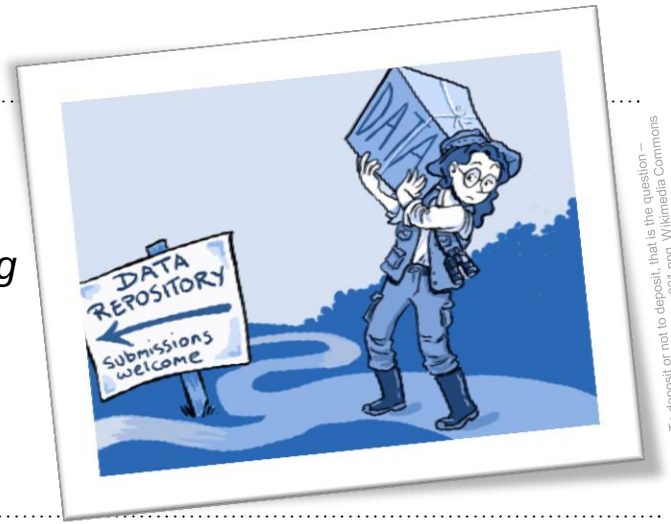
Services und Workflows einer metadatenbasierten Forschungsdateninfrastruktur für die Wirtschaftswissenschaften

Ein Werkstattbericht aus der ZBW

Dr. Timo Borst
Informationssysteme und Publikationstechnologien
ZBW – Leibniz-Informationszentrum Wirtschaft, Kiel / Hamburg



VHB

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

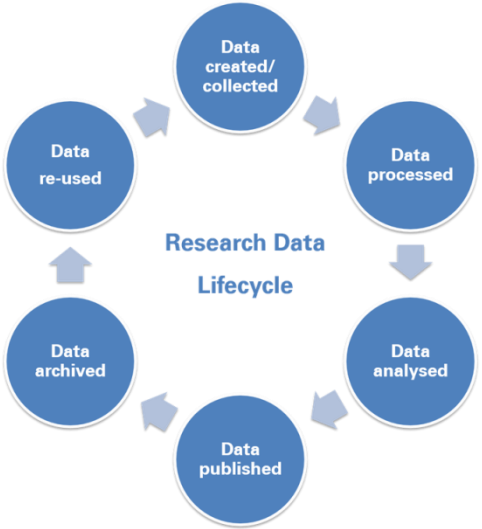
ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Mitglied der Leibniz Gemeinschaft

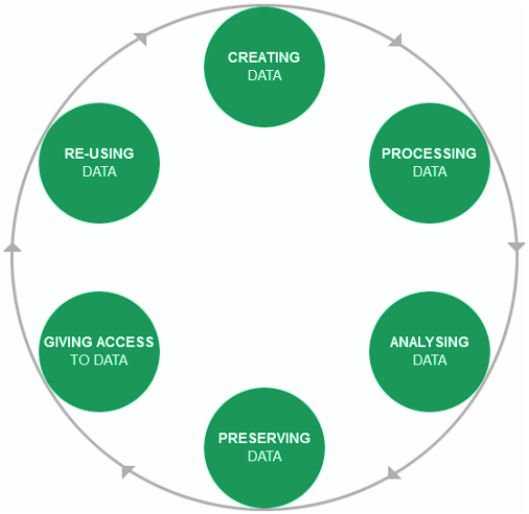
Übersicht

- 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



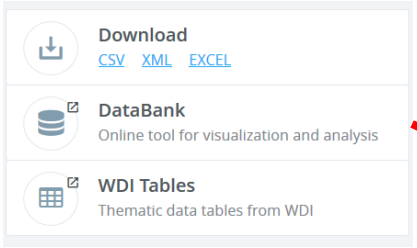
https://ukdataservice.ac.uk/media/132177/data_lifecycle_recolour.png

