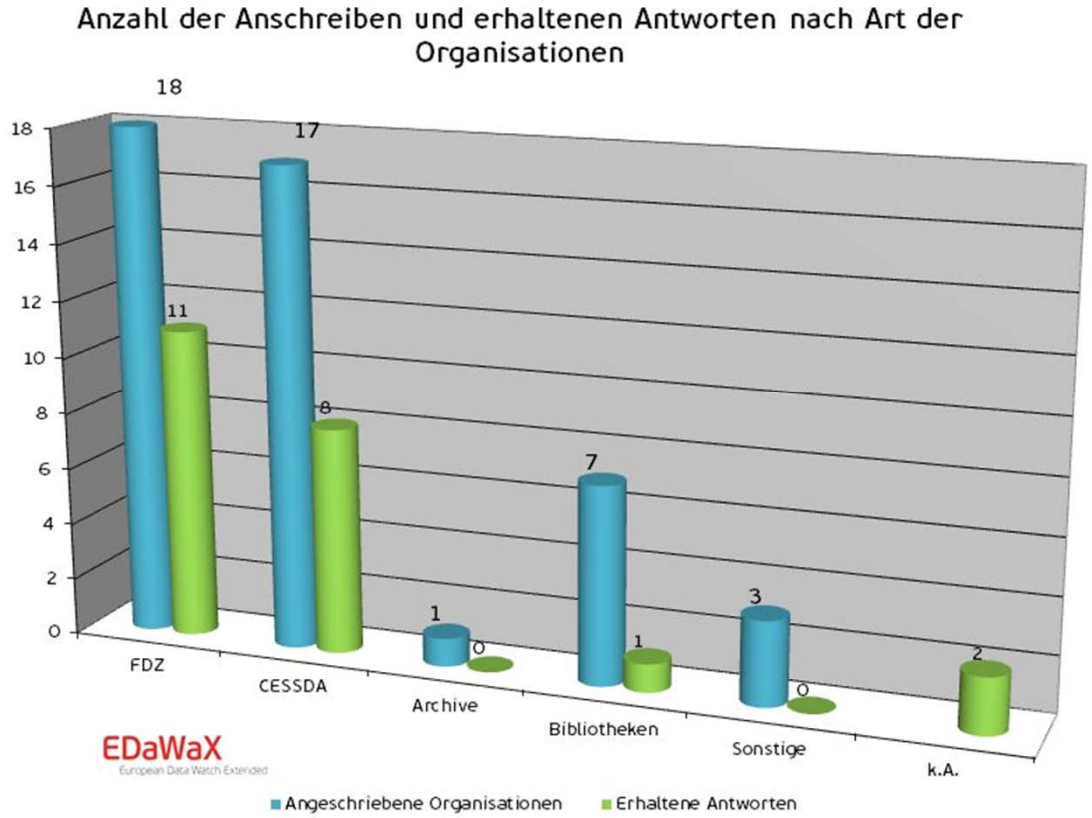
Ziele der Onlinebefragung

- Annahme: Forschungsdatenzentren = Experten im Bereich Datenhaltung, besitzen umfangreiche Kenntnisse von Sowi/Wiwi-Forschungsdaten.
- => Ideale Orte für Speicherung / Hosting eines Forschungsdatenarchivs

- Prüfung ob Datenzentren über Services für publikationsbezogene Forschungsdaten verfügen (u.a. Speicherung / Bereitstellung).

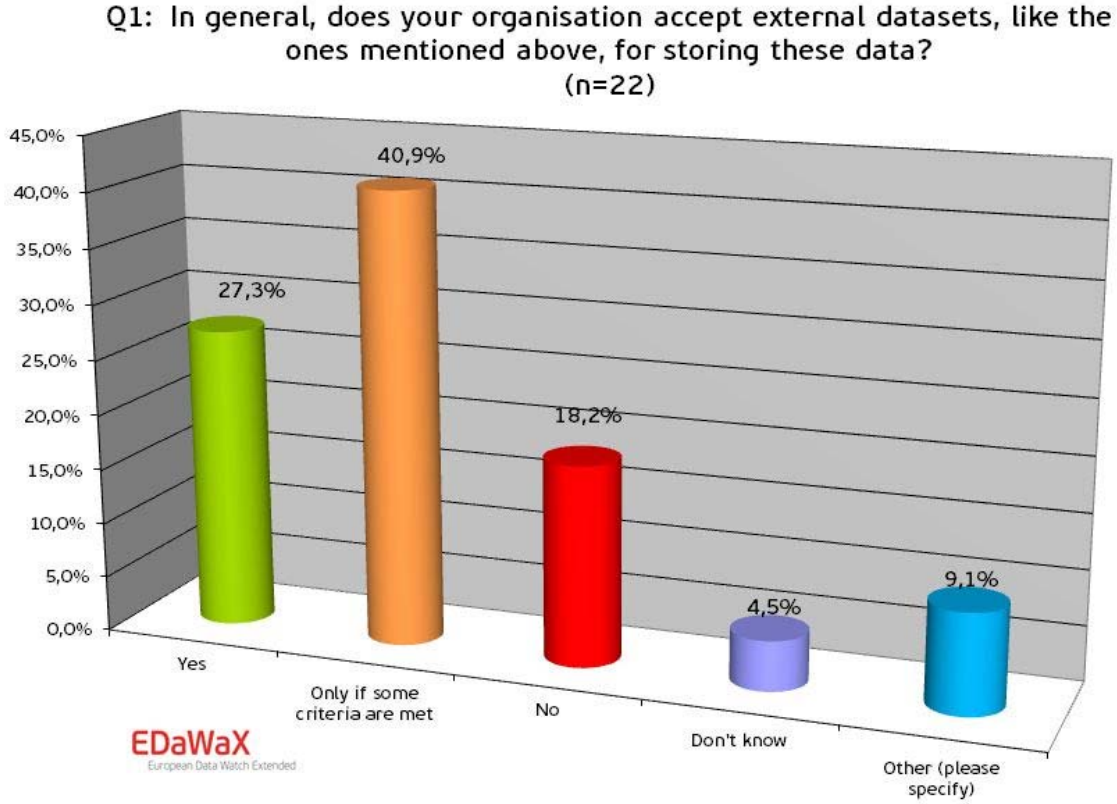
- Evaluierung...
  - der Richtlinien für die Abgabe von externen Datensätzen, und weiterer (für Replikationen wichtige) Daten.
  - der eingesetzten Software (-> Nutzerunterstützung),
  - der genutzten Metadatenstandards,
  - des Zugangs zu solchen Forschungsdaten.

