Data Provenance in the Research Lifecycle: Report from the Trenches

+ some thoughts on reproducibility when data are not publishable

Lars Vilhuber

Cornell University

February 2021

The opinions expressed in this talk are solely the authors, and do not represent the views of the U.S. Census Bureau, the American Economic Association, or any of the funding agencies.

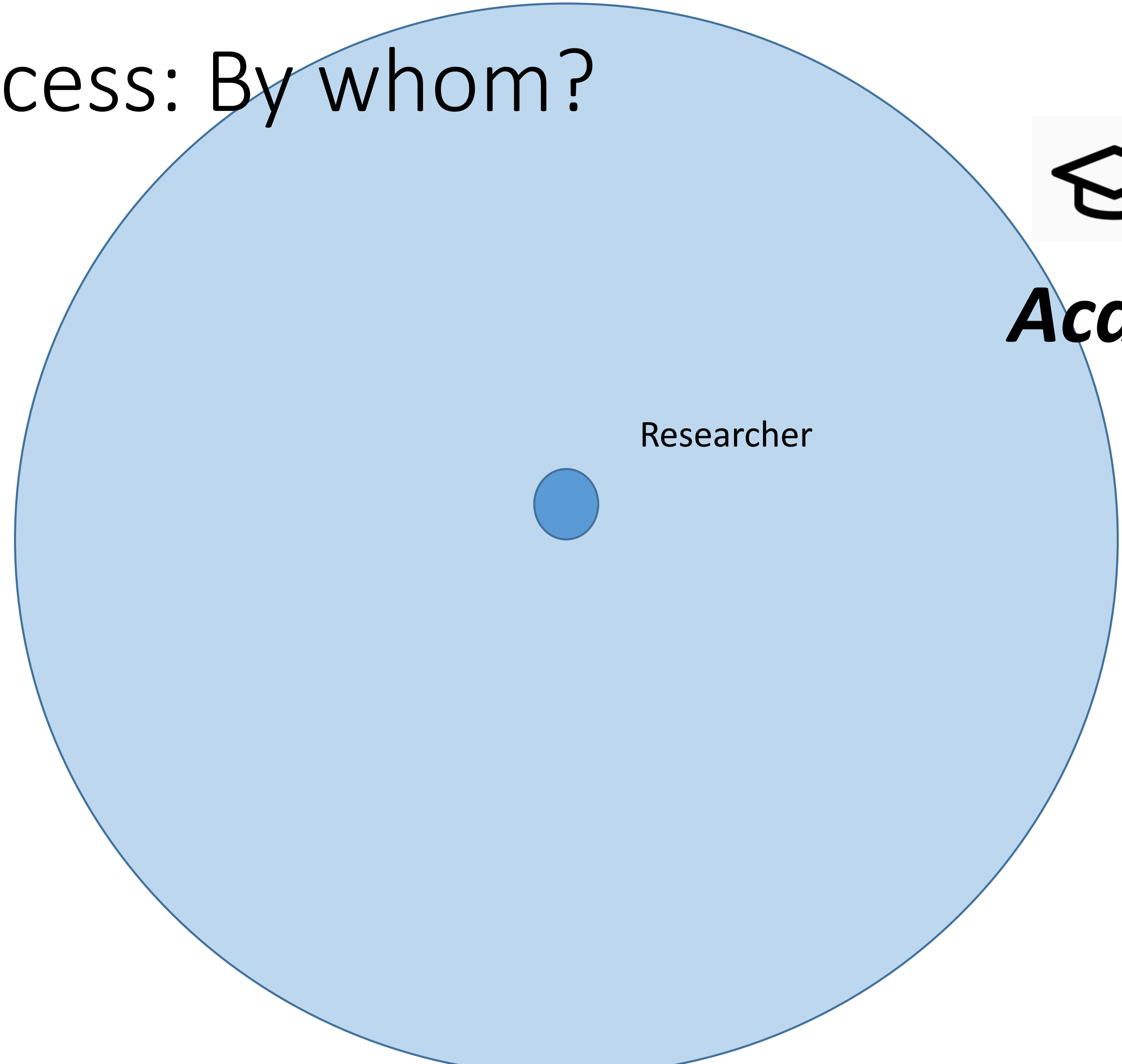
© Lars Vilhuber (cc) BY-NC

Access: By whom?

Researcher

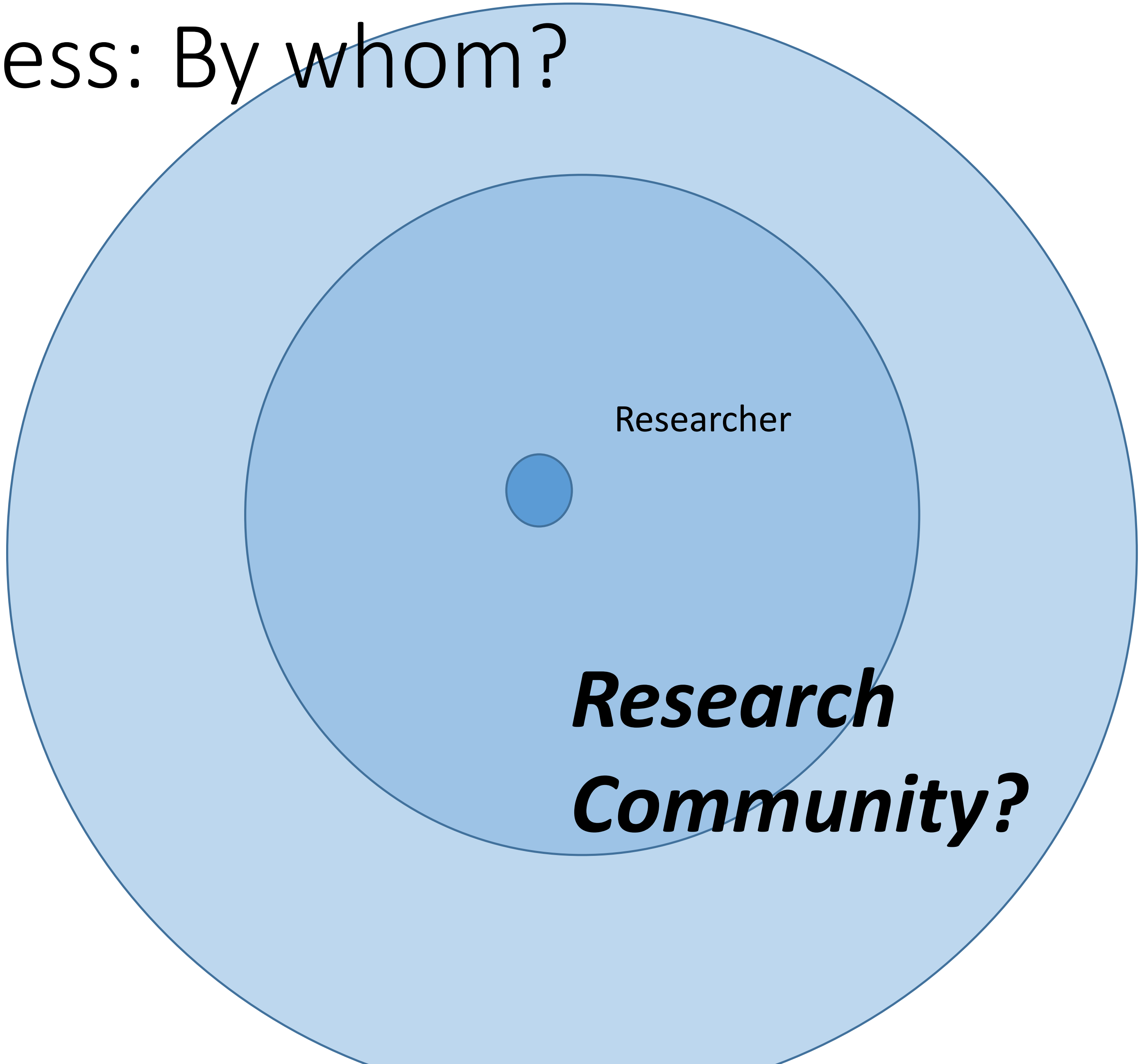
Access: By whom?

Access: By whom?