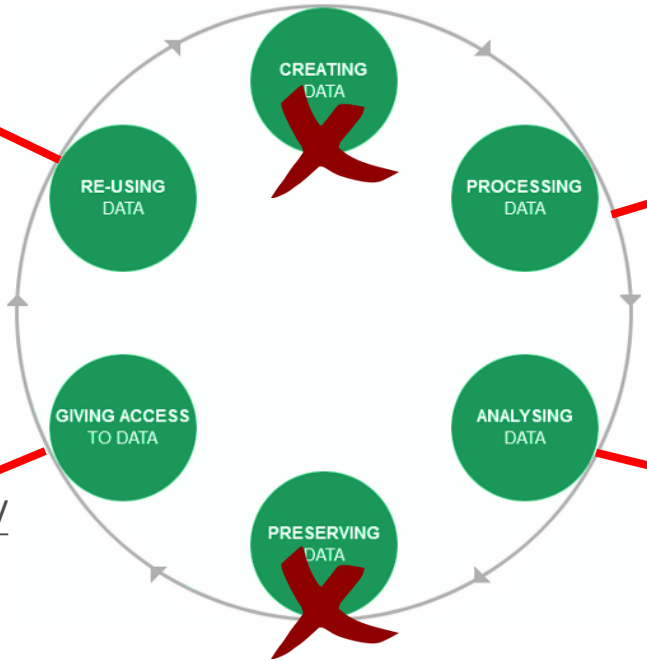
https://library.sydney.edu.au/research/data-management/images/Research_data_lifecycle_final_20150827.png

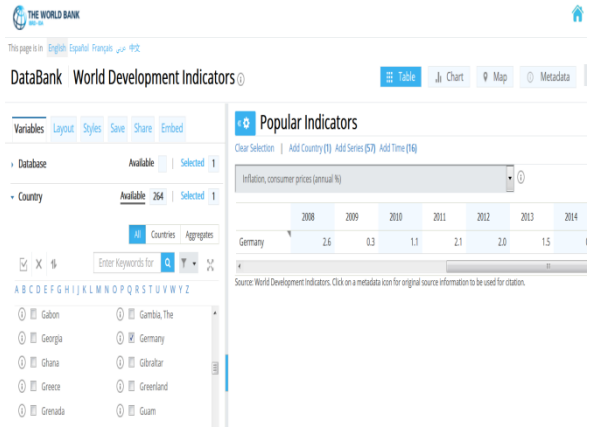
Datenlebenszyklus

- „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







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

Generelle Infrastrukturanforderungen

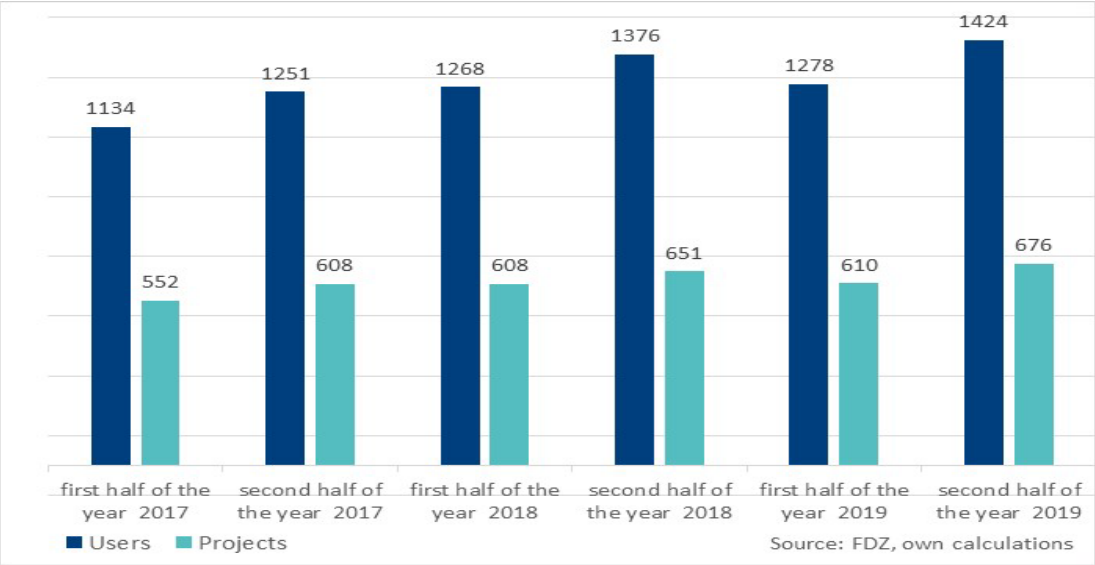
Durch wen und in welchem Ausmaß wird eine FD-Infrastruktur (mutmaßlich) genutzt?

