Transparency and Reproducibility in Economics:

Lessons learned from 1,000 papers

Lars Vilhuber

Cornell University

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.

3 Lessons (and many solutions)

- Lesson 1: Computational empathy
- Lesson 2: Data acumen
- Lesson 3: Role of institutions

Let me expand that a bit…

For students and researchers

0. *Do not necessarily learn from previous papers*
1. Have computational empathy for …
2. Track data whenever used
3. Learn the basic … of programming
4. Learn to automate
5. Preserve it all (and version it too)

For institutions

(departments, schools, libraries, IT, universities)

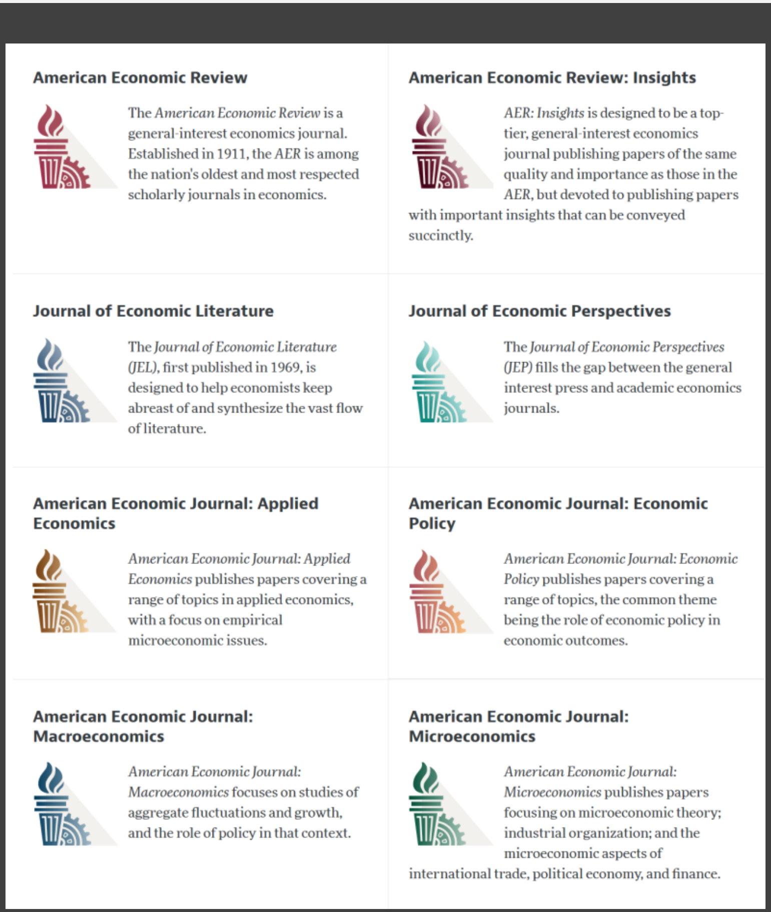
1. Offer training in adapted tools
2. Highlight appropriate community or university sites
3. Provide streamlined access to some frequently used (open/commercial) tools

For faculty

1. Encourage students to learn skills you don’t know
2. Demand reproducibility when reviewing
   (*articles, theses, intermediate reports from students, etc.*)
3. Incentivize reproducibility

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.

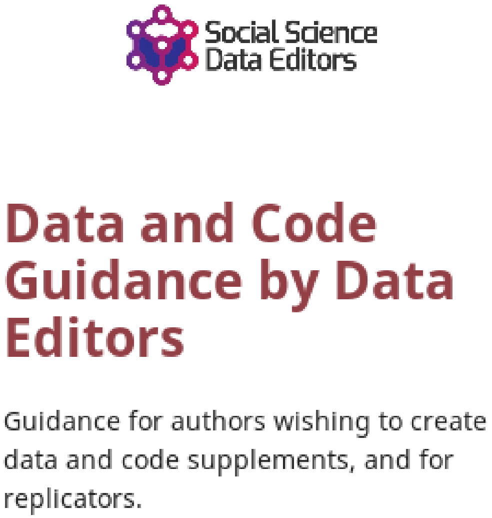
American Economic Journal: Microeconomics

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.**

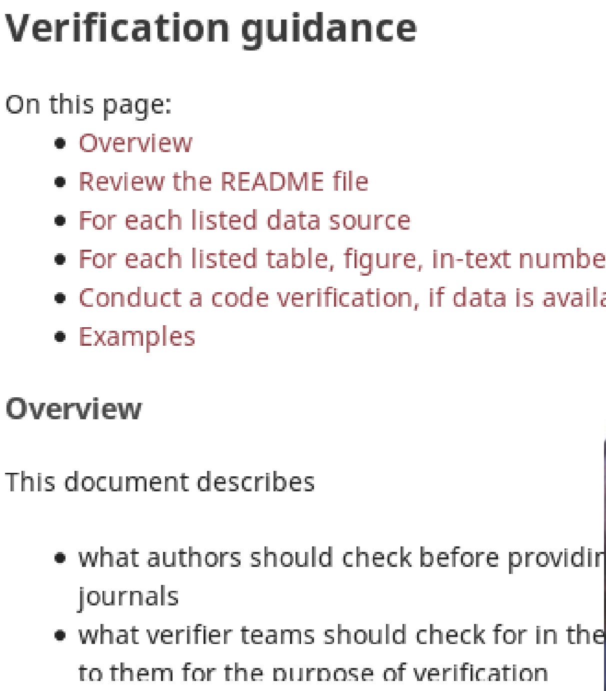
Action: Reproducibility Check



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

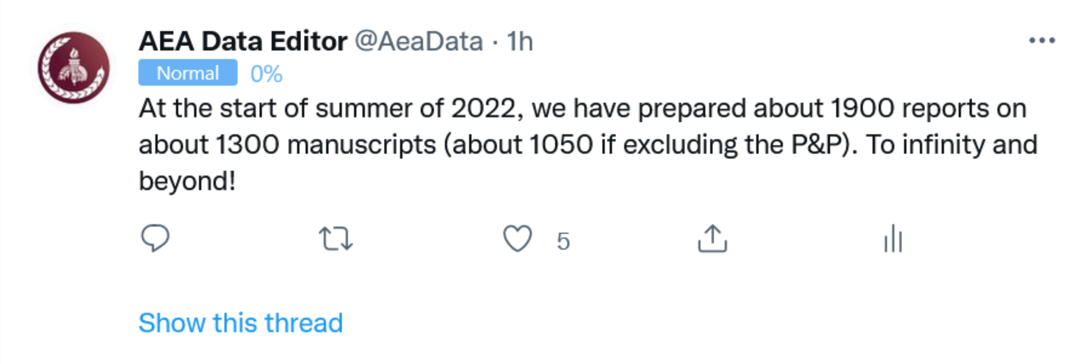
This document describes

- what authors should check before providi 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)



AEA Data Editor @AeaData · 1h

At the start of summer of 2022, we have prepared about 1900 reports on about 1300 manuscripts (about 1050 if excluding the P&P). To infinity and beyond!

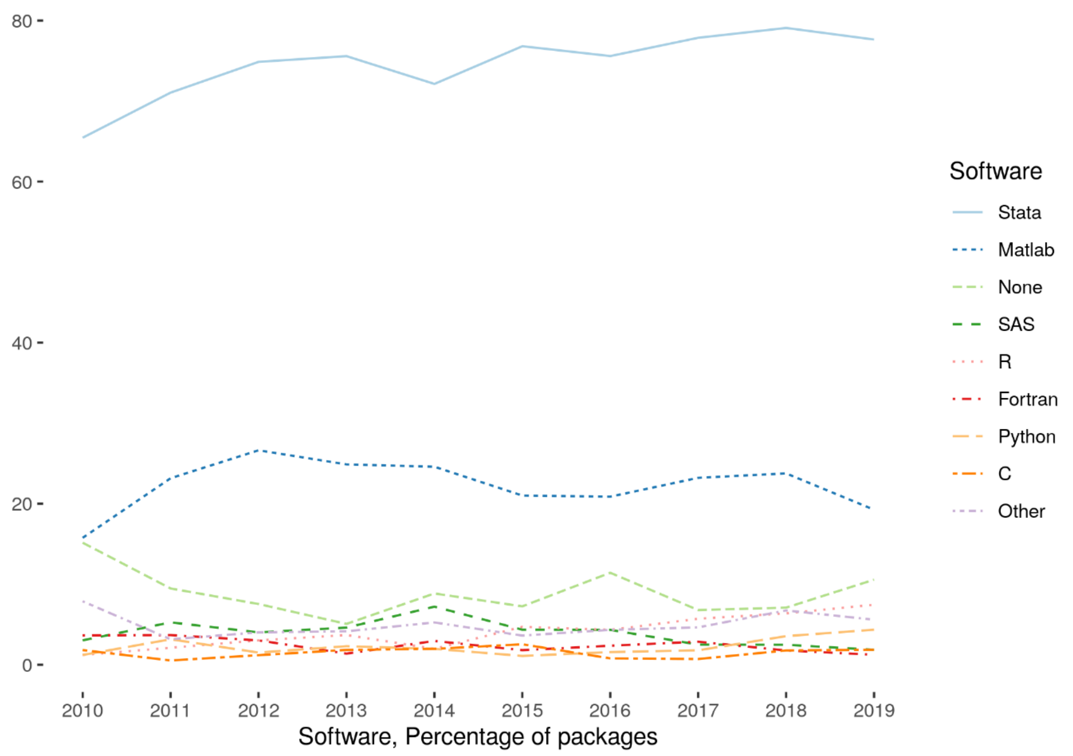
A bit of my background

My experience

- Creating statistical production system from research code (*still running 15+ years later*)
- Working with confidential data *(creating reproducible analyses, but also seeing how where others fail to do so)*
- Helped create and analyze synthetic data at scale *(including configuring and managing the server to induce reproducible programming…)*
- Comfortable on Linux systems *(since 1993)*, but also versed in MacOS and Windows *(to see what others do…)*
- Comfortable running Stata, R, Python, Matlab, Julia, compiling Fortran and C with and without Makefiles, etc.
- Configured departmental clusters *(for myself, for colleagues, accommodating different usage patterns)*

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%**)



Defining “reproducible research”

> “Reproducibility” refers to the ability of a researcher to duplicate the results of a prior study using the **same materials** and **procedures** as were used by the original investigator.

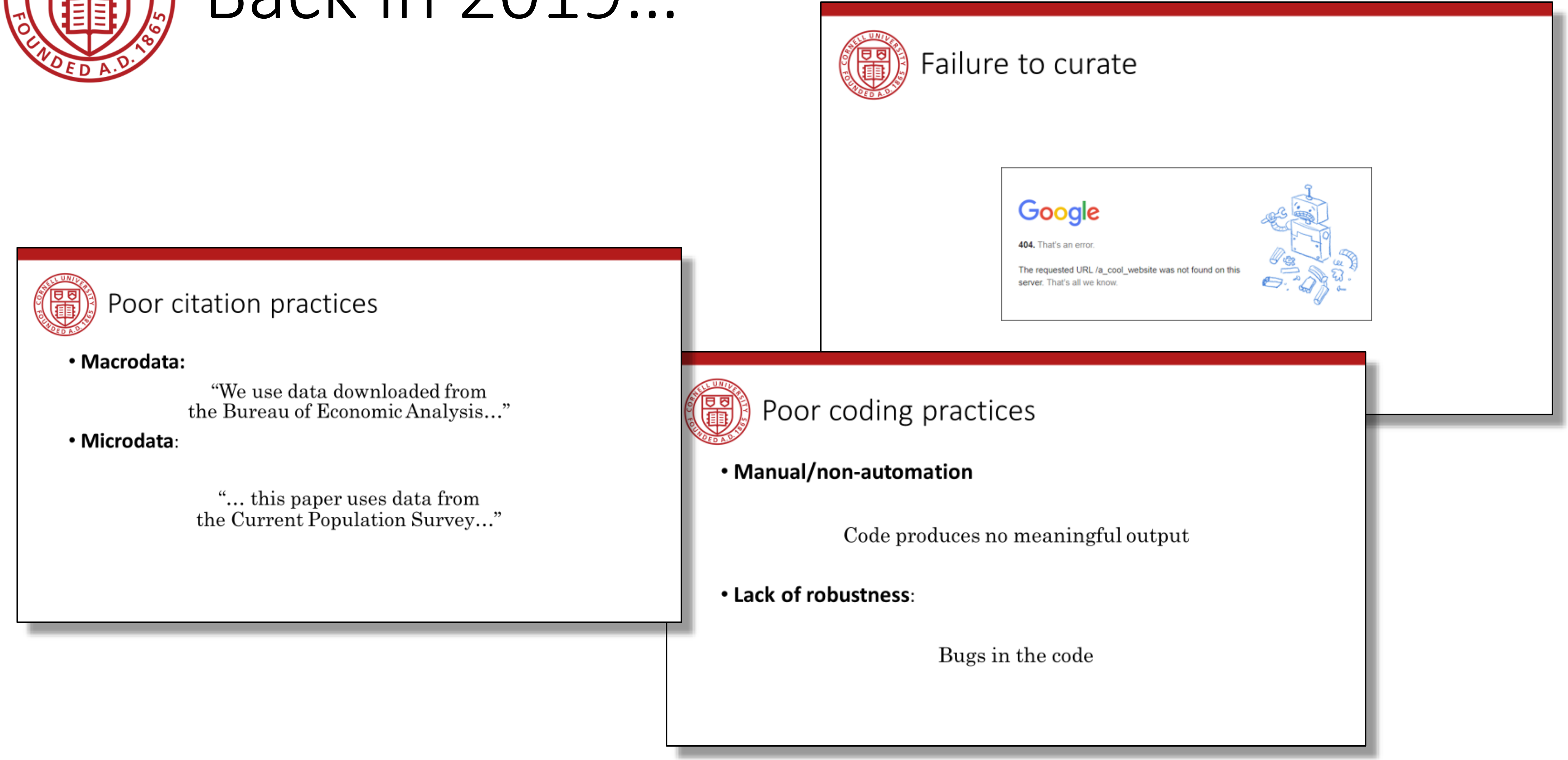
Bollen et al. 2015. “Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.” National Science Foundation.https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

Ingredients of “research”

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

Lessons?

Back in 2019…



Poor citation practices

- **Macrodata:**

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

- **Microdata**:

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

Failure to curate

Google

404. That's an error.

The requested URL /a_cool_website was not found on this server. That's all we know.

Poor coding practices

- **Manual/non-automation**

  Code produces no meaningful output

- **Lack of robustness**:

  Bugs in the code

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

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

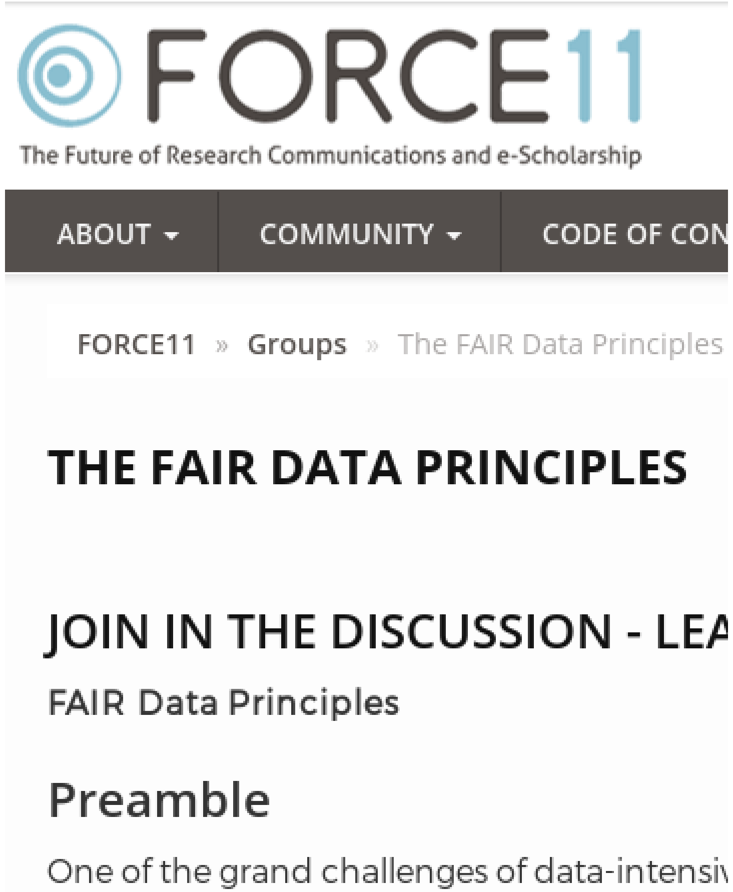
- Data provenance
  - Where did the data come from which I used?
- Data preservation
  - Where do I put the data I generated?
  - What if the data I used are not “robustly preserved”?
  - What do you mean by that?