Anschreiben und Rücklauf zur Befragung



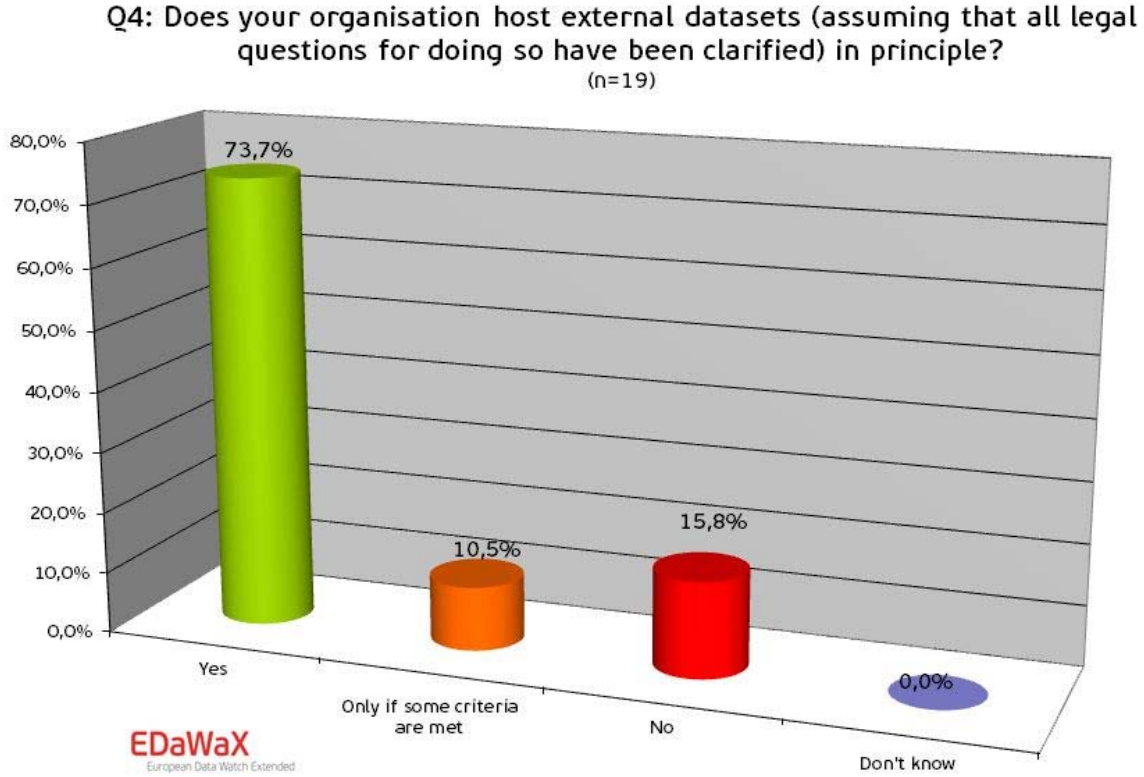
> Es haben sich fast ausschließlich nationale und europäische Forschungsdatenzentren an der Befragung beteiligt

Speicherung externer Forschungsdaten



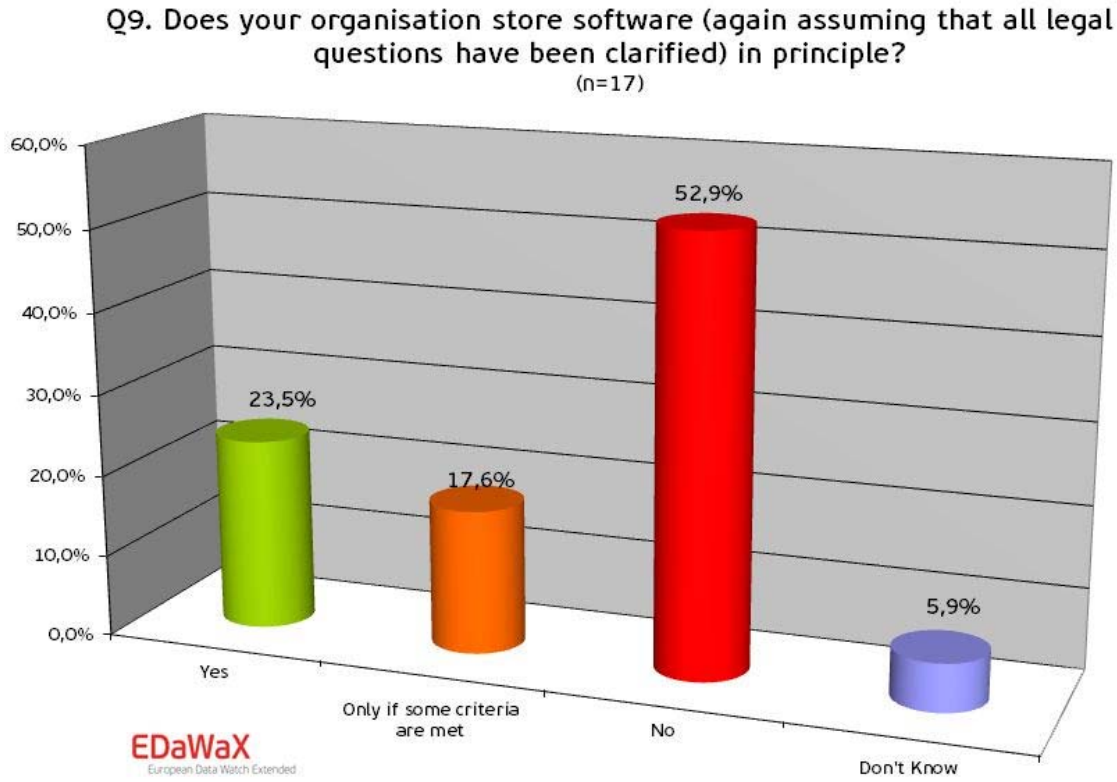
> Fast 70% speichern grundsätzlich externe Datensätze, meist aber nur wenn bestimmte Bedingungen erfüllt sind (fachlich/regional/national).

Hosting externer Forschungsdaten



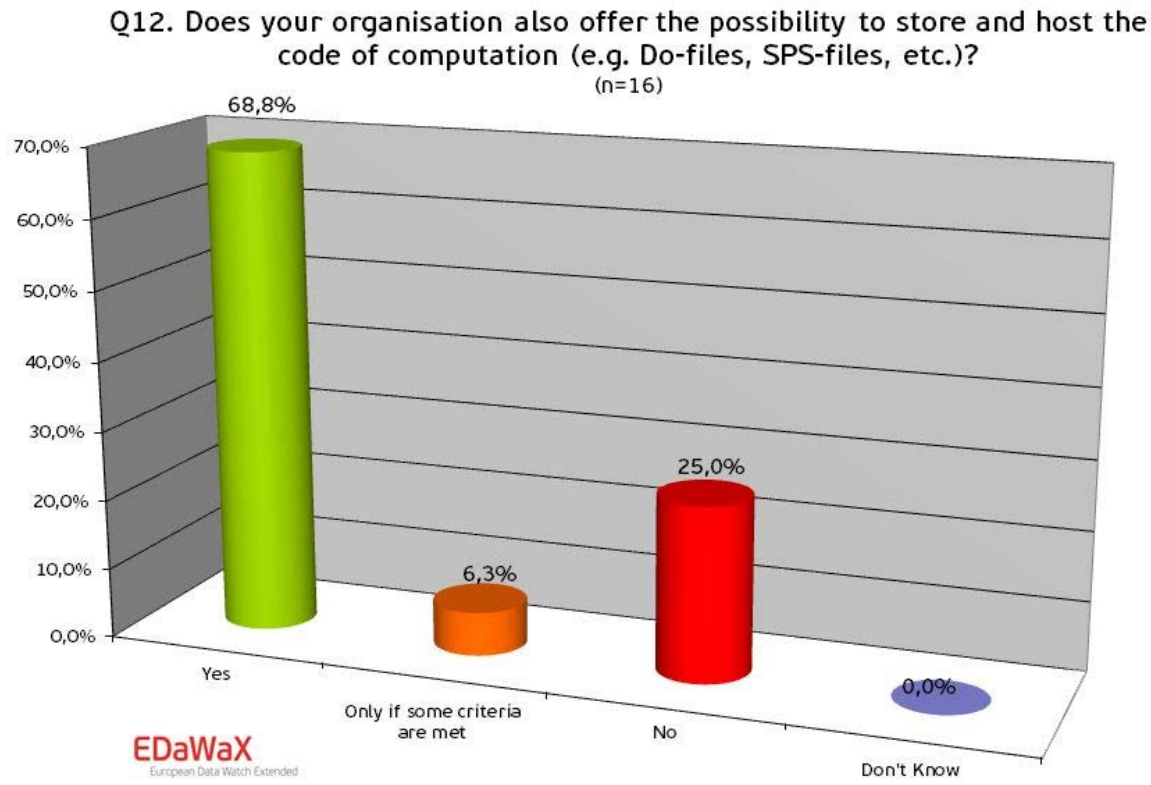
> Mehr als 80% hosten externe Datensätze, nur wenige Institutionen nennen dafür Kriterien.

Speicherung von Software



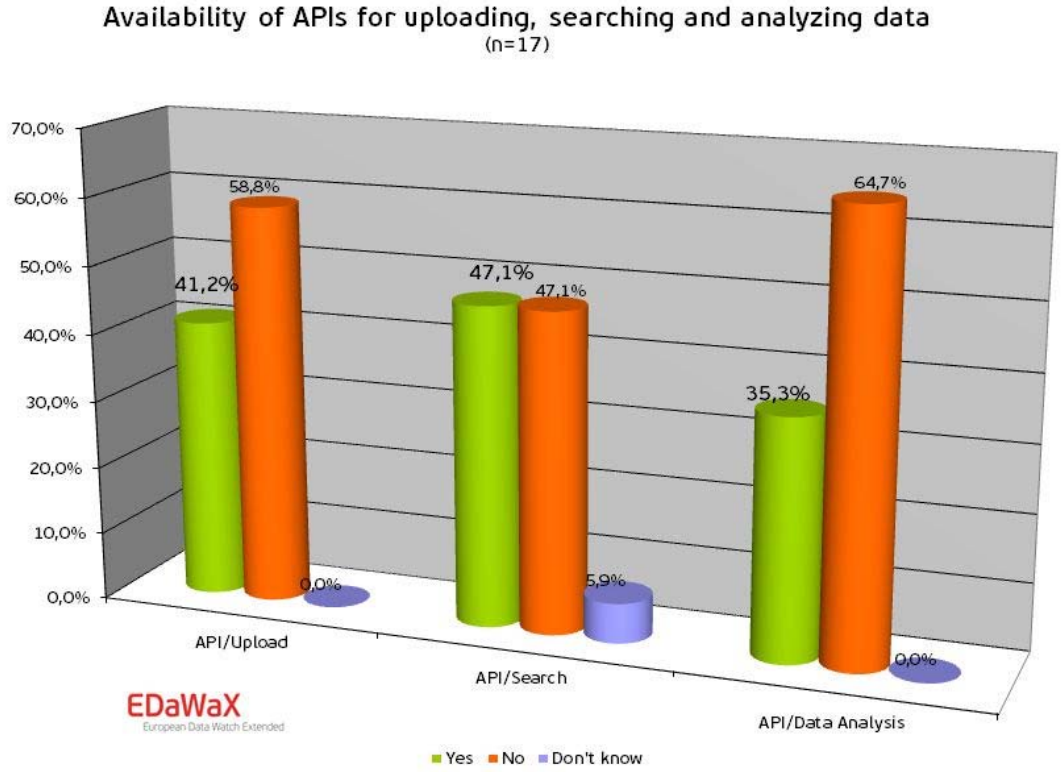
> Software wird von meisten Befragten nicht gespeichert und bereit gestellt.

Speicherung und Hosting von Berechnungscode



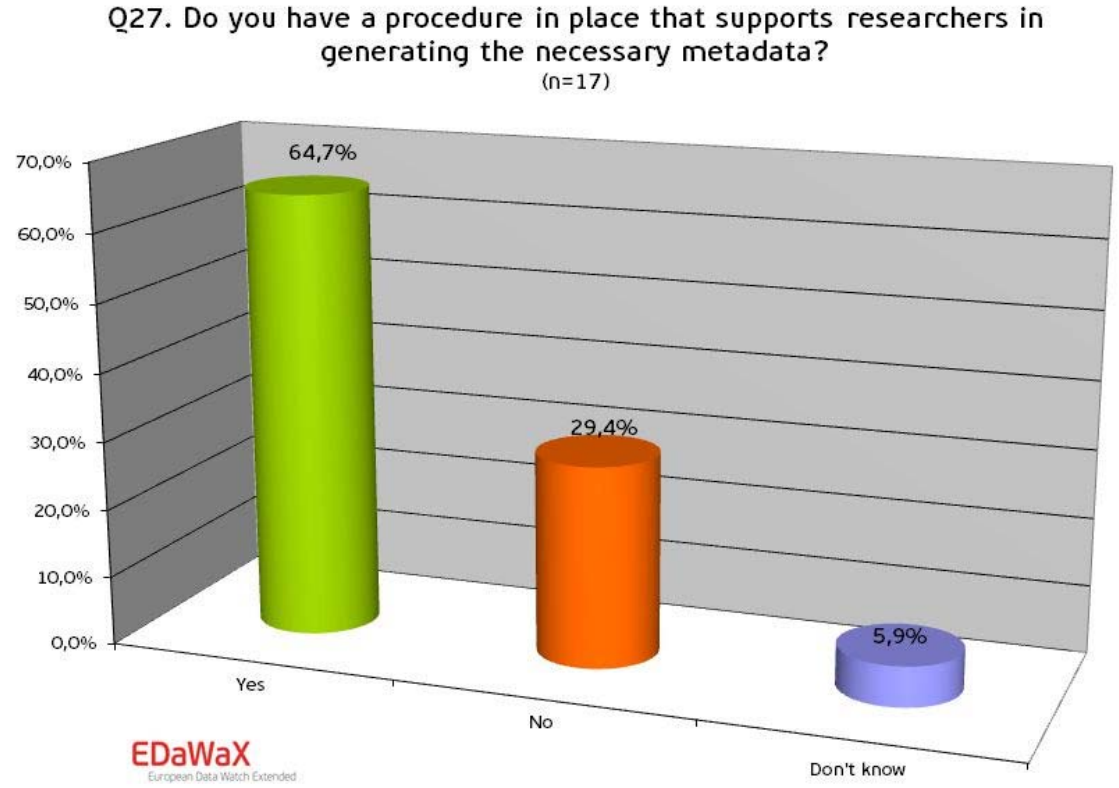
> Berechnungscode wird von gut als 75% der Befragten grundsätzlich gespeichert und bereit gestellt.

Verfügbar Schnittstellen (APIs)



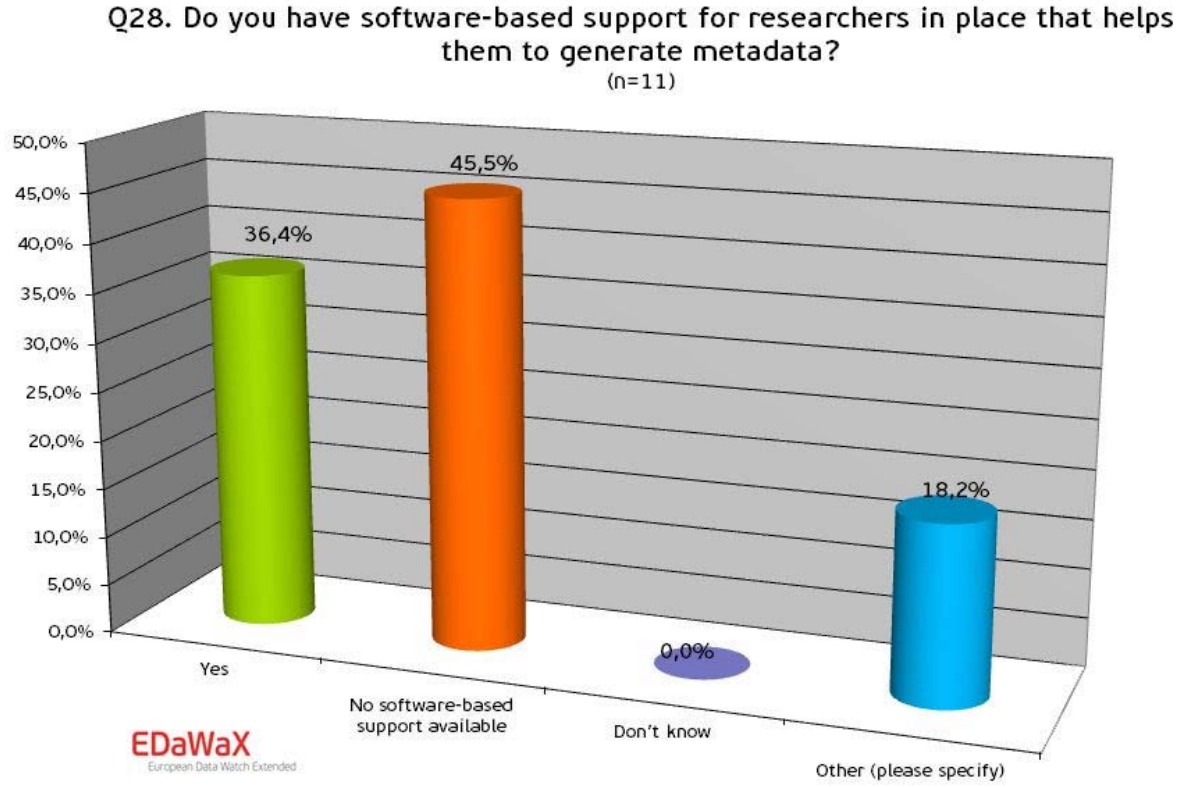
> 35-47% der Befragten geben an, über Schnittstellen zu externen Anwendungen zu verfügen. Unsere Prüfung ergab jedoch, dass solche Schnittstellen NICHT existieren.

User Support bei der Metadatenerstellung



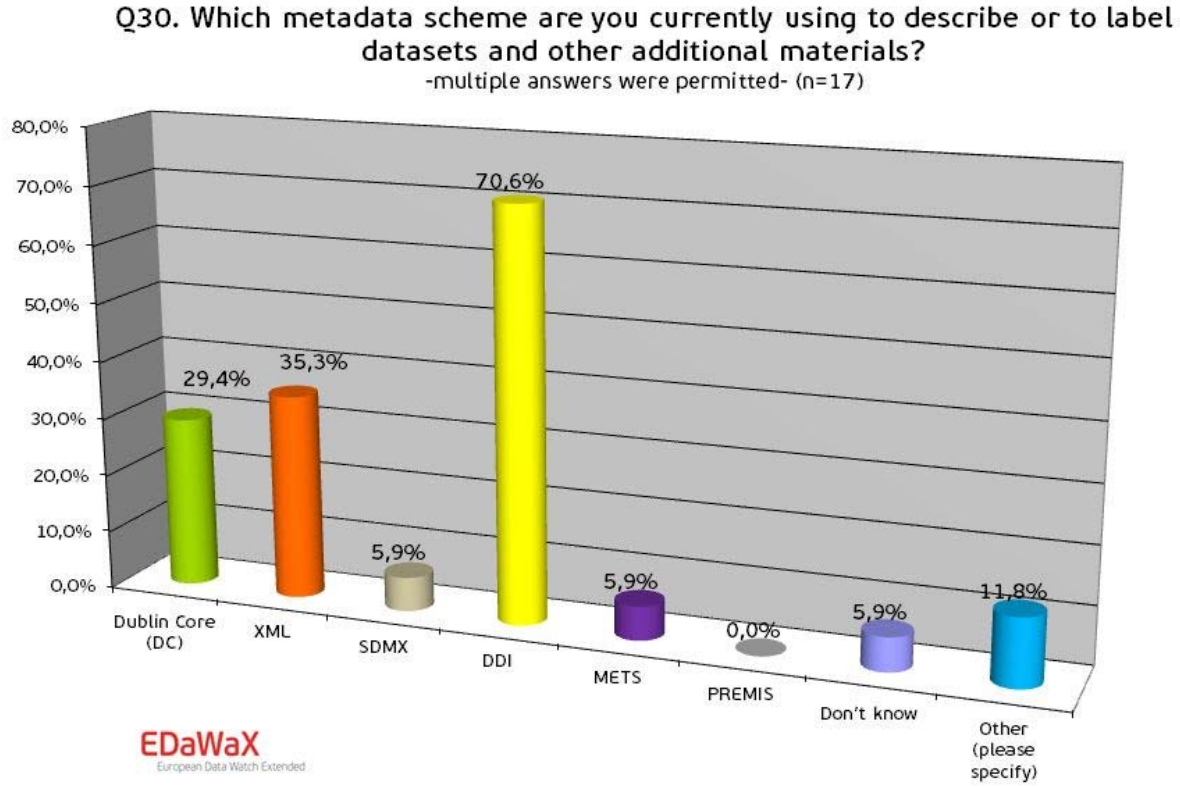
> Fast ⅔ der Befragten unterstützen ihre Nutzer/innen bei der Metadatenerstellung.

Software zur Metadatenerstellung



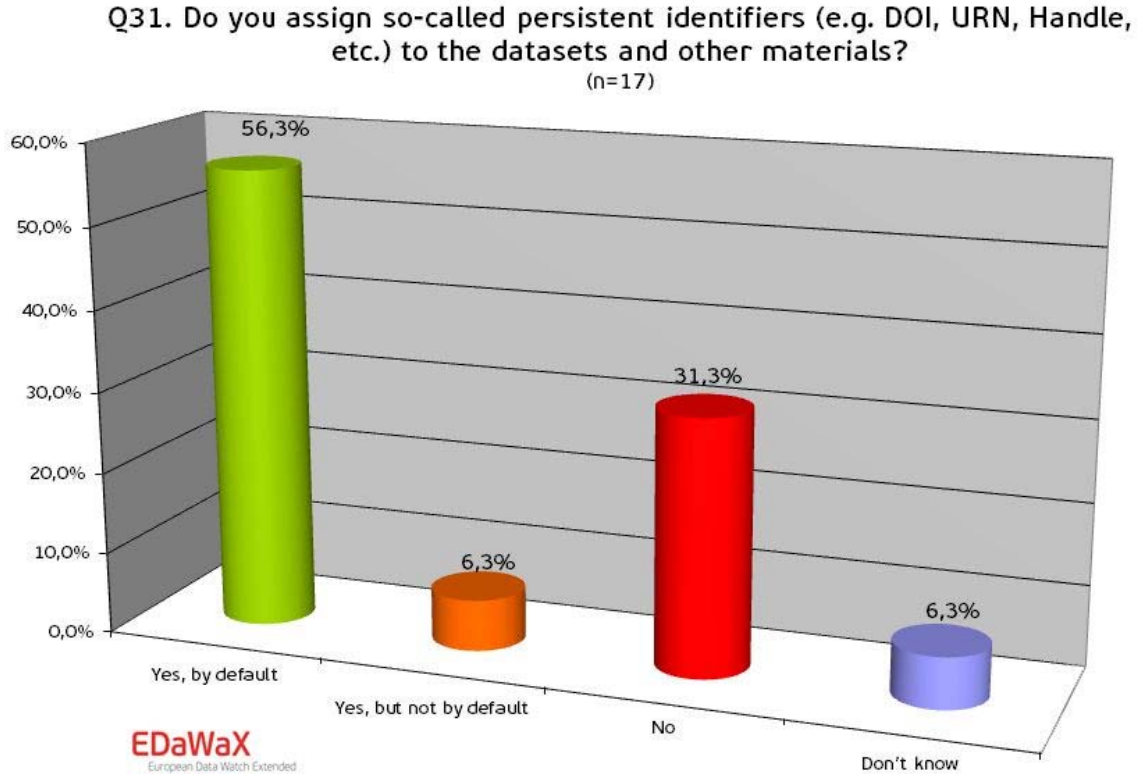
> ...und mehr als 36% nutzen dazu Software.

Genutzte Metadatenschemata



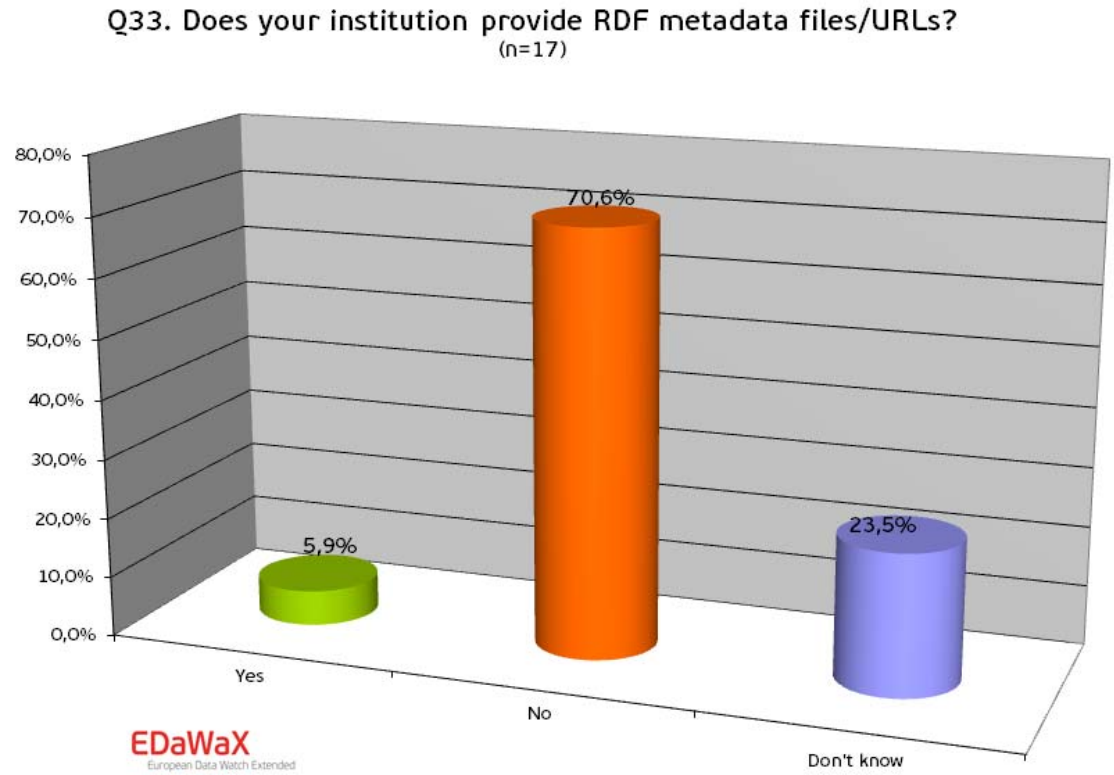
> Es wird v.a. DDI (>70%) als Metadatenschema verwendet, XML und DC folgen mit erheblichem Abstand. Andere Schemata sind marginal.