Beispiel: Forschungsdatenzentren (FDZ) des RatSWD

- Ø 1.450 externe Datennutzende pro FDZ

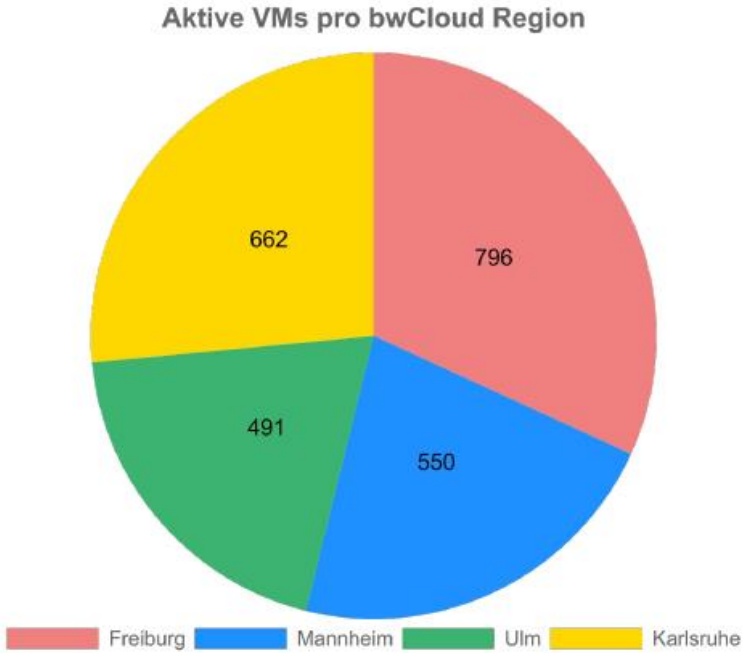
Generelle Infrastrukturanforderungen

Beispiel FDZ des IAB



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

Generelle Infrastrukturanforderungen – Virtuelle Maschinen (VM)



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

Quelle: bwCloud Nutzungs- und Auslastungsstatistik, Januar 2020

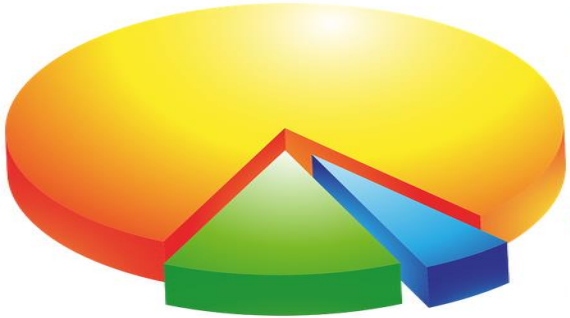
Statistik-Software

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

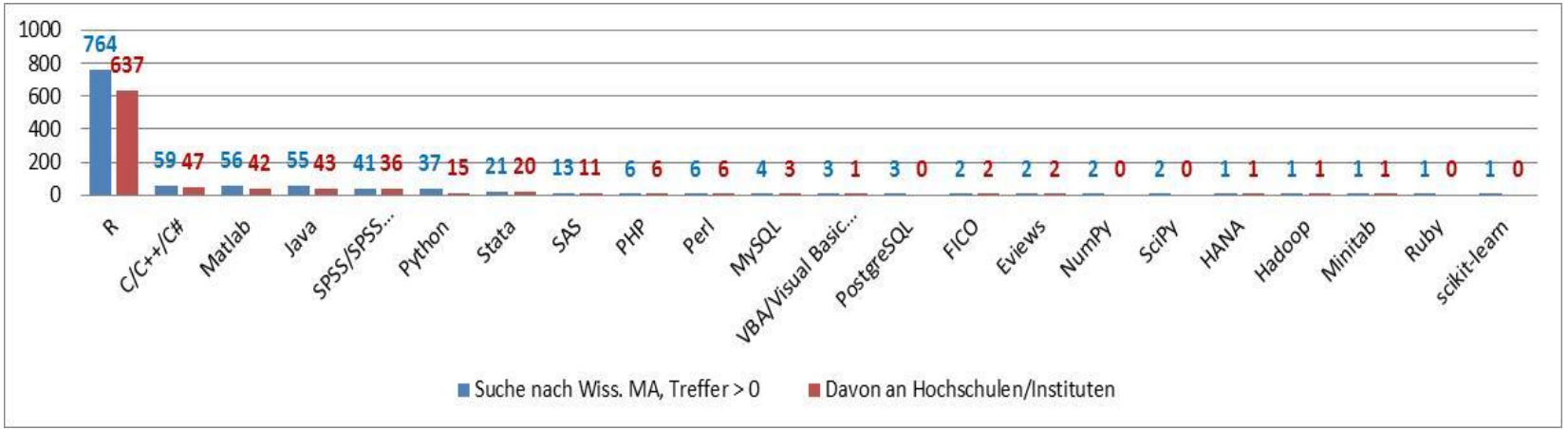
https://www.edawax.de/2017/08/statistical-software-its-use-and-popularity-in-economics/