Data provenance

Action: Data citations and metadata

What is **FAIR**?

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



FORCE11

The Future of Research Communications and e-Scholarship

FORCE11 » Groups » The FAIR Data Principles

THE FAIR DATA PRINCIPLES

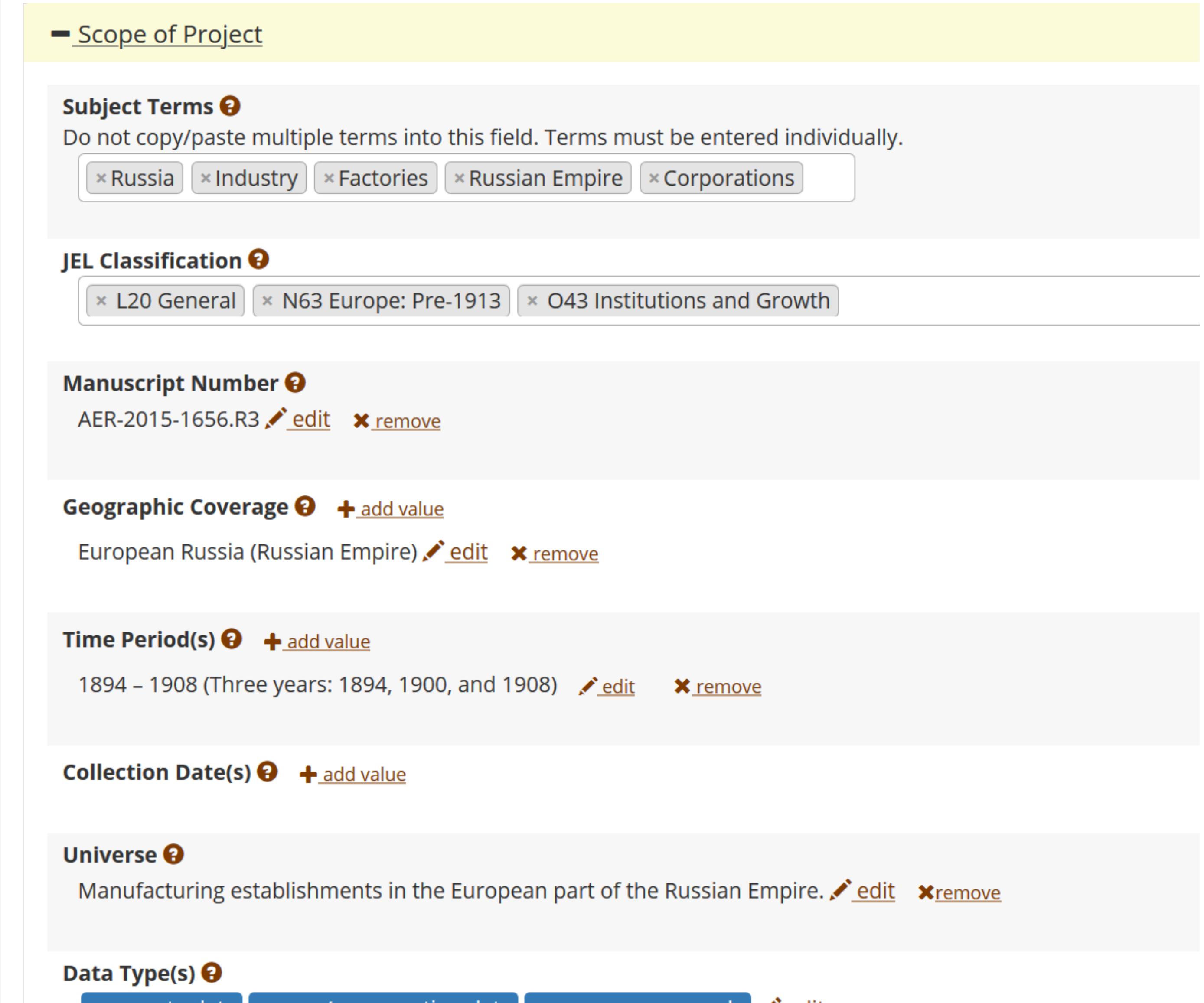
JOIN IN THE DISCUSSION - LEA

FAIR Data Principles

Preamble

One of the grand challenges of data-intensiv

FAIR data principles rely on metadata



Scope of Project

Subject Terms

Do not copy/paste multiple terms into this field. Terms must be entered individually.

× Russia × Industry × Factories × Russian Empire × Corporations

JEL Classification

× L20 General × N63 Europe: Pre-1913 × O43 Institutions and Growth

Manuscript Number

AER-2015-1656.R3

Geographic Coverage

European Russia (Russian Empire)

Time Period(s)

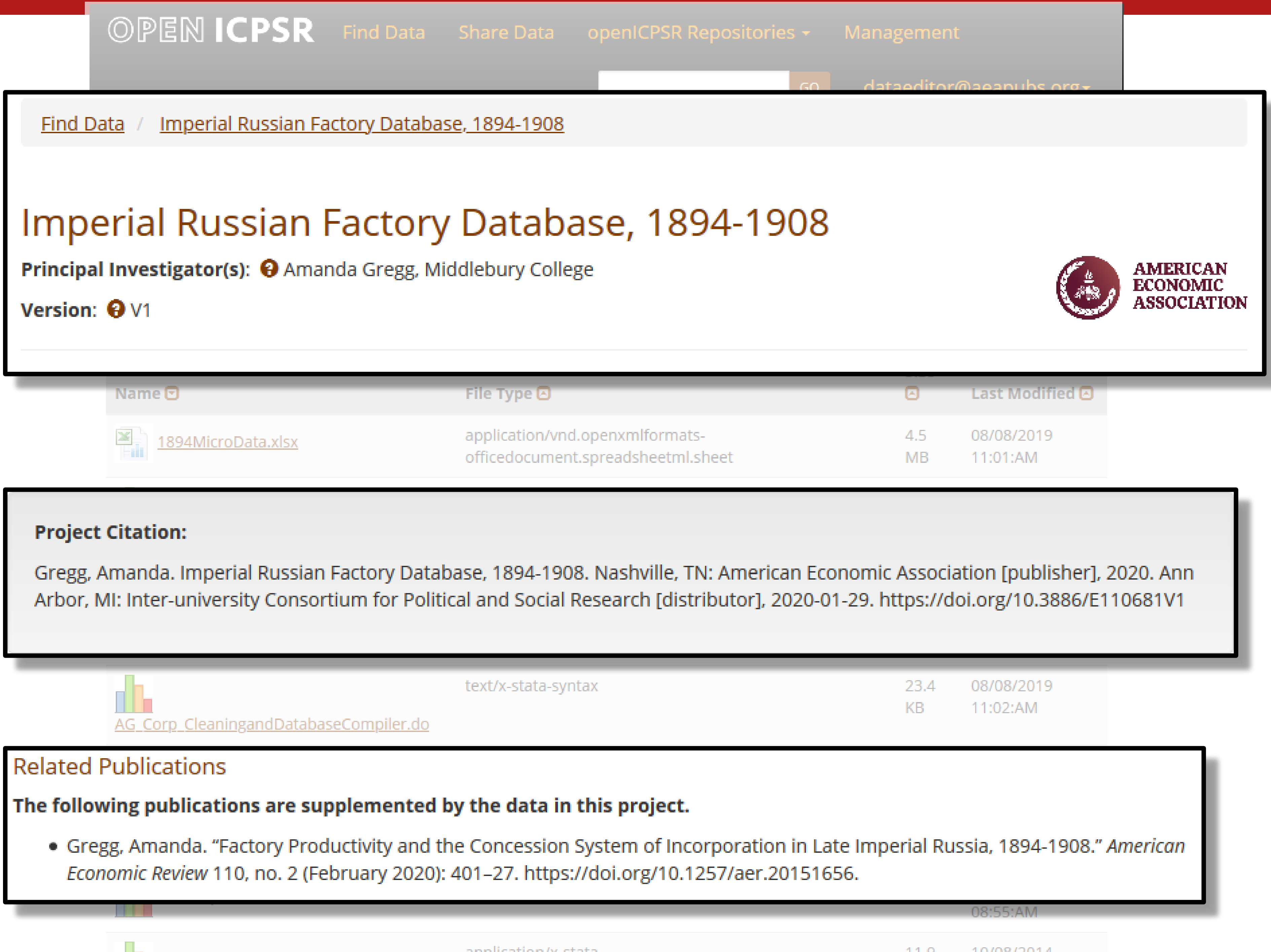
1894 – 1908 (Three years: 1894, 1900, and 1908)

Collection Date(s)

Universe

Manufacturing establishments in the European part of the Russian Empire.

Data Type(s)



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

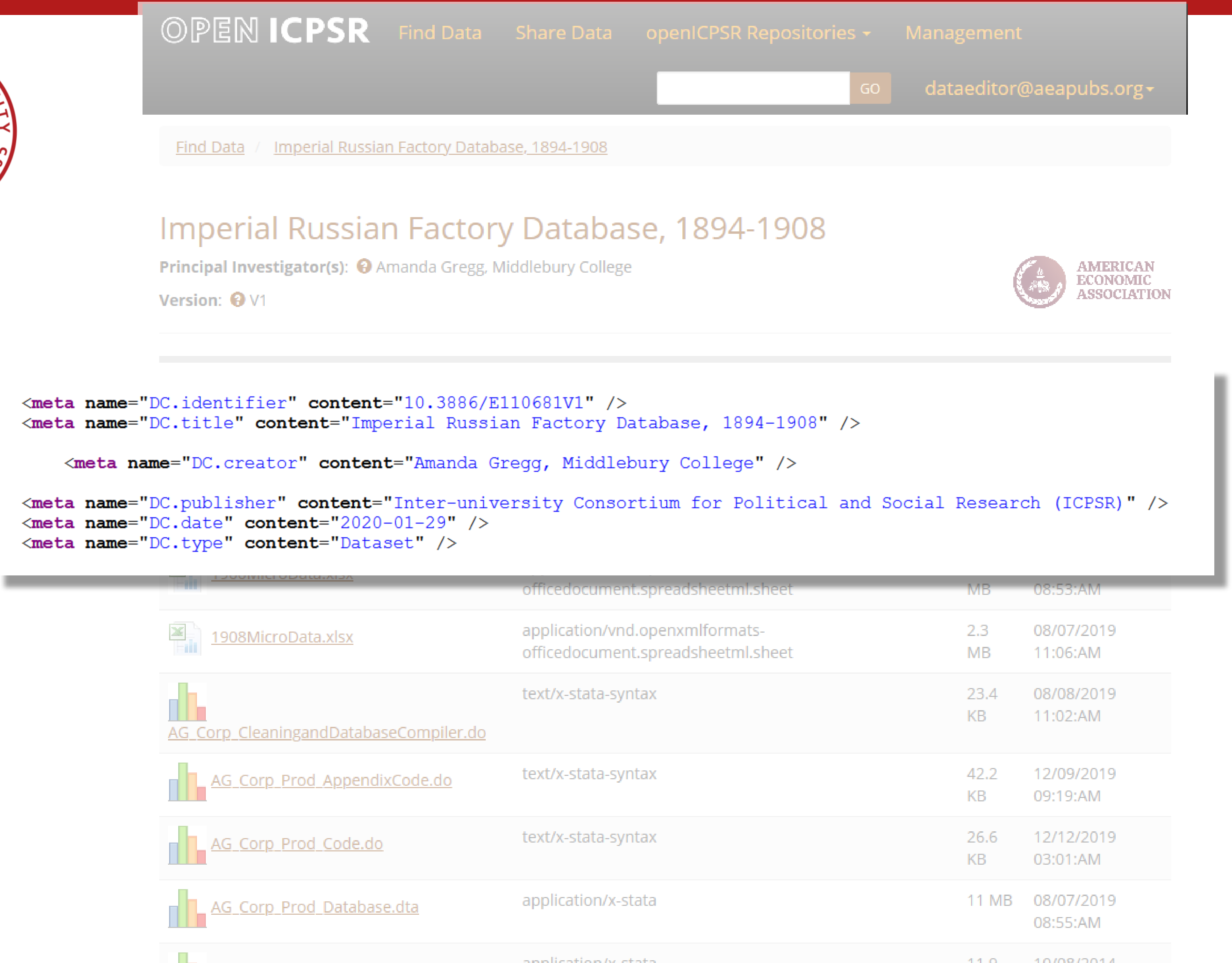
Project Citation:

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

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.



Imperial Russian Factory Database, 1894-1908

Principal Investigator(s): Amanda Gregg, Middlebury College

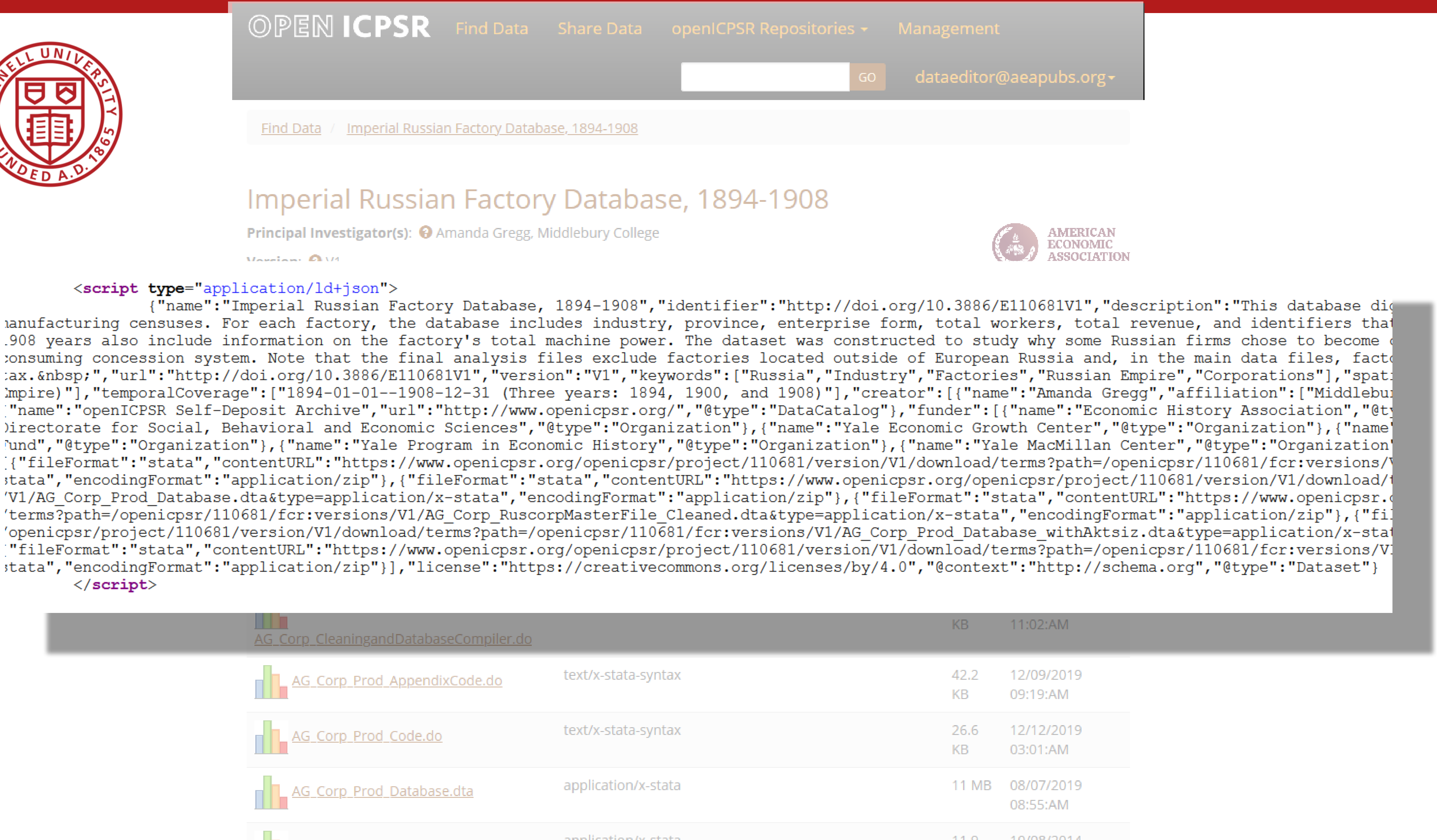
Version: V1

```
<meta name="DC.identifier" content="10.3886/E110681V1" />
<meta name="DC.title" content="Imperial Russian Factory Database, 1894-1908" />

    <meta name="DC.creator" content="Amanda Gregg, Middlebury College" />

<meta name="DC.publisher" content="Inter-university Consortium for Political and Social Research (ICPSR)" />
<meta name="DC.date" content="2020-01-29" />
<meta name="DC.type" content="Dataset" />
```

	officedocument.spreadsheetml.sheet	MB	08:53:AM
1908MicroData.xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	2.3 MB	08/07/2019 11:06:AM
AG_Corp_CleaningandDatabaseCompiler.do	text/x-stata-syntax	23.4 KB	08/08/2019 11:02:AM
AG_Corp_Prod_AppendixCode.do	text/x-stata-syntax	42.2 KB	12/09/2019 09:19:AM
AG_Corp_Prod_Code.do	text/x-stata-syntax	26.6 KB	12/12/2019 03:01:AM
AG_Corp_Prod_Database.dta	application/x-stata	11 MB	08/07/2019 08:55:AM



Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

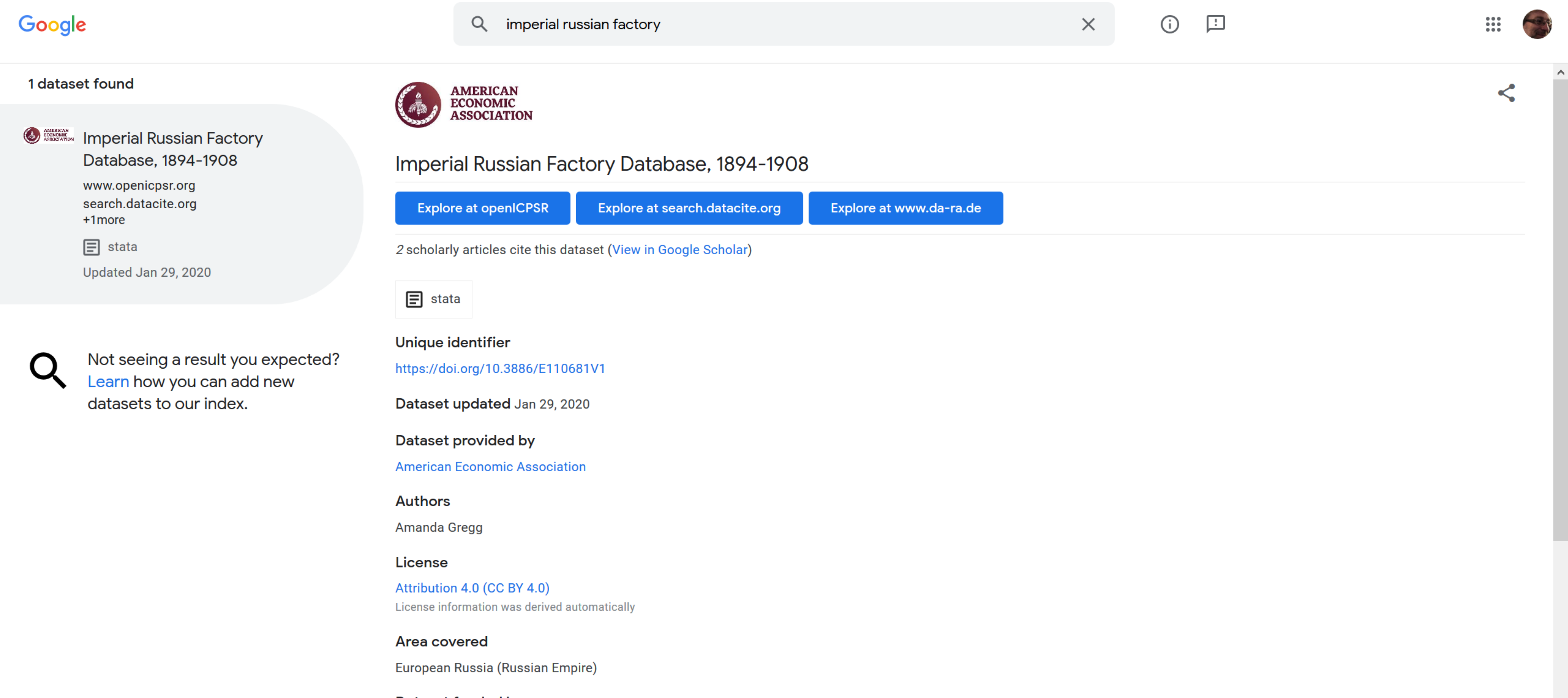
Principal Investigator(s): Amanda Gregg, Middlebury College

Version: V1

```
        <script type="application/ld+json">
                {"name":"Imperial Russian Factory Database, 1894-1908","identifier":"http://doi.org/10.3886/E110681V1","description":"This database dig
 anufacturing censuses. For each factory, the database includes industry, province, enterprise form, total workers, total revenue, and identifiers that
 908 years also include information on the factory's total machine power. The dataset was constructed to study why some Russian firms chose to become c
 onsuming concession system. Note that the final analysis files exclude factories located outside of European Russia and, in the main data files, facto
 ax.&nbsp;","url":"http://doi.org/10.3886/E110681V1","version":"V1","keywords":["Russia","Industry","Factories","Russian Empire","Corporations"],"spati
 mpire)"],"temporalCoverage":["1894-01-01--1908-12-31 (Three years: 1894, 1900, and 1908)"],"creator":[{"name":"Amanda Gregg","affiliation":["Middlebur
 "name":"openICPSR Self-Deposit Archive","url":"http://www.openicpsr.org/","@type":"DataCatalog"},"funder":[{"name":"Economic History Association","@ty
 irectorate for Social, Behavioral and Economic Sciences","@type":"Organization"},{"name":"Yale Economic Growth Center","@type":"Organization"},{"name"
 und","@type":"Organization"},{"name":"Yale Program in Economic History","@type":"Organization"},{"name":"Yale MacMillan Center","@type":"Organization"
 {"fileFormat":"stata","contentURL":"https://www.openicpsr.org/openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V
 tata","encodingFormat":"application/zip"},{"fileFormat":"stata","contentURL":"https://www.openicpsr.org/openicpsr/project/110681/version/V1/download/t
 V1/AG_Corp_Prod_Database.dta&type=application/x-stata","encodingFormat":"application/zip"},{"fileFormat":"stata","contentURL":"https://www.openicpsr.o
 terms?path=/openicpsr/110681/fcr:versions/V1/AG_Corp_RuscorpMasterFile_Cleaned.dta&type=application/x-stata","encodingFormat":"application/zip"},{"fil
 openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V1/AG_Corp_Prod_Database_withAktsiz.dta&type=application/x-stat
 "fileFormat":"stata","contentURL":"https://www.openicpsr.org/openicpsr/project/110681/version/V1/download/terms?path=/openicpsr/110681/fcr:versions/V1
 tata","encodingFormat":"application/zip"}],"license":"https://creativecommons.org/licenses/by/4.0","@context":"http://schema.org","@type":"Dataset"}
        </script>
```

AG_Corp_CleaningandDatabaseCompiler.do		KB	11:02:AM
AG_Corp_Prod_AppendixCode.do	text/x-stata-syntax	42.2 KB	12/09/2019 09:19:AM
AG_Corp_Prod_Code.do	text/x-stata-syntax	26.6 KB	12/12/2019 03:01:AM
AG_Corp_Prod_Database.dta	application/x-stata	11 MB	08/07/2019 08:55:AM
	application/x-stata	11.9	10/08/2014

… and findability relies on metadata



Google

imperial russian factory

1 dataset found

Imperial Russian Factory Database, 1894-1908

www.openicpsr.org

search.datacite.org

+1more

stata

Updated Jan 29, 2020

Not seeing a result you expected? Learn how you can add new datasets to our index.

AMERICAN ECONOMIC ASSOCIATION

Imperial Russian Factory Database, 1894-1908

Explore at openICPSR | Explore at search.datacite.org | Explore at www.da-ra.de

2 scholarly articles cite this dataset (View in Google Scholar)

stata

Unique identifier

https://doi.org/10.3886/E110681V1

Dataset updated Jan 29, 2020

Dataset provided by

American Economic Association

Authors

Amanda Gregg

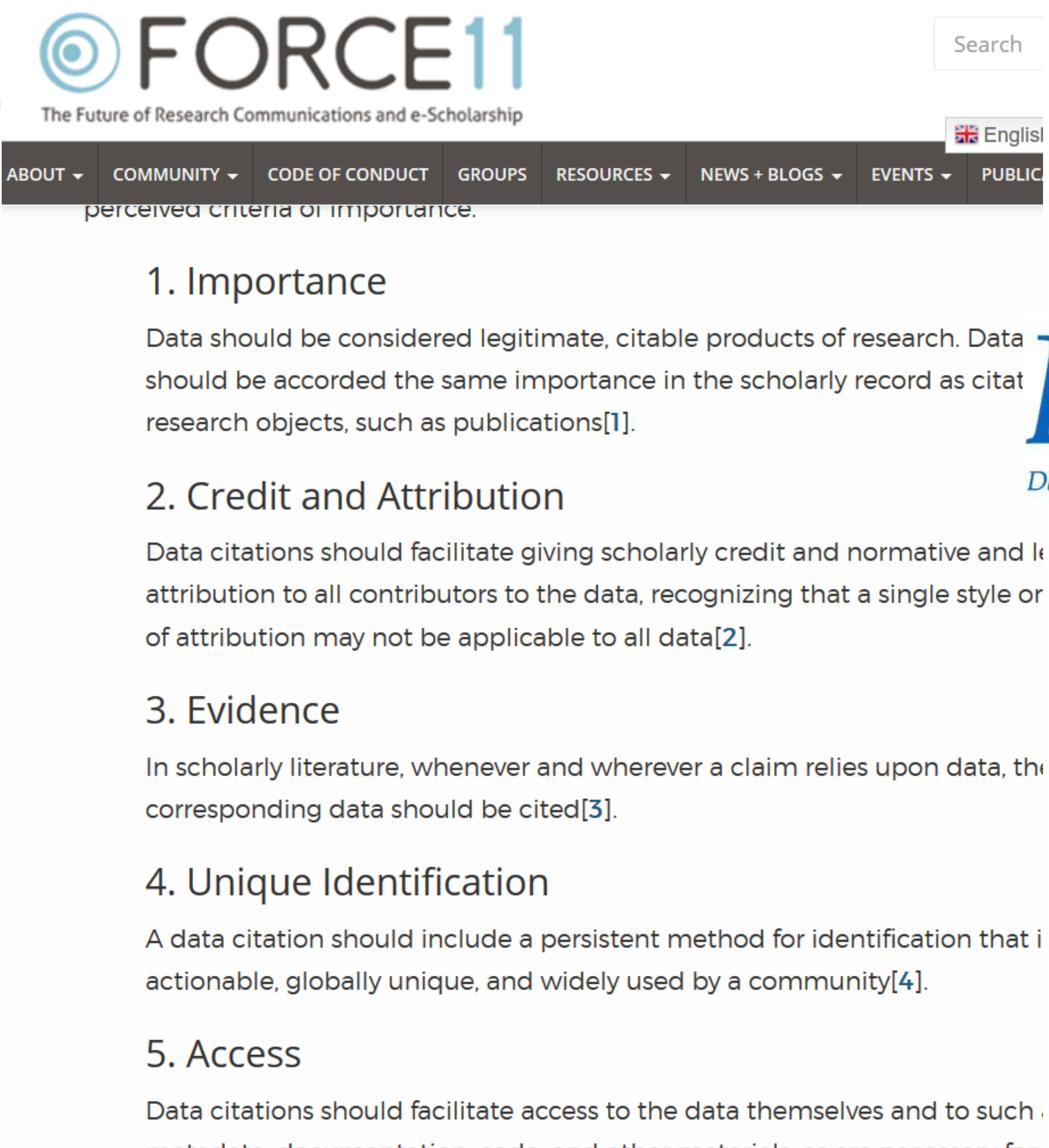
License

Attribution 4.0 (CC BY 4.0)

License information was derived automatically

Area covered

European Russia (Russian Empire)



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 l 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, th corresponding data should be cited[3].

4. Unique Identification

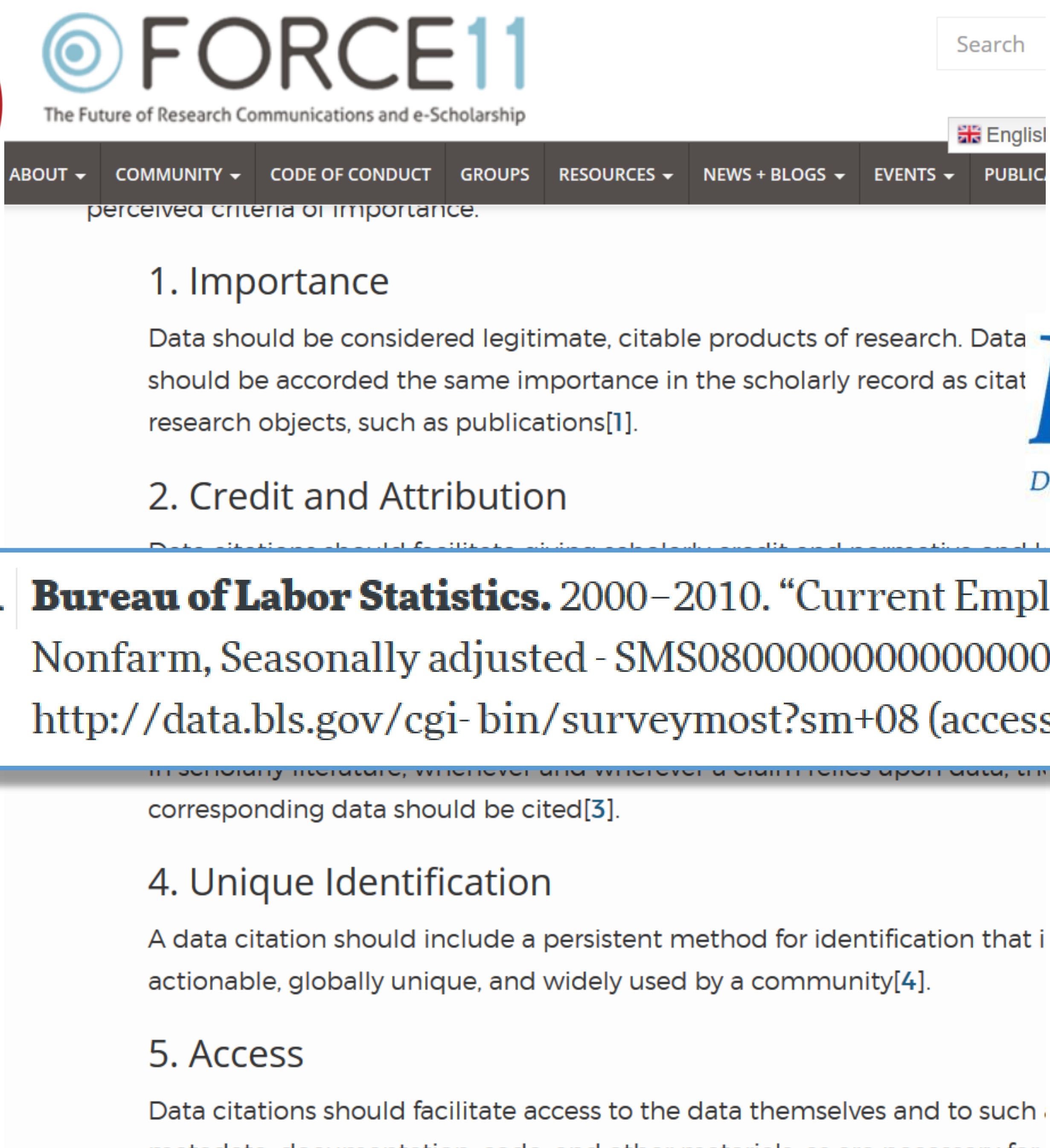
A data citation should include a persistent method for identification that i actionable, globally unique, and widely used by a community[4].

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].



