

Services und Workflows einer metadatenbasierten Forschungsdateninfrastruktur für die Wirtschaftswissenschaften

Ein Werkstattbericht aus der ZBW

Dr. Timo Borst

Informationssysteme und Publikationstechnologien

ZBW – Leibniz-Informationszentrum Wirtschaft, Kiel / Hamburg



**Jahrestagung
Frankfurt / Main, 17.-20. März 2020**



Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics



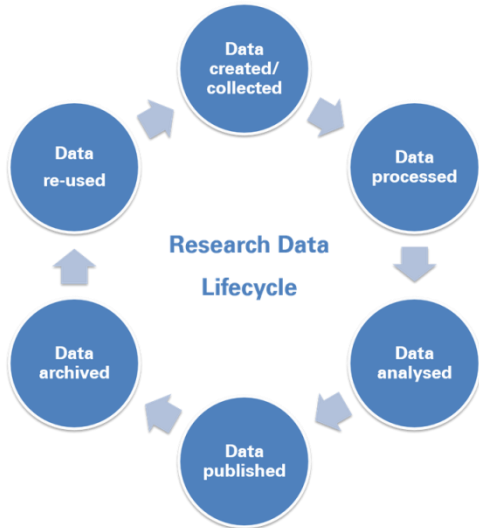
To deposit or not to deposit... that is the question –
journal.pbc.1001779.g001.png, Wikimedia Commons

Mitglied der
Leibniz
Leibniz
Gemeinschaft

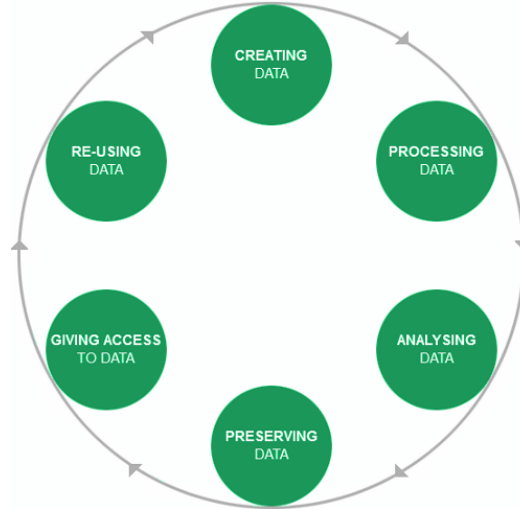
Übersicht

- Der „Datenlebenszyklus“ ...
- Generelle Infrastrukturanforderungen
- Statistik: Software und Umgebungen
- Unstrukturierte Daten
- Beispiele aus der ZBW

Datenlebenszyklus



<https://tu-dresden.de/zih/dienste/service-katalog/zusammenarbeiten-und-forschen/forschungsdatenmanagement>



https://ukdataservice.ac.uk/media/132177/data_lifecycle_recolour.png






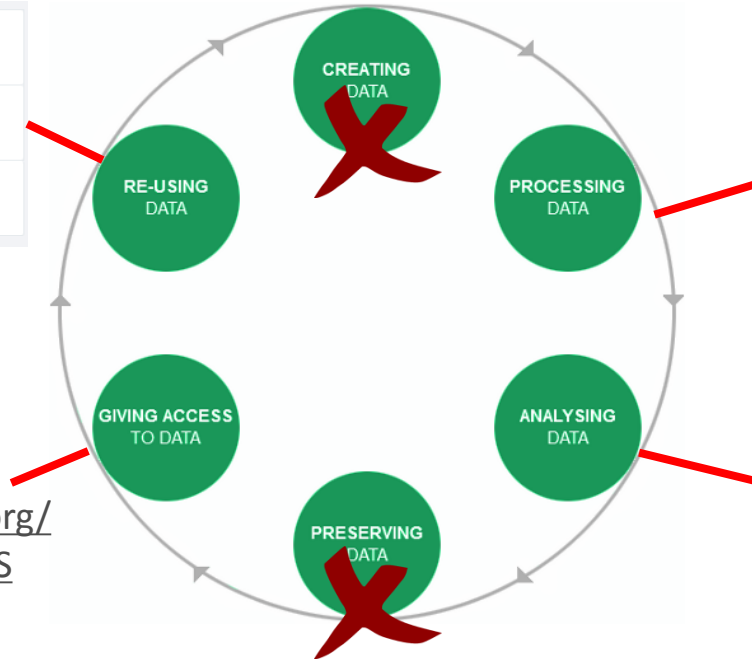
https://library.sydney.edu.au/research/data-management/images/Research_data_lifecycle_final_20150827.png