Verwendung von Persistenten Identifikatoren



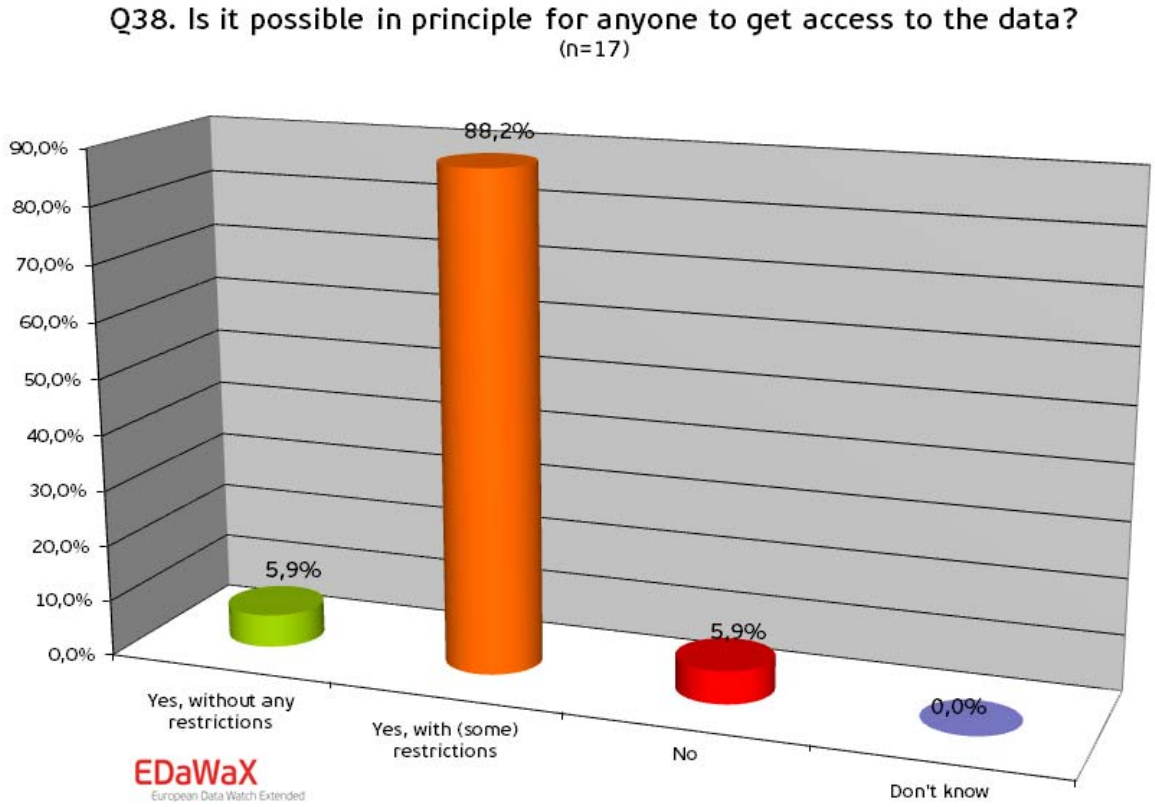
> Fast ⅔ der Befragten vergeben persistente Identifikatoren, mehr als die Hälfte sogar per default.

Nutzung von RDF (Linked Data)



> RDF wird kaum genutzt – und scheint wenig bekannt zu sein.

Zugriffsmöglichkeiten auf Forschungsdaten



> Fast alle befragten Institutionen ermöglichen –mit meist erheblichen Einschränkungen- den Zugriff auf ihre Daten für die Wissenschaft.

Fazit der Befragung

Lässt sich derzeit ein publikationsbezogenes Datenarchiv an einem FDZ speichern und hosten?

Zusammenfassung der Ergebnisse (I)

- Befragungsergebnisse zeigen, dass Forschungsdatenzentren ein relevanter Speicherort für publikationsbezogene Forschungsdaten sein können.
  - externe Forschungsdaten werden häufig angenommen und gehostet, ebenso Syntax,
  - Software/Sourcecode wird eher selten angenommen,
  - es existieren erhebliche Einschränkungen durch fachliche, regionale / überregionale Zuständigkeiten.

Zusammenfassung der Ergebnisse (II)

- Metadatenschema: v.a. DDI, teils auch XML und DC; wenig Support von Linked Data.
- User werden oft bei Erstellung von Metadaten unterstützt – meist ohne spezielle Software-Unterstützung (z.B. NESSTAR).
- Schnittstellen für die externe Suche / den Upload von Datensätzen werden bislang nicht durch die befragten Einrichtungen angeboten.
- Ein barrierefreier Zugang zu Daten ist nur in seltenen Fällen möglich.

➔ Gegenwärtig erfüllt kein FDZ alle unsere Anforderungen.

Die zukünftigen Herausforderungen sind groß...

"It is important that libraries and data centres act in conformance with the requirements of the research community, which they serve. [...]

Libraries and data centres have important, partly overlapping, but mostly complementary roles to fulfill."

Quelle: ODE-REPORT ON INTEGRATION OF DATA AND PUBLICATIONS, 10/2011

> ...durch die gemeinsame Expertise kann diesen Herausforderungen begegnet werden!

Vielen Dank für Ihr Interesse!

Gibt es Anregungen, Fragen oder Kommentare ?

Kontakt:

Sven Vlaeminck | s.vlaeminck@zbw.eu
ZBW – Leibniz Informationszentrum Wirtschaft
Neuer Jungfernstieg 21
20354 Hamburg

Dr. Hendrik Bunke| h.bunke@zbw.eu
ZBW – Leibniz Informationszentrum Wirtschaft
Neuer Jungfernstieg 21
20354 Hamburg

EDaWaX
European Data Watch Extended

RatSWD.

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics