Transparency and Reproducibility in Economics: Context, and Lessons learned from 1,000 papers

Lars Vilhuber

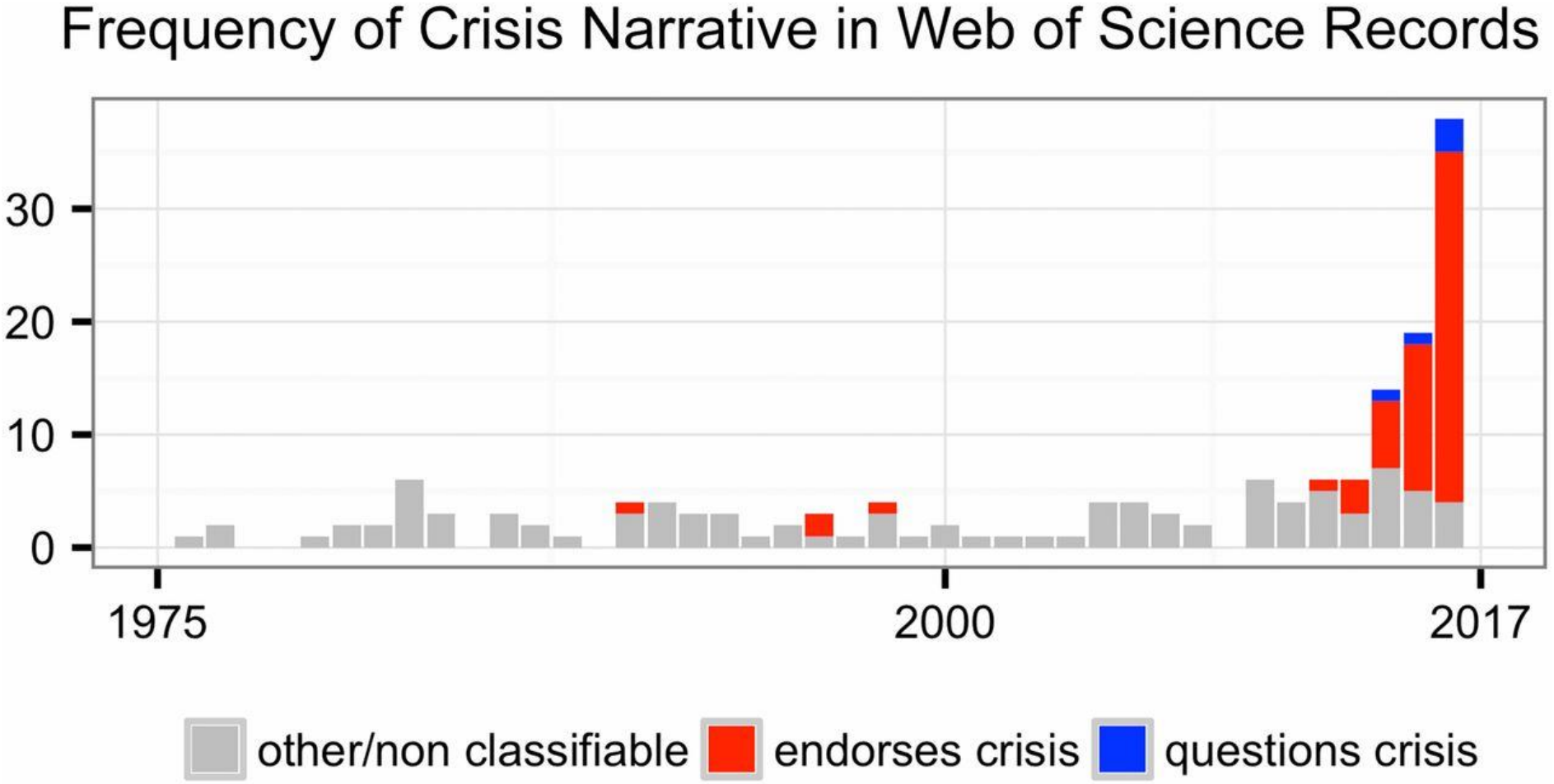
Cornell University

World Bank, 2022-10-04

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.

Context

This reproducibility crisis thing….



https://doi.org/10.1073/pnas.1708272114

The “crisis” in the 60s and 70s

Sterling, 1959; Cohen, 1962; Lykken, 1968; Tukey, 1969; Greenwald, 1975; Meehl, 1978; Rosenthal, 1979

Low power

Flexibility in analysis

Selective reporting

Ignoring nulls

Lack of replication

Misuse of statistics

Source: Nosek Sackler talk 2017

But some things HAVE changed!

A bit of history…

Progress

- Replication archives and Data (Code) Availability policies

1985!

Progress

- Replication archives and Data (Code) Availability policies
- Shared open source software

Statistical Software Components

From Boston College Department of Economics
Boston College, 140 Commonwealth Avenue, Chestnut Hill MA 02467 U
Contact information at EDIRC.
Bibliographic data for series maintained by Christopher F Baum (baum@

Access Statistics for this software series.
Track citations for all items by RSS feed
Is something missing from the series or not right? See the RePEc data
series.

GAPPORT: Stata module to calculates seats in party-list representation downloads
Ulrich Kohler

GCLSORT: Stata module to sort a single variable via ege
Philippe Van Kerm

GPROD: Stata module to extend egen for product of obs
Philip Ryan

Progress

- Replication archives and Data (Code) Availability policies
- Shared open source software
- Better public-use and shared data

CASD

CRDCN Canadian Research Data Centre Network

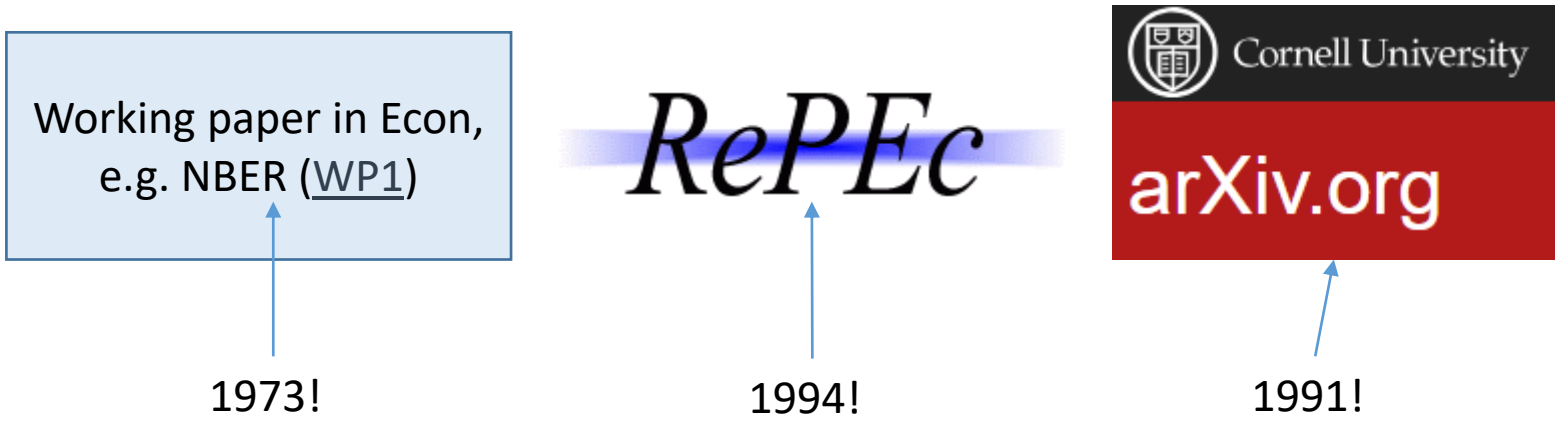
FSRDC

IAB INSTITUT FÜR ARBEITSMARKT- UND BERUFSFORSCHUNG
Die Forschungseinrichtung der Bundesagentur für Arbeit

CIQSS | QICSS
CENTRE INTERUNIVERSITAIRE QUÉBÉCOIS DE STATISTIQUES SOCIALES
QUEBEC INTER-UNIVERSITY CENTER FOR SOCIAL STATISTICS

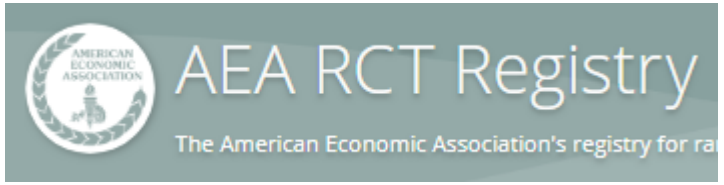
Progress

- Replication archives and Data (Code) Availability policies
- Shared open source software
- Better public-use and shared data
- Better ways of accessing preprints/ grey literature



Progress

- Replication archives and Data (Code) Availability policies
- Shared open source software
- Better public-use and shared data
- Better ways of accessing preprints/ grey literature
- Pre-registration of trials, experiments, and analyses

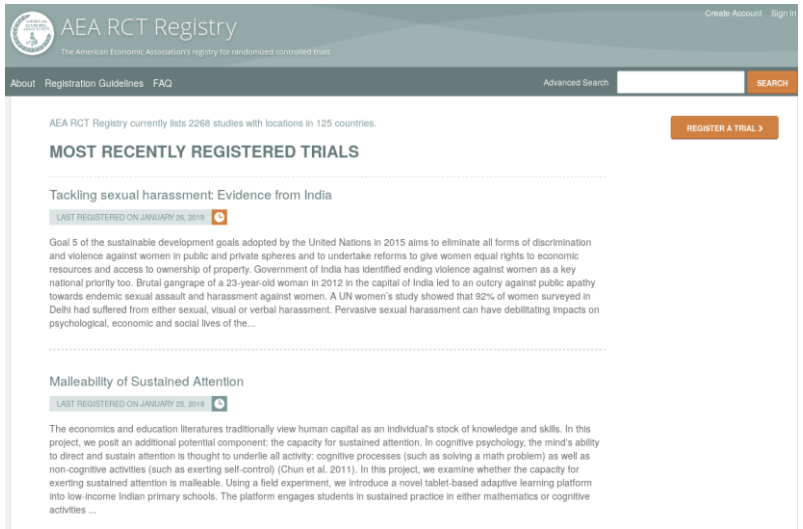


Second round (2012-)

- **Greater enforcement of data (and code) availability**
  - 2015, AJ Political Science
  - 2016, Data Editor for ASA Software Section
  - 2016, Statistical review added Science
  - 2017: AEA appoints Data Editor, with mandate to do similar activities
    (also EJ, Restud)

Pre-registration

- "That information is especially helpful in research that emphasizes **null hypothesis significance testing**.
- A thorough preregistration promotes transparency and openness and **protects researchers from suspicions of p-hacking**."



https://www.psychologicalscience.org/publications/psychological_science/preregistration

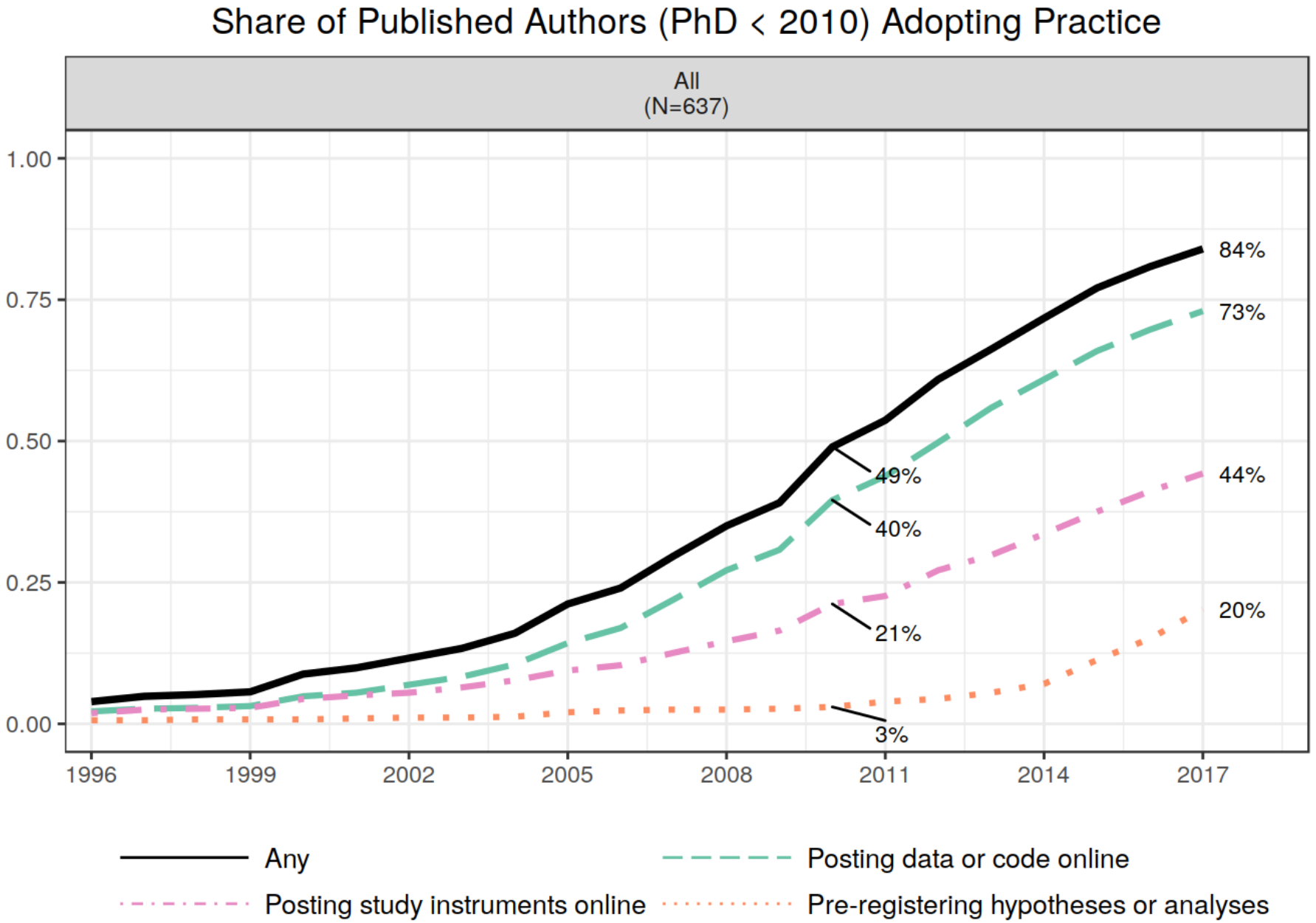
Registered Reports

- https://cos.io/rr
- Chambers (2014)
- Nosek & Lakens (2014)

- Close cousin: Results-blind review

Preprints in other sciences

- bioRxiv (2013)
- PsyArXiv (2016)



Paluck (2018) https://osf.io/kvbnh/

So we’re good, right?

Hm….

A bit of background

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.

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.**

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

Action: Reproducibility Check

Social Science Data Editors

Data and Code Guidance by Data Editors

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

Verification guidance

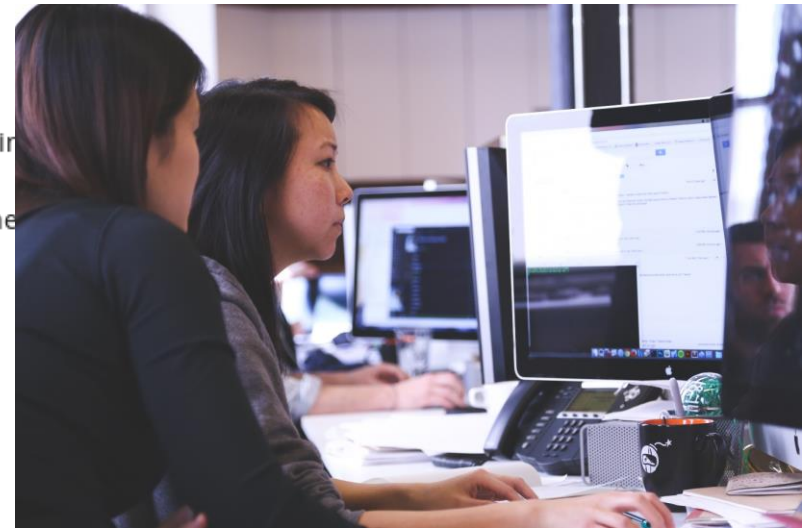
On this page:

- Overview
- Review the README file
- For each listed data source
- For each listed table, figure, in-text number
- Conduct a code verification, if data is available
- Examples

Overview

This document describes

- what authors should check before providir journals
- what verifier teams should check for in the to them for the purpose of verification



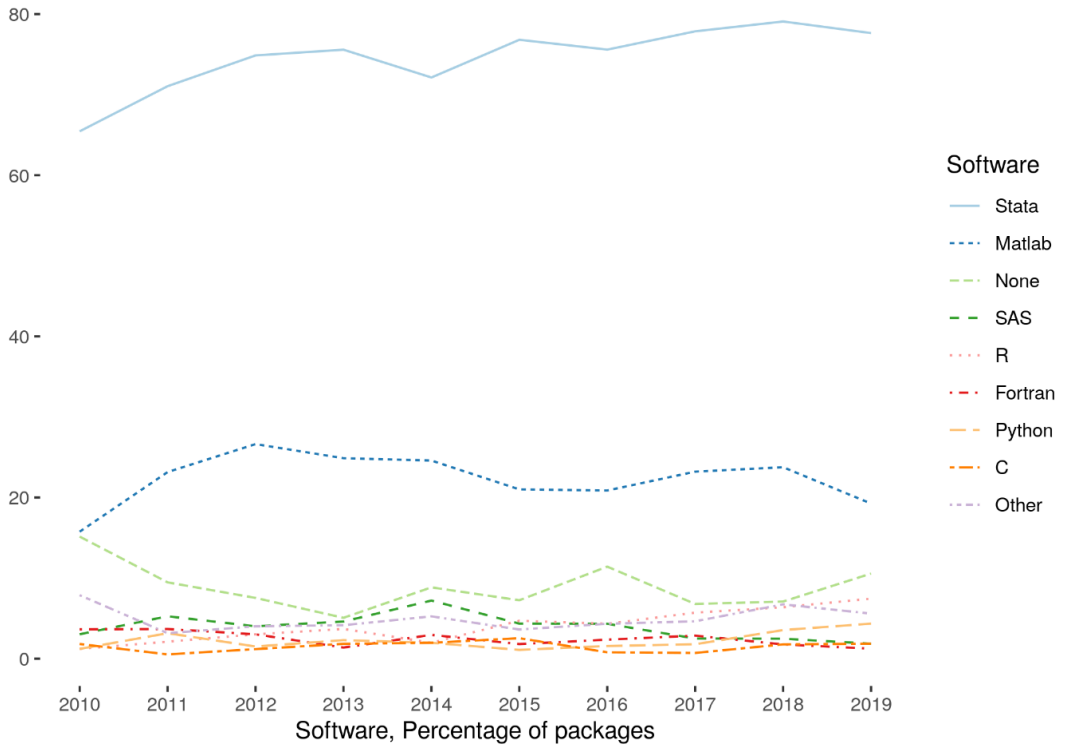
Stats on reproduced articles

Between July 16, 2019, and June 20, 2022, the AEA Data Editor team conducted

- **1900 assessments**
- for **1050 manuscripts** (full papers)

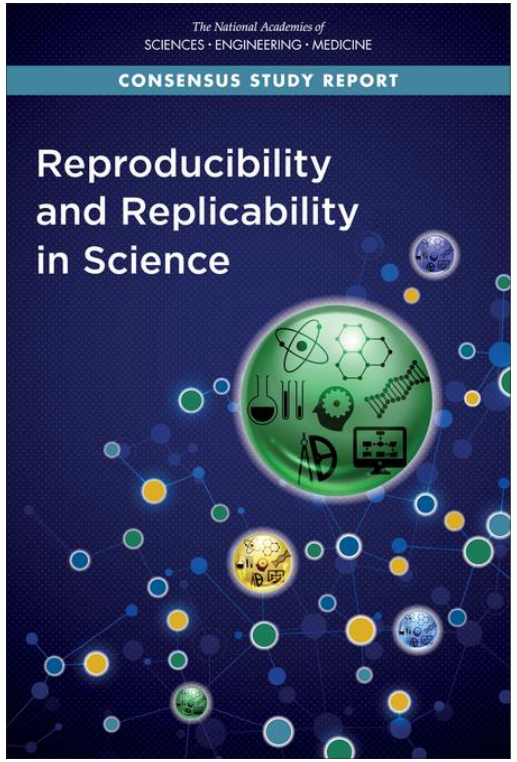
Very little diversity in software

- **Stata** is the most popular statistical software in the journals of the AEA (**72.96%** of all supplements, 2010-2019)
- followed by **Matlab** (**22.45%**)



Replication continuum

https://doi.org/10.17226/25303



Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replication continuum

Same data	Same code	Same methods	Same context

Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

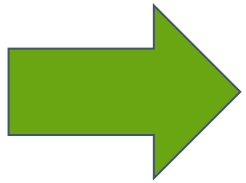
What does reproducibility buy you?

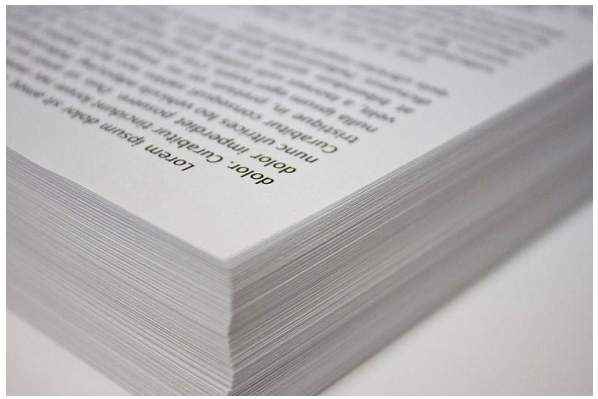
- Results in the paper were actually produced by the code (as claimed)

- Replication package has a complete and exhaustive description of what is needed to produce those results

Ingredients of “research”

1. “Procedures” = computer code
2. “Materials (1)” = data
3. “Materials (2)” = computers





What does reproducibility NOT buy you?

- Results in the paper are “right”

- Replication package will work in two months time

Replication continuum

Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

Replication continuum

Same data	**Different code or software**	Same methods	Same context

Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

Replication continuum

New data	Same code	Same methods	Same context
collection			

Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

Replication continuum

Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

Generalizability

- Wider Replication (Pesaran 2003)
- Scientific Replication (Hamermesh 2007)
- Reanalysis/Robustness (Clemens 2015)

Replication continuum

Different data	Different code	**Different**	**Different**
	or software	**methods**	**context or**
			country

Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

Generalizability

- Wider Replication (Pesaran 2003)
- Scientific Replication (Hamermesh 2007)
- Reanalysis/Robustness (Clemens 2015)

What does reproducibility ALSO buy you?

- It may be worth your while to start thinking about replicating or generalizing the results! (*Plausibility*)

Lessons?

Observation 0

Researchers don’t…

- Re-run their code before submitting
- Don’t streamline (automate) enough
- Are not careful about how they document data sources
- Fail to curate their own data

Back in 2019…

Failure to curate

Poor citation practices

- **Macrodata:**

"We use data downloaded from the Bureau of Economic Analysis…"

- **Microdata**:

"… this paper uses data from the Current Population Survey…"

Poor coding practices

- **Manual/non-automation**

Code produces no meaningful output

- **Lack of robustness**:

Bugs in the code

Lessons!

Computational empathy

Lesson 1: Computational empathy

In the words of the slogan popularized by Buckheit and Donoho (1995),

> ***"a scientific publication is [...] merely advertising of the scholarship: [...] the complete software development environment and the complete set of instructions which generated the figures."***

Lesson 1: Computational empathy

Put yourself in the position of the reader of the research compendium:

- Can they understand those instructions?
- Under what premises/ shared common knowledge?
- What might they assume about the computing environment?
- How concise or diffuse are the instructions?

Lesson 1: Computational empathy

Potential readers

- **You** *(in 4 years, between prepping 2 new courses, an R&R, a new child, and tenure coming up in 2 years)*
- Your RA *(in 4 years, because you are… see above)*
- Your future readers who will cite you *(in 4-10 years, who may want to extend or replicate your study, but won't if it is too complex)*

Lesson 1: Computational empathy

= “Pity the poor replicator”

Intermezzo

Observation 1

Social scientists do not read the manual

(beyond the first few pages)

Observation 1: Please read the manual

Persistent misconceptions

- About setting **working directories**
- How to record **pathnames**
- How to leverage **loops**
- How to leverage **subroutines**
- How to pass **parameters**
- How (and if) to use **controller scripts**

Observation 2

Social scientists
love
point-and-click interfaces
(which are hard to reproduce)

Observation 2: point-and-click interfaces

This is reflected in

- **GIS (maps)** that appear in papers
- **Data extraction** tools
- How to **run software** (any software)

Observation 1 and 2 are the result of a
lack of Computational Empathy,
and lead to
high burden
of reproducibility and replicability

Solutions?

Hold that thought, we will get there.

Data acumen

Data acumen

"the ability to understand data, to make good judgments about and good decisions with data, and to use data analysis tools responsibly and effectively"

National Academies of Sciences, Engineering, and Medicine. 2018. Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104.

Lesson 2: Data acumen in the context of reproducibility

Two key components

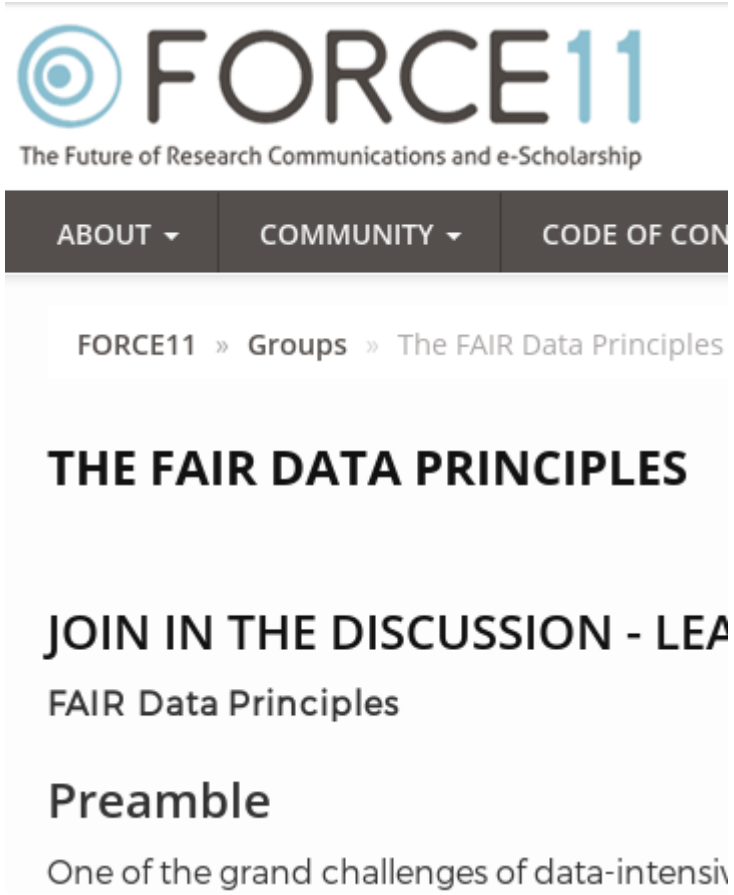
- Data provenance
  - Where did the data come from which I used?
- Data preservation
  - Where do I put the data I generated?
  - What if the data I used are not “robustly preserved”?
  - What do you mean by that?

Data provenance

Action: Data citations and metadata

What is **FAIR**?

- **F**indable,
- **A**ccessible,
- **I**nteroperable, and
- **R**e-usable



OPEN ICPSR Find Data Share Data openICPSR Repositories Management

dataeditor@aeapubs.org

Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

Principal Investigator(s): Amanda Gregg, Middlebury College

Version: V1

AMERICAN ECONOMIC ASSOCIATION

Name	File Type		Last Modified
1894MicroData.xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	4.5 MB	08/08/2019 11:01:AM

Project Citation:

Gregg, Amanda. Imperial Russian Factory Database, 1894-1908. Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-01-29. https://doi.org/10.3886/E110681V1

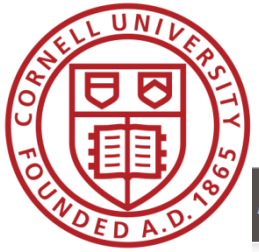
AG_Corp_CleaningandDatabaseCompiler.do	text/x-stata-syntax	23.4 KB	08/08/2019 11:02:AM

Related Publications

The following publications are supplemented by the data in this project.

- Gregg, Amanda. "Factory Productivity and the Concession System of Incorporation in Late Imperial Russia, 1894-1908." *American Economic Review* 110, no. 2 (February 2020): 401–27. https://doi.org/10.1257/aer.20151656.

08:55:AM



FORCE11

The Future of Research Communications and e-Scholarship

Search

English

ABOUT | COMMUNITY | CODE OF CONDUCT | GROUPS | RESOURCES | NEWS + BLOGS | EVENTS | PUBLIC

perceived criteria of importance.

1. Importance

Data should be considered legitimate, citable products of research. Data should be accorded the same importance in the scholarly record as citat research objects, such as publications[1].

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and le attribution to all contributors to the data, recognizing that a single style or of attribution may not be applicable to all data[2].

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited[3].

4. Unique Identification

A data citation should include a persistent method for identification that i actionable, globally unique, and widely used by a community[4].

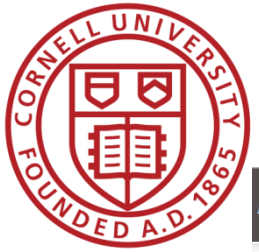
5. Access

Data citations should facilitate access to the data themselves and to such

DC1

Data Citation Principles

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].



FORCE11

The Future of Research Communications and e-Scholarship

Search

English

ABOUT | COMMUNITY | CODE OF CONDUCT | GROUPS | RESOURCES | NEWS + BLOGS | EVENTS | PUBLIC

perceived criteria of importance.

1. Importance

Data should be considered legitimate, citable products of research. Data should be accorded the same importance in the scholarly record as citat research objects, such as publications[1].

2. Credit and Attribution

DC1

Data Citation Principles

1 **Bureau of Labor Statistics.** 2000–2010. "Current Employment Statistics: Colorado, Total Nonfarm, Seasonally adjusted - SMS08000000000000001." United States Department of Labor. http://data.bls.gov/cgi- bin/surveymost?sm+08 (accessed February 9, 2011).

corresponding data should be cited[3].

4. Unique Identification

A data citation should include a persistent method for identification that i actionable, globally unique, and widely used by a community[4].

5. Access

Data citations should facilitate access to the data themselves and to such

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final].

Observation 3

Social scientists are not trained to cite data

Would you buy a car from this guy?

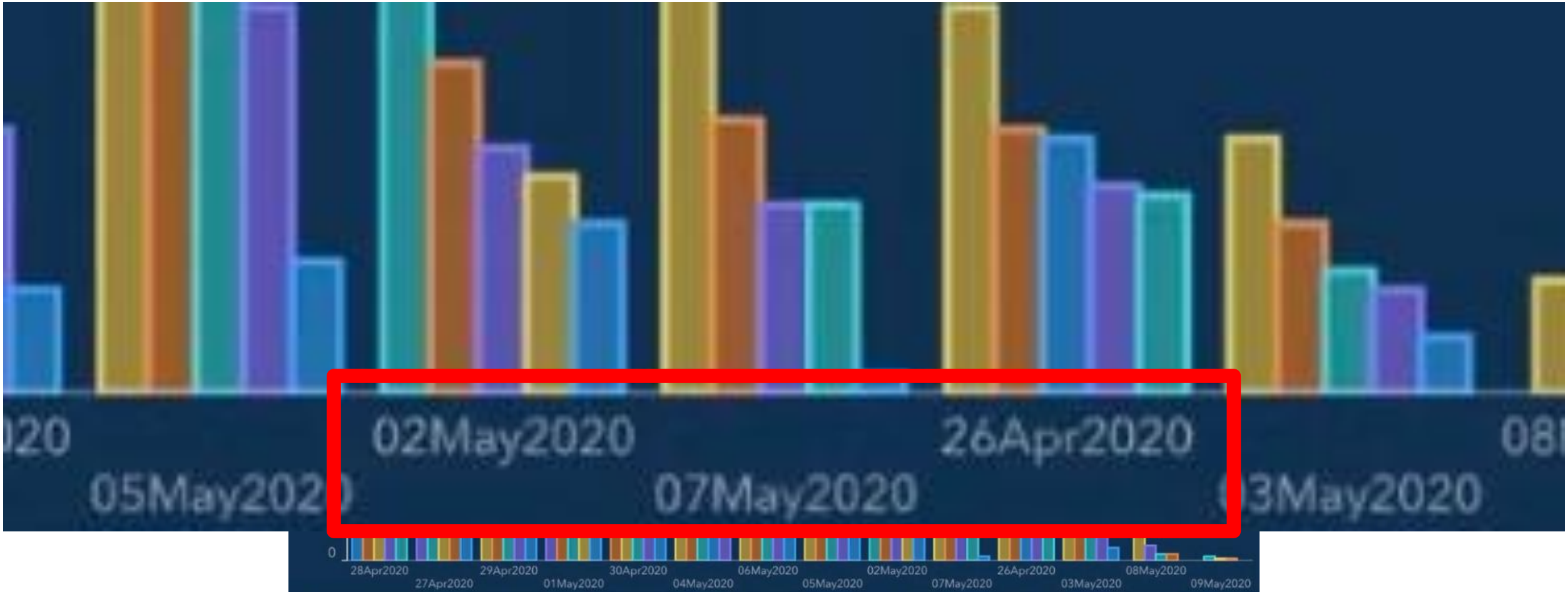
Provenance!

- Does the sales person have a good record?
- Where does the car come from?
- What do we know about the car?

Would you use this data?

```
0000000 cfd0 e011 b1a1 e11a 0000 0000 0000 0000
0000010 0000 0000 0000 0000 003e 0003 fffe 0009
0000020 0006 0000 0000 0000 0000 0000 0004 0000
0000030 008f 0000 0000 0000 1000 0000 fffe ffff
0000040 0000 0000 fffe ffff 0000 0000 008b 0000
0000050 008c 0000 008d 0000 008e 0000 ffff ffff
0000060 ffff ffff ffff ffff ffff ffff ffff ffff
*
0000200 0809 0010 0600 0005 209a 07cd c0c9 0000
0000210 0306 0000 00e1 0002 04b0 00c1 0002 0000
0000220 00e2 0000 005c 0070 0001 4c00 2020 2020
0000230 2020 2020 2020 2020 2020 2020 2020 2020
*
0000290 2020 2020 2020 2020 0042 0002 04b0 0161
00002a0 0002 0000 013d 0002 0001 009c 0002 000e
00002b0 0019 0002 0000 0012 0002 0000 0013 0002
00002c0 0000 01af 0002 0000 01bc 0002 0000 003d
00002d0 0012 0000 000f 3f1b 27f6 0038 0000 0000
00002e0 0001 0258 0040 0002 0000 008d 0002 0000
00002f0 0022 0002 0000 000e 0002 0001 01b7 0002
```

Or would you trust this data?



Provenance!

- Does the provider have a good record?
- Where do the data come from?
- What do we know about the data?

Metadata!

Would you use this data?

```
0000000 cfd0 e011 b1a1 e11a 0000 0000 0000 0000
0000010 0000 0000 0000 0000 003e 0003 fffe 0009
0000020 0006 0000 0000 0000 0000 0000 0004 0000
0000030 008f 0000 0000 0000 1000 0000 fffe ffff
0000040 0000 0000 fffe ffff 0000 0000 008b 0000
0000050 008c 0000 008d 0000 008e 0000 ffff ffff
0000060 ffff ffff ffff ffff ffff ffff ffff ffff
*
0000200 0809 0010 0600 0005 209a 07cd c0c9 0000
0000210 0306 0000 00e1 0002 04b0 00c1 0002 0000
0000220 00e2 0000 005c 0070 0001 4c00 2020 2020
0000230 2020 2020 2020 2020 2020 2020 2020 2020
*
0000290 2020 2020 2020 2020 0042 0002 04b0 0161
00002a0 0002 0000 013d 0002 0001 009c 0002 000e
00002b0 0019 0002 0000 0012 0002 0000 0013 0002
00002c0 0000 01af 0002 0000 01bc 0002 0000 003d
00002d0 0012 0000 000f 3f1b 27f6 0038 0000 0000
00002e0 0001 0258 0040 0002 0000 008d 0002 0000
00002f0 0022 0002 0000 000e 0002 0001 01b7 0002
```

“It’s a file called stockmarket.xlsx”

2101.49
2057.64
2063.11
2077.42
2076.78
0
2068.76
2081.34
2046.68
2051.31
2076.62
2099.60
2108.95
2107.40
2124.29
2126.64
2128.28
2119.21
2114.15
2102.15
2079.65
2067.64
2093.25
2108.57
2108.63
2103.84

“It’s a file called SP500.xlsx”

SP500	S&P 500, Index, Daily, Not Seasonally Adjusted

Frequency: Daily, Close

observation_date	SP500
2015-06-26	2101.49
2015-06-29	2057.64
2015-06-30	2063.11
2015-07-01	2077.42
2015-07-02	2076.78
2015-07-03	0
2015-07-06	2068.76
2015-07-07	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
2015-07-13	2099.60
2015-07-14	2108.95
2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

“It’s a file called SP500.xlsx, downloaded from FRED.”

SP500	S&P 500, Index, Daily, Not Seasonally Adjusted

Frequency: Daily, Close

observation_date	SP500
2015-06-26	2101.49
2015-06-29	2057.64
2015-06-30	2063.11
2015-07-01	2077.42
2015-07-02	2076.78
2015-07-03	0
2015-07-06	2068.76
2015-07-07	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
2015-07-13	2099.60
2015-07-14	2108.95
2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

“It’s a file called SP500.xlsx, downloaded from FRED.”

SP500	S&P 500, Index, Daily, Not Seasonally Adjusted

Frequency: Daily, Close

observation_date	SP500
2015-06-26	2101.49
2015-06-29	2057.64
2015-06-30	2063.11
2015-07-01	2077.42
2015-07-02	2076.78
2015-07-03	0
2015-07-06	2068.76
2015-07-07	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
2015-07-13	2099.60
2015-07-14	2108.95
2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

S&P Dow Jones Indices LLC. 2020. “*S&P 500 [SP500] [dataset]*”, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

“SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. ”

SP500	S&P 500, Index, Daily, Not Seasonally Adjusted

Frequency: Daily, Close

observation_date	SP500
2015-06-26	2101.49
2015-06-29	2057.64
2015-06-30	2063.11
2015-07-01	2077.42
2015-07-02	2076.78
2015-07-03	0
2015-07-06	2068.76
2015-07-07	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
2015-07-13	2099.60
2015-07-14	2108.95
2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

S&P Dow Jones Indices LLC. 2020. “*S&P 500 [SP500] [dataset]”*, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

Data Availability Statements

"SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. "

Describes data file, where to get it, how to get it, and any conditions of obtaining it

S&P 500, Index, Daily,
Not Seasonally Adjusted

2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

S&P Dow Jones Indices LLC. 2020. "*S&P 500 [SP500] [dataset]*", retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

Data Citation

"SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. "

Attributes the file to the proper source

SP500	S&P 500, Index, Daily, Not Seasonally Adjusted
	2101.49
	2057.64
	2063.11
	[illegible]
	2076.78
	0
	2068.76
	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
2015-07-13	2099.60
2015-07-14	2108.95
2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

S&P Dow Jones Indices LLC. 2020. "*S&P 500 [SP500] [dataset]*", retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

Some practical tips
(based on 1000 articles)

1. Computational empathy

- *Focal reader:* your next RA in 4 years
- *Interaction*: you hand them your README, but don't have time to go through all the details…
- *Budget constraint*: It shouldn't take too many RA hours
- *Time constraint*: It shouldn't take more than 1 week to "get it"


Social Science Data Editors

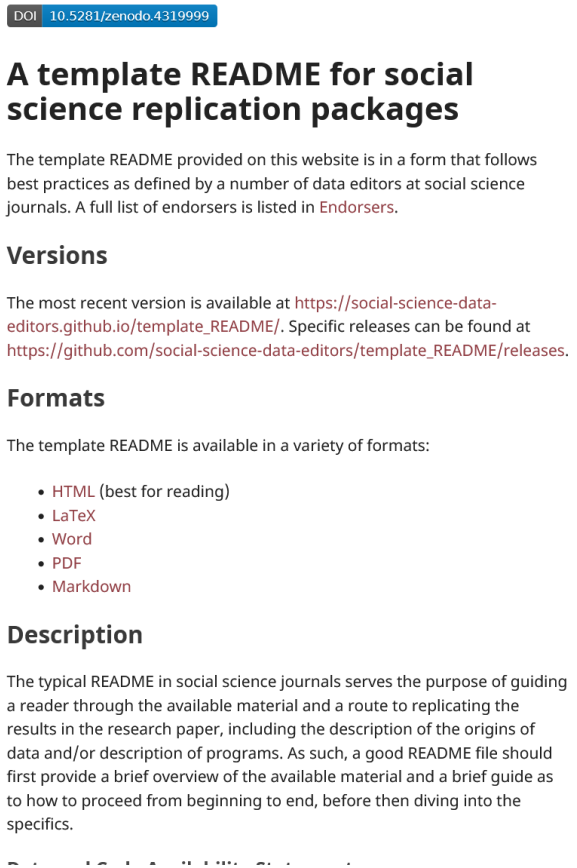
A template README for social science replication packages.

The template README provided on this website is in a form that follows best practices as defined by a number of data editors at social science journals.

Authors: Lars Vilhuber, Miklos Kóren, Joan Llull, Marie Connolly, Peter Morrow

This project is maintained at social-science-data-editors/template_README

Disclaimer


DOI 10.5281/zenodo.4319999

A template README for social science replication packages

The template README provided on this website is in a form that follows best practices as defined by a number of data editors at social science journals. A full list of endorsers is listed in Endorsers.

Versions

The most recent version is available at https://social-science-data-editors.github.io/template_README/. Specific releases can be found at https://github.com/social-science-data-editors/template_README/releases.

Formats

The template README is available in a variety of formats:

- HTML (best for reading)
- LaTeX
- Word
- PDF
- Markdown

Description

The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper, including the description of the origins of data and/or description of programs. As such, a good README file should first provide a brief overview of the available material and a brief guide as to how to proceed from beginning to end, before then diving into the specifics.

Solution 1: Computational Empathy

Use the Social Science Data Editors'

template README

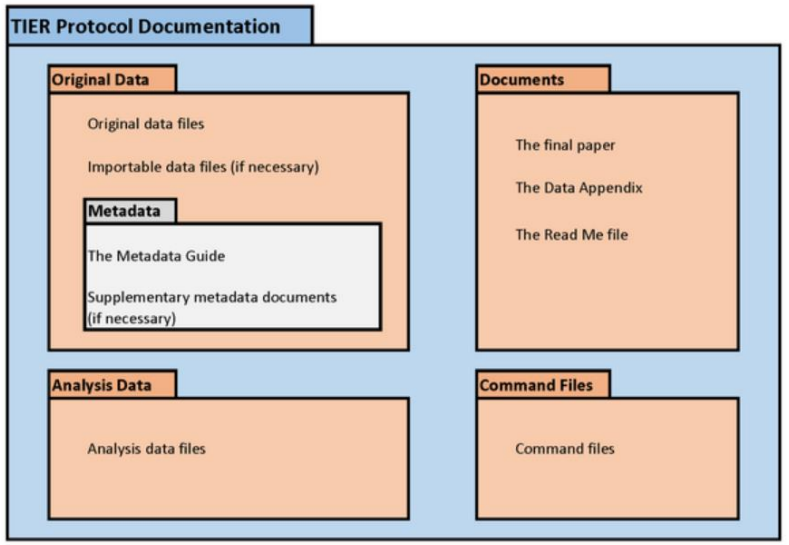
https://doi.org/10.5281/zenodo.4319999

Keeping track: **Students and Researchers**

1. Computational empathy
   Consider the next person to run the analysis, and don't assume too much

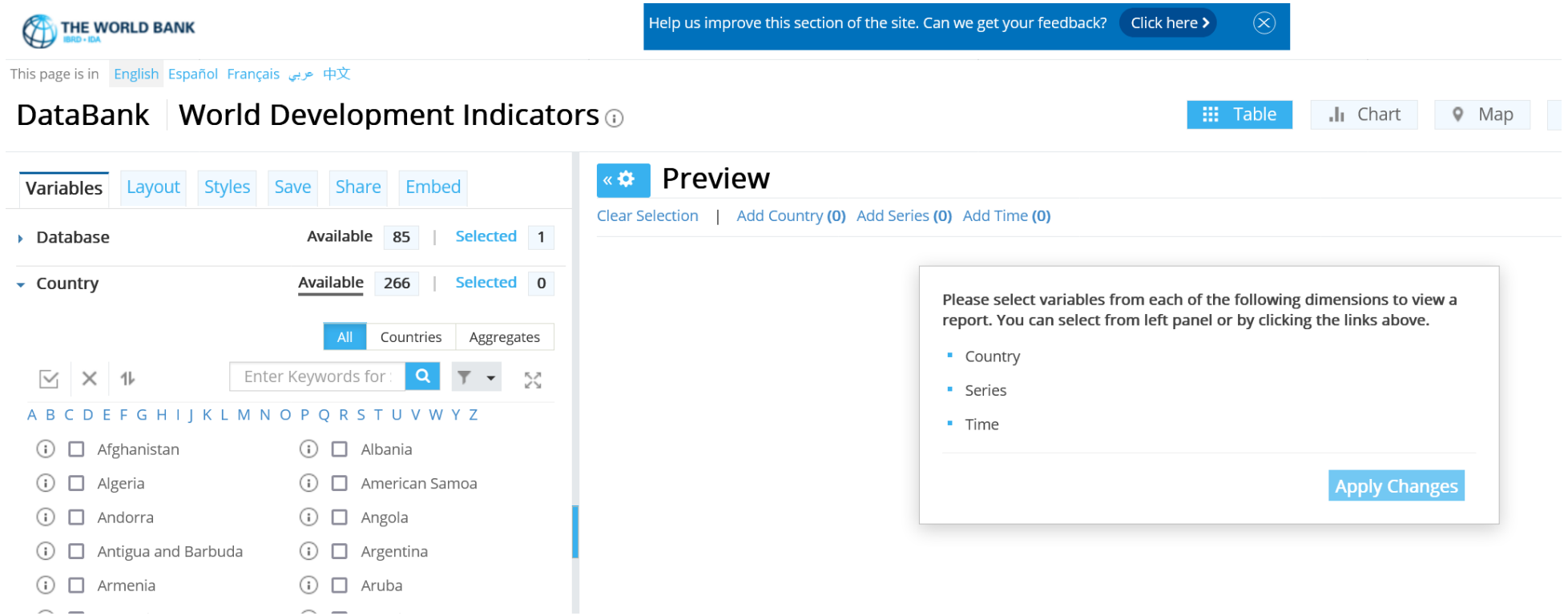
2. Keeping track of data: Data provenance

- Keep all information as you collect data
  - See TIER Protocol for good and simple guidance
- If you must use a point-and-click tool, keep detailed instructions
  - Also: obsolescence
- Try to use API, bulk download, or packages that allow for extraction
  - Also: obsolescence of API



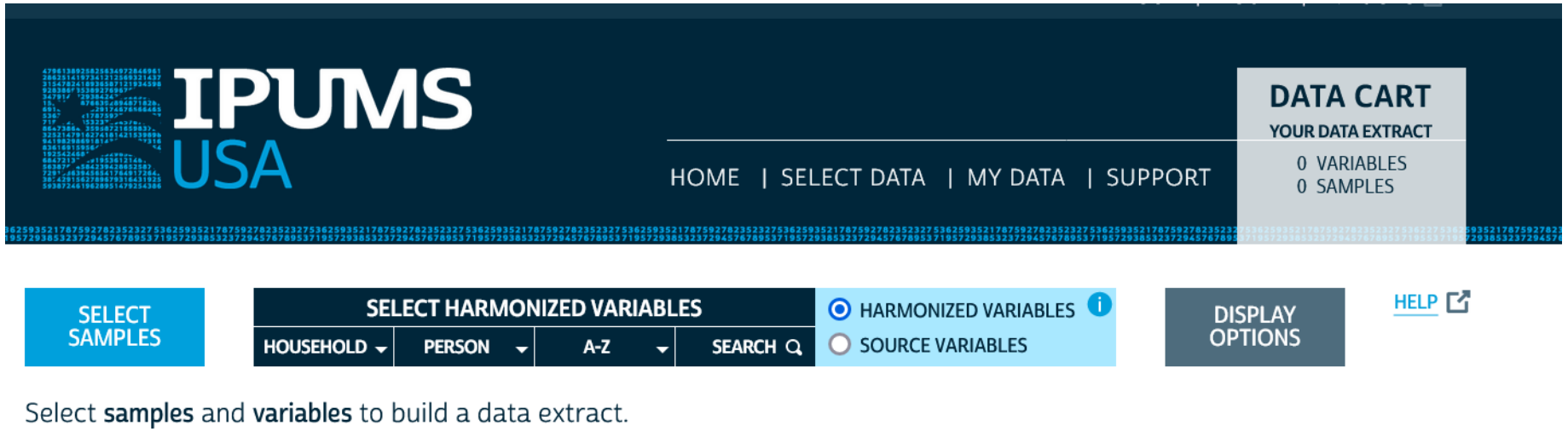
API? But the interface is so cool!

- World Development Indicators



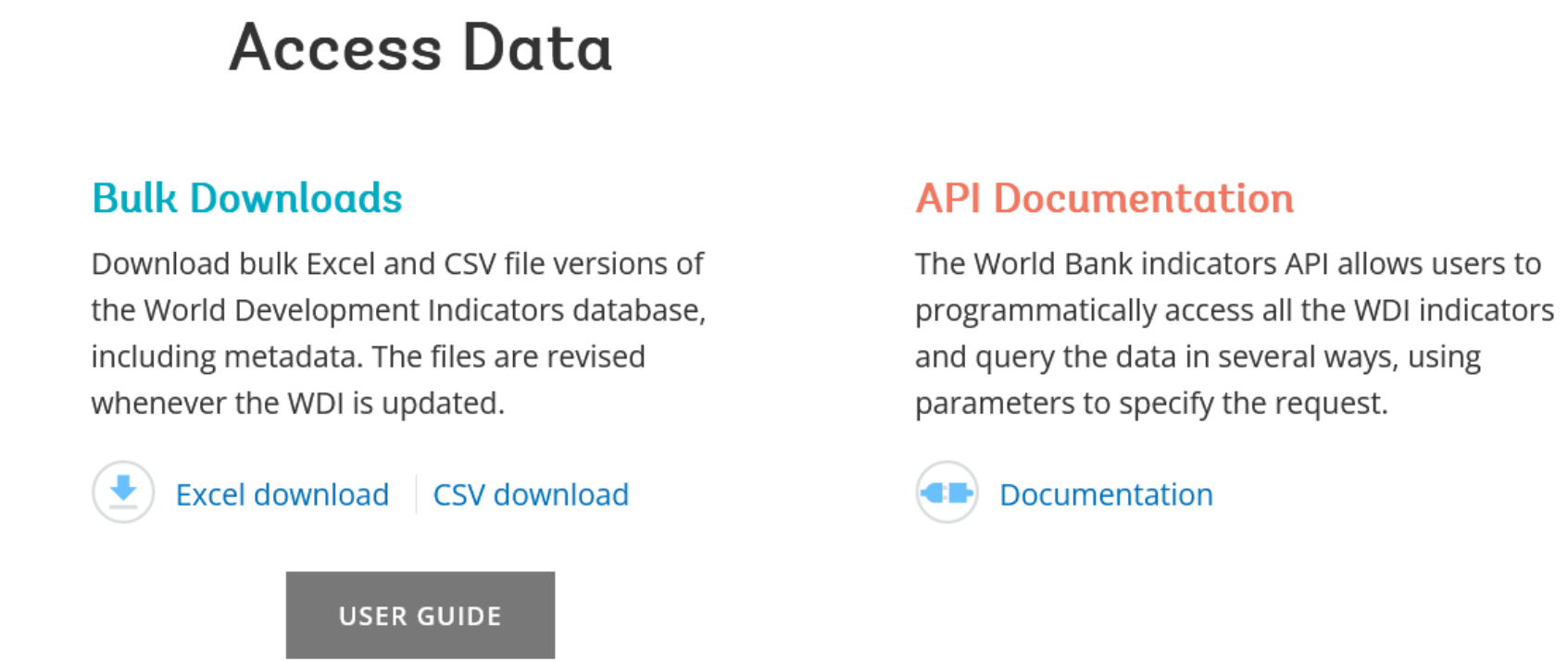
API? But the interface is so cool!

- IPUMS



API or Bulk Download

- World Development Indicators



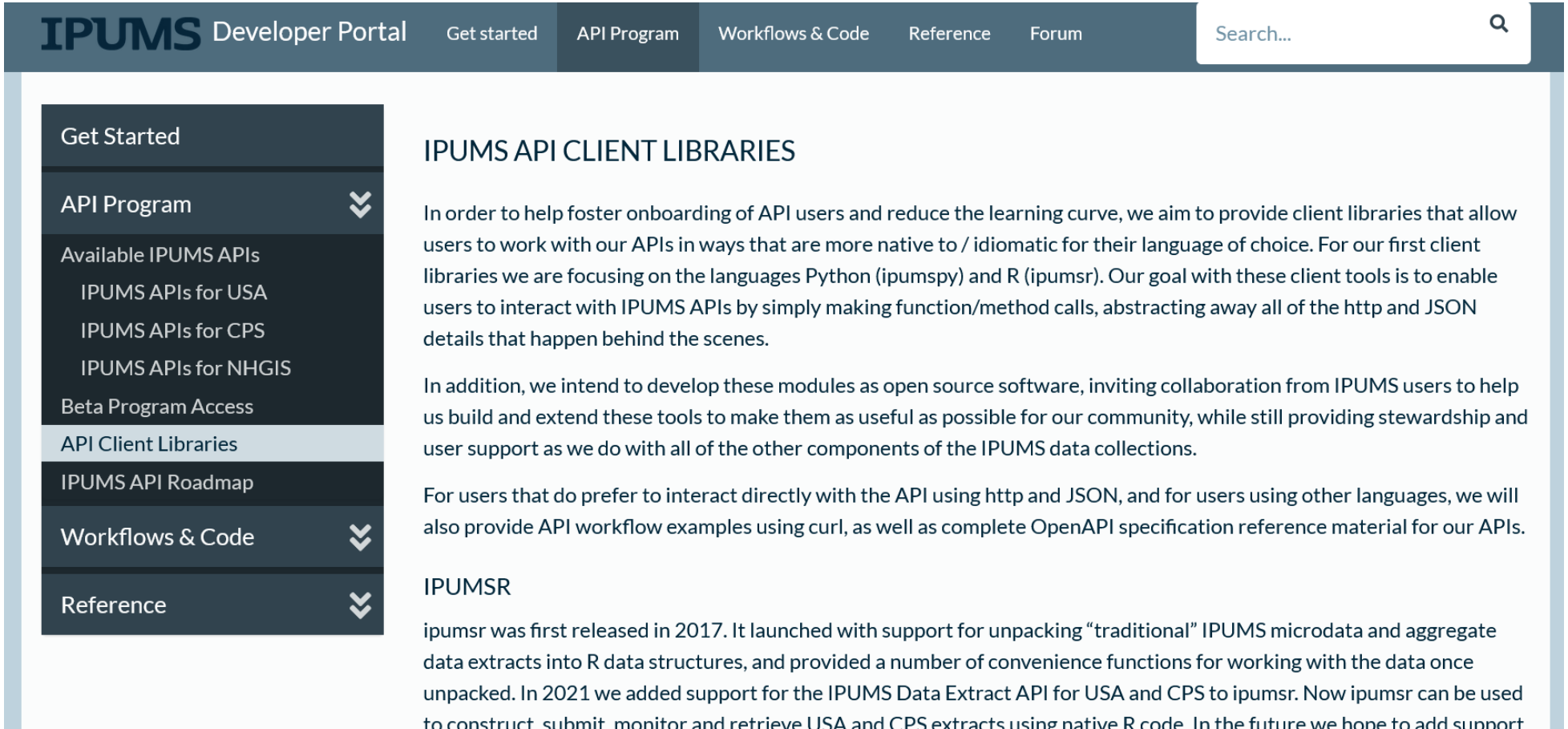
API or Bulk Download

- World Development Indicators

```
. ssc install wbopendata
. wbopendata, country(ago;bdi;chi;dnk;esp) indicator(sp.pop.0610.fe.un) ///
> year(2000:2010) clear long
```

API or Bulk Download

- IPUMS (beta)

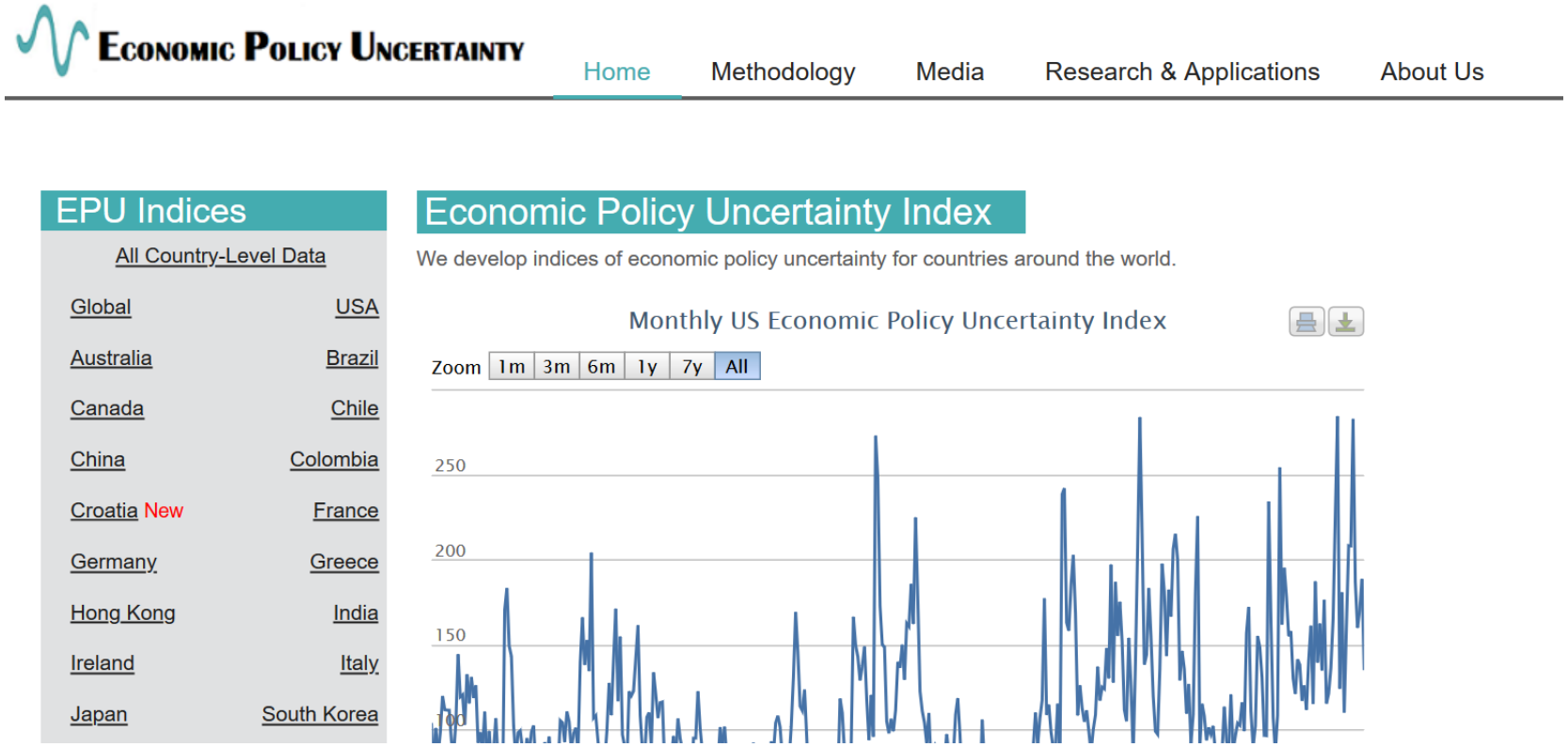


2. Keeping track of data:

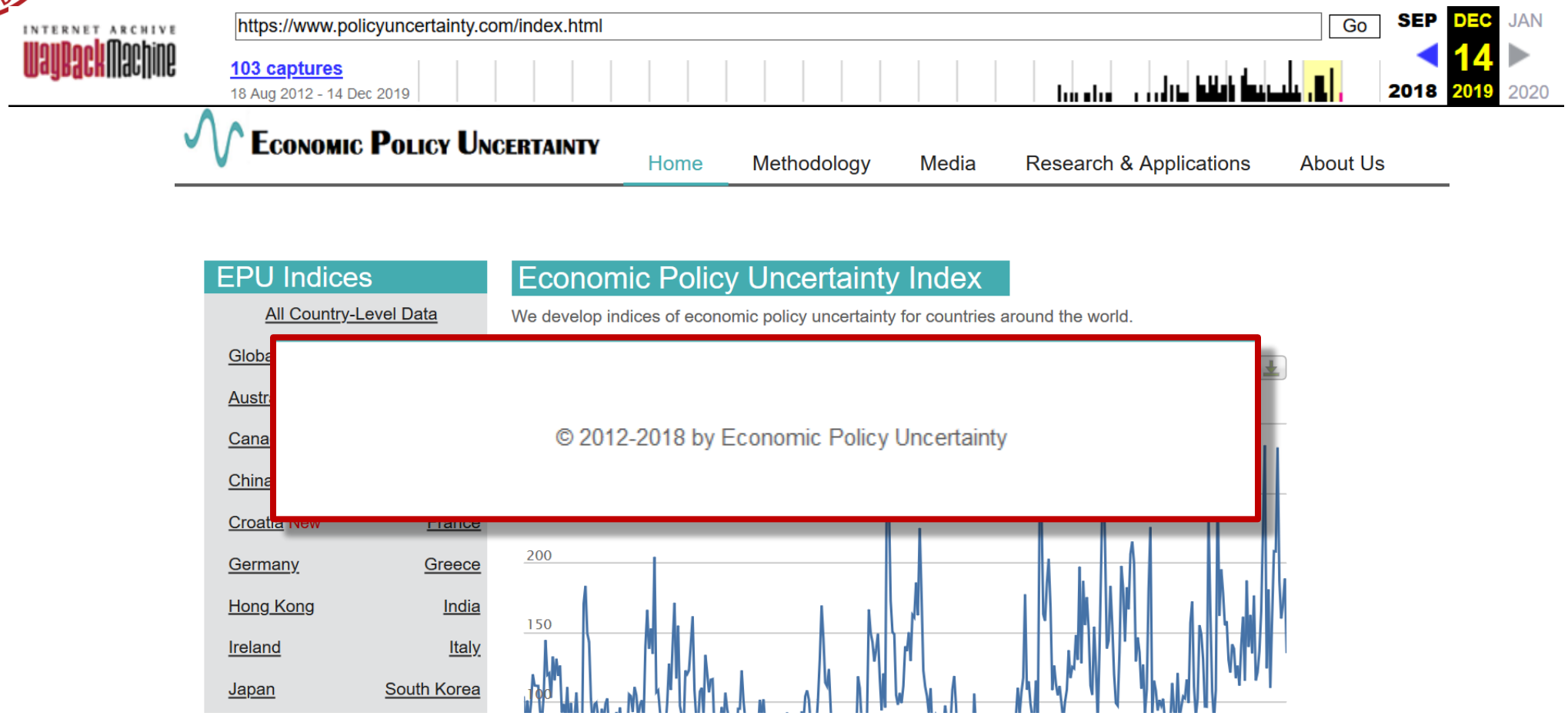
Don’t forget to check the TERMS of USE!

Because you may not be able to provide others with a copy of the data (legally)…

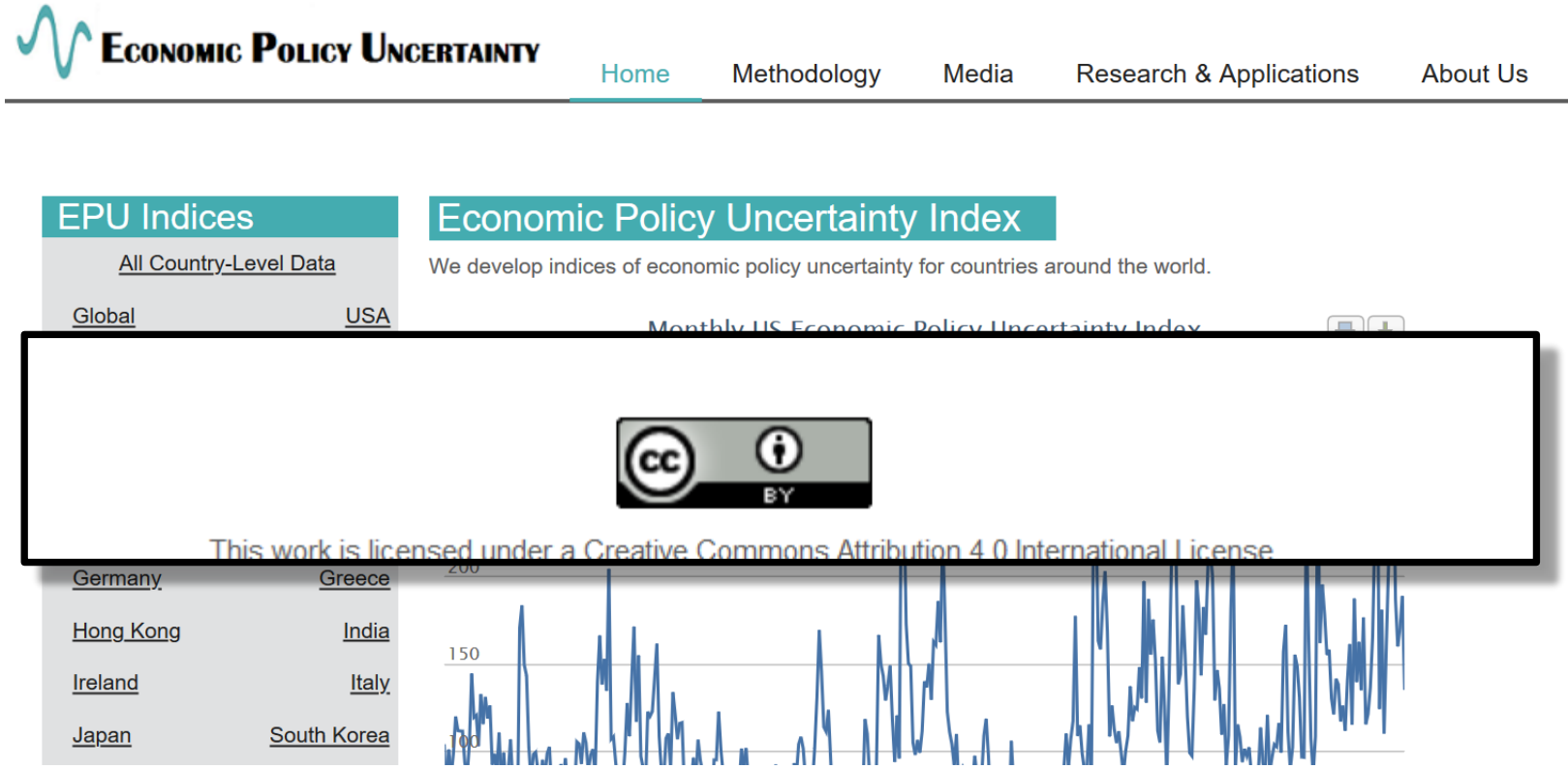
Example 2: Academic data publisher



Example 2: Academic data publisher



Example 2: Academic data publisher-new!



Rights to **use** data

- You browsed a website
- You purchased the data
- You signed a data use agreement
- You created the data (lab experiment)
- You had survey respondents consent to use (IRB approval!)

Rights to **distribute** the data

- If you created the data, you decide.
- If you got it from somewhere else:

READ THE TERMS OF USE / DATA USE AGREEMENT / CLICK-THROUGH / ETC.

Example 4: German Restricted-access

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

Home | Newsletter | Jobs | Contact | Data Privacy | Imprint

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

- External data
- Data Archive
- Data Access
- Campus Files
- Publications
- Events
- Projects of FDZ users
- FDZ Projects
- Complaint point of the RatSWD
- Figures of the FDZ

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

Example 4: German Restricted-access

Establishment History Panel (BHP) – Version 7518 v1

DOI: 10.5164/IAB.BHP7518.de.en.v1

Summary

Data source:

Data Access

The IAB Establishment Panel is available via the following ways of access:

- On-site use at the FDZ. Further information on Applying for on-site use.
- Remote data Access. Further information on Applying for remote data access.

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.

Dataset Descriptions and Frequencies

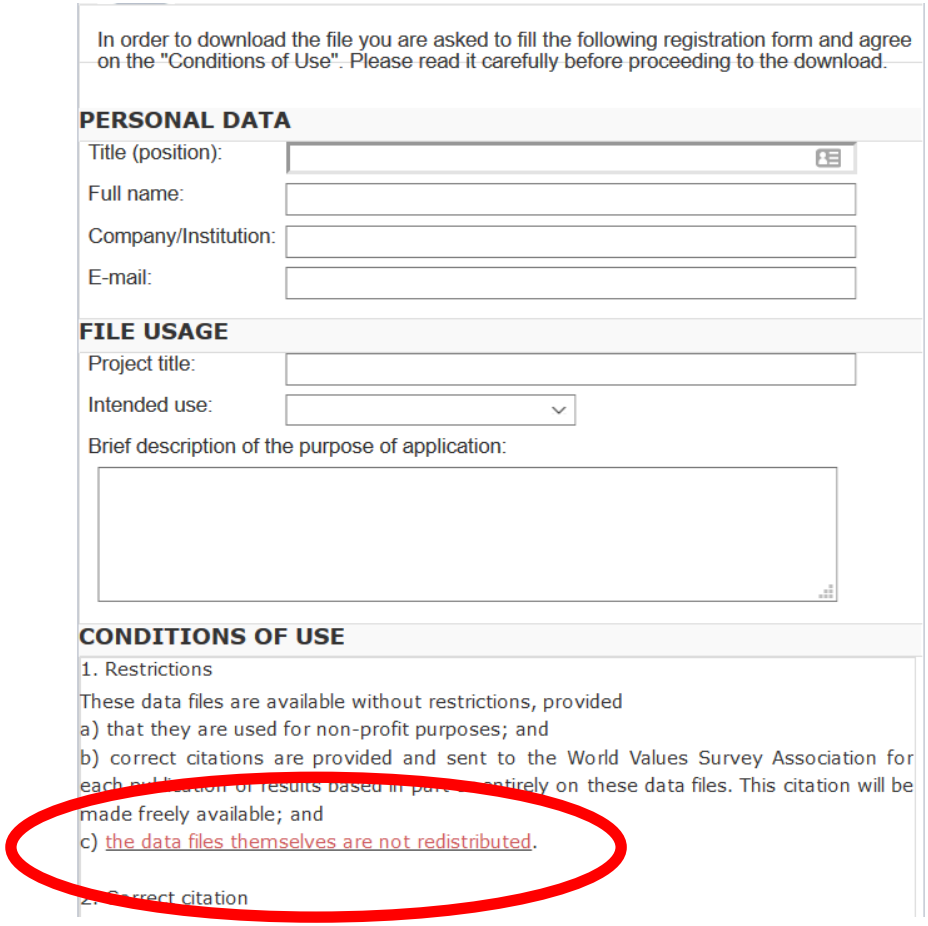
German

- DOI: 10.5164/IAB.FDZD.2001.de.v1
- FDZ-Datenreport 01/2020
- Fallzahlen und Labels

English

- DOI: 10.5164/IAB.FDZD.2001.en.v1

And we check them!



In order to download the file you are asked to fill the following registration form and agree on the "Conditions of Use". Please read it carefully before proceeding to the download.

PERSONAL DATA

Title (position):

Full name:

Company/Institution:

E-mail:

FILE USAGE

Project title:

Intended use:

Brief description of the purpose of application:

CONDITIONS OF USE

1. Restrictions

These data files are available without restrictions, provided
a) that they are used for non-profit purposes; and
b) correct citations are provided and sent to the World Values Survey Association for each publication of results based in part or entirely on these data files. This citation will be made freely available; and
c) the data files themselves are not redistributed.

2. Correct citation

- What does the site say?

Please use the following citation when referring to this file in the different versions:
Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Six - Country-Pooled Datafile Version: www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Madrid: JD Systems Institute.

- Is that in the README / Paper/ Appendix?
- Are all the conditions met/described?

Solution 2: Data Provenance

- Keep detailed notes
- script as much as possible
- (also: Use the Social Science Data Editors' template README)

Keeping track: **Students and Researchers**

1. Computational empathy
   Consider the next person to run the analysis, and don't assume too much
2. Track data (provenance)
   even when using API, especially when manually downloading, keep in mind what the next downloader may see/find/receive, terms of use

Coding for Reproducibility

Lesson 1: Computational empathy

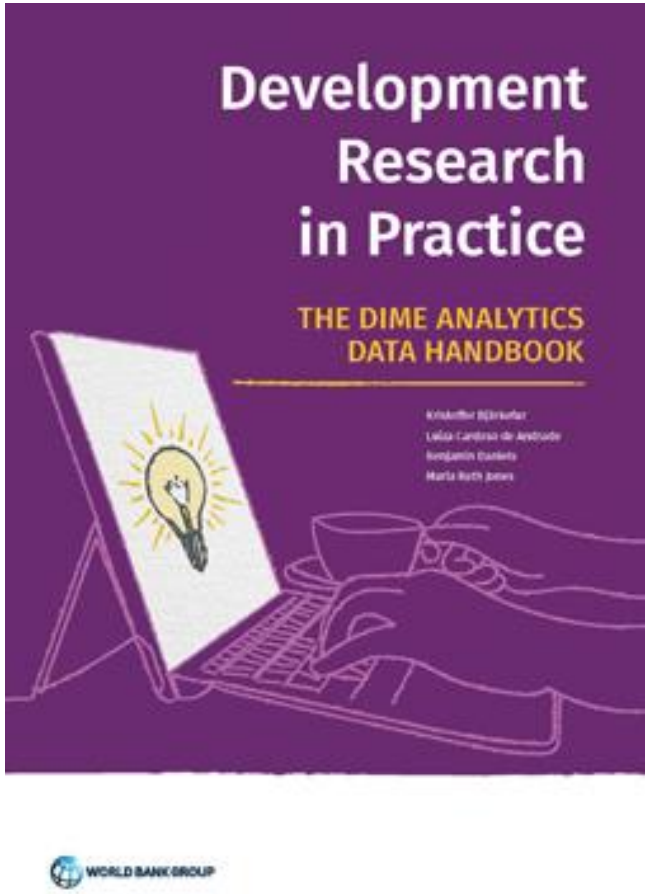
= “Pity the poor replicator”

Streamlining replication packages

- Master script preferred
  - Least amount of manual effort
- No manual manipulation
  - "Change the parameter to 0.2, then run the code again"
- No manual copying of results
  - Write out/save tables and figures using packages
  - Compute all numbers in package
- No manual install of packages
  - Use a script to create all directories, install all necessary packages/requirements/etc.
- Clear instructions!

But I don’t need to tell you that

- DIME Wiki, Handbook, this course!
- https://www.worldbank.org/en/research/dime/data-and-analytics



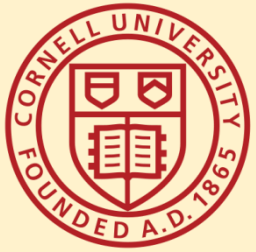
Assume replicators can access the data

Sometimes we (=AEA) cannot

- We will still check if the code seems complete
- We will still verify that all data that *can* be provided have been provided
- Plausibility checks

Sometimes we can:

- In the past, we have worked with
  - French, Brazilian, and US confidential admin data
  - Purchased commercial data (Twitter, Indian GDP)
  - Proprietary data under NDA/DUA (Ebay)
  - Data with application procedure (Chinese Panel, Demographic and Health Survey, European establishment data)
  - Remotely or locally



Solution 3: Learn basics of programming

Code reproducibly

(and do so right from the start)

(also: way easier to describe in the
Social Science Data Editors' template README)

Keeping track: **Students and Researchers**

1. Computational empathy
   Consider the next person to run the analysis, and don't assume too much
2. Track data
   even when using API, especially when manually downloading, keep in mind what the next downloader may see/find/receive, terms of use
3. Learn the basics of programming
   code reproducibly, use parameter files, re-usable code, robust file structure

That’s a lot of stuff to learn and remember…
I want to focus on the economics!

Keeping track: **Students and Researchers**

1. Computational empathy
   Consider the next person to run the analysis, and don't assume too much
2. Track data
   even when using API, especially when manually downloading, keep in mind what the next downloader may see/find/receive, terms of use
3. Learn the basics of programming
   code reproducibly, use parameter files, re-usable code, robust file structure
4. Learn to automate
   Run all code again and again, use APIs to download, use conditional processing to handle various aspects
5. Preserve it all
   Use version control, tag releases, preserve data (separately), understand the difference between sharing and preserving

That’s a lot of stuff to learn and remember…

I want to focus on the economics!

Support by Institutions

Lesson 3: Support by institutions is insufficient

- When should these skills be taught?
  - These are core “tools of the trade”!
  - Undergrad, core part of graduate curricula
  - In other disciplines: students learn how to collect use a pipette, how to tag field mice in the wild…

Lesson 3: Support by institutions is insufficient

- Should all of these skills be taught?
  - How to deposit data
  - How to set up a compute cluster
- Some of these skills fall into other categories, but
  - Data librarians are understaffed, and not trained in discipline-specific practices
  - Campus IT has highly varying funding and consulting time
  - Cross-campus IT practices are nowhere close to compatible

Lesson 3: Support by institutions is insufficient

Institutional funding and mandates are not adequate

- Most grant funding in social sciences does not require (or allow!) for this kind of budgeting
  - Earmarked portions of funds would be great!
  - (This is slowly coming, NSF and NIH are making progress)
- (Most) Universities consider this an external mandate, not part of their “overhead”
  - “Provide us with an account, and we will do it”
  - Leads to highly scattershot infrastructure

Luckily, you have DIME

DIME Analytics
Reproducibility Package Checklist
v1.4

Required Content and Practices

The reproducibility package must be shared as a .zip file, GitHub link or shared folder containing the items listed below.

1. Files

A filled version of this checklist.

A README file describing the folders and datasets included in the package, and any general or ad-hoc instructions about the directory structure or for the code to run.

A PDF version of the research output being reviewed (paper, brief, report).

Solution 6: Institutional support

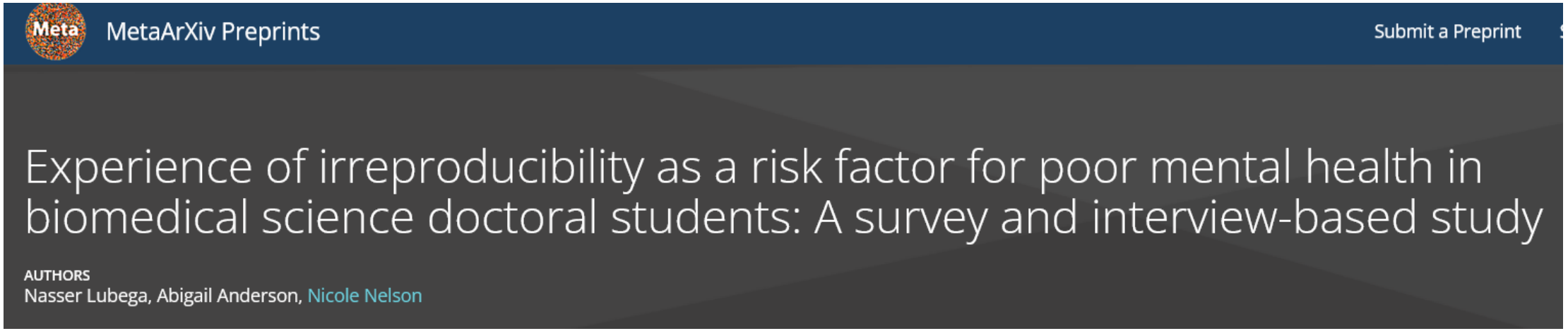
(departments, schools, libraries, IT, universities)

1. Offer training in adapted tools
   (not sufficient to just show how to do a Rmarkdown document)
2. Highlight appropriate community *(Zenodo, Dataverse, others)* or university sites
3. Provide streamlined access to some frequently used (open/commercial) tools
   (AWS/GCS/Azure, CI on Github/others, etc.)

Some thoughts for the role of Faculty

1. Encourage students to learn new skills outside of economics
   Even or especially if you do not have the time to do so
2. Demand reproducibility when reviewing (*articles, theses, intermediate reports from students, etc.*)
   automated re-runs on Github, which particular release to review, refusing emailed copies
3. Incentivize reproducibility
   For robustness, for efficiency, but also training for exposure to the discipline.
   Example: Replication Challenges

Please don’t produce irreproducible articles!



The role for journals

Goal: Transportability

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

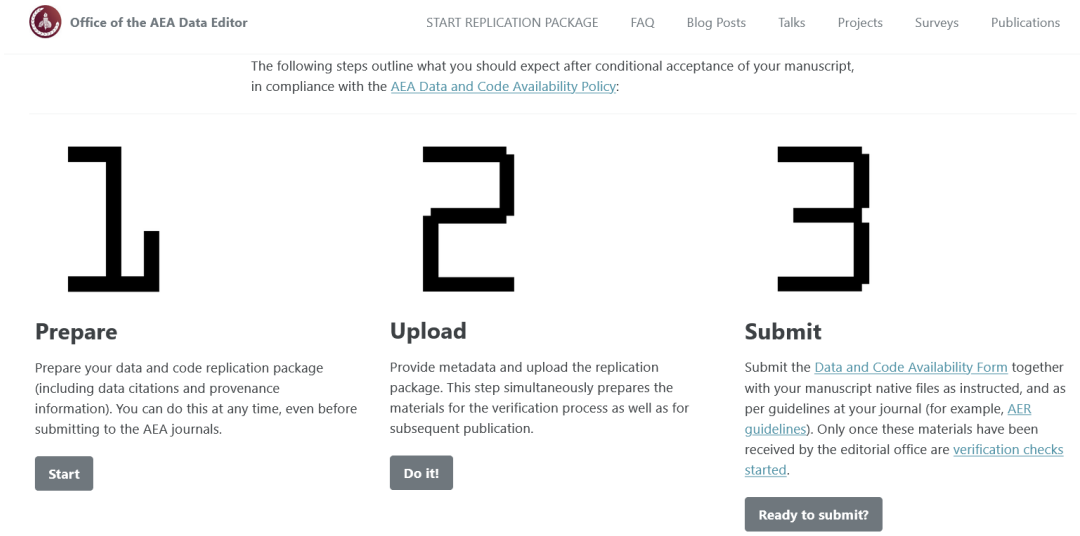
Social science “guild”

Social Science Data Editors

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

Some resources

- https://social-science-data-editors.github.io/guidance/
  - **template README**
  - discussion of licensing
  - data citation guidance
- https://aeadataeditor.github.io/



Thank you!

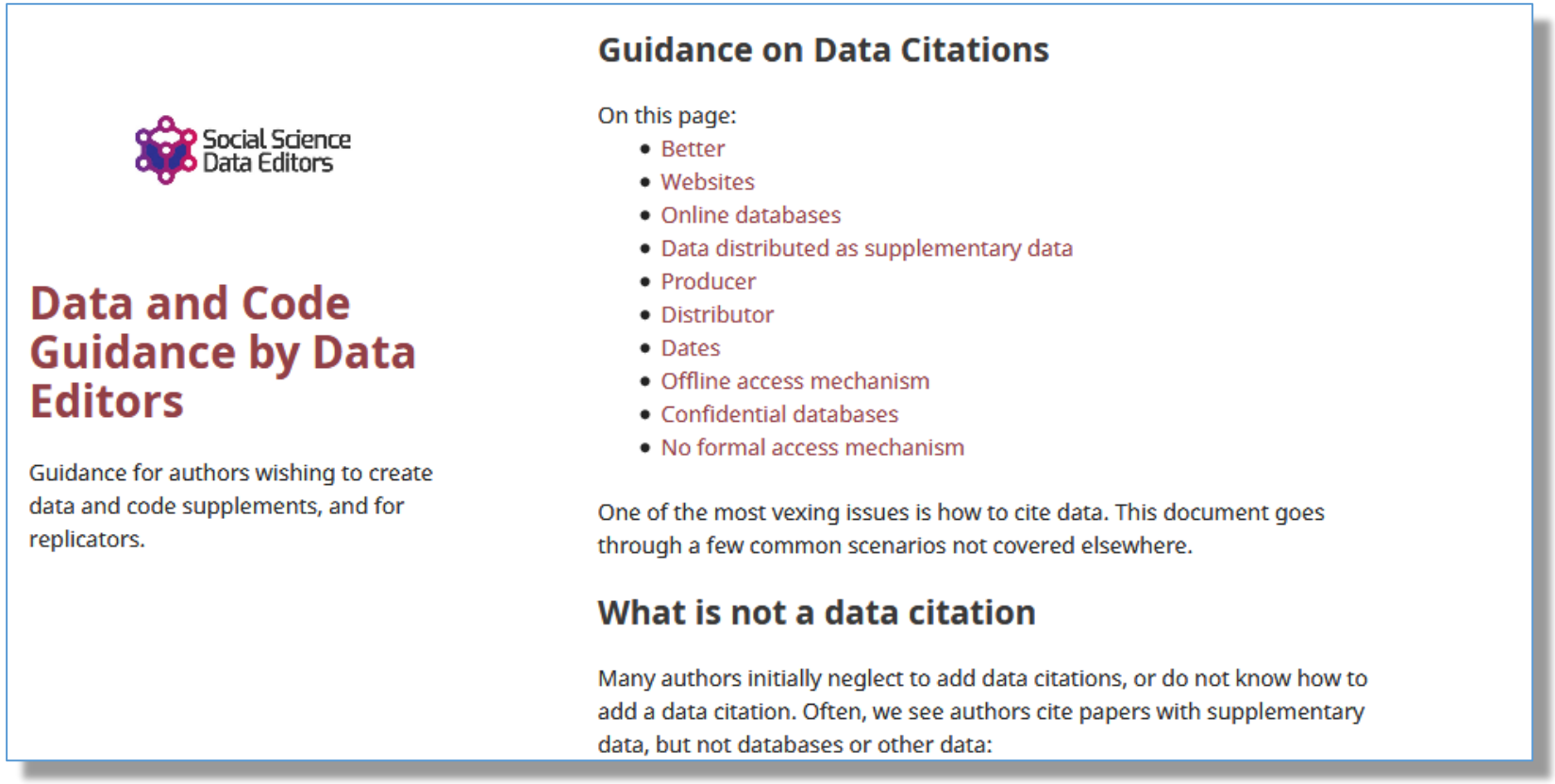
Reminder: **Students and Researchers**

1. Computational empathy
   Consider the next person to run the analysis, and don't assume too much
2. Track data
   even when using API, especially when manually downloading, keep in mind what the next downloader may see/find/receive, terms of use
3. Learn the basics of programming
   code reproducibly, use parameter files, re-usable code, robust file structure
4. Learn to automate
   Run all code again and again, use APIs to download, use conditional processing to handle various aspects
5. Preserve it all
   Use version control, tag releases, preserve data (separately), understand the difference between sharing and preserving

Data Citations

Data citations

- Creating specific guidance in the absence of strong discipline-specific guidance



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

Example 4: German Restricted-access

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

Home | Newsletter | Jobs | Contact | Data Privacy | Imprint

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

- External data
- Data Archive
- Data Access
- Campus Files
- Publications
- Events
- Projects of FDZ users
- FDZ Projects
- Complaint point of the RatSWD
- Figures of the FDZ

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

Data Citation

"SP500.xlsx, from S&P (2020). Not provided as part of replication package because © S&P. "

Attributes the file to the proper source

SP500	S&P 500, Index, Daily, Not Seasonally Adjusted
	2101.49
	2057.64
	2063.11
	[illegible]
	2076.78
	0
	2068.76
	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
2015-07-13	2099.60
2015-07-14	2108.95
2015-07-15	2107.40
2015-07-16	2124.29
2015-07-17	2126.64
2015-07-20	2128.28

S&P Dow Jones Indices LLC. 2020. "*S&P 500 [SP500] [dataset]*", retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

Element of a (data) citation

ICPSR notes that a citation should include the following items:

- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

Suggested Citation:

S&P Dow Jones Indices LLC, *S&P 500 [SP500]*, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.

Element of a (data) citation

ICPSR notes that a citation should include the following items:

- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

Constructed Citation:

Institute for Employment Research (IAB), Establishment History Panel 1975-2018. Accessed via the Research Data Centre (FDZ) of the German Federal Employment Agency DOI: 10.5164/IAB.BHP7518.de.en.v1 June 26, 2020.

Element of a (data) citation

ICPSR notes that a citation should include the following items:

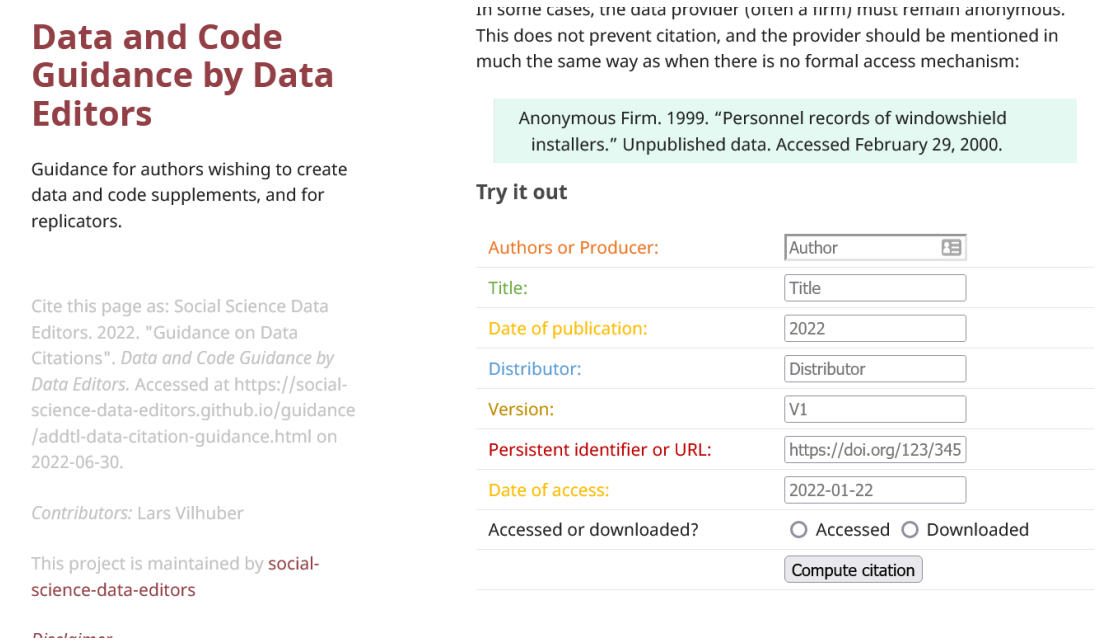
- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

Constructed Citation:

US Census Bureau, Longitudinal Business Database (LBD) 1975-2018. Last accessed via the Federal Statistical Research Data Centre (FSRDC) June 26, 2020.

Try it out yourself

- Construct an (approximate) data citation
- https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html#try-it-out



Data and Code Guidance by Data Editors

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

Cite this page as: Social Science Data Editors. 2022. "Guidance on Data Citations". *Data and Code Guidance by Data Editors*. Accessed at https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html on 2022-06-30.

Contributors: Lars Vilhuber

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

In some cases, the data provider (often a firm) must remain anonymous. This does not prevent citation, and the provider should be mentioned in much the same way as when there is no formal access mechanism:

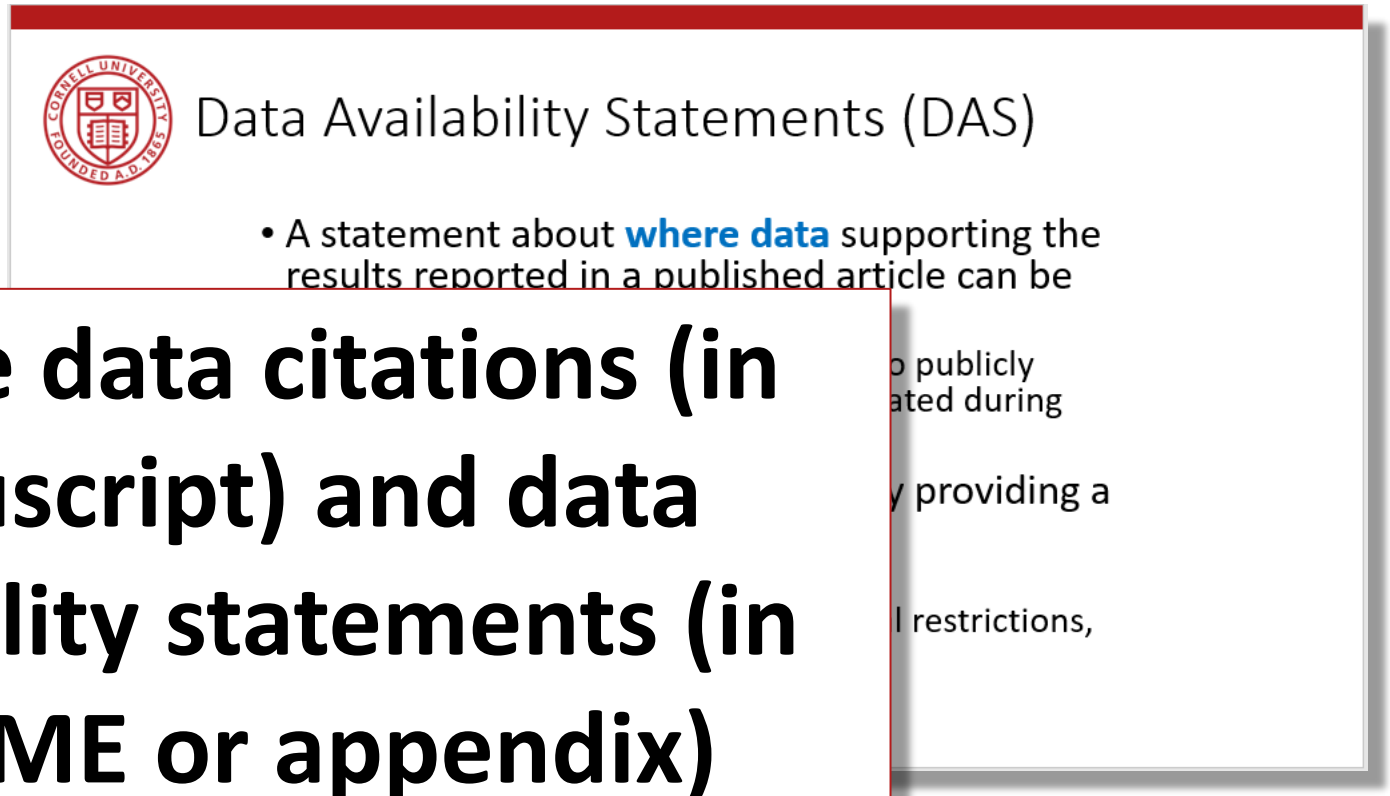
> Anonymous Firm. 1999. "Personnel records of windowshield installers." Unpublished data. Accessed February 29, 2000.

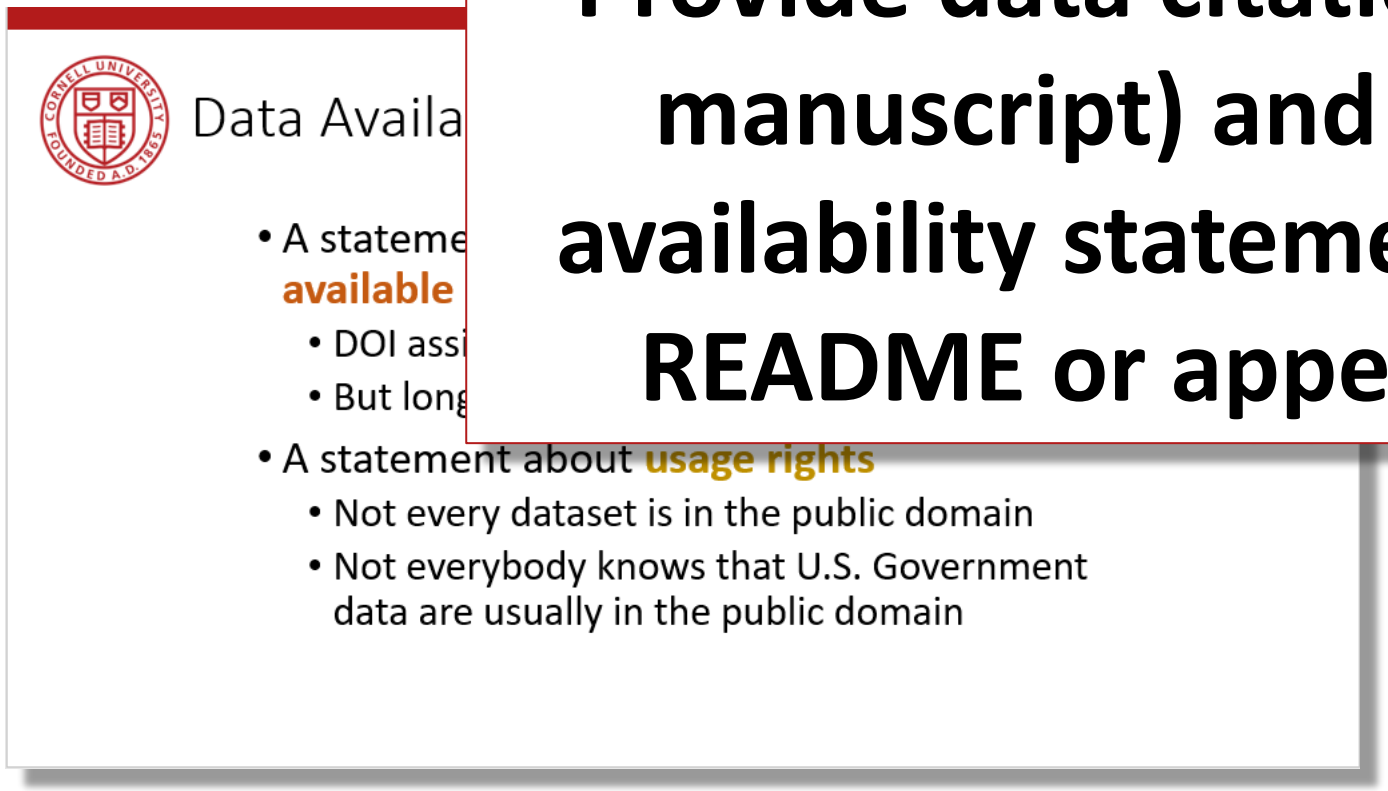
Try it out

Authors or Producer:	Author
Title:	Title
Date of publication:	2022
Distributor:	Distributor
Version:	V1
Persistent identifier or URL:	https://doi.org/123/345
Date of access:	2022-01-22
Accessed or downloaded?	○ Accessed ○ Downloaded
	Compute citation

Data: Citations, Access, Rights

- Any data can be cited – even if you can't download it
- Any data that you accessed … can have that access be described
  - But caution: It should be such that others can also repeat the access!
- Just because you "have" the data does not mean you can give it to others
  - Also: distinguish between "sharing" and "publishing"
  - Know your terms of use!





Provide data citations (in manuscript) and data availability statements (in README or appendix)

Solution 3: Data Citations

Cite every data source

(not only the paper that describes the source!)

(also: add them to the
Social Science Data Editors' template README)