Datenlebenszyklus

- „Kreislaufmetapher“
- Offenkundig: Daten durchlaufen verschiedene Stadien
(Generierung/Erhebung – Bearbeiten / Transformieren – Prozessieren – Analysieren – Veröffentlichen - Suchen / Auffinden)
- Im Unterschied zu „früher“ passiert dies zunehmend in einer webbasierten Umgebung
- *Was fehlt?*
 - Flexible „Einstiegspunkte“, Schleifen, (Rück-)Sprünge und Exit-Optionen
 - Rückbindung an Forschungsfragen, die sich ggf. dynamisch verändern
 - Die einzelnen Schritte finden i.a.R. in verschiedenen (auch kommerziellen) Umgebungen statt

Datenlebenszyklus – verschiedene Einstiegspunkte

-  **Download**
[CSV](#) [XML](#) [EXCEL](#)
-  **DataBank**
Online tool for visualization and analysis
-  **WDI Tables**
Thematic data tables from WDI



<https://data.worldbank.org/indicator/AG.LND.AGRI.ZS>

DataBank World Development Indicators

Popular Indicators

Inflation, consumer prices (annual %)

	2008	2009	2010	2011	2012	2013	2014
Germany	2.6	0.3	1.1	2.1	2.0	1.5	

Generelle Infrastrukturanforderungen

Durch wen und in welchem Ausmaß wird eine FD-Infrastruktur (mutmaßlich) genutzt?