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Mitglied der Leibniz Gemeinschaft

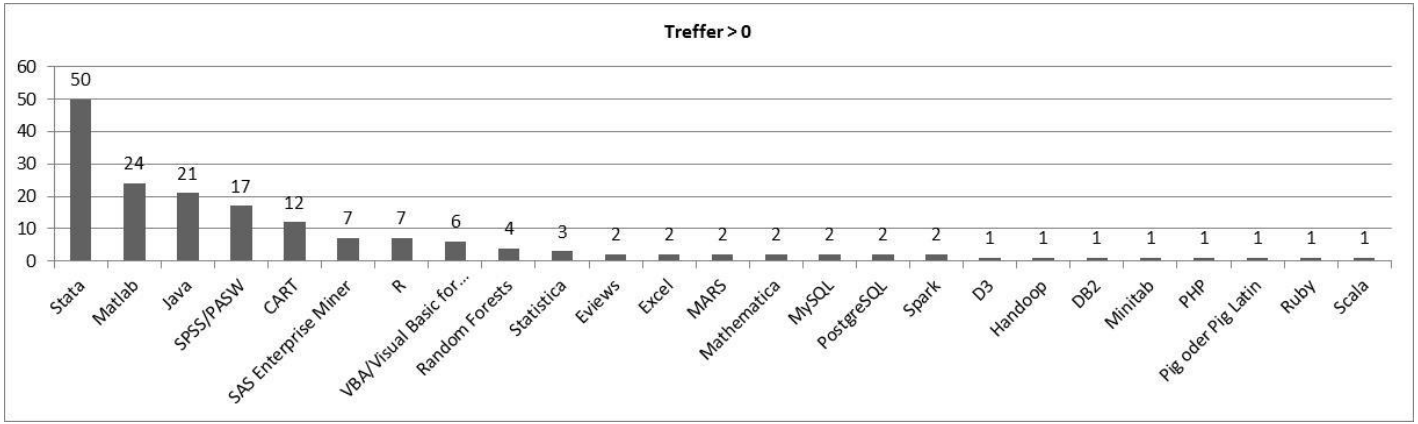
Statistik-Software

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



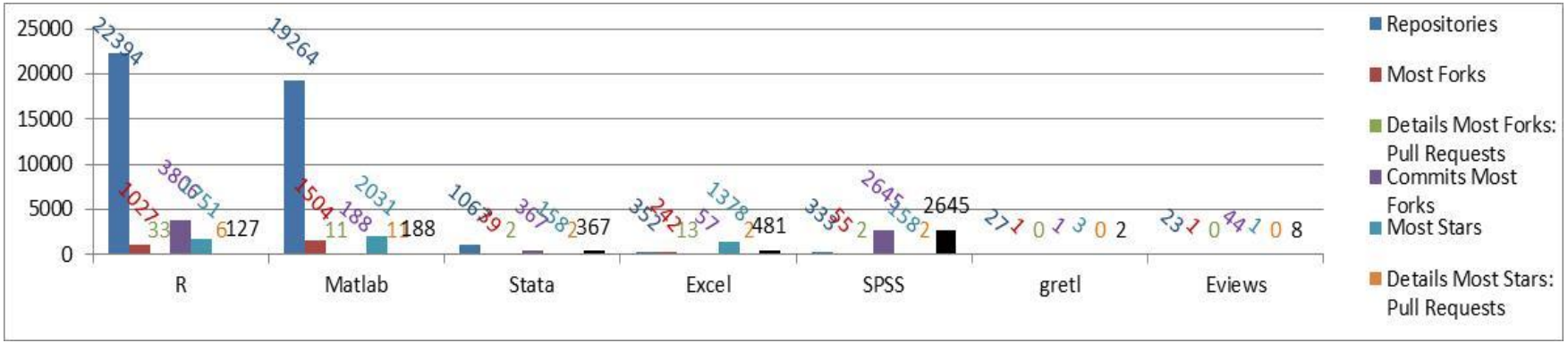
Statistik-Software

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



Statistik-Software

Statistik-Pakete in GitHub-Repositorien (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

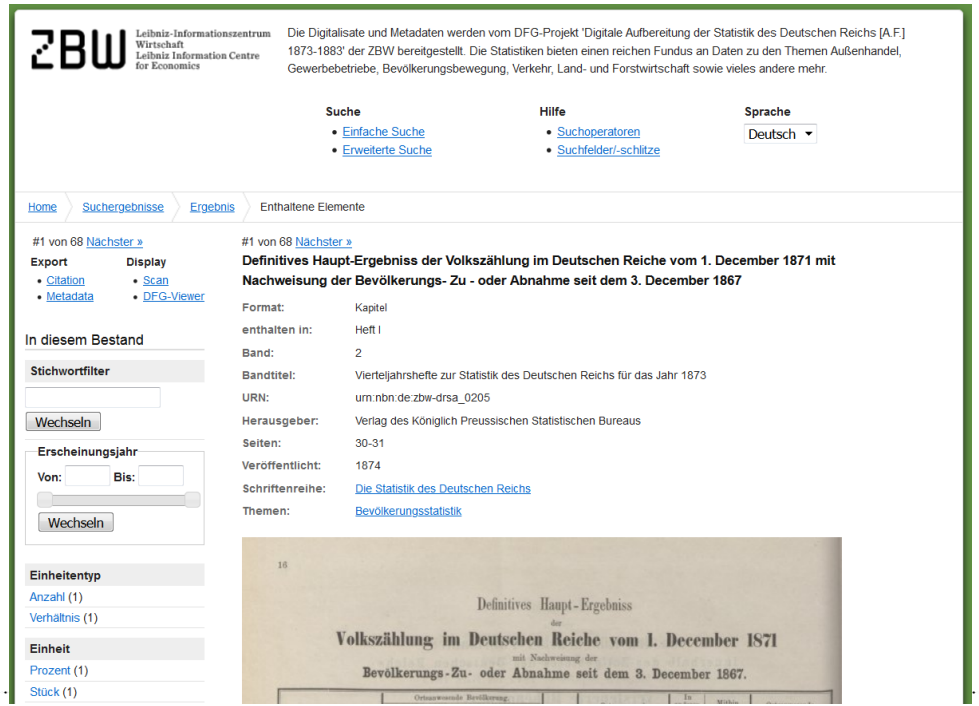
Entwicklung disziplinärer Forschungsdatenservices bei der ZBW

da|ra Registrierungsagentur für Sozial- und Wirtschaftsdaten

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

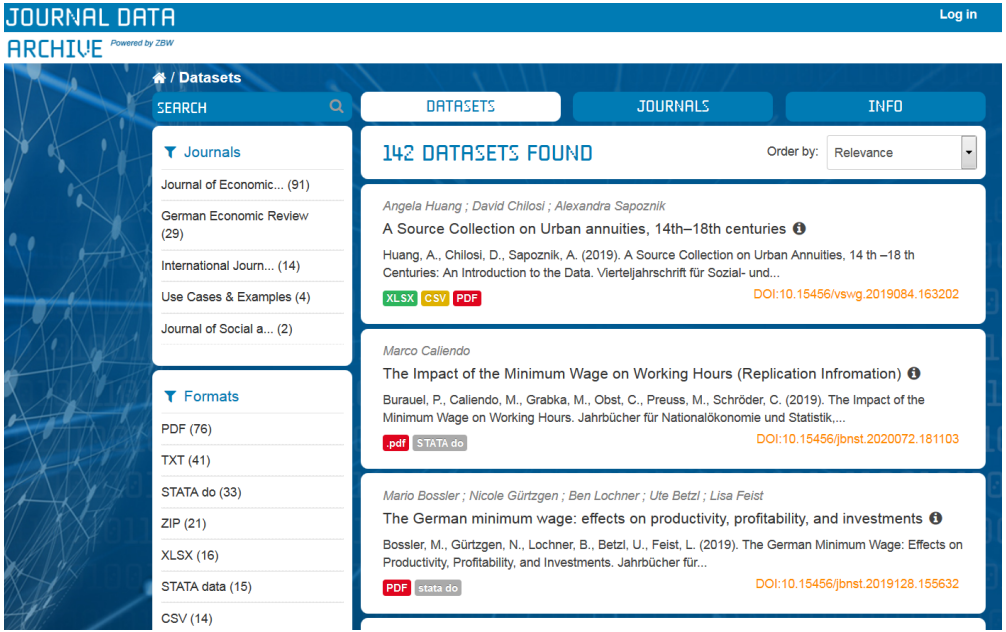
Mitglied der Leibniz Gemeinschaft

Publikationsdienst I: Digitale Reichsstatistik



- 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



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

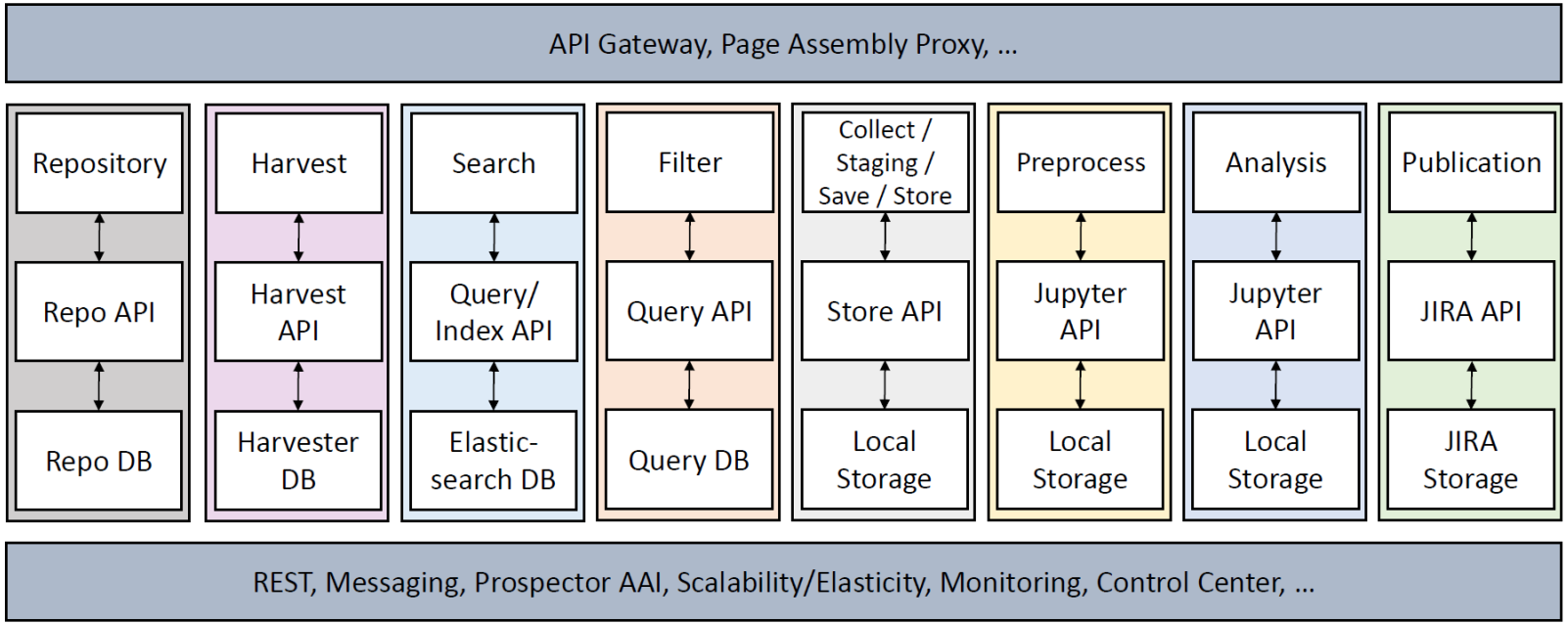
Dienstebasierte FD-Infrastruktur: GeRDI

- Bisher: Publikationsunterstützende Dienste für spezifische Kontexte und Datenbestände (Analoge Statistiken, Journal-Publikationen, kleinere Forschungsgruppen)

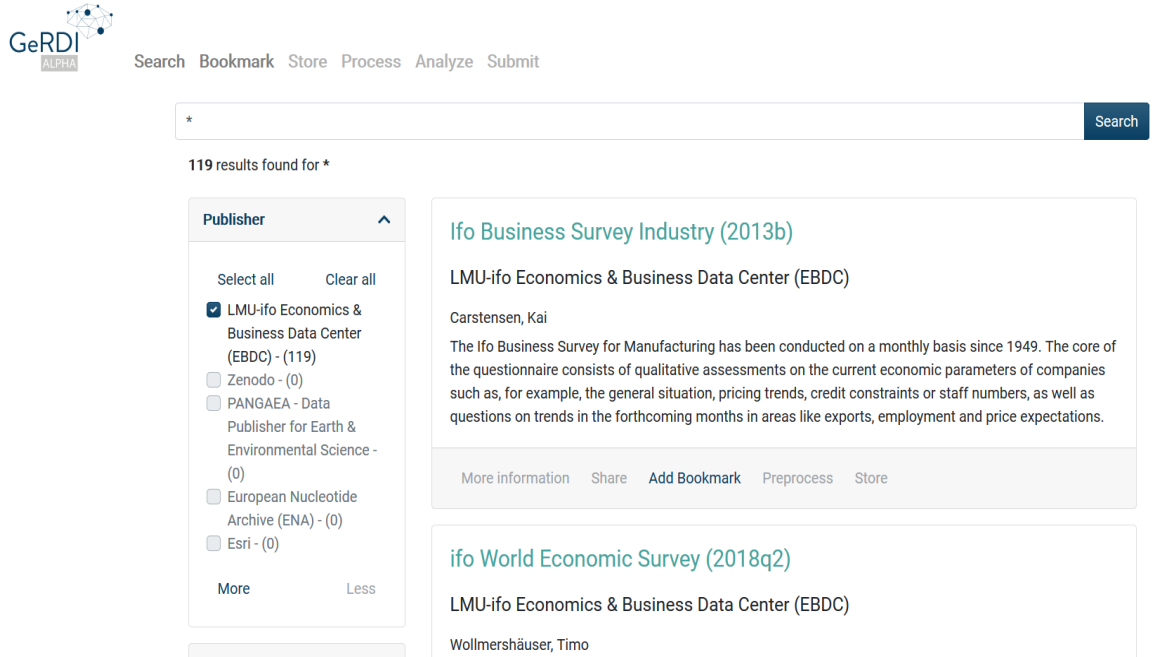
- Nunmehr: Softwarebasierte Dienstearchitektur („Microservices“), die nicht die Datenbestände selbst zum Gegenstand hat, sondern die Unterstützung prinzipiell gleichförmiger Prozessketten und Einzelschritte wie z.B. das Auffinden, Bookmarken, Teilen, Herunterladen oder Analysieren eines Datensatzes

Dienstebasierte FD-Infrastruktur: GeRDI

GeRDI
Generic Research Data Infrastructure



Dienstebasierte FD-Infrastruktur: GeRDI



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

t.borst@zbw.eu
https://twitter.com/redsys
http://orcid.org/0000-0002-2481-029X