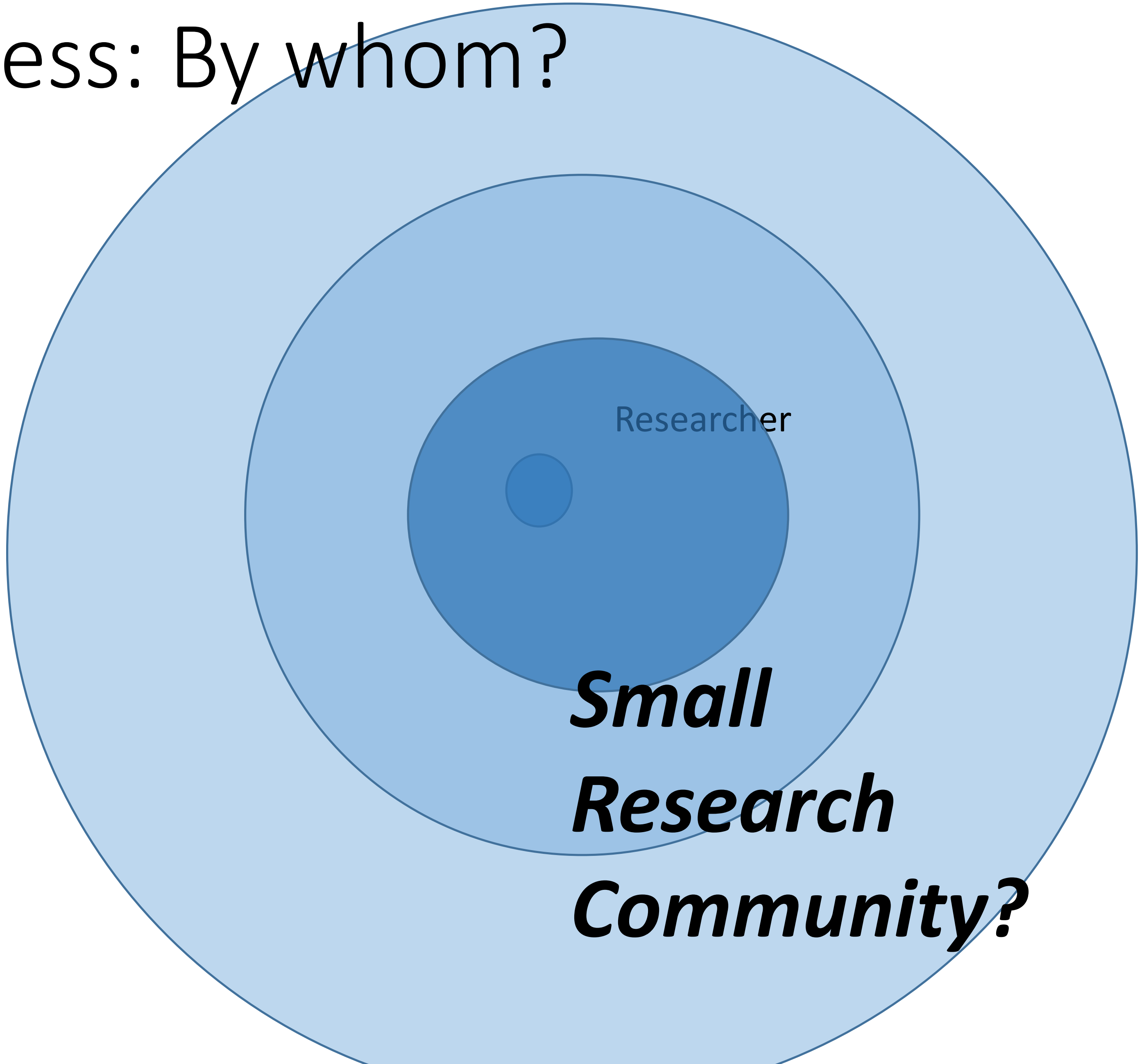


Academia?

Access: By whom?



Access: By whom?

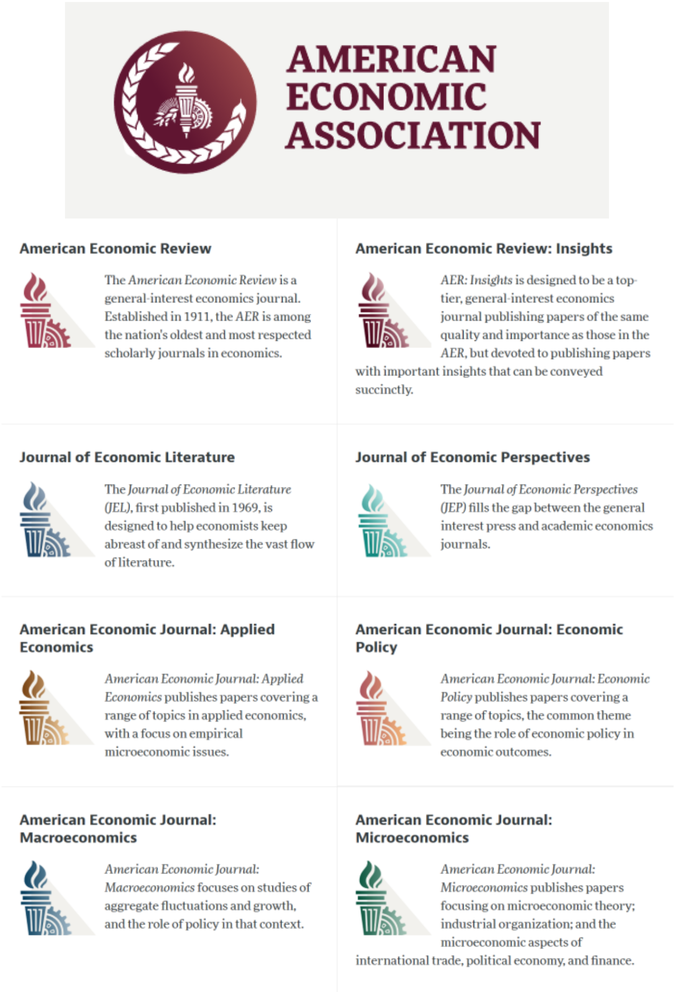


Themes

- Importance of tiered access
  - Recognition that not all data can be made “public”
  - How to better handle access requests
- Support for private datasets
  - Importance of non-academic data providers (commercial, ad-hoc)
  - Importance of university data retention policies and infrastructure
- Meta-repositories
  - Or: how to see all deposits that relate to a journal’s articles
- APIs:
  - Better information about deposits
  - Easier computational integration

Current efforts at the AEA

- **Pre-emptively improve code archives**
  - By conducting reproducibility checks when we can
  - By working with groups that conduct reproducibility checks when we cannot
- **Better archives**
  - Greater transparency of the code and data archives
- **Better provenance tracking**
  - Leave code where it is when appropriate
  - Leave data where it is almost always
  - Display that information



AMERICAN ECONOMIC ASSOCIATION

American Economic Review

The *American Economic Review* is a general-interest economics journal. Established in 1911, the *AER* is among the nation's oldest and most respected scholarly journals in economics.

American Economic Review: Insights

AER: Insights is designed to be a top-tier, general-interest economics journal publishing papers of the same quality and importance as those in the *AER*, but devoted to publishing papers with important insights that can be conveyed succinctly.

Journal of Economic Literature

The *Journal of Economic Literature (JEL)*, first published in 1969, is designed to help economists keep abreast of and synthesize the vast flow of literature.

Journal of Economic Perspectives

The *Journal of Economic Perspectives (JEP)* fills the gap between the general interest press and academic economics journals.

American Economic Journal: Applied Economics

American Economic Journal: Applied Economics publishes papers covering a range of topics in applied economics, with a focus on empirical microeconomic issues.

American Economic Journal: Economic Policy

American Economic Journal: Economic Policy publishes papers covering a range of topics, the common theme being the role of economic policy in economic outcomes.

American Economic Journal: Macroeconomics

American Economic Journal: Macroeconomics focuses on studies of aggregate fluctuations and growth, and the role of policy in that context.

American Economic Journal: Microeconomics

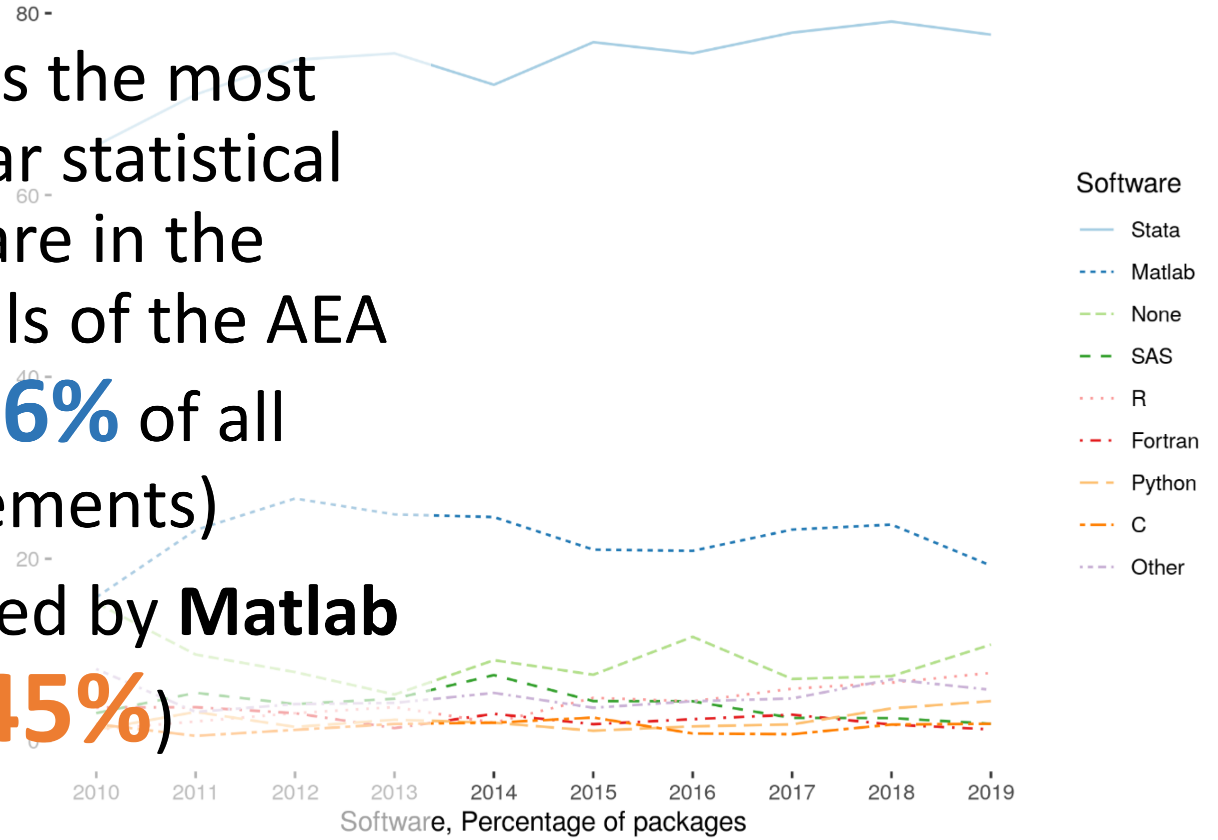
American Economic Journal: Microeconomics publishes papers focusing on microeconomic theory; industrial organization; and the microeconomic aspects of international trade, political economy, and finance.

Stats on reproduced articles

Between July 16, 2019, and yesterday, the AEA Data Editor team conducted

- **~1156 assessments**
- for ~**675 manuscripts**

- **Stata** is the most popular statistical software in the journals of the AEA (**72.96%** of all supplements)
- followed by **Matlab** (**22.45%**)



Themes

- Importance of tiered access
  - Recognition that not all data can be made “public”
  - How to better handle access requests
- Support for private datasets
  - Importance of non-academic data providers (commercial, ad-hoc)
  - Importance of university data retention policies and infrastructure

Meta-repositories

- Or: how to see all deposits that relate to a journal’s articles

APIs:

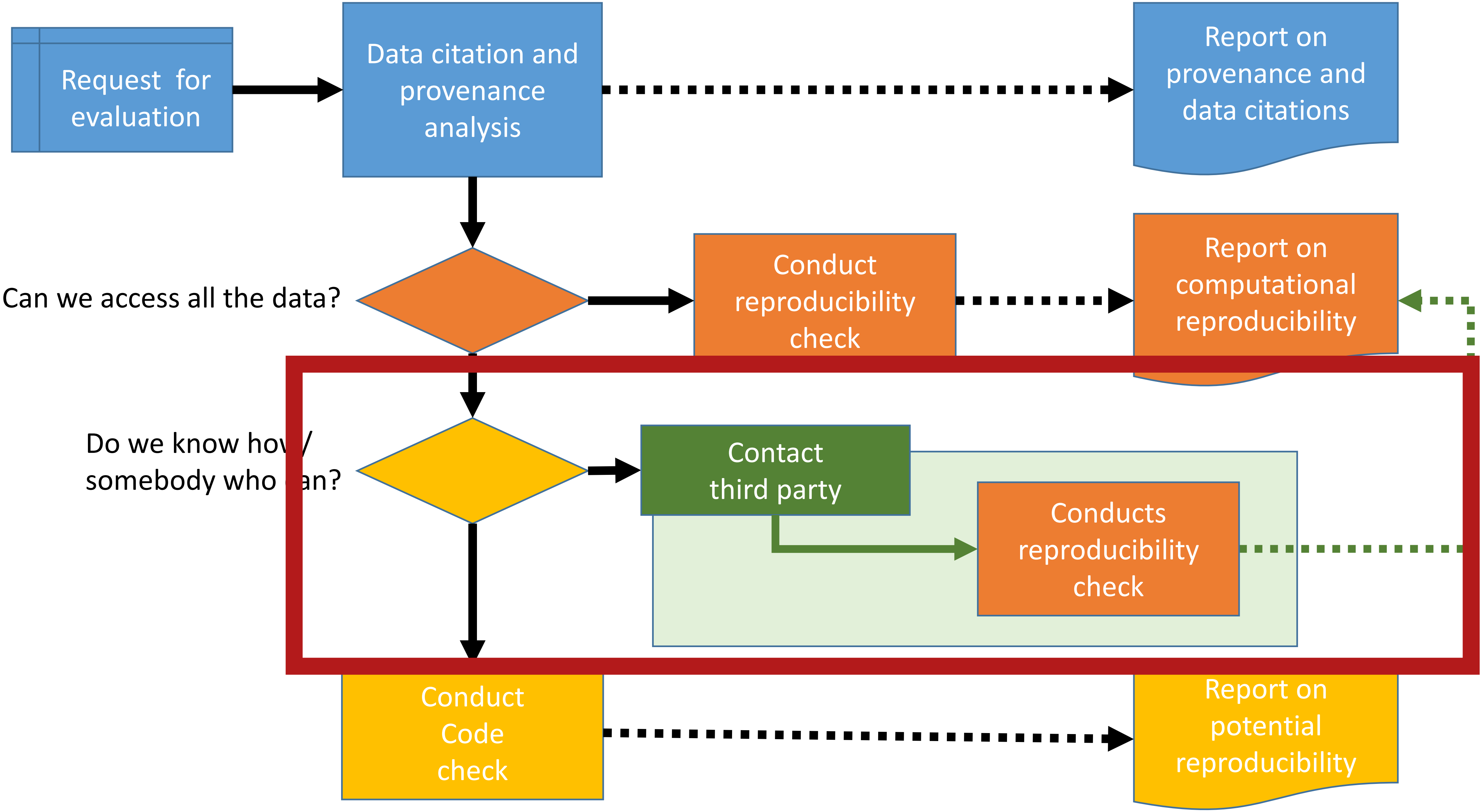
- Better information about deposits
- Easier computational integration

AEA Data & Code Availability Policy (2019)

- It is the policy of the American Economic Association to publish papers only if the data used in the analysis are **clearly and precisely** documented and **access to the data and code is clearly and precisely documented and is non-exclusive to the authors.**
- Authors of accepted papers that contain empirical work, simulations, or experimental work must **provide**, **prior to acceptance**, the data, programs, and other details of the computations **sufficient to permit replication**, as well as **information about access to data and programs.**

AEA Pre-Publication Verification

- Every paper that receives a “conditional acceptance” is verified
  - *Data citations*
  - *Quality of README (Data Availability Statement)*
  - *Quality of code*
  - *Reproducibility of code*
  - *Quality of metadata in the repository*



Request for evaluation

Data citation and provenance analysis

Report on provenance and data citations

Can we access all the data?

Conduct reproducibility check

Report on computational reproducibility

Do we know how / somebody who can?

Contact third party

Conducts reproducibility check

Conduct Code check

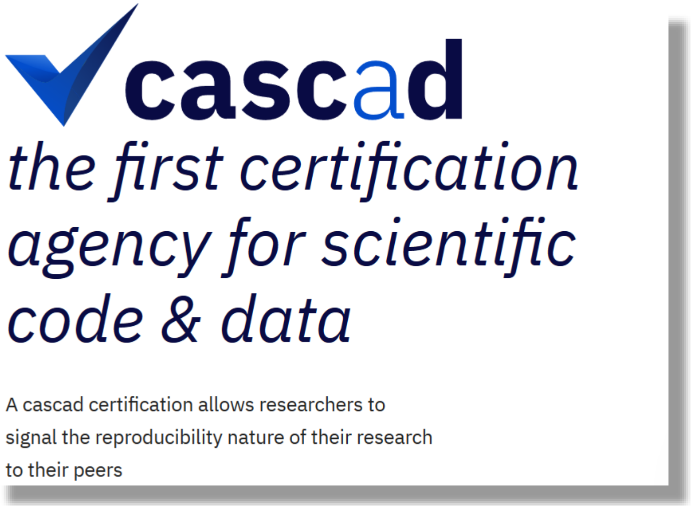
Report on potential reproducibility

Restricted-access data

- Between 20% and 40% of papers have some sort of access restrictions
  - NDA (but Data Editor allowed)
  - DUA + IRB (in principle possible, but not quickly)
  - Click-through license with redistribution restriction (often not respected)
  - Commercial data

We work with 3rd parties

- cascad – Using confidential data!
  - DADS
  - French customs data
- CISER (R-Squared)
- Various contributors with access to confidential data
  - Authors
  - Their institutions
  - Network of known groups with access (less frequent)

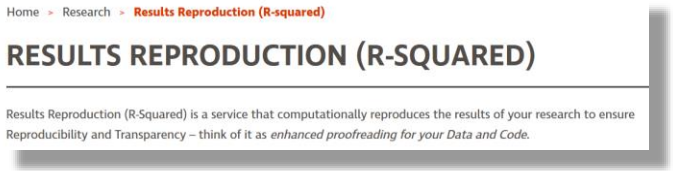


cascad

the first certification agency for scientific code & data

A cascad certification allows researchers to signal the reproducibility nature of their research to their peers

CISER CORNELL INSTITUTE for Social and Economic Research



Home > Research > Results Reproduction (R-squared)

RESULTS REPRODUCTION (R-SQUARED)

Results Reproduction (R-Squared) is a service that computationally reproduces the results of your research to ensure Reproducibility and Transparency – think of it as *enhanced proofreading for your Data and Code.*

Themes

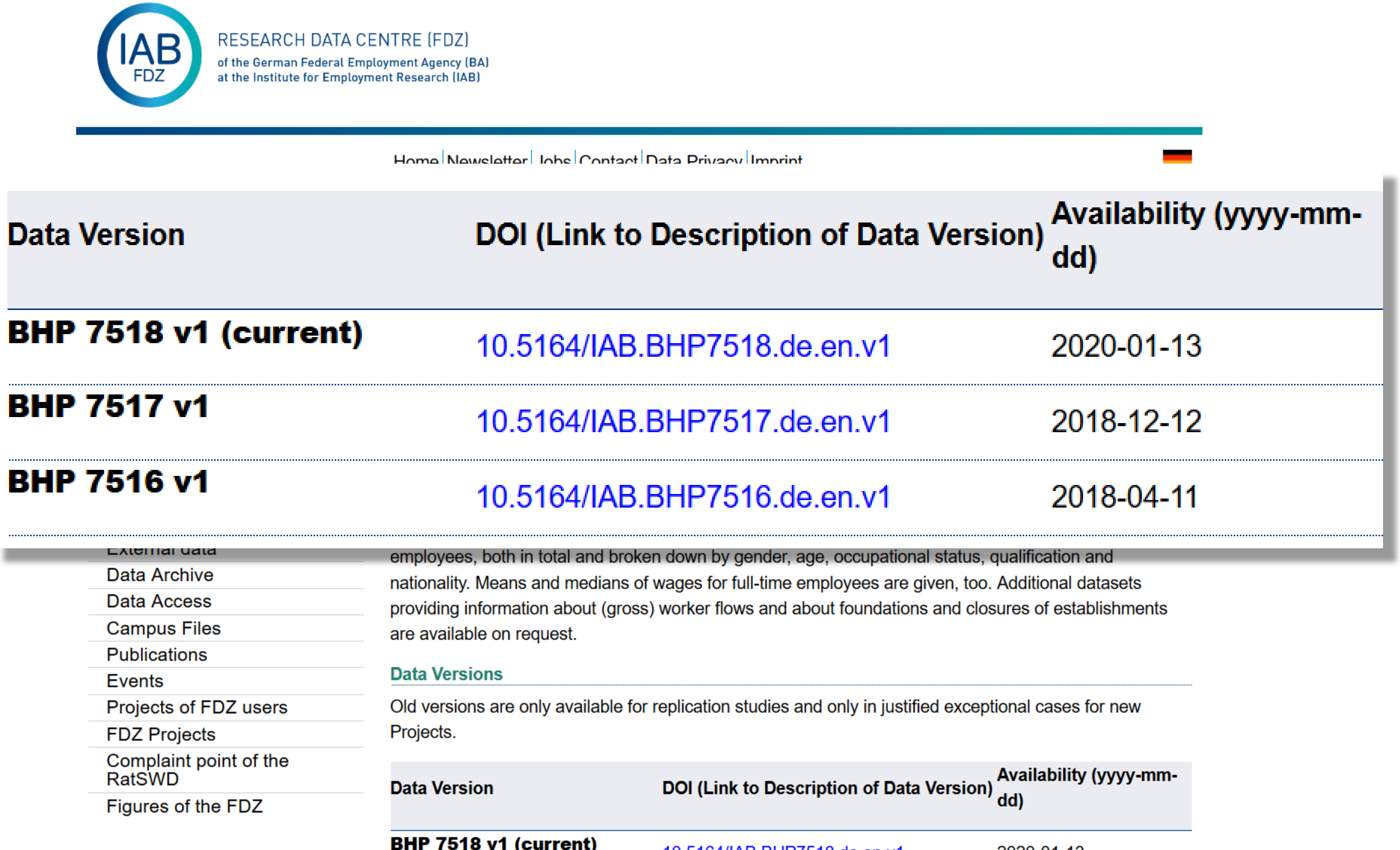
- Importance of tiered access
  - Recognition that not all data can be made “public”
  - How to better handle access requests
- Support for private datasets
  - Importance of non-academic data providers (commercial, ad-hoc)
  - Importance of university data retention policies and infrastructure

- **IRBs** need to recognize role of replicators
- **Agreements (DUA, NDA, Licenses)** need to incorporate the possibility of legitimate 3rd party access
- **Repositories** need to efficiently handle legitimate access requests
  - Possibly pre-publication
  - Not just for editors

Themes

- Importance of tiered access
  - Recognition that not all data can be made "public"
  - How to better handle access requests
- Support for private datasets
  - Importance of non-academic data providers (commercial, ad-hoc)
  - Importance of university data retention policies and infrastructure
- Meta-repositories
  - Or: how to see all deposits that relate to a journal's articles
- APIs:
  - Better information about deposits
  - Easier computational integration

Non-repository institution: German Restricted-access



RESEARCH DATA CENTRE (FDZ)
of the German Federal Employment Agency (BA)
at the Institute for Employment Research (IAB)

Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13
BHP 7517 v1	10.5164/IAB.BHP7517.de.en.v1	2018-12-12
BHP 7516 v1	10.5164/IAB.BHP7516.de.en.v1	2018-04-11

employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are available on request.

Data Versions

Old versions are only available for replication studies and only in justified exceptional cases for new Projects.

Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13

Citing restricted-access data

“Well, I can’t download the data, so I can’t cite it.”

A role for institutions?

Pitch:

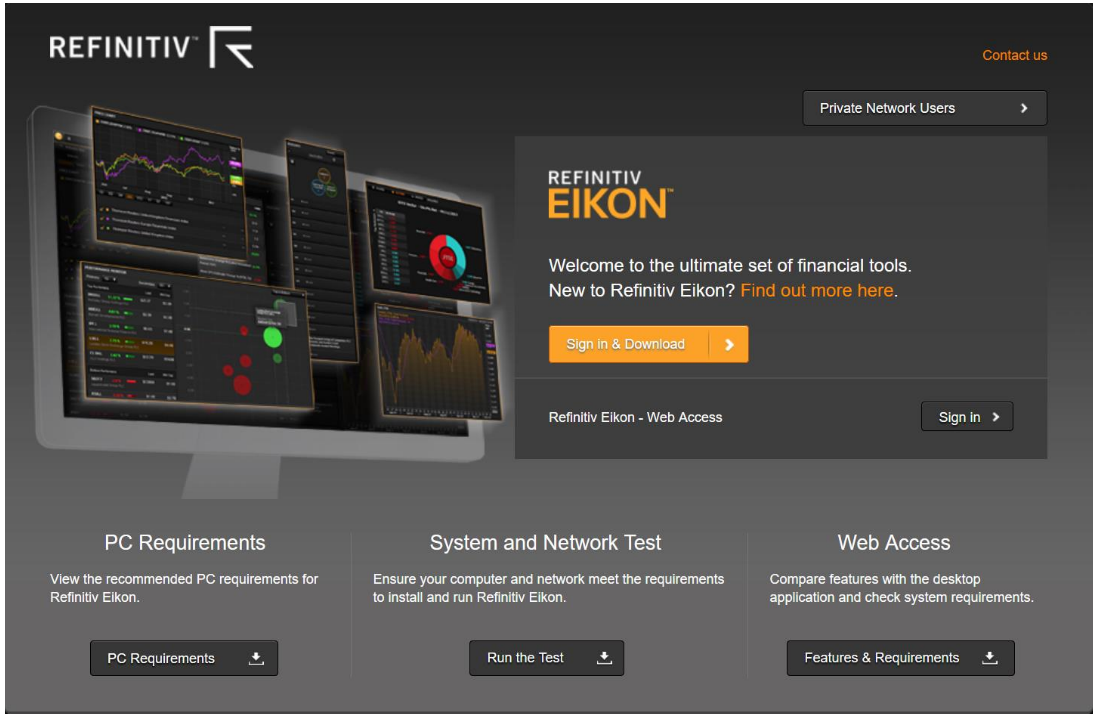
- Institutions can (must?) provide infrastructure for archives
  - **Internal-only (dark)**
  - **Internal with public metadata** (FAIR!)
  - Public
  - Mix of the above
- Institutions provide researchers with information (citation, access protocols)

University
Data Retention
Policies!

How did you get the data in first place?

- You **applied** for the data **through a process**
- You **purchased** the data from a provider
- You signed an **Non-Disclosure Agreement (NDA)** with a company
- Your **university** has an **agreement** with a data provider

…



Welcome to the ultimate set of financial tools.
New to Refinitiv Eikon? Find out more here.

PC Requirements

View the recommended PC requirements for Refinitiv Eikon.

System and Network Test

Ensure your computer and network meet the requirements to install and run Refinitiv Eikon.

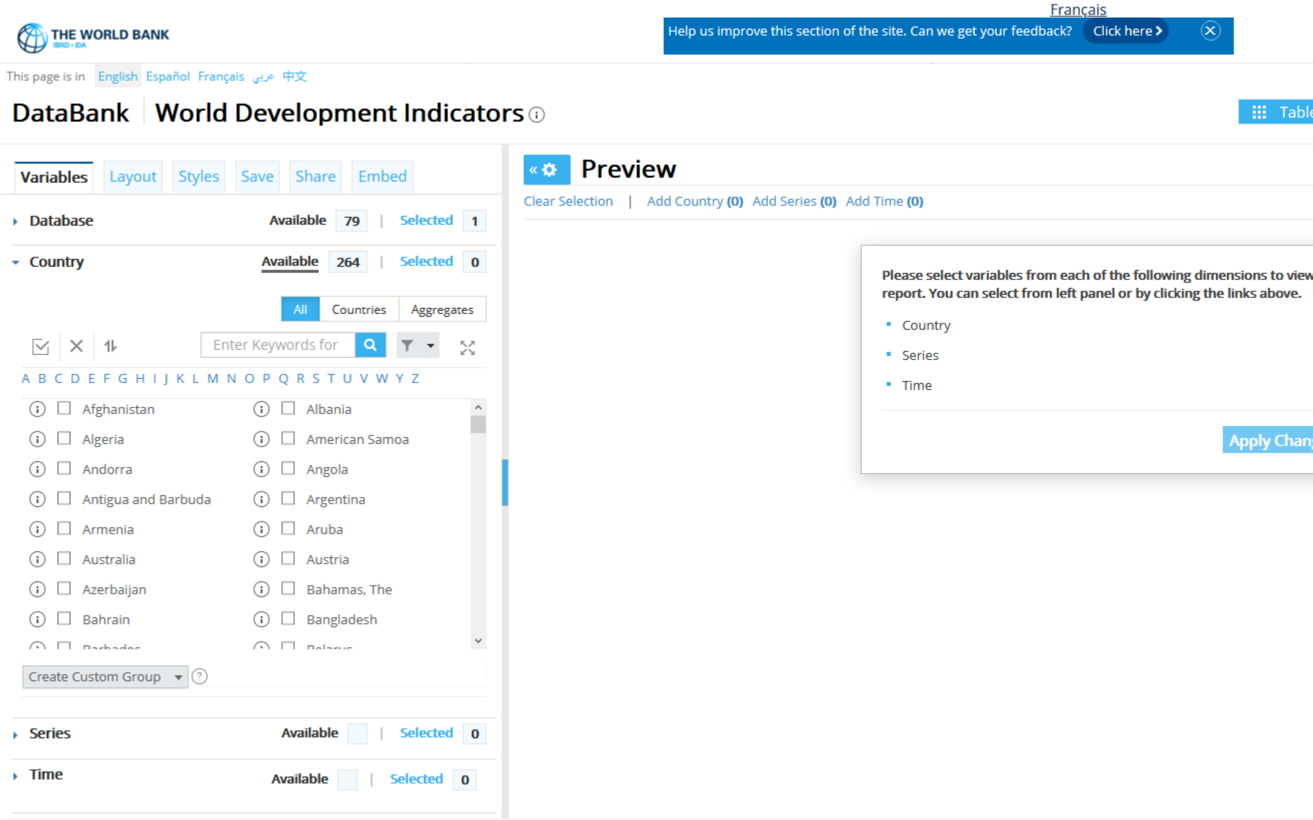
Web Access

Compare features with the desktop application and check system requirements.

You must have described the data

- You must have **named** the dataset you wanted
- You downloaded the data from from an **online query system**
- You **specified the extract** from a company database
(in words, in SQL, etc.)

…



Français

THE WORLD BANK

Help us improve this section of the site. Can we get your feedback? Click here ›

This page is in English Español Français عربي 中文

DataBank World Development Indicators ⓘ

Table

Variables Layout Styles Save Share Embed

› Database	Available	79	Selected	1
▾ Country	Available	264	Selected	0

All Countries Aggregates

Enter Keywords for

A B C D E F G H I J K L M N O P Q R S T U V W Y Z

ⓘ ☐ Afghanistan	ⓘ ☐ Albania
ⓘ ☐ Algeria	ⓘ ☐ American Samoa
ⓘ ☐ Andorra	ⓘ ☐ Angola
ⓘ ☐ Antigua and Barbuda	ⓘ ☐ Argentina
ⓘ ☐ Armenia	ⓘ ☐ Aruba
ⓘ ☐ Australia	ⓘ ☐ Austria
ⓘ ☐ Azerbaijan	ⓘ ☐ Bahamas, The
ⓘ ☐ Bahrain	ⓘ ☐ Bangladesh
ⓘ ☐ Barbados	ⓘ ☐ Belarus

Create Custom Group ▾ ?

› Series	Available		Selected	0
› Time	Available		Selected	0

Preview

Clear Selection | Add Country (0) Add Series (0) Add Time (0)

Please select variables from each of the following dimensions to view report. You can select from left panel or by clicking the links above.

- Country
- Series
- Time

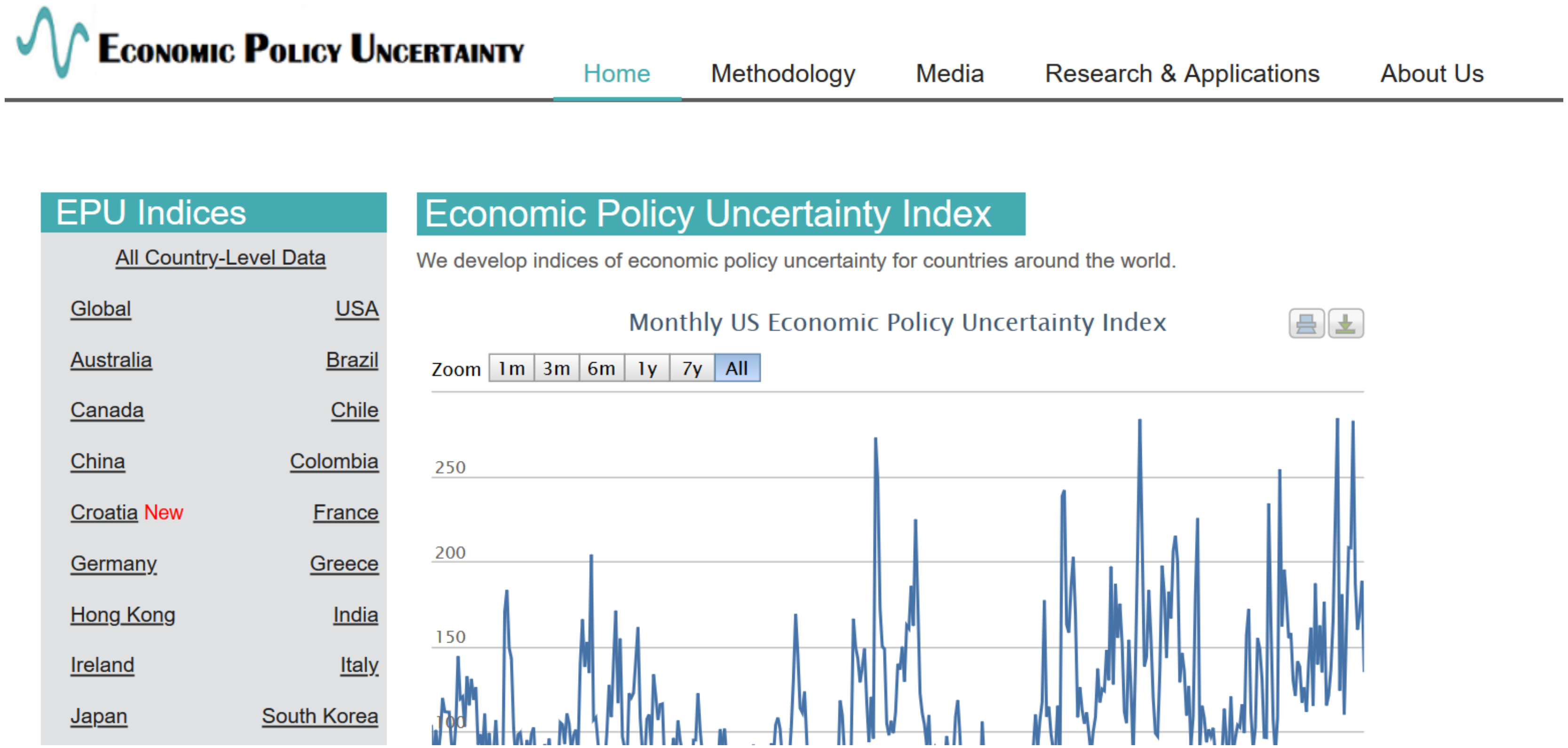
Apply Chan

A role for institutions?

VOTRE PROBLÈME CONCERNE UN SPA! ALLEZ VOIR LE JUGE ADMINISTRATIF
JUDICIAIRE
NON! C'EST UN SPIC! JE NE SUIS PAS COMPÉTENT
ADMINISTRATIF
IL FAUDRAIT DEMANDER AU TRIBUNAL DES CONFLITS
S'IL N'A PAS FAIT DE RECOURS PRÉALABLE, JE NE PEUX RIEN FAIRE
ADMINISTRATIF
JUDICIAIRE

- Institutions (**data librarian!**) can provide researchers with
  - The necessary information about original sources
  - Conveying the necessary **access information**
  - **Arrange for permission** to redistribute extracts
- Data providers can
  - **Provide suggested citations**
  - Be clear about original sources!
  - **Information about persistence**
  - Potentially be clearer about the rights to redistribute

(Semi-)Academic data publishers



ECONOMIC POLICY UNCERTAINTY

EPU Indices

All Country-Level Data

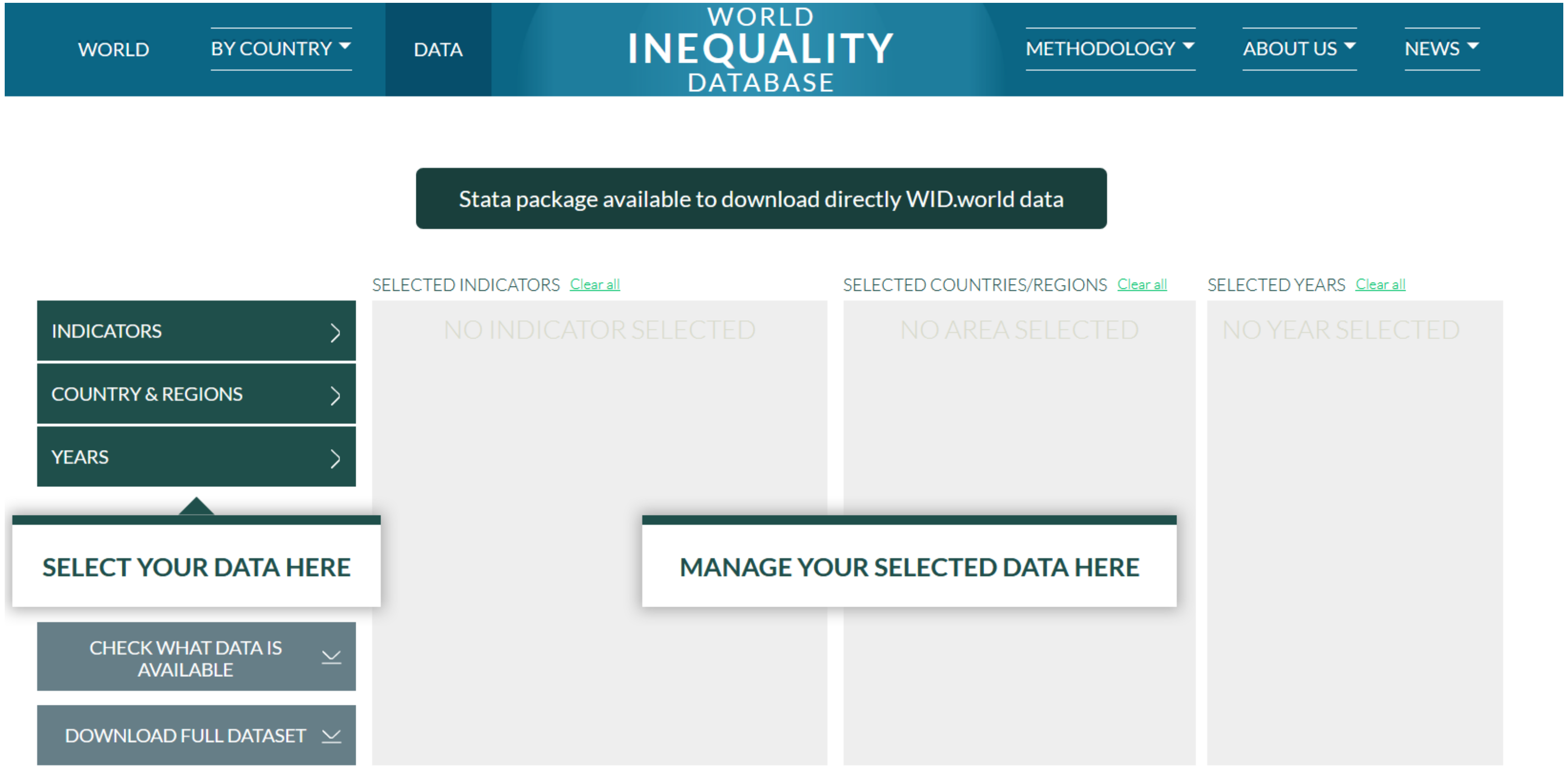
- Global
- Australia
- Canada
- China
- Croatia New
- Germany
- Hong Kong
- Ireland
- Japan
- USA
- Brazil
- Chile
- Colombia
- France
- Greece
- India
- Italy
- South Korea

Economic Policy Uncertainty Index

We develop indices of economic policy uncertainty for countries around the world.

Monthly US Economic Policy Uncertainty Index

(Semi-)Academic data publishers

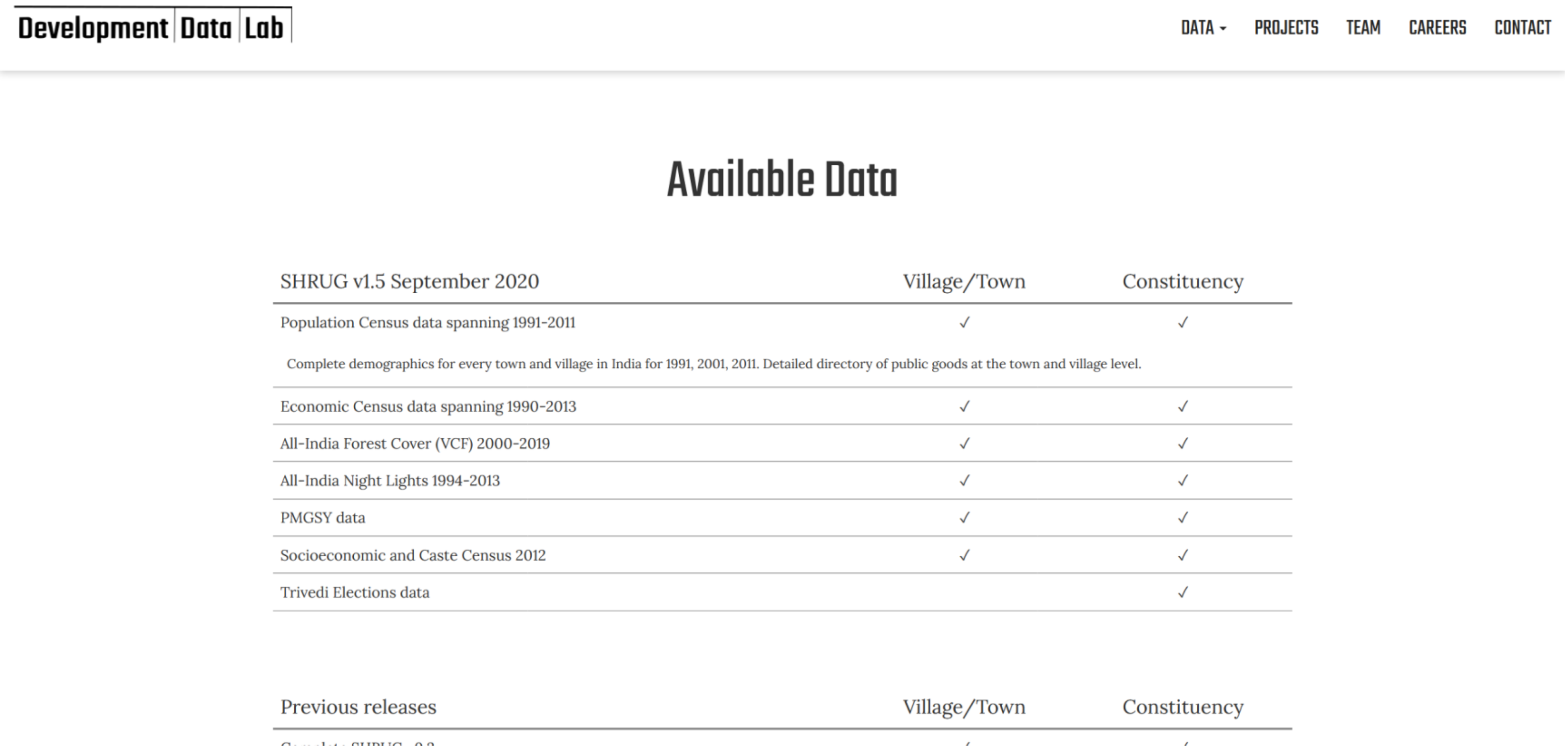


Stata package available to download directly WID.world data

SELECT YOUR DATA HERE

MANAGE YOUR SELECTED DATA HERE

(Semi-)Academic data publishers



Available Data

SHRUG v1.5 September 2020	Village/Town	Constituency
Population Census data spanning 1991-2011	✓	✓
Complete demographics for every town and village in India for 1991, 2001, 2011. Detailed directory of public goods at the town and village level.		
Economic Census data spanning 1990-2013	✓	✓
All-India Forest Cover (VCF) 2000-2019	✓	✓
All-India Night Lights 1994-2013	✓	✓
PMGSY data	✓	✓
Socioeconomic and Caste Census 2012	✓	✓
Trivedi Elections data		✓

Previous releases	Village/Town	Constituency

(Semi-)Academic data publishers

- Only one of those three uses a **trusted repository** for preservation
  - And it is not the largest/best-funded…
- All are lead by university (economists)

Bring them into the fold?

Themes

- Importance of tiered access
  - Recognition that not all data can be made "public"
  - How to better handle access requests
- Support for private datasets
  - Importance of non-academic data providers (commercial, ad-hoc)
  - Importance of university data retention policies and infrastructure

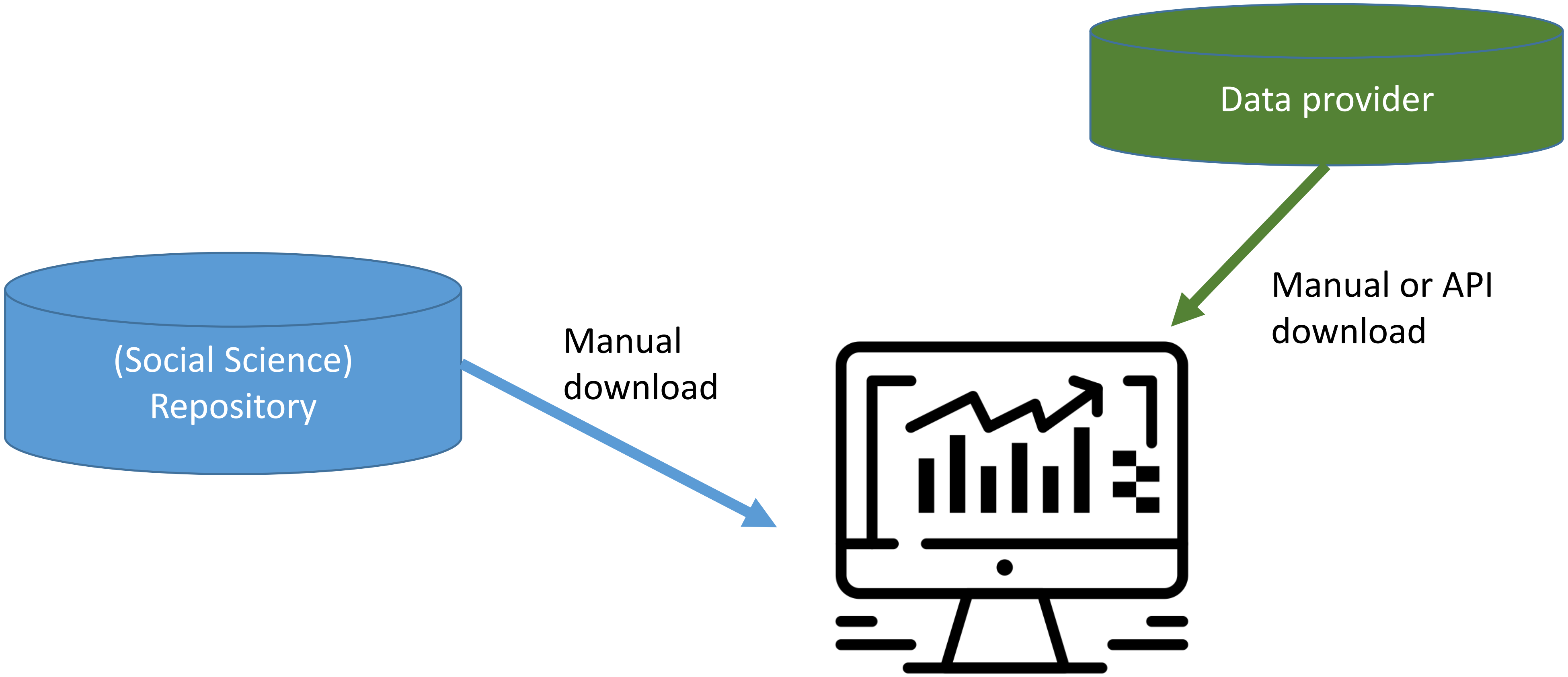
- **Universities** need to create/support this
- **SOMEBODY** needs to more consistently message to commercial/ non-archive repositories the need to
  - Preserve
  - Describe
- **Non-Repositories** that provide widely used data need to get their act together (some already have!)

Themes

- Importance of tiered access
  - Recognition that not all data can be made “public”
  - How to better handle access requests
- Support for private datasets
  - Importance of non-academic data providers (commercial, ad-hoc)
  - Importance of university data retention policies and infrastructure

- Meta-repositories
  - Or: how to see all deposits that relate to a journal’s articles
- APIs:
  - Better information about deposits
  - Easier computational integration

API integration



Data provider

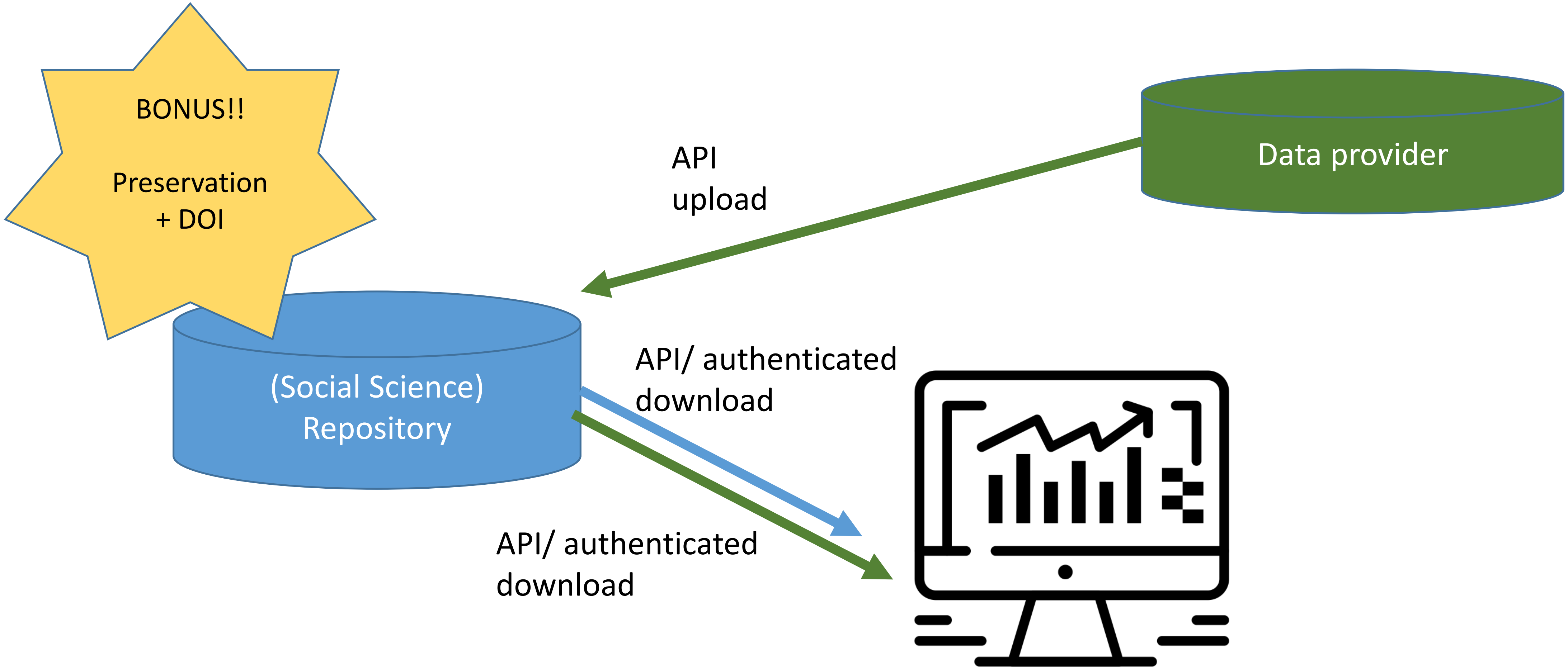
Manual or API download

(Social Science) Repository

Manual download

Created by Flatart from Noun Project

API integration



BONUS!!

Preservation + DOI

Data provider

API upload

(Social Science) Repository

API/ authenticated download

API/ authenticated download

Created by Flatart from Noun Project

Themes

You have to have APIs

You have to have person-level machine-authentication

Ability to list individual files

Ability to interact with computational systems

- Meta-repositories
  - Or: how to see all deposits that relate to a journal's articles
- APIs:
  - Better information about deposits
  - Easier computational integration

Themes

Better support for "mixed" deposits (data + code)

- Meta-repositories
  - Or: how to see all deposits that relate to a journal's articles
- APIs:
  - Better information about deposits
  - Easier computational integration

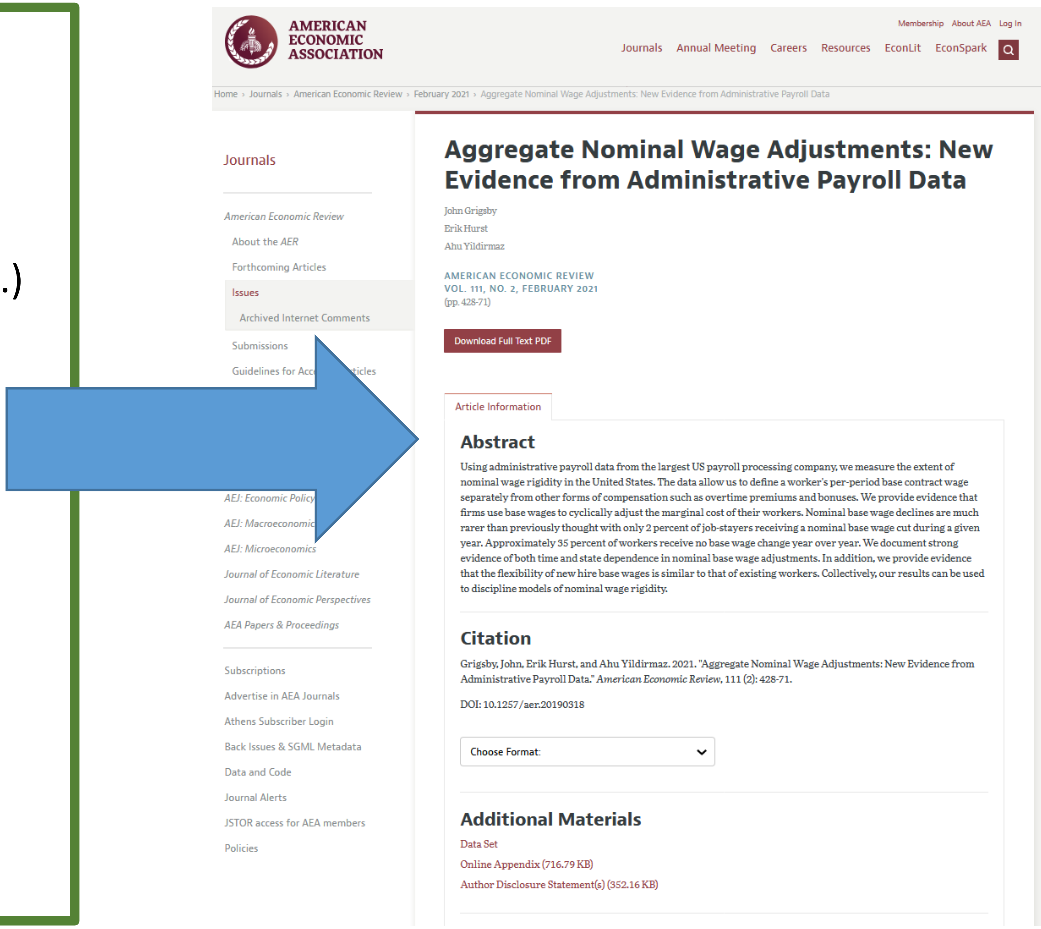
Themes

- Importance of tiered access
  - Recognition that not all data can be made "public"
  - How to better handle access requests
- Support for private datasets
  - Importance of non-academic data providers (commercial, ad-hoc)
  - Importance of university data retention policies and infrastructure

- Meta-repositories
  - Or: how to see all deposits that relate to a journal's articles
- APIs:
  - Better information about deposits
  - Easier computational integration

Dispersed

• Private data at University X	• DOI?
• German data at IAB	• DOI!
• IPUMS data	• DOI! (Mmmh…)
• Survey data at PSID	• No DOI
• Survey data at GSOEP	• DOI!
• Code at openICPSR	• DOI!



Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data

John Grigsby
Erik Hurst
Ahu Yildirmaz

AMERICAN ECONOMIC REVIEW
VOL. 111, NO. 2, FEBRUARY 2021
(pp. 428-71)

Abstract

Using administrative payroll data from the largest US payroll processing company, we measure the extent of nominal wage rigidity in the United States. The data allow us to define a worker's per-period base contract wage separately from other forms of compensation such as overtime premiums and bonuses. We provide evidence that firms use base wages to cyclically adjust the marginal cost of their workers. Nominal base wage declines are much rarer than previously thought with only 2 percent of job-stayers receiving a nominal base wage cut during a given year. Approximately 35 percent of workers receive no base wage change year over year. We document strong evidence of both time and state dependence in nominal base wage adjustments. In addition, we provide evidence that the flexibility of new hire base wages is similar to that of existing workers. Collectively, our results can be used to discipline models of nominal wage rigidity.

Citation

Grigsby, John, Erik Hurst, and Ahu Yildirmaz. 2021. "Aggregate Nominal Wage Adjustments: New Evidence from Administrative Payroll Data." *American Economic Review*, 111 (2): 428-71.

DOI: 10.1257/aer.20190318

Additional Materials

Data Set
Online Appendix (716.79 KB)
Author Disclosure Statement(s) (352.16 KB)

Fake example!

This is what data citations should do!

```
Select *

Where (
C.Relation = “isSupplementedBy”,

C.RelationID = D.DOI,

C.Journal = “AER” )

from CrossRef  as C,

      DataCite as D,

Order by C.DOI
```

Supplements for AER:

- Dataset A (for Article 1)
- Code C (for Article 1)
- Private data (for Article 2)
- German data (for Article 2, 5 ,7)
- …

But standard data citations don’t

- Citation link is not supplement link

Themes

- Journals need to code (and report) richer relations
- Repositories need to display dynamic relations (reward good journal reporting behavior)
- Authors may need to contribute to linking
- Non-repository data providers need to create DOIs, support data citations

- Meta-repositories
  - Or: how to see all deposits that relate to a journal’s articles
- APIs:
  - Better information about deposits
  - Easier computational integration

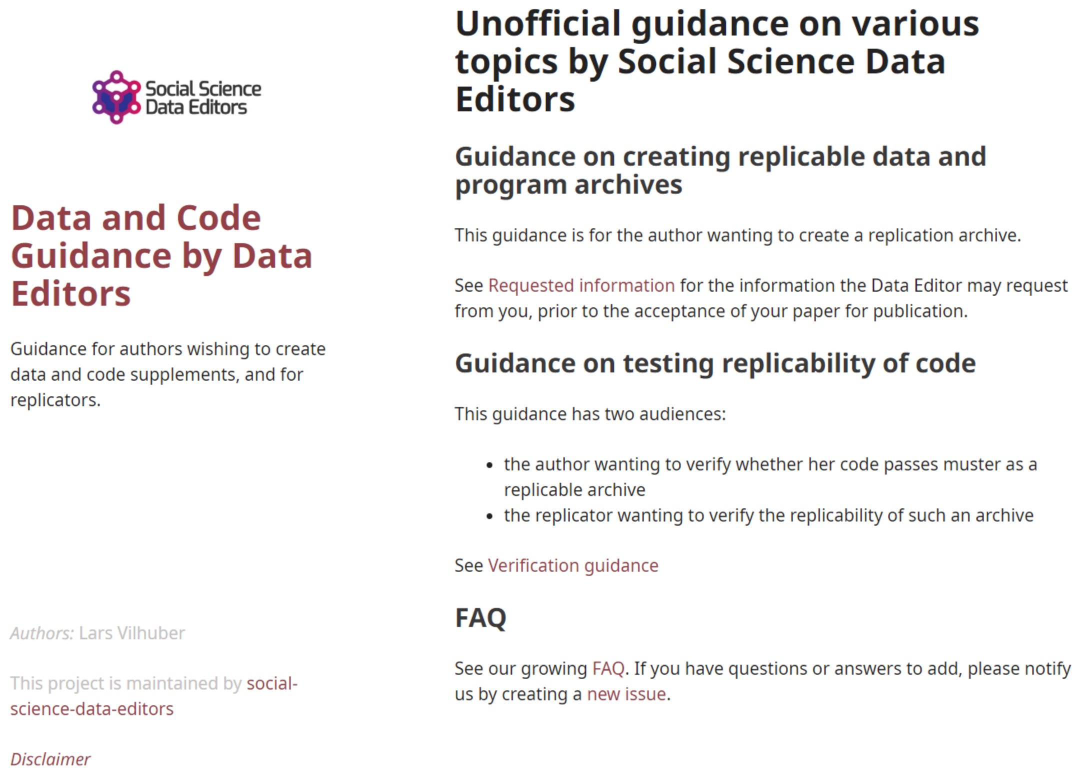
The role for journals

Goal: Transportability

Any standards, tools, methods: must be transportable across journals (no custom solutions)

Social science “guild”

Social Science Data Editors



Social Science Data Editors

Data and Code Guidance by Data Editors

Guidance for authors wishing to create data and code supplements, and for replicators.

Authors: Lars Vilhuber

This project is maintained by social-science-data-editors

Disclaimer

Unofficial guidance on various topics by Social Science Data Editors

Guidance on creating replicable data and program archives

This guidance is for the author wanting to create a replication archive.

See Requested information for the information the Data Editor may request from you, prior to the acceptance of your paper for publication.

Guidance on testing replicability of code

This guidance has two audiences:

- the author wanting to verify whether her code passes muster as a replicable archive
- the replicator wanting to verify the replicability of such an archive

See Verification guidance

FAQ

See our growing FAQ. If you have questions or answers to add, please notify us by creating a new issue.

https://social-science-data-editors.github.io/guidance/

Thank you!