Beispiel: Forschungsdatenzentren (FDZ) des RatSWD

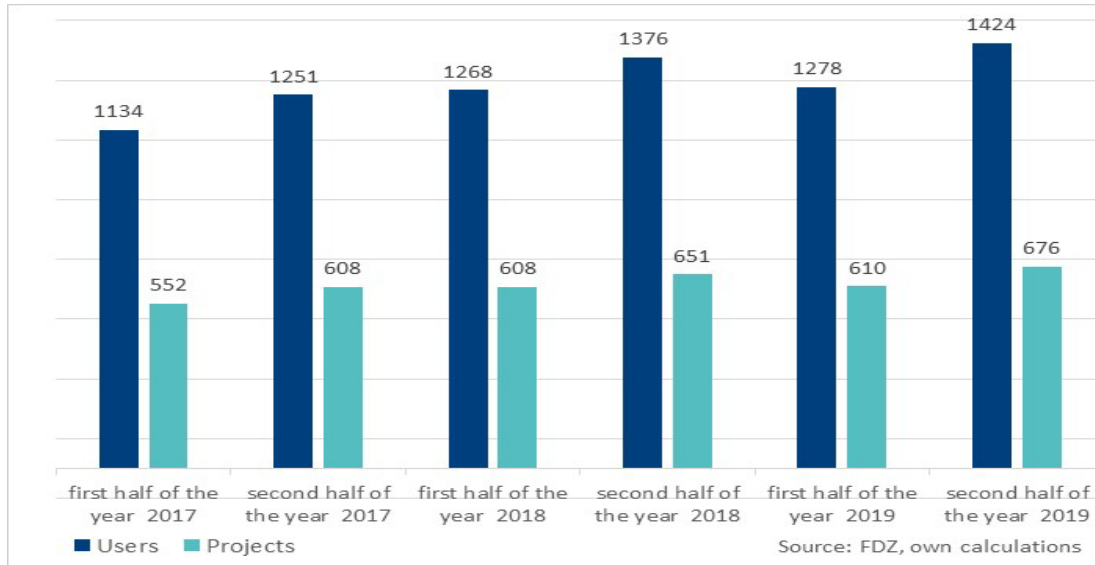


- Ø 1.450 externe Datennutzende pro FDZ

© RatSWD 2019

Generelle Infrastrukturanforderungen

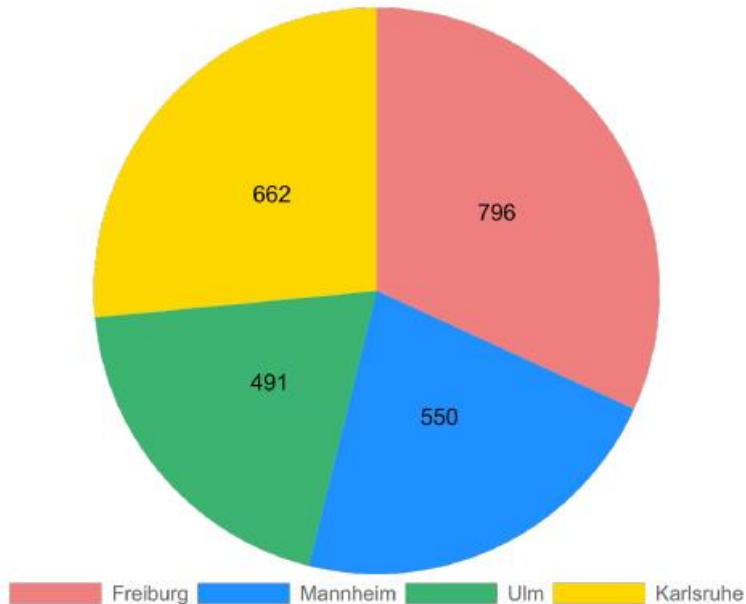
Beispiel FDZ des IAB



- ~ 650 Forschungsprojekte
- ~ 1.400 Nutzer*innen =
niedriger vierstelliger Bereich

Generelle Infrastrukturanforderungen – Virtuelle Maschinen (VM)

Aktive VMs pro bwCloud Region



- Verteilung an virtuellen Maschinen für WiWi-Forschung in Baden-Württemberg
- ~ 2.500 VMs über vier Standorte verteilt

Quelle: bwCloud Nutzungs- und Auslastungsstatistik, Januar 2020

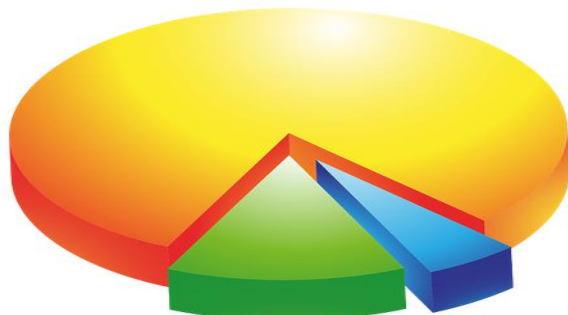
Statistik-Software

EDaWaX
European Data Watch Extended

[About](#) [Partners](#) [Events](#) [Downloads](#)

Statistical software: its use and popularity in Economics

Posted: August 7th, 2017 | **Author:** Timo | **Filed under:** Report | **Tags:** economics, repositories, Software, Statistics
| Comments Off on Statistical software: its use and popularity in Economics



by Christina Kläre & Timo Borst

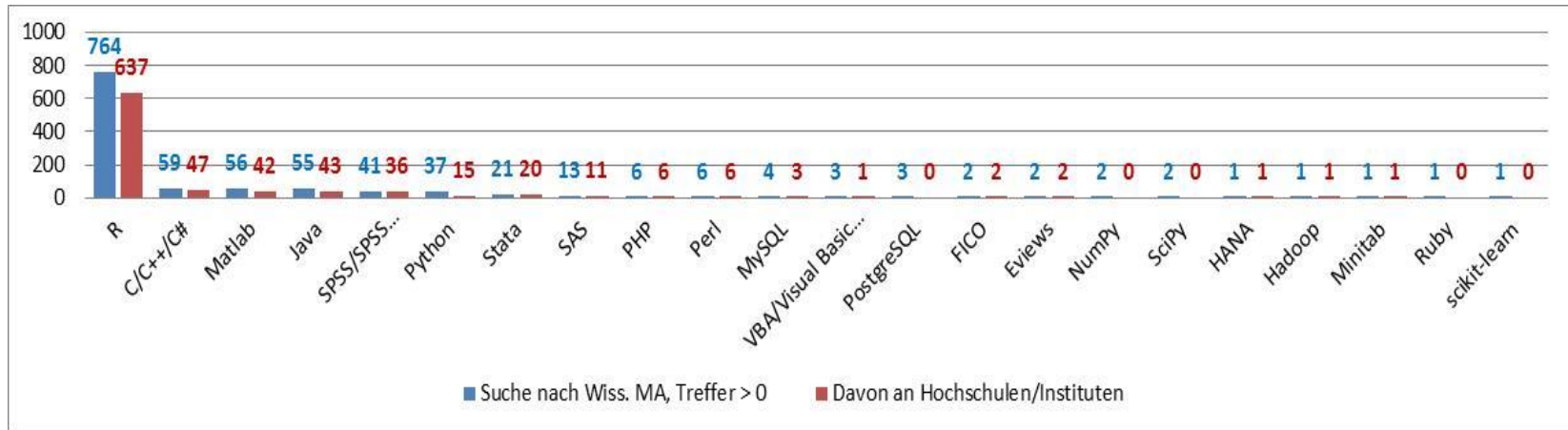
During a four weeks project at ZBW's Department for Information Systems and Publishing Technologies, we collected some publicly available information about statistical software packages being used in research in Economics. This work is inspired by a **constantly updated blog post from Robert A. Muenchen**, who examined information sources like job announcements, scientific articles, reports from IT companies, questionnaires, sales statistics from

software textbooks, blogposts, forums, polls measuring popularity of programming languages, sales

<https://www.edawax.de/2017/08/statistical-software-its-use-and-popularity-in-economics/>

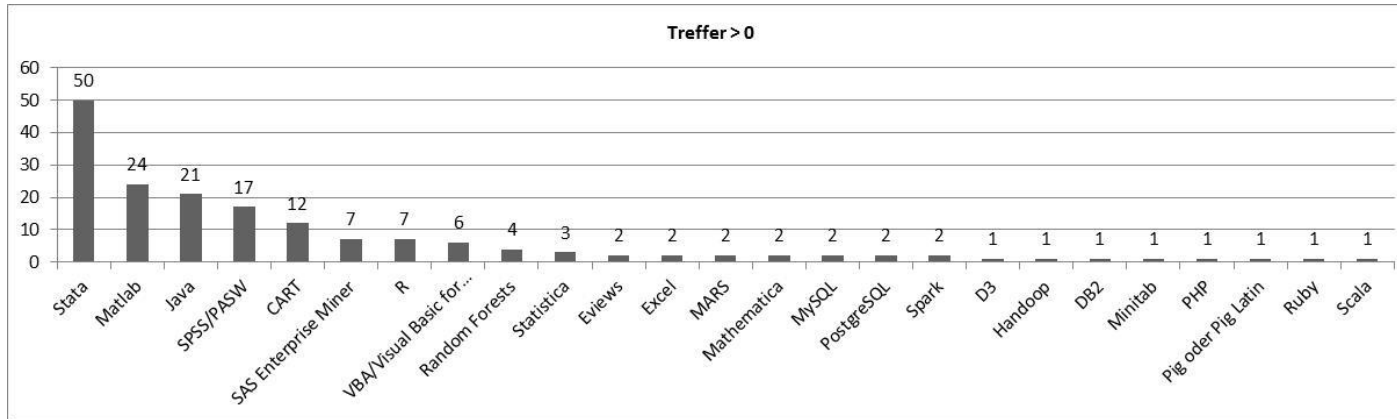
Statistik-Software

Erforderliche Skills bei Stellenausschreibungen für wiss. Personal (gemäß indeed.com, Stand 2017)



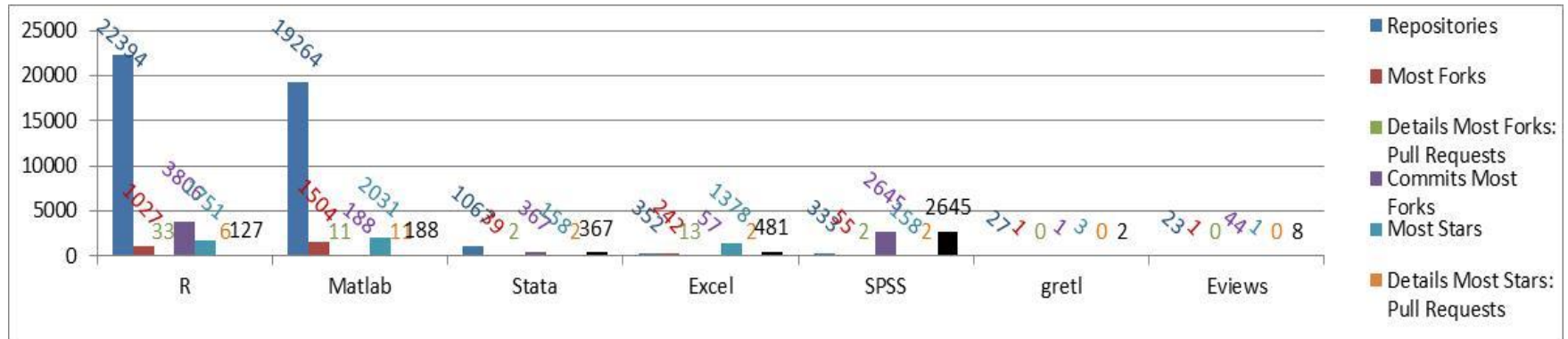
Statistik-Software

Erforderliche Skills bei Stellenausschreibungen speziell für Wirtschaftswissenschaftler*innen
(gemäß indeed.com, Stand 2017)



Statistik-Software

Statistik-Pakete in GitHub-Repositoryn (Stand 2017)



Fazit: R (als Statistikprogramm bzw. als Programmier- und Skriptsprache) ist die meistgenutzte frei verfügbare und quelloffene Software

Statistikumgebungen

R / RStudio / R Markdown aka Notebooks als integrative Analyseumgebung



Programmier-/Skriptsprache



Entwicklungsumgebung



Dokumentation / Präsentation /
Veröffentlichung

Statistikumgebungen

Integration mit einer übergreifenden (dezentralen) Forschungsdateninfrastruktur

- ursprünglich: lokale Datenerzeugung und -haltung, lokale Auswertung
- mittlerweile: (remote oder on-site) Zugriff auf zentral gehostete Daten (Datenzentren, statistische Ämter)
- künftig: Cloud-Dienste und Online-Umgebungen zur Erzeugung von Daten, Bereitstellung von Daten + Auswertungsumgebungen – aber auch Mischformen wie z.B. ein Cloud-Speicher, den ich in meine lokale Arbeitsumgebung einbinden kann

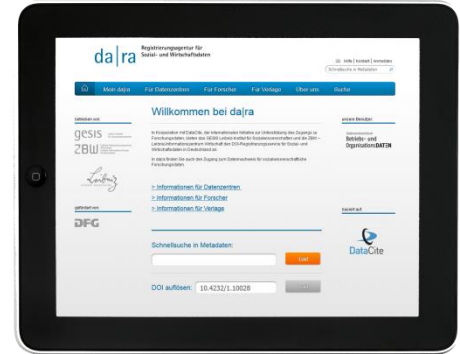
Unstrukturierte Daten und FD-Infrastruktur

- Klassischer Weise hochgradig strukturierte Daten in Form von Statistiken und Tabellen
- Angesichts von „Big Data“ – d.h. großen, dynamisch veränderlichen und unstrukturierten Datenbeständen – verändern sich auch Forschungsgegenstände und Fragestellungen auch in den WiWis
 - Analyse von Twitter-Streams (Ranco, Aleksovski, Caldarelli et al., 2015)
 - Bildanalysen (Heitmann, Siebert, Hartmann & Schamp, 2020)
 - Finanzdaten-Analysen (Lin, Hu & Tsai, 2011)

Unstrukturierte Daten und FD-Infrastruktur

- Größere Datenmengen können nicht mehr lokal gehalten werden (Stichwörter: Lizenzen, Mehrfachzugriffe, Updates), sondern sollten als zentraler Cloud-Dienst Forscher*innen oder Forschungsgruppen bereit gestellt werden
 - Wichtige und arbeitsaufwändige Schritte beim Preprocessing (z.B. Tokenizing von Texten oder Bereinigen von Finanzdaten, Trainingsdaten für Modelle) können an zentraler Stelle erfolgen bzw. zur Verfügung gestellt werden (als Package oder gleich als aufbereiteter Datensatz)
 - Erweiterte Statistikanwendungen auf dem Gebiet von Machine Learning benötigen Rechenkapazitäten (z.B. GPUs), die in einer lokalen Umgebung häufig nicht mehr spontan bereit gestellt werden können
-

Entwicklung disziplinärer Forschungsdatenservices bei der ZBW

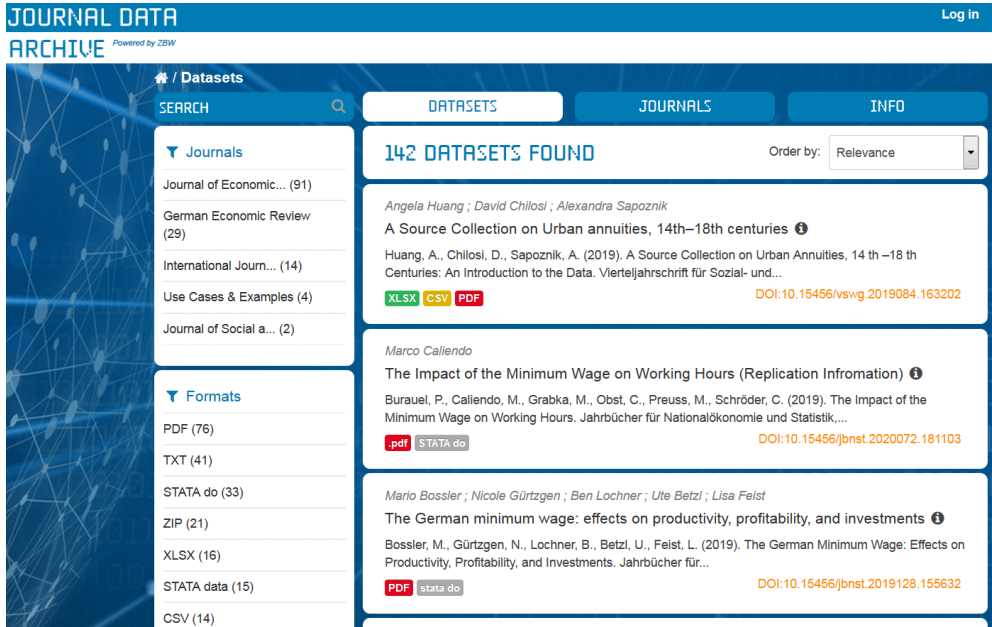


Publikationsdienst I: Digitale Reichsstatistik

The screenshot shows the ZBW (Leibniz-Informationzentrum) website interface. At the top, the ZBW logo and name are displayed. Below it, a navigation bar includes 'Home', 'Suchergebnisse', and 'Ergebnis'. The main content area shows search results for a specific document. The document title is 'Definitives Haupt-Ergebniss der Volkszählung im Deutschen Reiche vom 1. December 1871 mit Nachweisung der Bevölkerungs- Zu - oder Abnahme seit dem 3. December 1867'. The interface includes various filters and options such as 'Suche' (Simple and Extended search), 'Hilfe' (Help), 'Sprache' (Language), 'Export' (Citation, Metadata), 'Display' (Scan, DFG-Viewer), and 'Einheitentyp' (Anzahl, Verhältnis, Einheit). A thumbnail of the document cover is visible at the bottom of the search result.

- Digitalisierung von Statistiken des Deutschen Reiches von 1873 – 1883
- Scannen und freihändige Erfassung der Tabellendaten („double-keying“)
- Umfangreiche (Meta-)Datensuche

Publikationsdienst II: Journal Data Archive



The screenshot shows the 'JOURNAL DATA ARCHIVE' website. The header includes 'JOURNAL DATA ARCHIVE' and 'Powered by ZBW' with a 'Log in' link. The main navigation bar has 'DATASETS', 'JOURNALS', and 'INFO' tabs. A search bar is present. On the left, there are filters for 'Journals' (listing 'Journal of Economic...' (91), 'German Economic Review' (29), 'International Journ...' (14), 'Use Cases & Examples' (4), 'Journal of Social a...' (2)) and 'Formats' (listing 'PDF' (76), 'TXT' (41), 'STATA do' (33), 'ZIP' (21), 'XLSX' (16), 'STATA data' (15), 'CSV' (14)). The main content area displays '142 DATASETS FOUND' with a dropdown menu set to 'Relevance'. Three dataset entries are visible:

- Angela Huang ; David Chilosi ; Alexandra Sapoznik**
A Source Collection on Urban annuities, 14th–18th centuries ⓘ
Huang, A., Chilosi, D., Sapoznik, A. (2019). A Source Collection on Urban Annuities, 14 th –18 th Centuries: An Introduction to the Data. Vierteljahrschrift für Sozial- und...
Formats: XLSX, CSV, PDF | DOI:10.15456/vswg.2019084.163202
- Marco Callendo**
The Impact of the Minimum Wage on Working Hours (Replication Information) ⓘ
Burauel, P., Callendo, M., Grabka, M., Obst, C., Preuss, M., Schröder, C. (2019). The Impact of the Minimum Wage on Working Hours. Jahrbücher für Nationalökonomie und Statistik...
Formats: PDF, STATA do | DOI:10.15456/jbnst.2020072.181103
- Mario Bossler ; Nicole Gürtzgen ; Ben Lochner ; Ute Betzl ; Lisa Feist**
The German minimum wage: effects on productivity, profitability, and investments ⓘ
Bossler, M., Gürtzgen, N., Lochner, B., Betzl, U., Feist, L. (2019). The German Minimum Wage: Effects on Productivity, Profitability, and Investments. Jahrbücher für...
Formats: PDF, stata do | DOI:10.15456/jbnst.2019128.155632

- Publikation und Verknüpfung von Datensätzen zu Journal-Veröffentlichungen
- Unterstützung von Replikationsstudien
- Mehrheitlich STATA-Dateien

Dienstebasierte FD-Infrastruktur: GeRDI

- Bisher: Publikationsunterstützende Dienste für spezifische Kontexte und Datenbestände (Analoge Statistiken, Journal-Publikationen, kleinere Forschungsgruppen)
- Nunmehr: Softwarebasierte Dienstarchitektur („Microservices“), die nicht die Datenbestände selbst zum Gegenstand hat, sondern die Unterstützung prinzipiell gleichförmiger Prozessketten und Einzelschritte wie z.B. das Auffinden, Bookmarken, Teilen, Herunterladen oder Analysieren eines Datensatzes

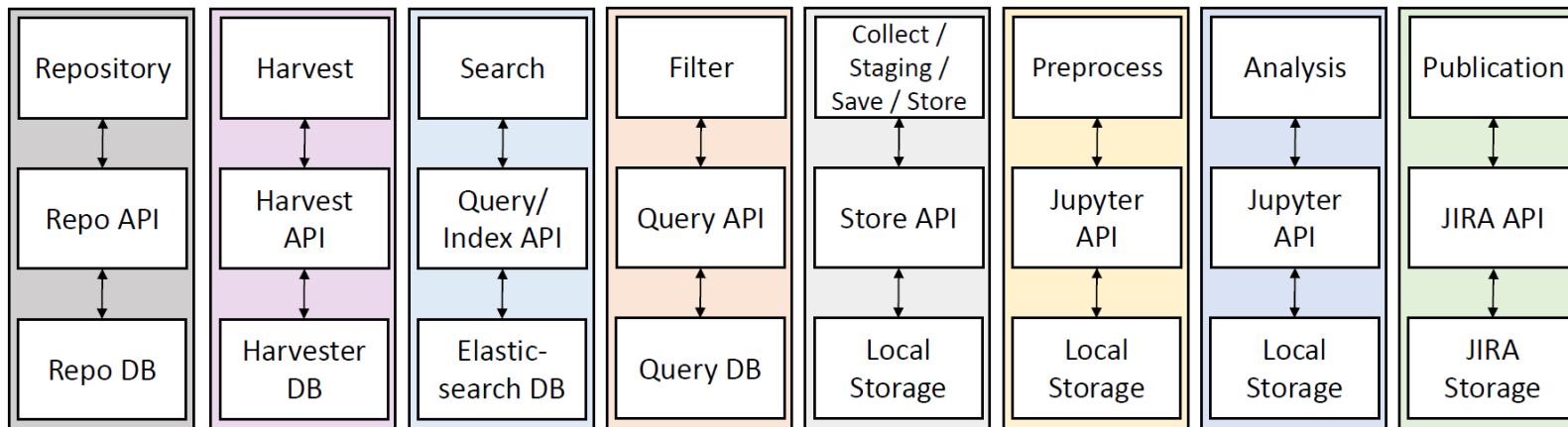
Dienstebasierte FD-Infrastruktur: GeRDI



GeRDI

Generic Research Data Infrastructure

API Gateway, Page Assembly Proxy, ...



REST, Messaging, Prospector AAI, Scalability/Elasticity, Monitoring, Control Center, ...



Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

Mitglied der
Leibniz
Leibniz
Gemeinschaft

Dienstebasierte FD-Infrastruktur: GeRDI



Search Bookmark Store Process Analyze Submit

*

119 results found for *

Publisher ^

Select all Clear all

- LMU-ifo Economics & Business Data Center (EBDC) - (119)
- Zenodo - (0)
- PANGAEA - Data Publisher for Earth & Environmental Science - (0)
- European Nucleotide Archive (ENA) - (0)
- Esri - (0)

More Less

Ifo Business Survey Industry (2013b)

LMU-ifo Economics & Business Data Center (EBDC)

Carstensen, Kai

The Ifo Business Survey for Manufacturing has been conducted on a monthly basis since 1949. The core of the questionnaire consists of qualitative assessments on the current economic parameters of companies such as, for example, the general situation, pricing trends, credit constraints or staff numbers, as well as questions on trends in the forthcoming months in areas like exports, employment and price expectations.

[More information](#) [Share](#) [Add Bookmark](#) [Preprocess](#) [Store](#)

ifo World Economic Survey (2018q2)

LMU-ifo Economics & Business Data Center (EBDC)

Wollmershäuser, Timo

Zwar letztlich auch eine Webseite, aber als Eintiegsseite für den (geschützten) Zugriff auf verteilte Diensten, die unabhängig voneinander operieren



Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics