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 l

in scholarly literature, whenever and wherever a claim relies upon data, th 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

DC^1

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).

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

Citing restricted-access data

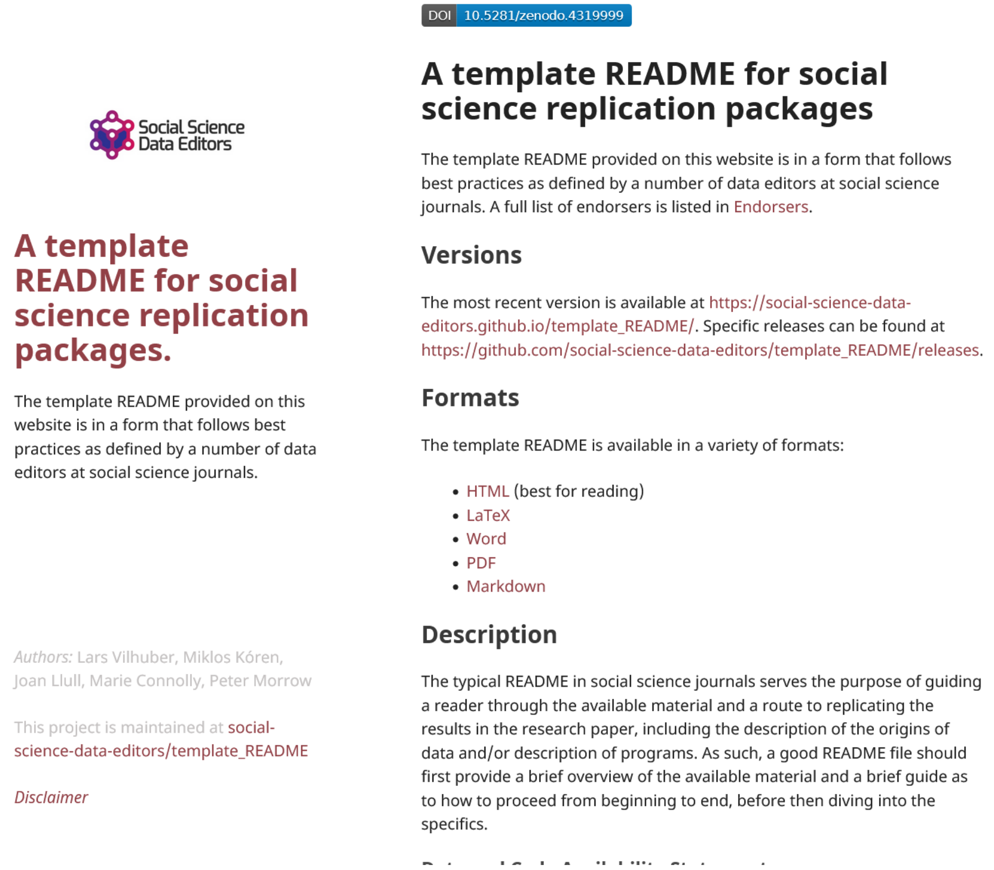
“Well, I can’t download the data, so I can’t cite it.”

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"



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

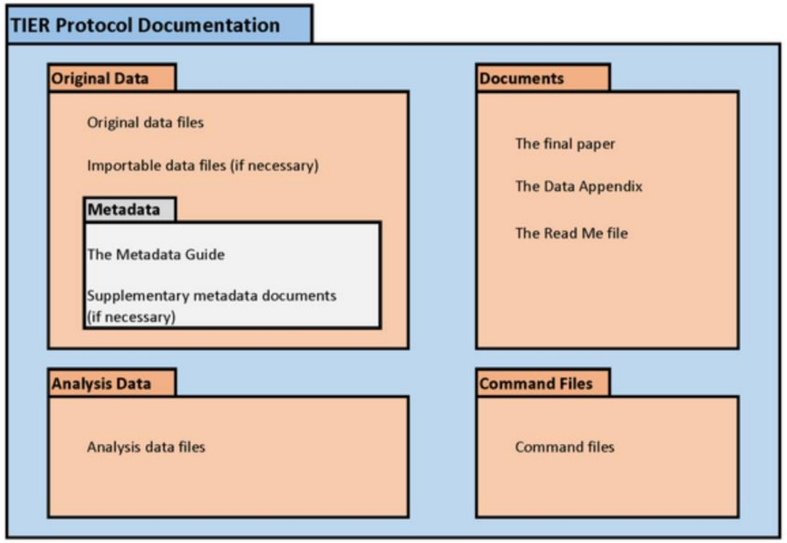
https://doi.org/10.5281/zenodo.4319999

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



TIER Protocol Documentation

Original Data

Original data files

Importable data files (if necessary)

Metadata

The Metadata Guide

Supplementary metadata documents (if necessary)

Documents

The final paper

The Data Appendix

The Read Me file

Analysis Data

Analysis data files

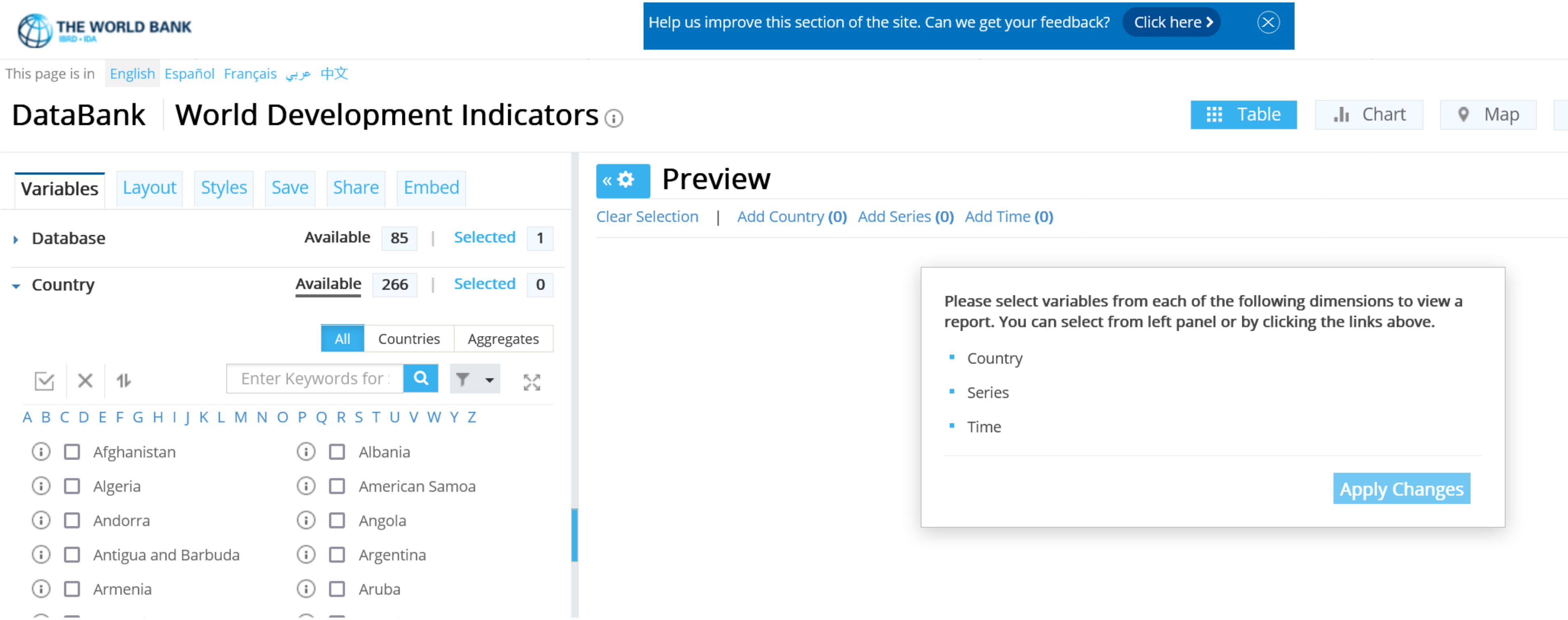
Command Files

Command files

[redacted]

API? But the interface is so cool!

- World Development Indicators



THE WORLD BANK
IBRD • IDA

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 | Chart | Map

Variables | Layout | Styles | Save | Share | Embed

Database	Available	85	Selected	1
Country	Available	266	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

Preview

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

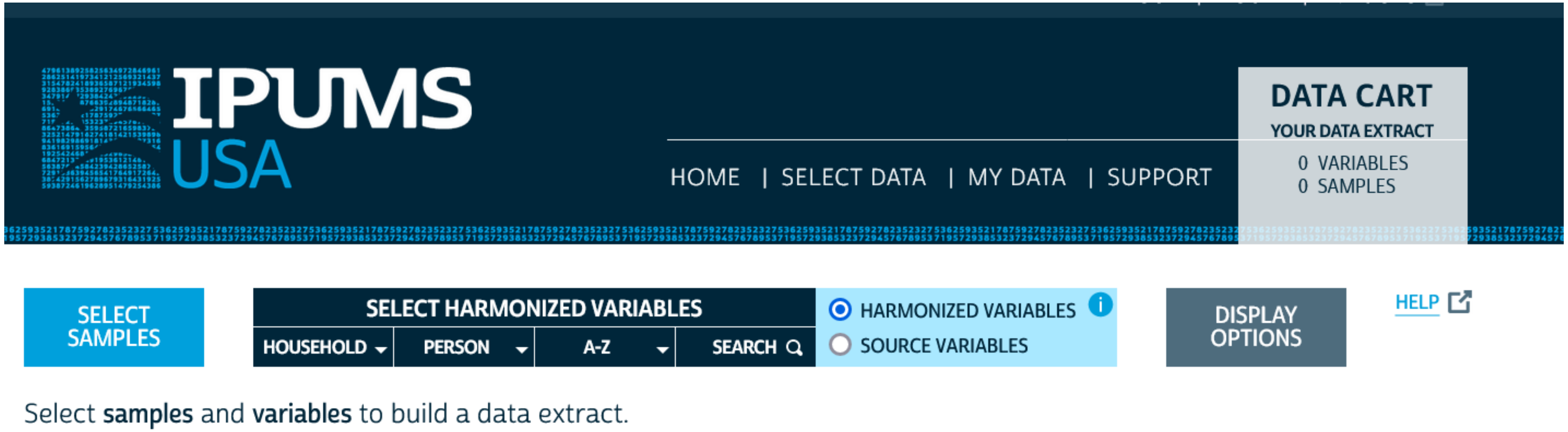
Please select variables from each of the following dimensions to view a report. You can select from left panel or by clicking the links above.

- Country
- Series
- Time

Apply Changes

API? But the interface is so cool!

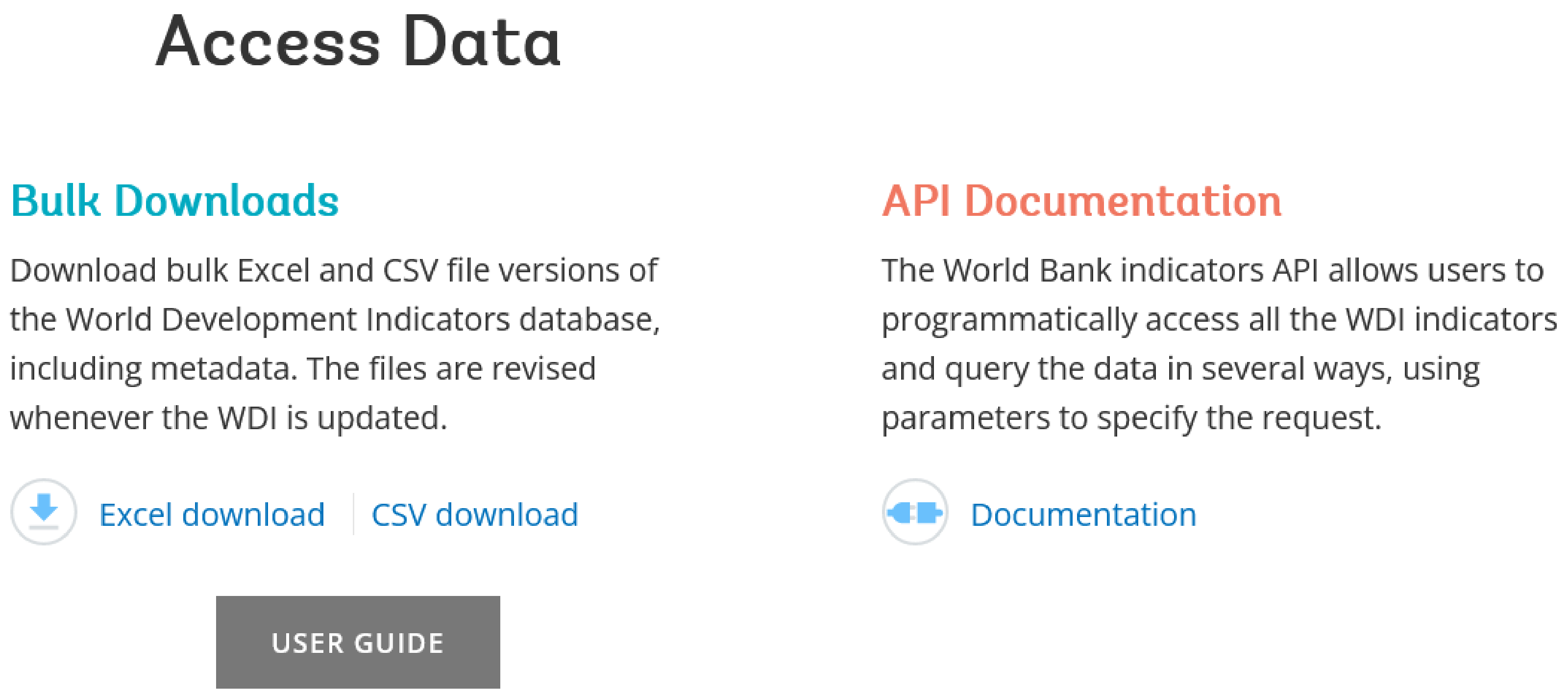
- IPUMS



Select **samples** and **variables** to build a data extract.

API or Bulk Download

- World Development Indicators



Access Data

Bulk Downloads

Download bulk Excel and CSV file versions of the World Development Indicators database, including metadata. The files are revised whenever the WDI is updated.

Excel download | CSV download

USER GUIDE

API Documentation

The World Bank indicators API allows users to programmatically access all the WDI indicators and query the data in several ways, using parameters to specify the request.

Documentation

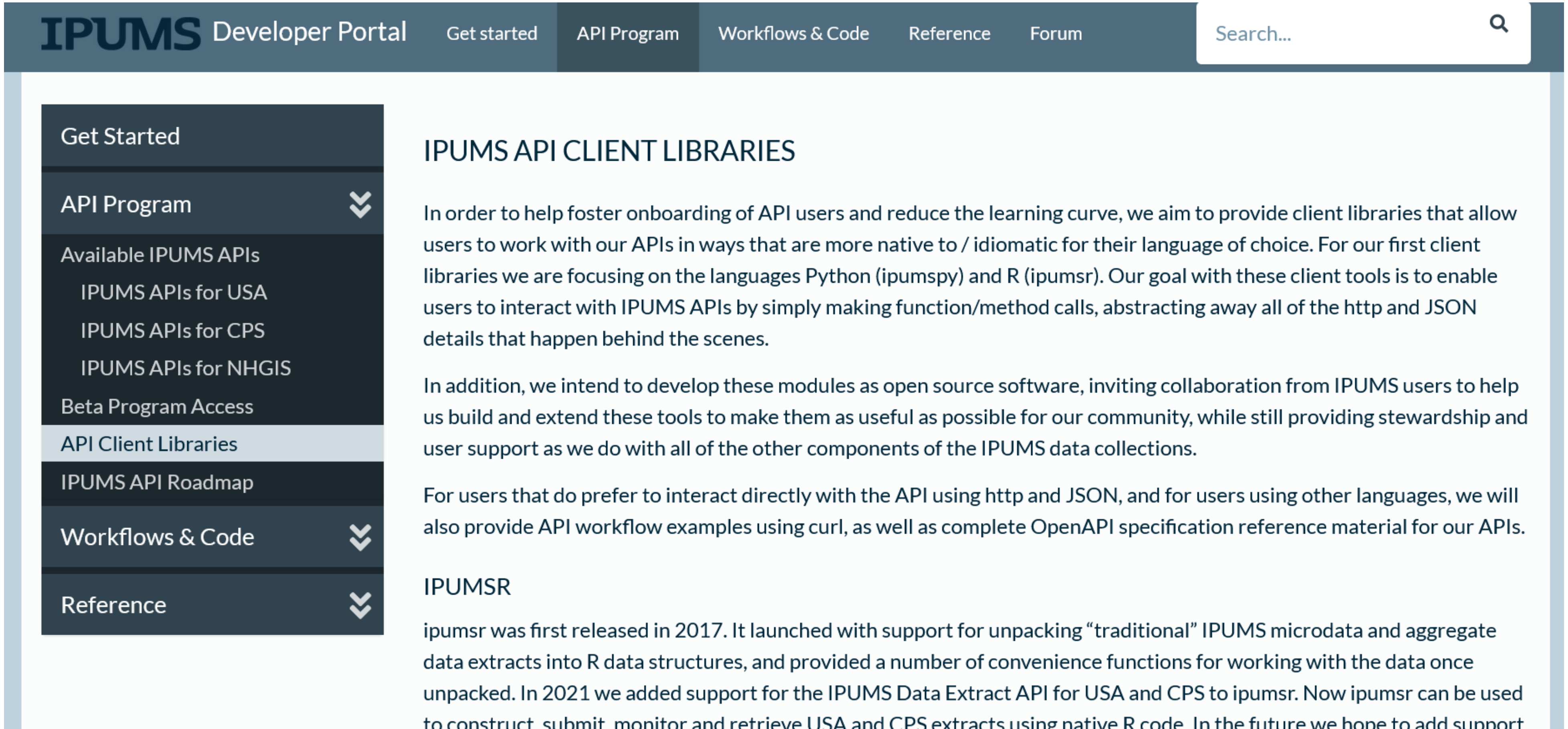
API or Bulk Download

- 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)



IPUMS API CLIENT LIBRARIES

In order to help foster onboarding of API users and reduce the learning curve, we aim to provide client libraries that allow users to work with our APIs in ways that are more native to / idiomatic for their language of choice. For our first client libraries we are focusing on the languages Python (ipumspy) and R (ipumsr). Our goal with these client tools is to enable users to interact with IPUMS APIs by simply making function/method calls, abstracting away all of the http and JSON details that happen behind the scenes.

In addition, we intend to develop these modules as open source software, inviting collaboration from IPUMS users to help us build and extend these tools to make them as useful as possible for our community, while still providing stewardship and user support as we do with all of the other components of the IPUMS data collections.

For users that do prefer to interact directly with the API using http and JSON, and for users using other languages, we will also provide API workflow examples using curl, as well as complete OpenAPI specification reference material for our APIs.

IPUMSR

ipumsr was first released in 2017. It launched with support for unpacking “traditional” IPUMS microdata and aggregate data extracts into R data structures, and provided a number of convenience functions for working with the data once unpacked. In 2021 we added support for the IPUMS Data Extract API for USA and CPS to ipumsr. Now ipumsr can be used to construct, submit, monitor and retrieve USA and CPS extracts using native R code. In the future we hope to add support

2. Keeping track of data:

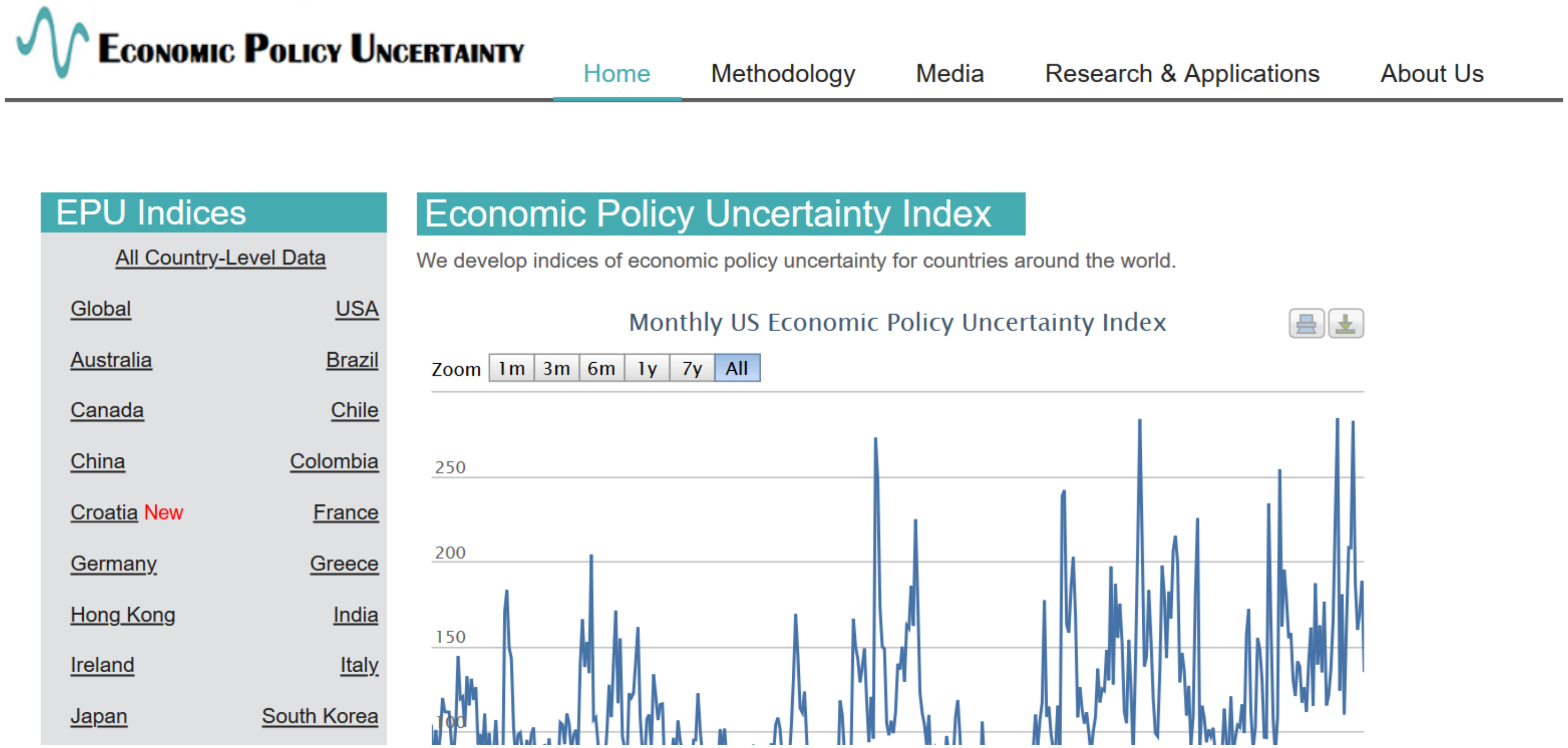
Don’t forget to check the TERMS of USE!

Observation 4

(Academic)
Social scientists
do not read
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



Economic Policy Uncertainty

Home Methodology Media Research & Applications About Us

EPU Indices

All Country-Level Data

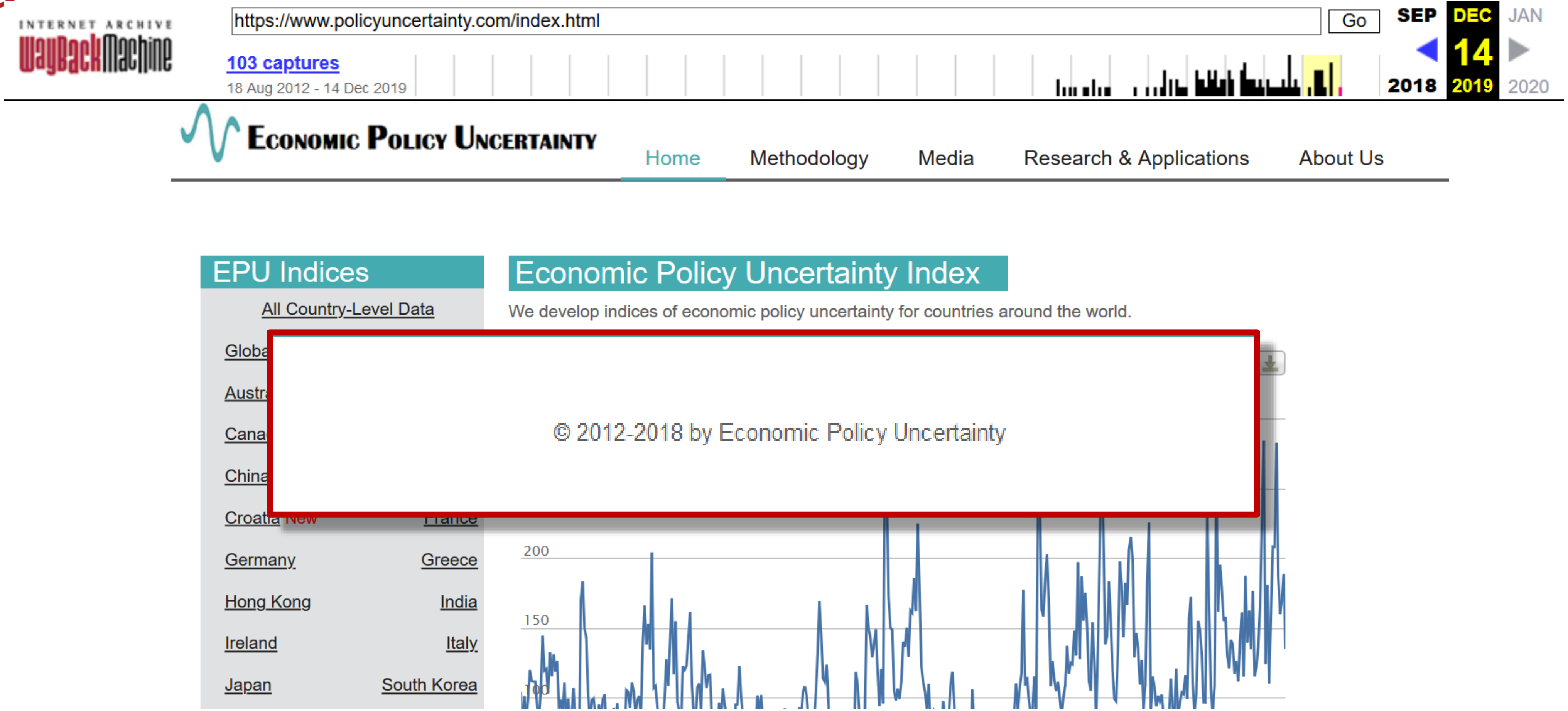
Global USA
Australia Brazil
Canada Chile
China Colombia
Croatia New France
Germany Greece
Hong Kong India
Ireland Italy
Japan South Korea

Economic Policy Uncertainty Index

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

Monthly US Economic Policy Uncertainty Index

Example 2: Academic data publisher



https://www.policyuncertainty.com/index.html

103 captures

18 Aug 2012 - 14 Dec 2019

Economic Policy Uncertainty

EPU Indices

All Country-Level Data

Globa

Austr

Cana

China

Croatia New

France

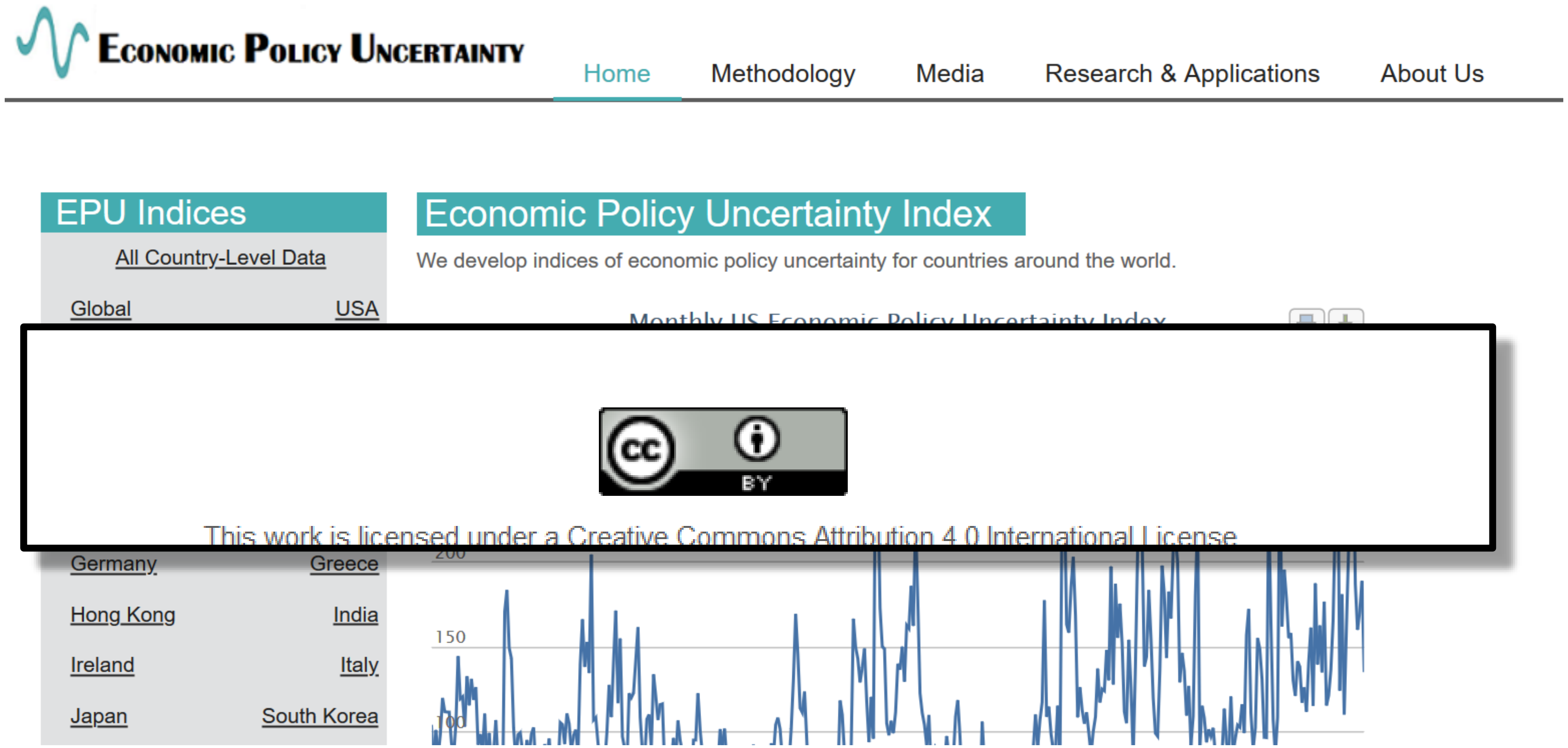
Germany	Greece
Hong Kong	India
Ireland	Italy
Japan	South Korea

Economic Policy Uncertainty Index

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

© 2012-2018 by Economic Policy Uncertainty

Example 2: Academic data publisher-new!


Economic Policy Uncertainty
Home
Methodology
Media
Research & Applications
About Us
Monthly US Economic Policy Uncertainty Index
200
150
100

EPU Indices

All Country-Level Data

Global	USA
Germany	Greece
Hong Kong	India
Ireland	Italy
Japan	South Korea

Economic Policy Uncertainty Index

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

Rights to <u>use</u> data

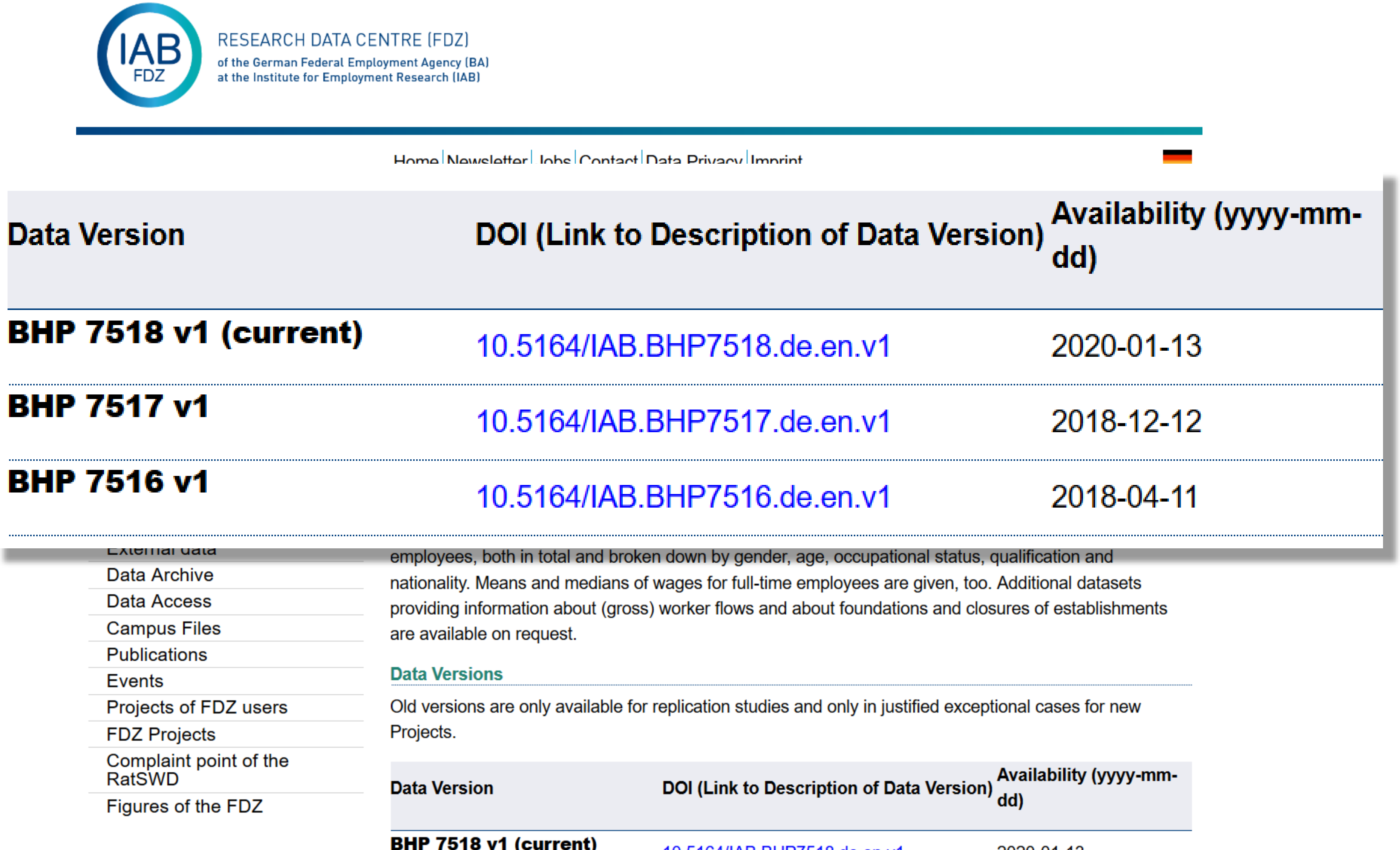
- You browsed a website
- You purchased the data
- You signed a data use agreement
- You created the data (lab experiment)
- You had survey respondents consent to use (IRB approval!)

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



IAB FDZ

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

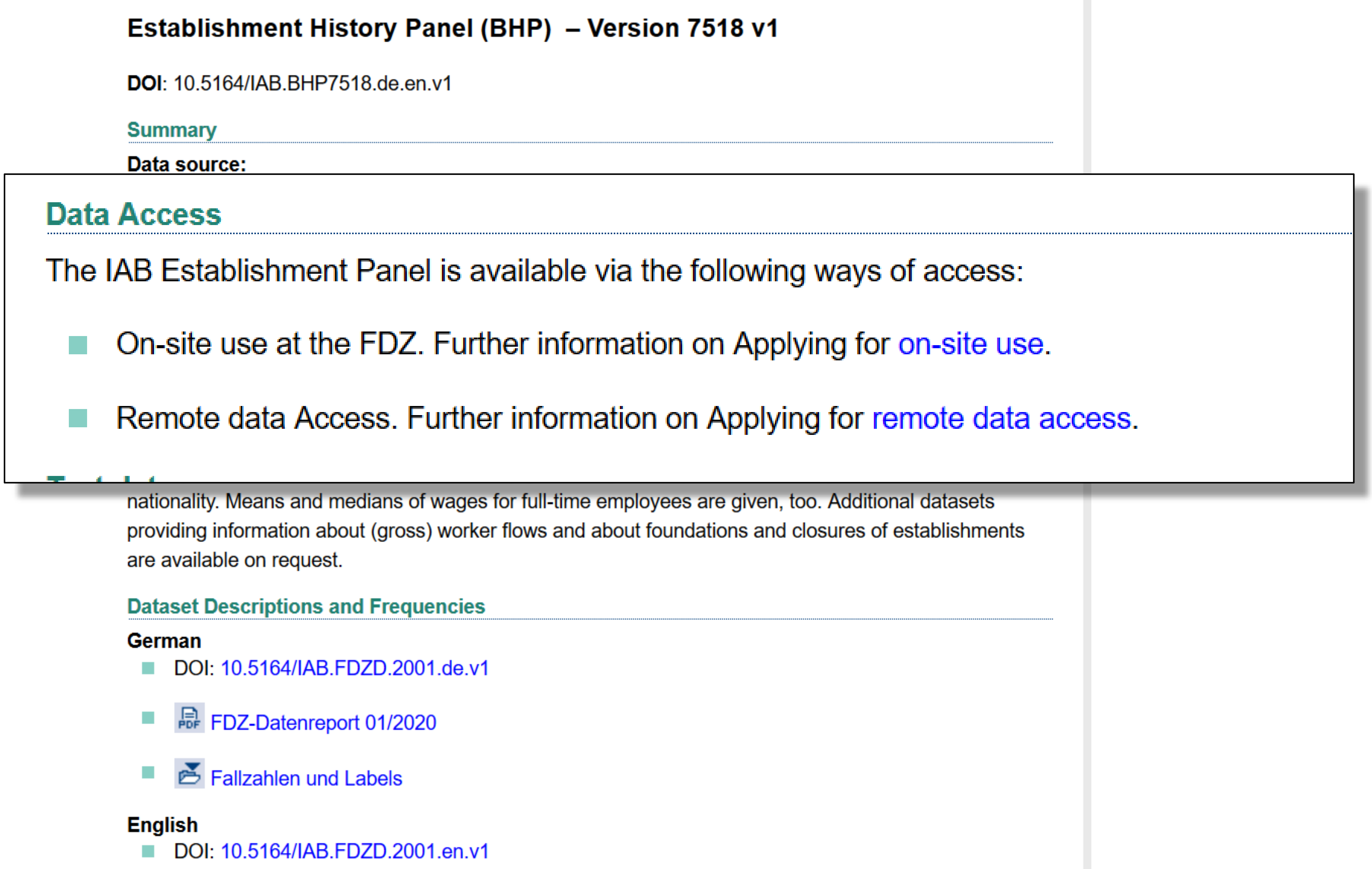
employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are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

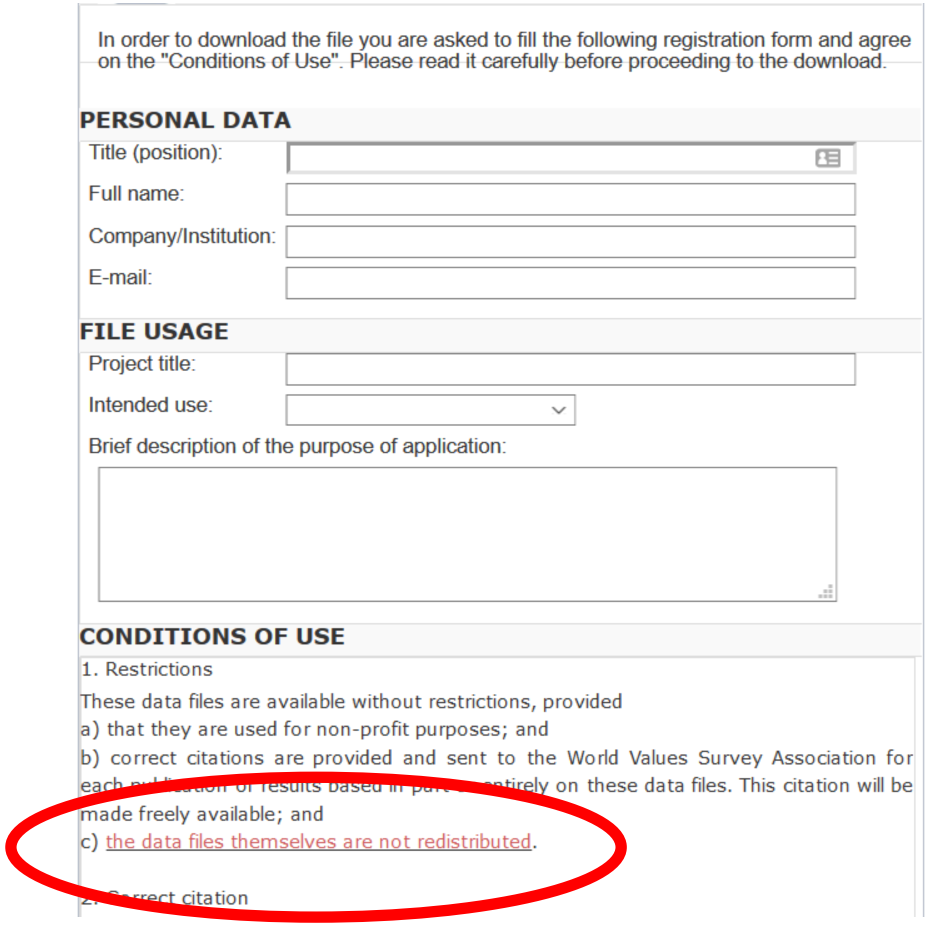
German

- 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!

Observation 5

(Academic)
Social scientists
do not read
the manual…

(or at least not some key parts)

Extreme examples

- Matlab-based simulation
- Real example, 10 figures, 4 panels each

- For Figure 5a, comment line 52, uncomment line 151, run the code, then copy the figure into your document.
- For Figure 5b, comment line 151 again, leave line 52 commented, and change the parameter on line 75 to “3”
- ….

Extreme examples

- Matlab-based simulation

- For Figure 5a, comment line 52, uncomment line 151, run the

- ....

Write re-usable code

Use primitive I/O to read parameter files

Extreme examples

- Stata-based estimation
- 4 variants

- Run the data creation programs, then copy the data to Folder A
- Copy programs “b.do” and “c.do” from Folder A to Folder B, but modify “c.do” on line 20
- Once done, convert the output from “d.do” to a Matlab file, and run the simulation in Folder B/C
- ….

Extreme examples

- Stata-based estimation
- 4 minutes

- Run the data creation programs, then copy the data to Folder A
- Once done, convert the output from “d.do” to a Matlab file, and run the simulation in Folder B/C
- ….

Re-use code files (ado)

Use relative or root-relative paths

Ideal setup

- 1 program to prepare the setup
  - Installs all packages
  - Creates all directories
- 1 program (or a very small number) that creates the rest
  - Possibly with macros/ ado files/ subroutines
  - Possibly with parameter files that might differ per directory
- All tables and figures are output programmatically

- Setting up can be done in all languages
  - Matlab, Stata, R, Python, Fortran
- Subroutines exist in all languages
  - You might need to learn how!
- Ability to output figures and tables (Excel, LaTeX) exist in all languages

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

Lesson 3: Support by institutions is insufficient

Disciplinary institutions need funding

- Direct funding: should archives be treated like infrastructure?
  - This is the case in many NIH archives
  - Not quite as ubiquitous in social sciences
  - Where do you want to preserve 1TB per user per day for the next 50 years? It's not free…
- Indirect funding (via grants)
  - See previous slide

Lesson 3: Support by institutions is insufficient

All institutions need to consider the user experience

- Data provenance and preservation would be a lot easier to implement if scripted
    - Need for APIs both for upload and download (usage)
    - Some progress: Zenodo and Dataverse (usually Python, sometimes R)

Lesson 3: Support by institutions is insufficient

Integration of computational resources with data resources is highly inadequate

- Most journals have “data” policies, but all research compendia have code – this is a problem with many publishers
- Most data repositories are optimized for … data. Support for computational code or actual execution is at best preliminary
  - See CodeOcean, WholeTale, efforts around Dataverse
  - See various “continuous integration” using Github, Travis CI, etc. but difficult to integrate data

Solution 6: Institutional support

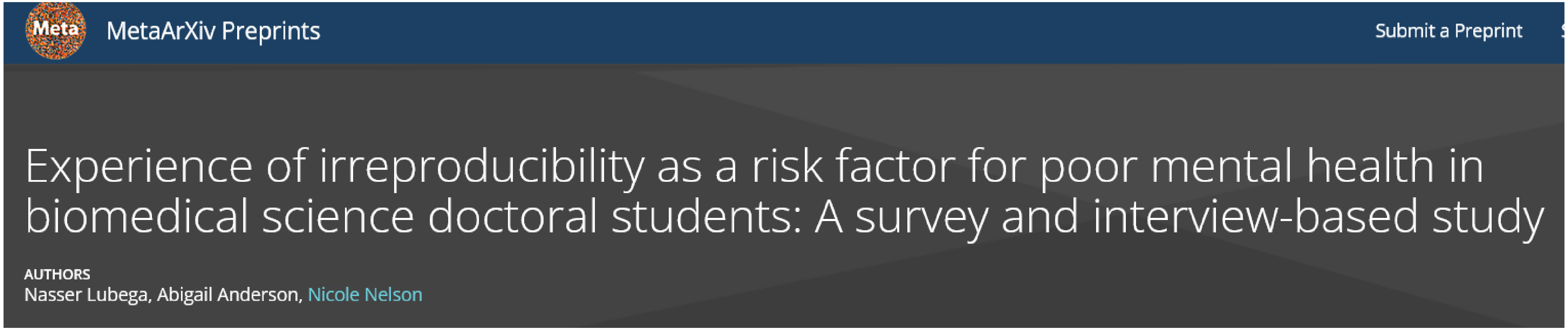
(departments, schools, libraries, IT, universities)

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!



Experience of irreproducibility as a risk factor for poor mental health in biomedical science doctoral students: A survey and interview-based study

AUTHORS
Nasser Lubega, Abigail Anderson, Nicole Nelson

The role for journals

Goal: Transportability

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

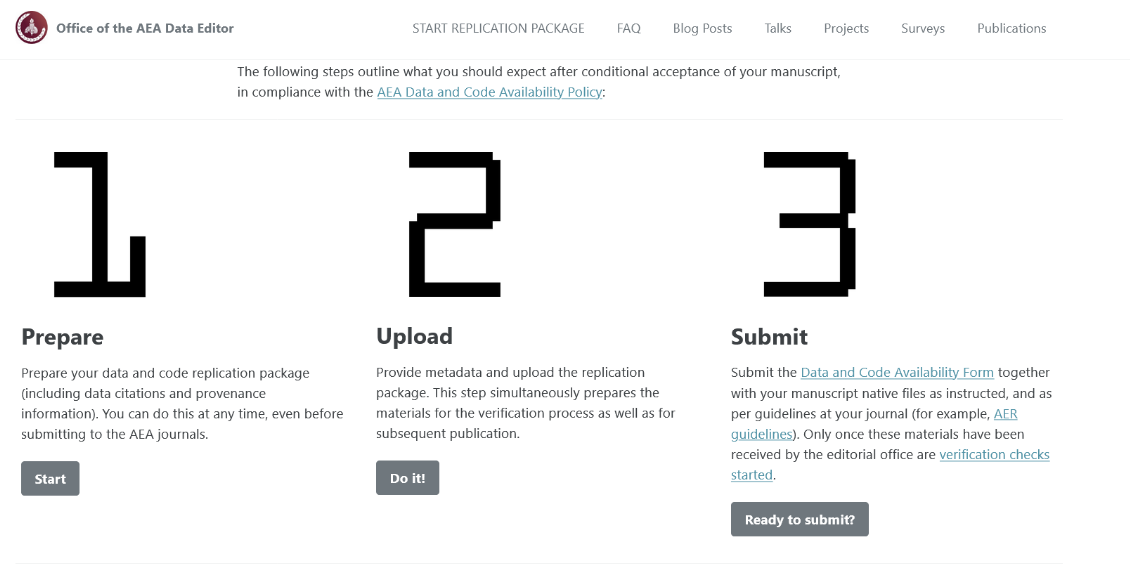
Social science "guild"

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/



The following steps outline what you should expect after conditional acceptance of your manuscript, in compliance with the AEA Data and Code Availability Policy:

1

Prepare

Prepare your data and code replication package (including data citations and provenance information). You can do this at any time, even before submitting to the AEA journals.

2

Upload

Provide metadata and upload the replication package. This step simultaneously prepares the materials for the verification process as well as for subsequent publication.

3

Submit

Submit the Data and Code Availability Form together with your manuscript native files as instructed, and as per guidelines at your journal (for example, AER guidelines). Only once these materials have been received by the editorial office are verification checks started.

Thank you!

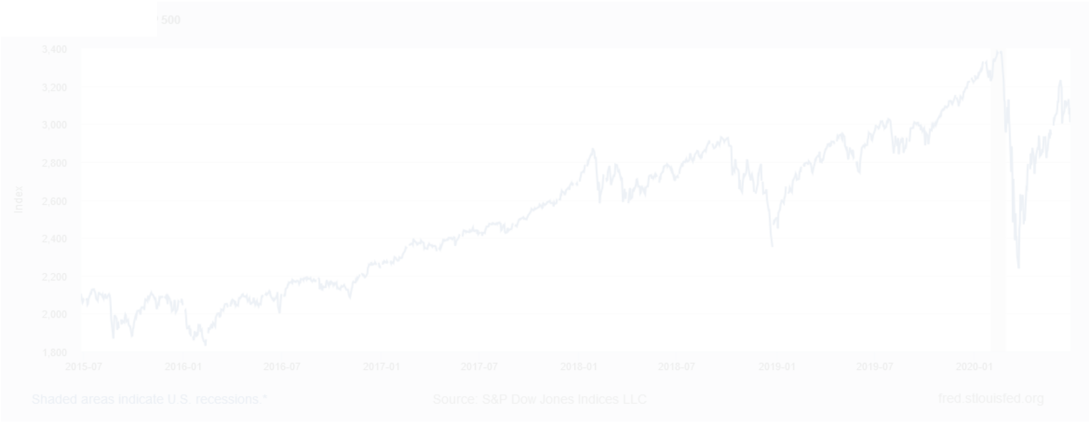
Reminder: **Students and Researchers**

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

Example of data provenance

“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



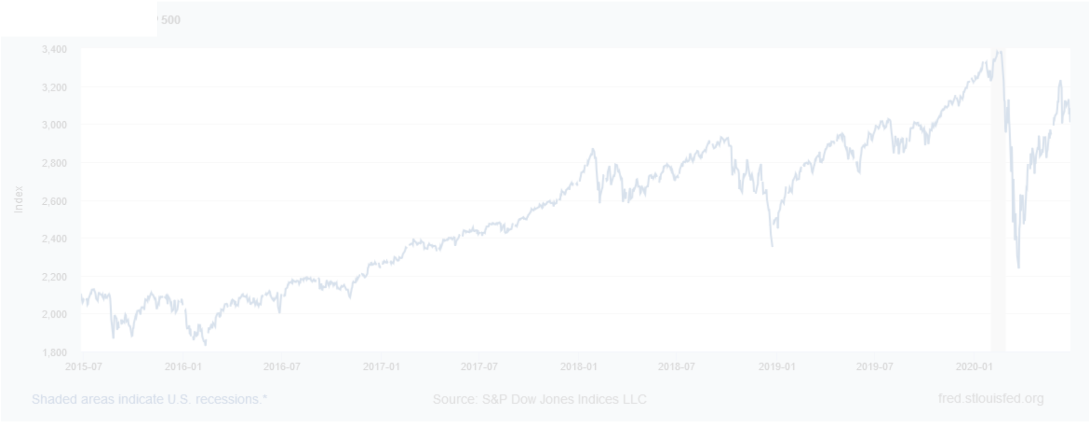
Shaded areas indicate U.S. recessions.*

Source: S&P Dow Jones Indices LLC

fred.stlouisfed.org

“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



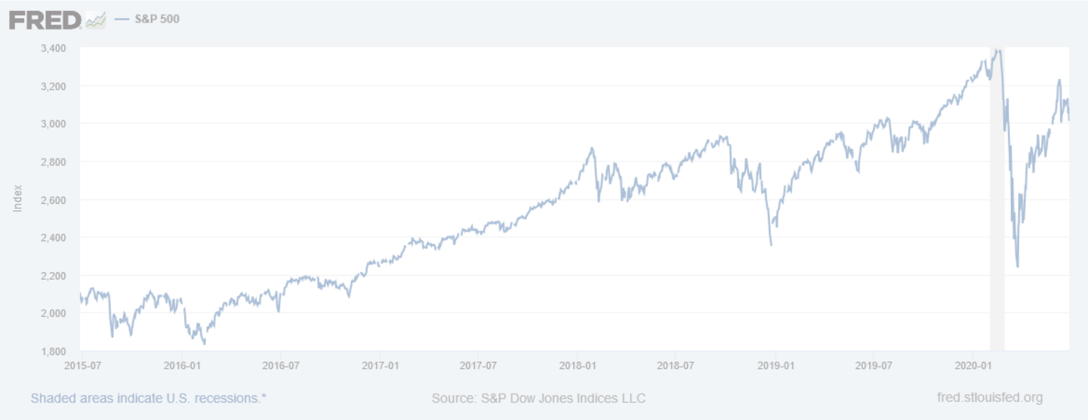
Shaded areas indicate U.S. recessions.*

Source: S&P Dow Jones Indices LLC

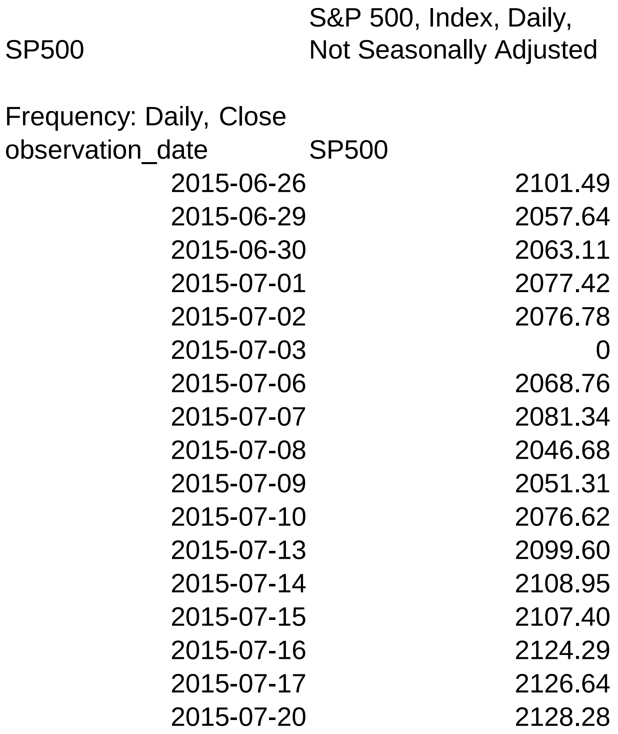
fred.stlouisfed.org

“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

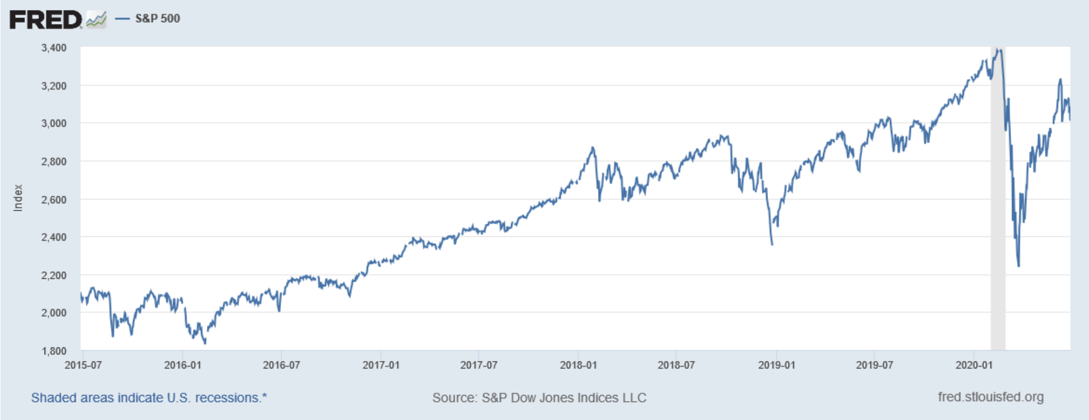


“It’s a file called SP500.xlsx, downloaded from FRED.”



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

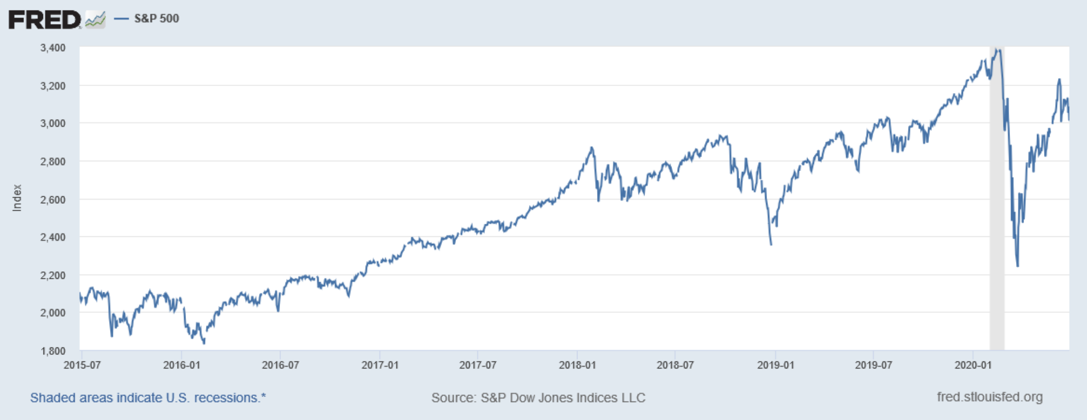
S&P Dow Jones Indices LLC. 2020. “*S&P 500 [SP500] [dataset]*”, retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SP500, June 26, 2020.



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

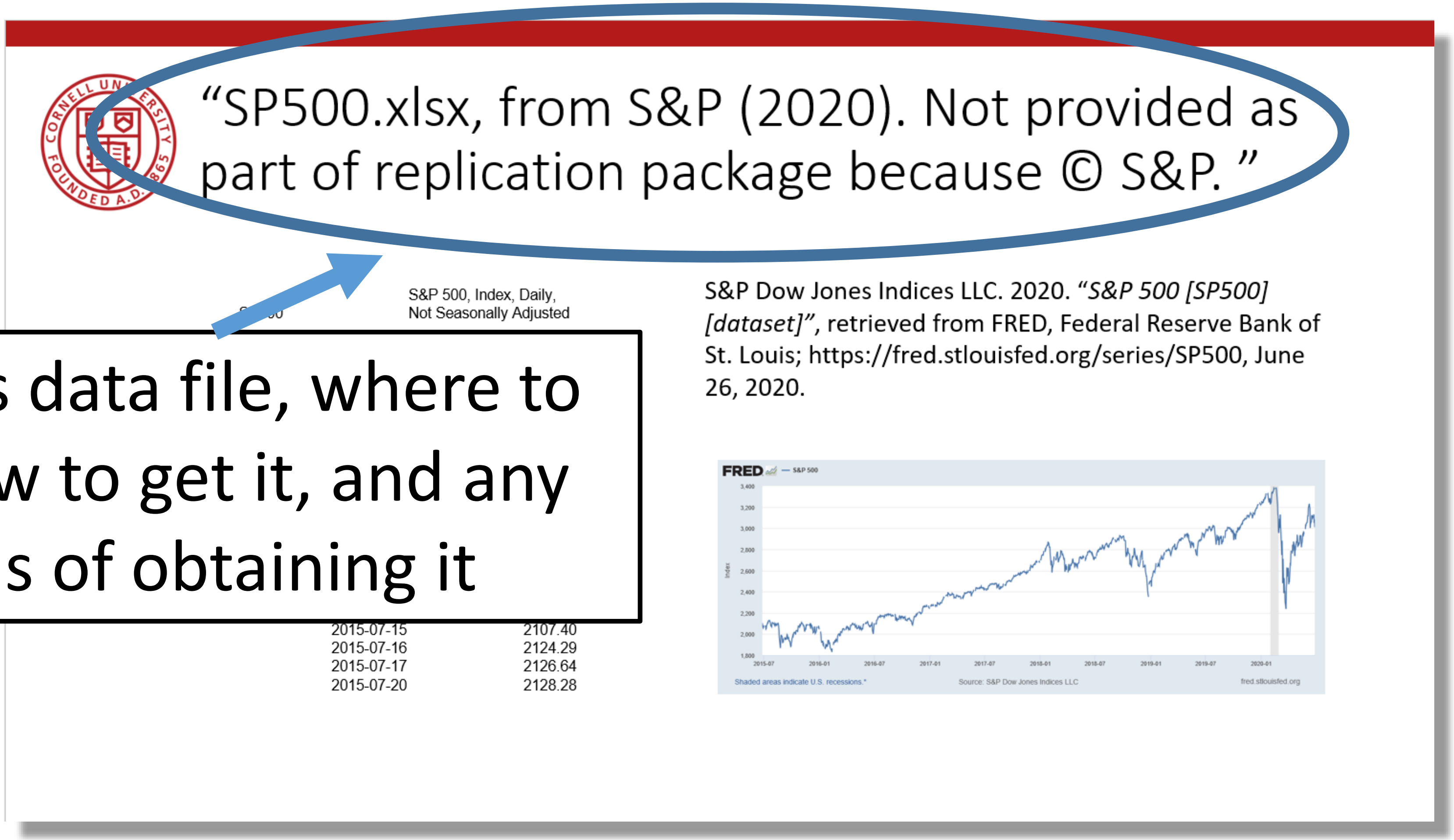
SP500	S&P 500, Index, Daily, Not Seasonally Adjusted
Frequency: Daily, Close	
observation_date	SP500
2015-06-26	2101.49
2015-06-29	2057.64
2015-06-30	2063.11
2015-07-01	2077.42
2015-07-02	2076.78
2015-07-03	0
2015-07-06	2068.76
2015-07-07	2081.34
2015-07-08	2046.68
2015-07-09	2051.31
2015-07-10	2076.62
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.



Source: S&P Dow Jones Indices LLC

Data Availability Statements



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

	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.

Shaded areas indicate U.S. recessions.*

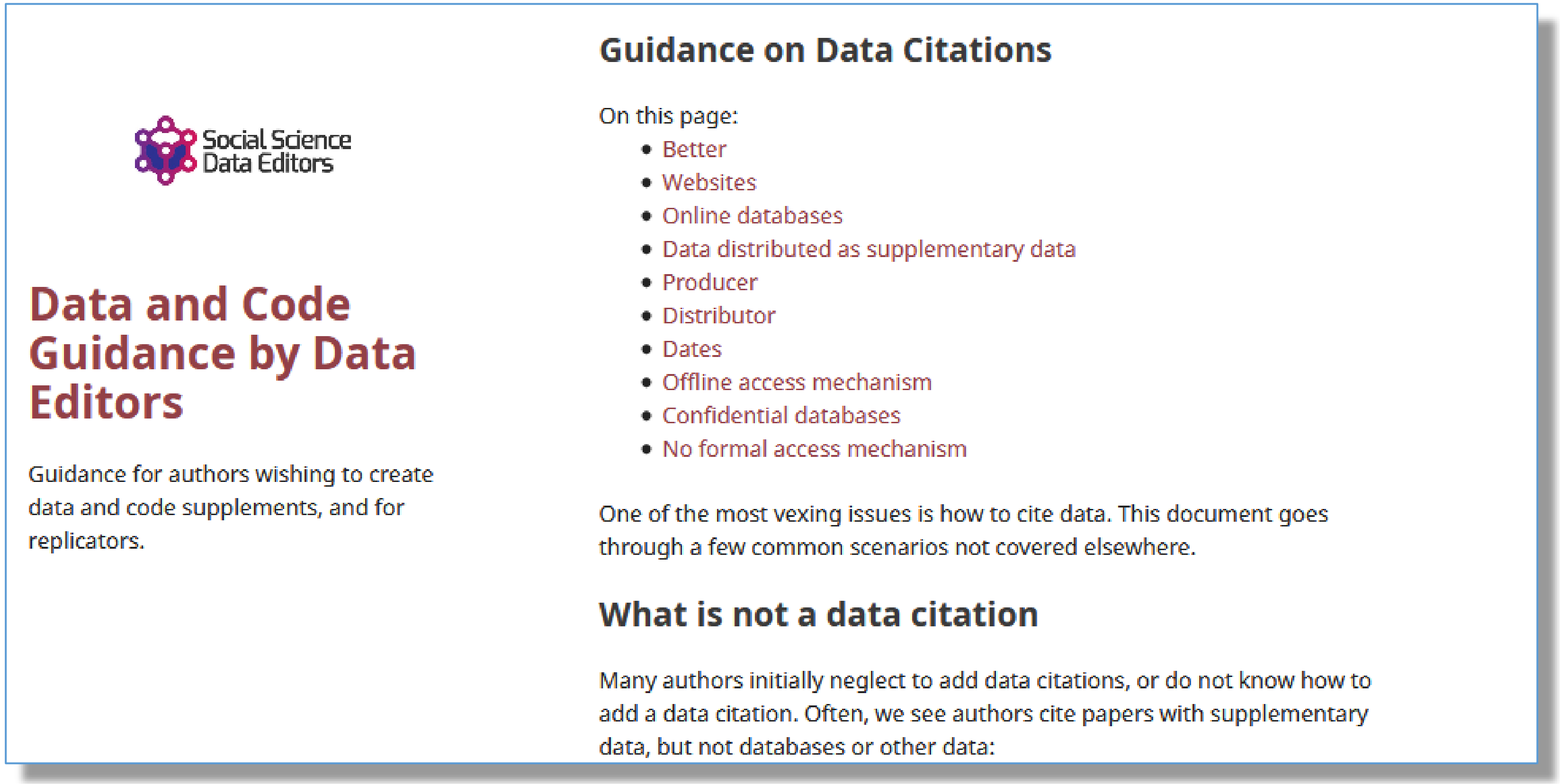
Source: S&P Dow Jones Indices LLC

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

Data Citations

Data citations

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



Social Science Data Editors

Data and Code Guidance by Data Editors

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

Guidance on Data Citations

On this page:

- Better
- Websites
- Online databases
- Data distributed as supplementary data
- Producer
- Distributor
- Dates
- Offline access mechanism
- Confidential databases
- No formal access mechanism

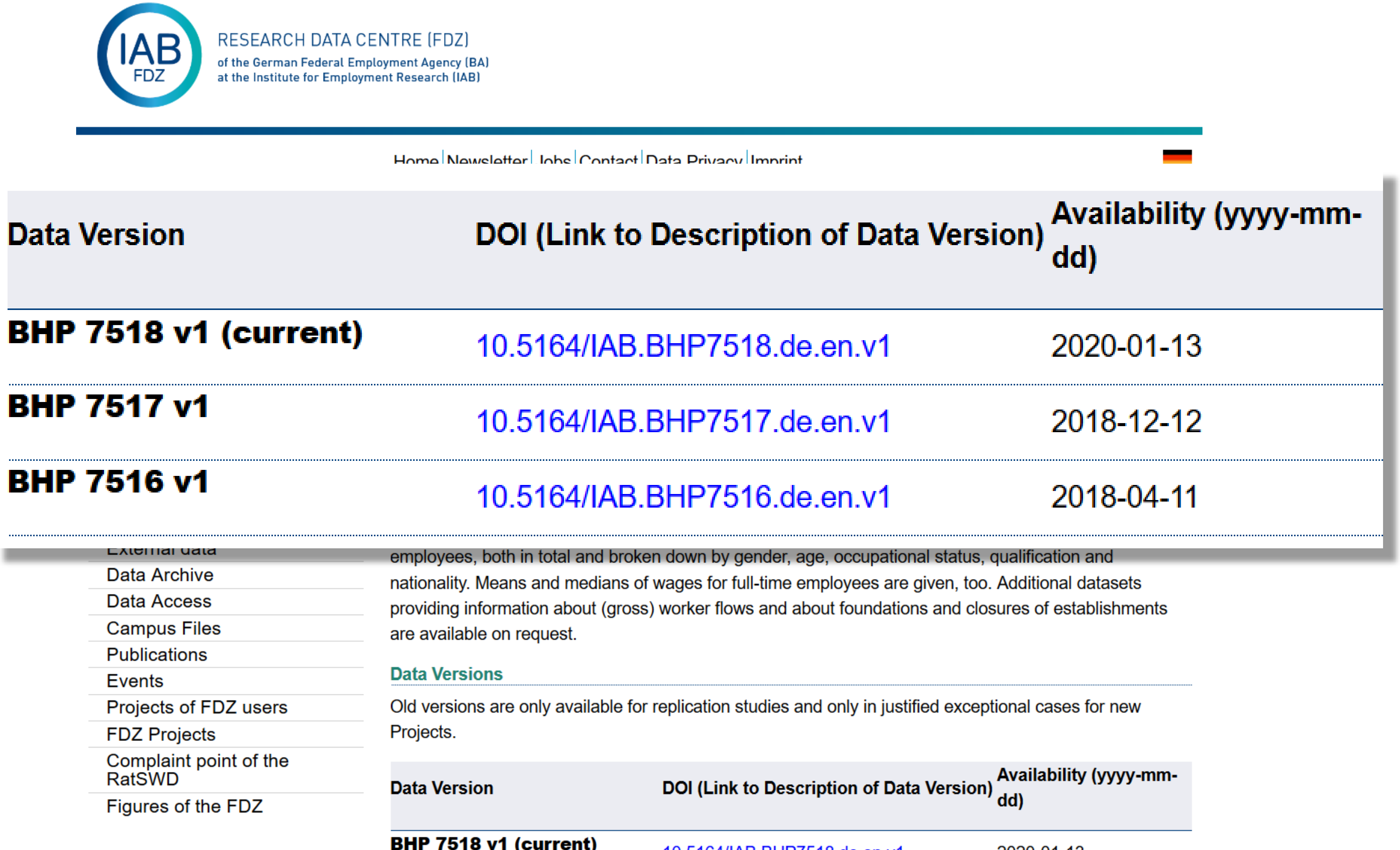
One of the most vexing issues is how to cite data. This document goes through a few common scenarios not covered elsewhere.

What is not a data citation

Many authors initially neglect to add data citations, or do not know how to add a data citation. Often, we see authors cite papers with supplementary data, but not databases or other data:

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

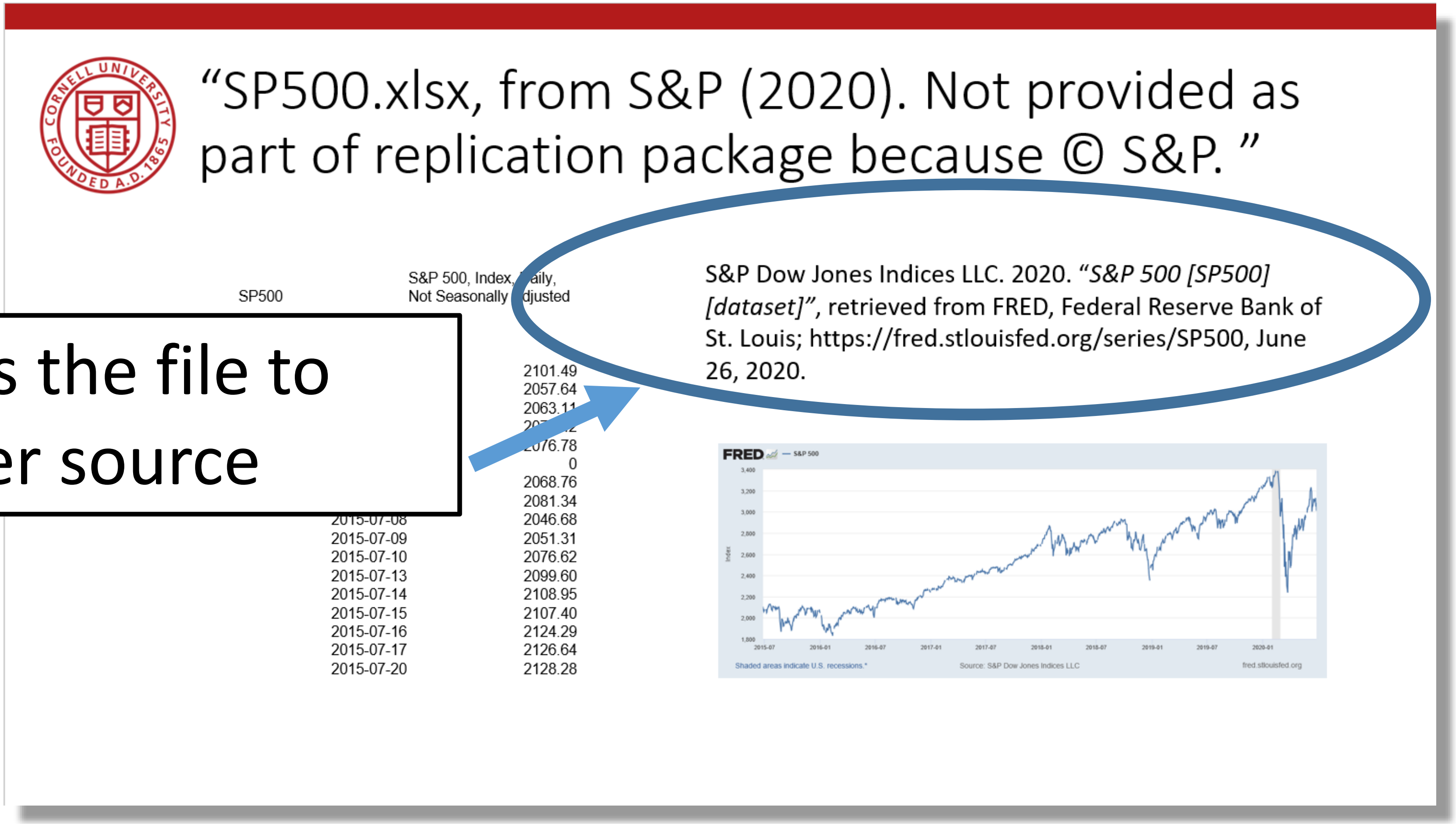
employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are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

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.

Attributes the file to the proper source

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:

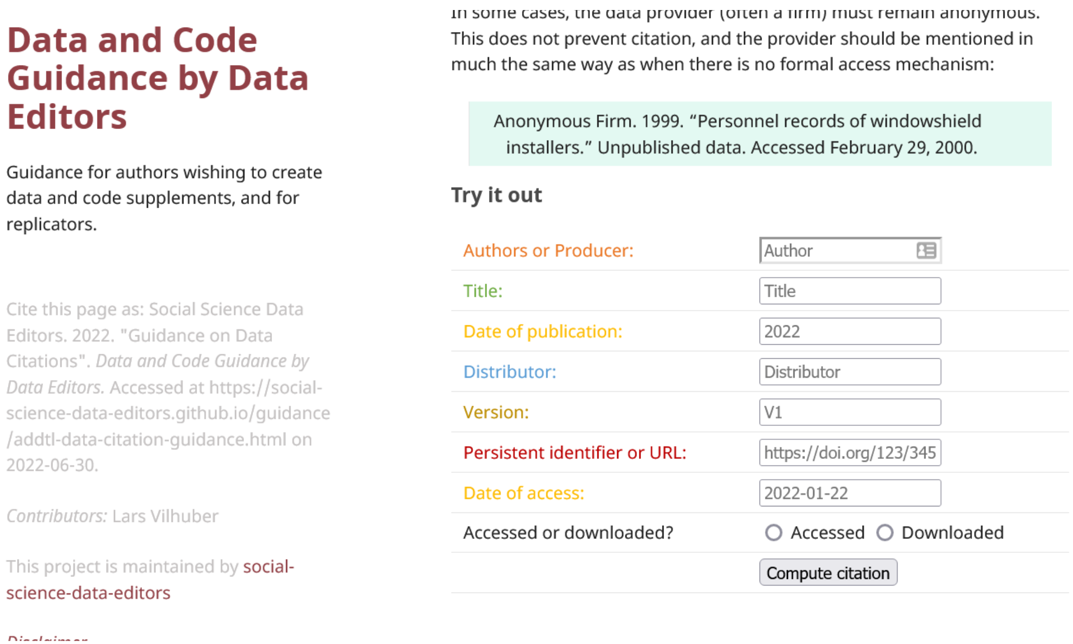
- Author
- Title
- Distributor
- Date
- Version
- Persistent identifier

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

Disclaimer

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.

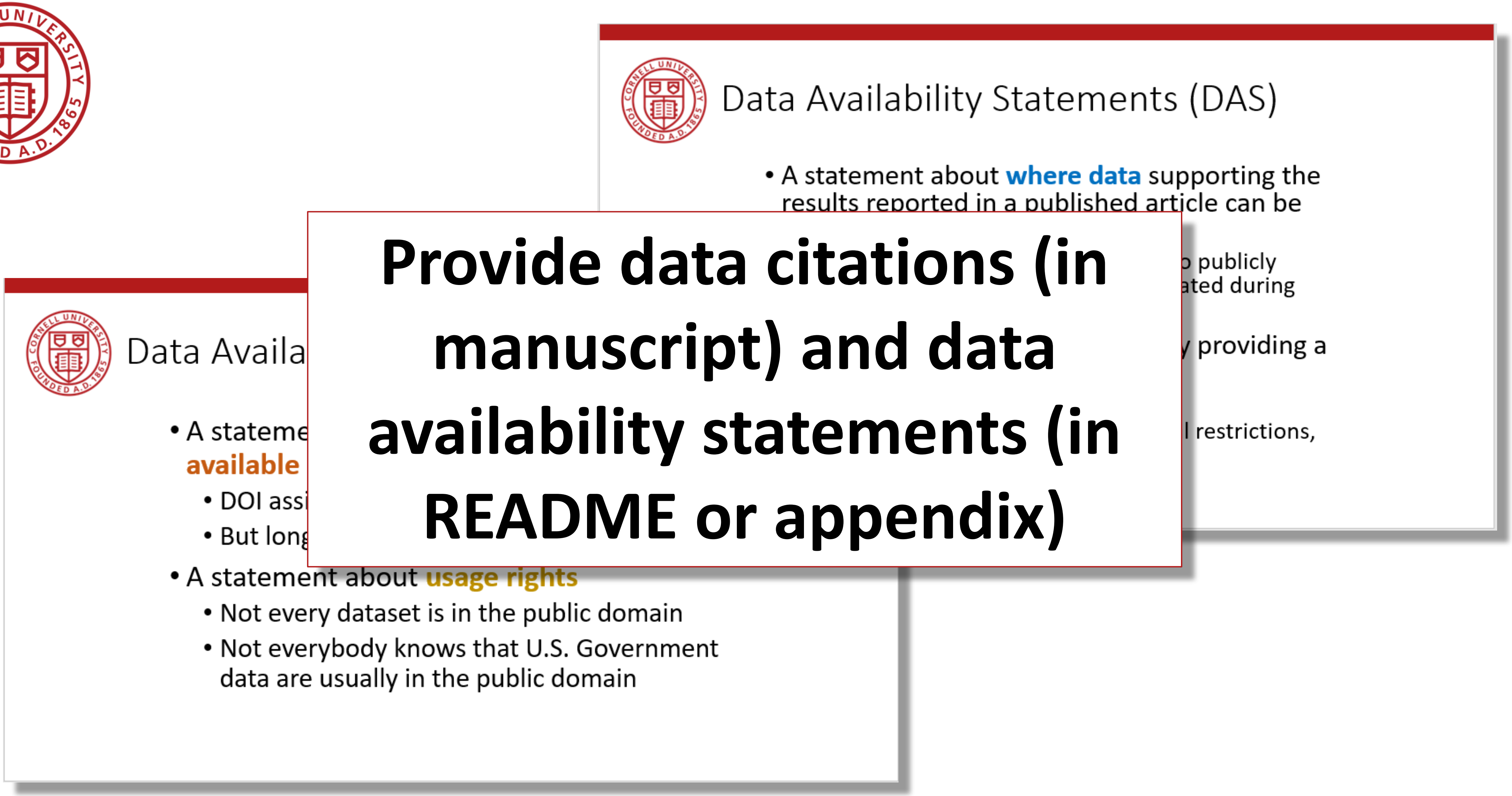
Try it out

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)



Data Availability Statements (DAS)

- A statement about **where data** supporting the results reported in a published article can be
- o publicly
  ated during
- y providing a
- l restrictions,

Data Availa

- A stateme
  available
  - DOI assi
  - But long
- A statement about **usage rights**
  - Not every dataset is in the public domain
  - Not everybody knows that U.S. Government data are usually in the public domain

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)