Implementing Increased Transparency and Reproducibility in Economics: Restricted Data

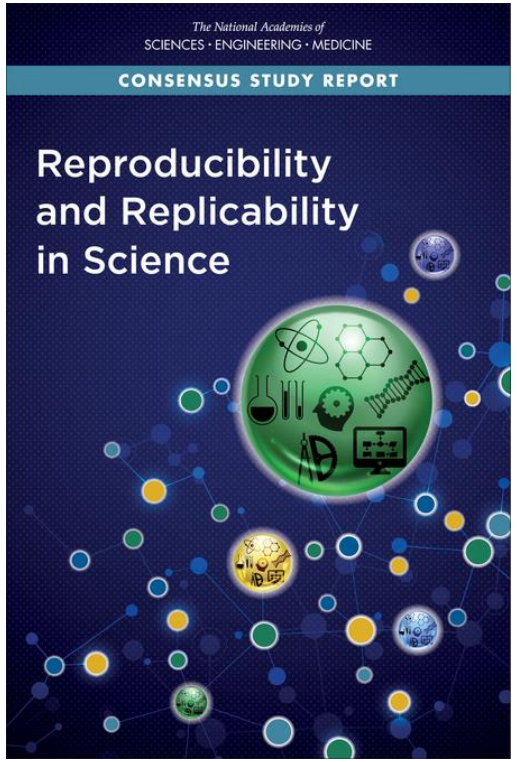
Lars Vilhuber

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.

Replication continuum

https://doi.org/10.17226/25303



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

Same data	Same code	Same methods	Same context

Reproducibility

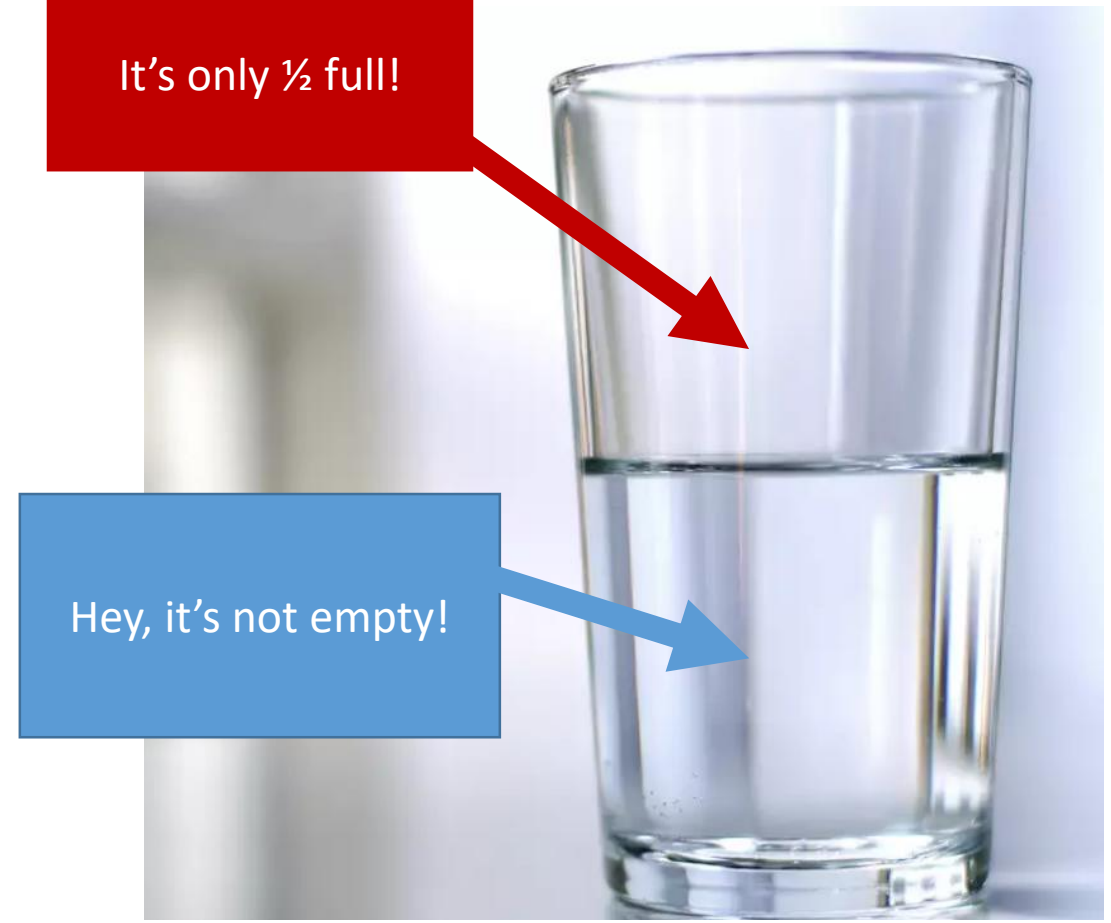
- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Progress

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

In a nutshell

- **40%** use restricted-access data
- **25%** use public-use data and are mostly or completely reproducible
- **25%** use public-use data and are only partially reproducible
- **10%** fail to yield useful results



Why?

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

Poor citation practices

- **Macrodata:**

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

- **Microdata**:

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

Actions

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)

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

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

RDCs (and similar) pose a challenge and an opportunity

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

RDCs (and similar) pose a challenge and an opportunity

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

RDCs (and similar) pose a challenge and an opportunity

How do you check code when the data access is complex?

How do you improve archives when you do not control data management?

How do you document data provenance when you cannot provide the data?

How do you document provenance when you cannot de the data?

Wrong question!

How did you get the data in first place?

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

…

About the

CRDCN Canadian Research Data Centre Network

Search Publications

Data | Research | Publications | Events | News | KT Corner

Home » Research

Research

Researchers use RDC microdata to investigate a wide range of social, economic and health-related issues, from child poverty to eldercare, from school success to workplace stress, from Aboriginal well-being to immigrant integration, from family-life balance to mental health. This section informs you about the application procedure for accessing these microdata.

To find out what datasets are available in the RDCs, refer to our data page. For past publications using these data, consult our online bibliography in the publications section.To learn more about the research currently underway throughout the Network, you can browse the list of projects on Statistics Canada's website.

How to access RDC data

Protecting confidentiality in the RDCs

Note that researchers whose projects are approved will be subject to a security check before being sworn in under the Statistics Act as a 'deemed employee' of Statistics Canada. Deemed employees are subject to all the conditions and penalties of regular Statistics Canada employees, including fines and/or imprisonment for breach of confidentiality. Effective **December 1, 2016**, the RCMP requires all federal public servants in Canada (including RDC researchers and staff) to undergo fingerprinting as part of their security clearance. Additionally, Treasury Board of Canada requires a credit check. Please contact your local RDC to learn more about the security clearance procedure.

All results to be physically removed from secure areas will be carefully screened for confidential data, whether as direct listings or as possible residual disclosures.

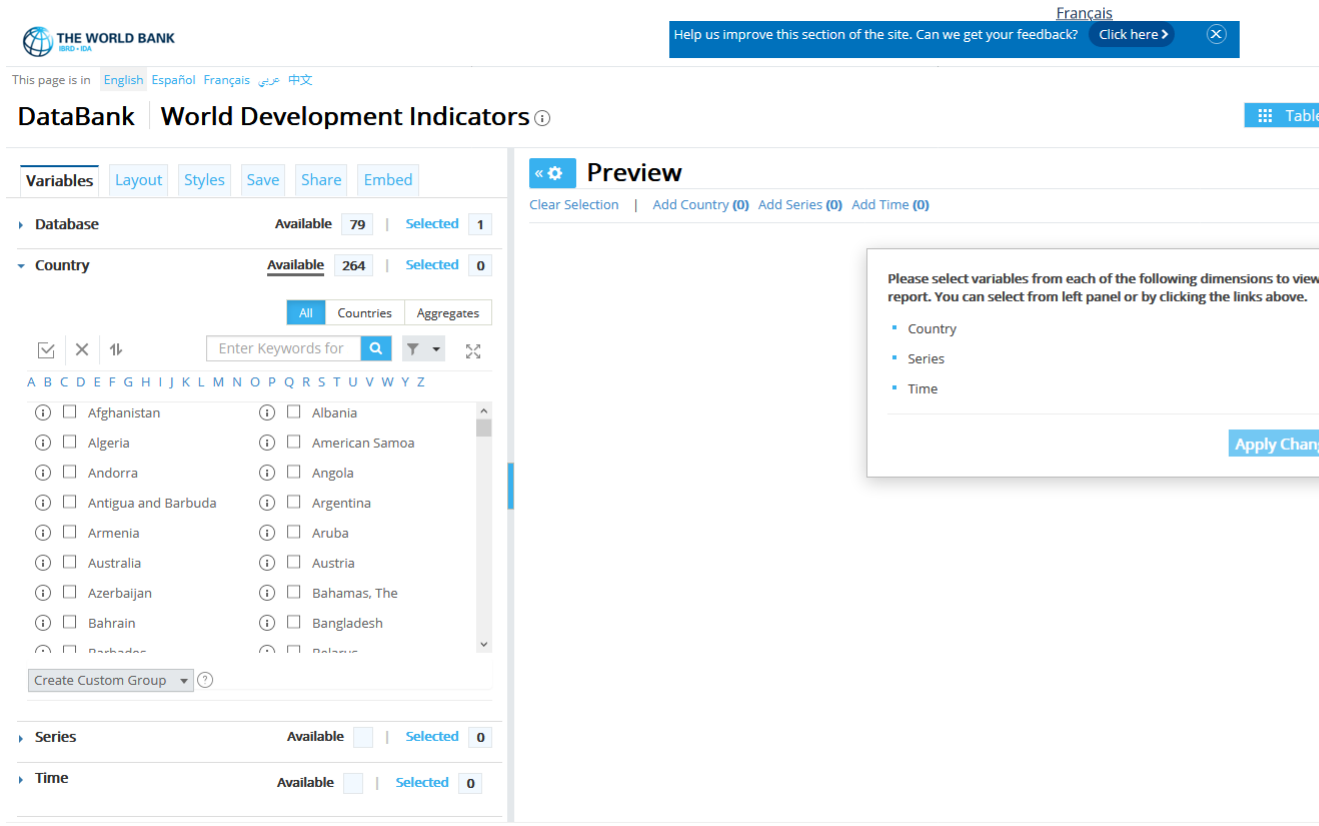
Application process for academic researchers

Researchers wishing to access the RDC should create an account on the Statistics Canada Microdata Access Platform and follow the steps to create a new proposal. The proposal is evaluated by Statistics Canada for feasibility before access can be granted. In addition, if you are a student, your thesis supervisor must write a letter in support of your RDC application and join the application as a co-investigator. For other academic users, a completed peer-review may be required. The review must be conducted by a tenured faculty-member at an accredited Canadian university. Researchers who are required to submit such a peer review can source their own peer reviewer, or contact CRDCN for assistance if they are unable to find a suitable candidate.

You must have described the data

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

…



How do you document data provenance when you don’t control the process?

How do you document data provenance?

- What do you need to request?
  - Name, specification, DOI, etc.
- Where do you need to request it?
  - Website, your local CRDCN, a Freedom of Information Act officer, etc.
- Details, details:
  - Copy of your request form?
  - Copy of your request letter?
  - Etc.
- Don’t assume (too much) prior knowledge!

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

RDCs (and similar) pose a challenge and an opportunity

How do you check code when the data access is complex?

How do you improve archives when you do not control data management?

How do you document data provenance when you don't control the process?

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

AEA Pre-Publication Verification

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

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

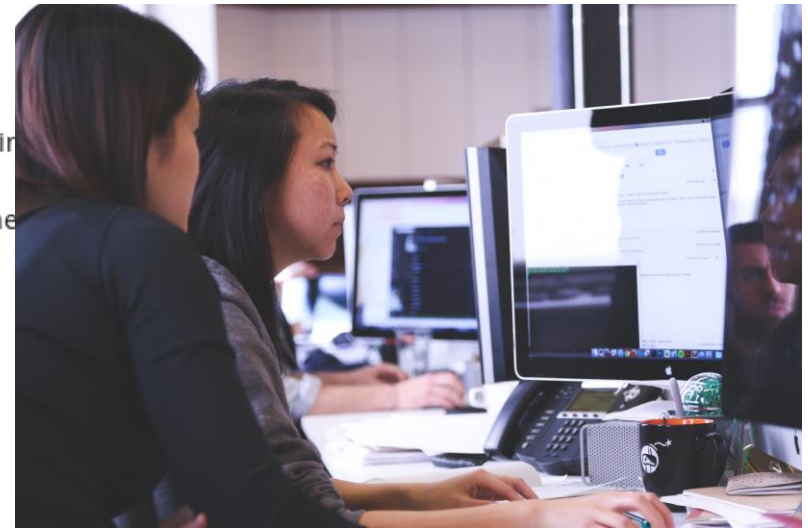
On this page:

- 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

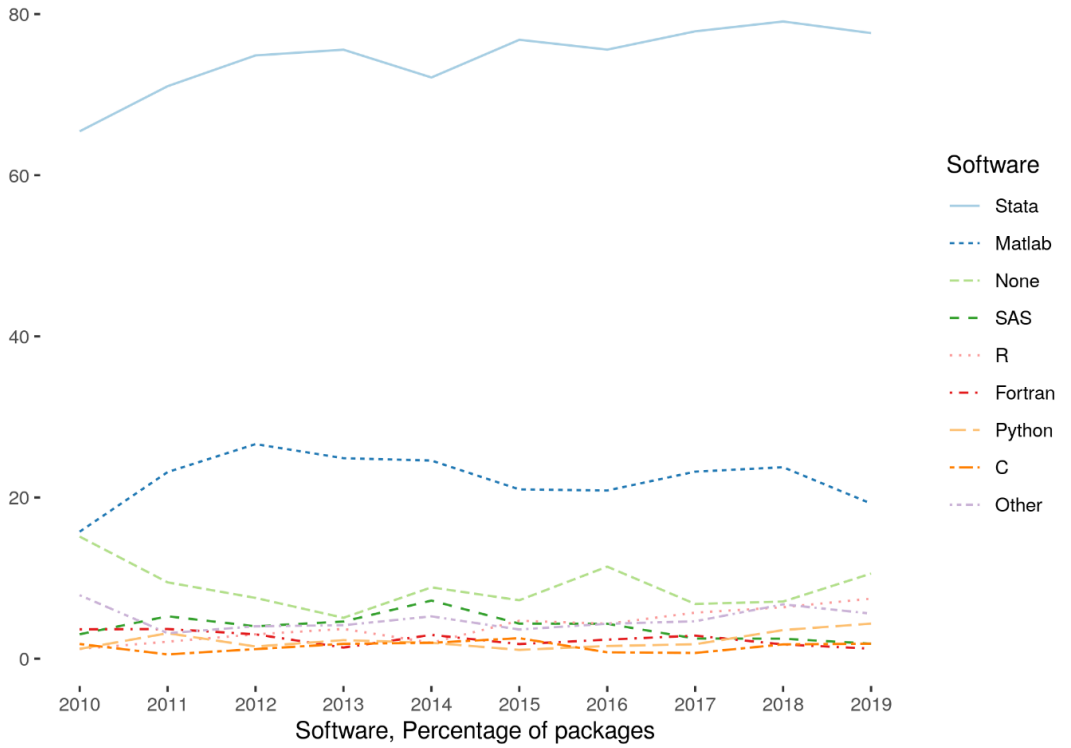


Who is doing that?

- Earlier reproducibility work: **Flavio Stanchi** (now at AirBnb), **Sylverie Herbert** (Banque de France), **Hautahi Kingi** (Impaq)
- Assistant: **Michael Darisse**
- Current lead graduate student: **Harry Son** (since Aug 2020)
- Current and past undergraduate students: Alexia Ge, Anthony Peraza, Craig Schulman, Elijah B. Ruiz, Gabriel Bond, Jason S. Katz, Jeong Hyun Lee, Jiayin Song, John Park, Joshua Passel, Kirubeal T. Wondimu, Linchen Zhang, Louis Liu, Luis Lopez Cabrera, Luke O'Leary, Mary-Jo Ajiduah, Naomi Li, Nicholas Swan, Nishat Peuly, Ryan Ali, Samuel Frey, Siyang (Elaine) Yu, Steve Yeh, Weilun Shi, William Hernandez, Yanyun (Iris) Chen, Yuan-Hsuan (Sharon) Lin, Zebang Xu, Xing Su, Jiazhen Tan , Xueshi Su, Vendela Norman, Anderson Park, Nehedin Juarez, Rubal Mistry, Syon Verma, William Silverman, Zechariah Karsana, Franklin Omullo, Liam P. Cushen, Ololade Omotoba, Lydia Reiner, Xiangru Li, Melanie Chen, Peter Rafael Sanchez, Jill Crosby, Matthew H. Wang, Daniella Pena, Julia Zimmerman, Kate Hofer, Tarangana Thapa
- Former graduate students: **David Wasser, Meredith Welch,** Aviv Caspi, Leah Kim

Very little diversity in software

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



larsvilhuber Minor edits to the report - clarifications about data preparation pro... e9ad1f8 10 days ago

1 contributor

195 lines (135 sloc) 10.3 KB

Raw Blame History

[MC number] [Manuscript Title] Validation and Replication results

> INSTRUCTIONS: Once you've read these instructions, DELETE THESE AND SIMILAR LINES. In the above title, replace [Manuscript Title] with the actual title of the paper, and [MC number] with the Manuscript Central number (e.g., AEJPol-2017-0097) Go through the steps to download and attempt a replication. Document your steps here, the errors generated, and the steps you took to alleviate those errors.

You may want to consult Unofficial Verification Guidance for additional tips and criteria.

SUMMARY

> INSTRUCTION: The Data Editor will fill this part out. It will be based on any [REQUIRED] and [SUGGESTED] action

Stats on reproduced articles

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

- **~1000 assessments**
- Of which ~**445 manuscripts** have been “accepted”

Current efforts at the AEA

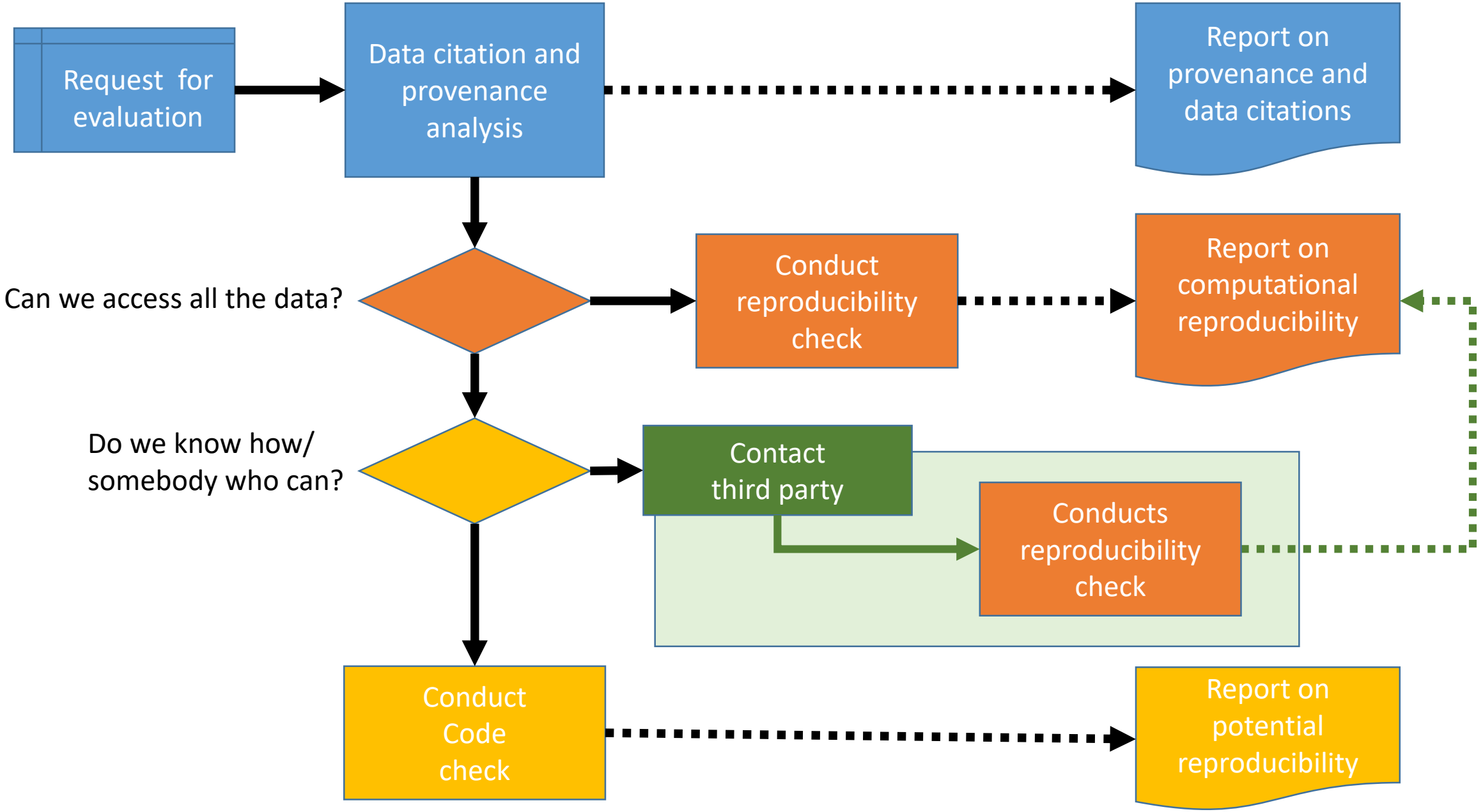
- **Pre-emptively improve code archives**
  - By conducting reproducibility checks when we can
  - By working with groups that conduct reproducibility checks when we cannot
- **Better archives**
  - Greater transparency of the code and data archives
- **Better provenance tracking**
  - Leave code where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

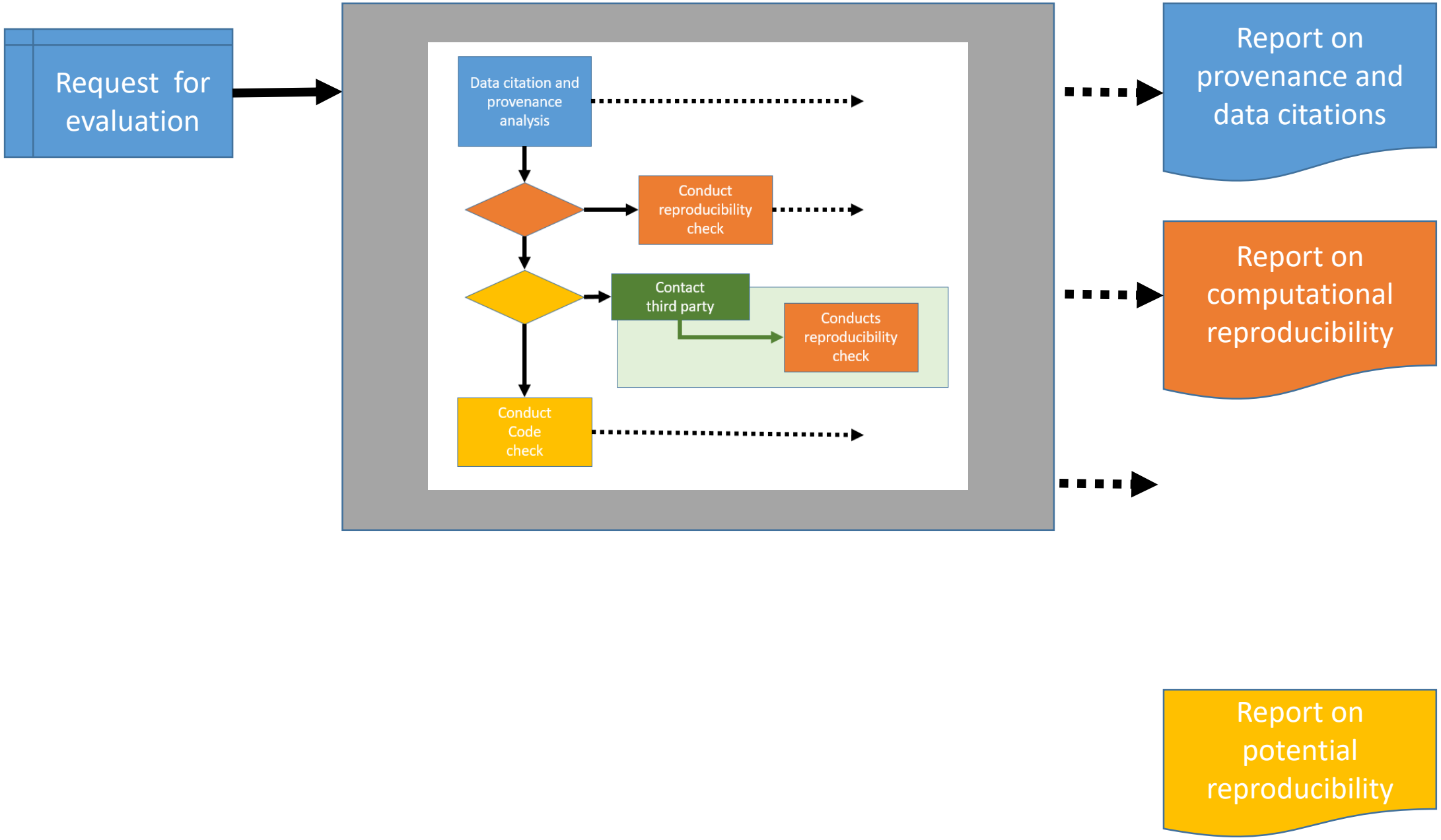
RDCs (and similar) pose a challenge and an opportunity

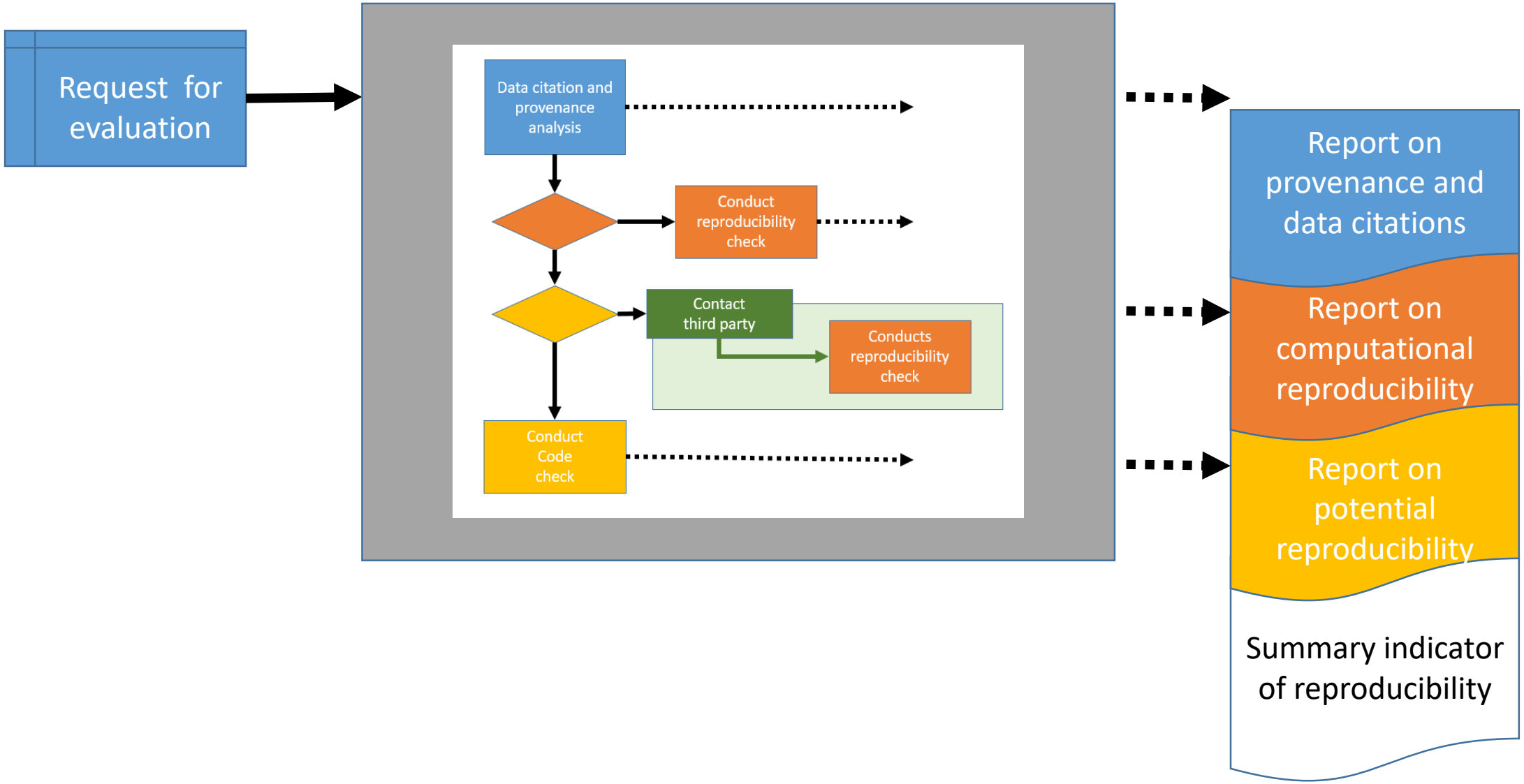
How do you check code when the data access is complex?

How do you improve archives when you do not control data management?

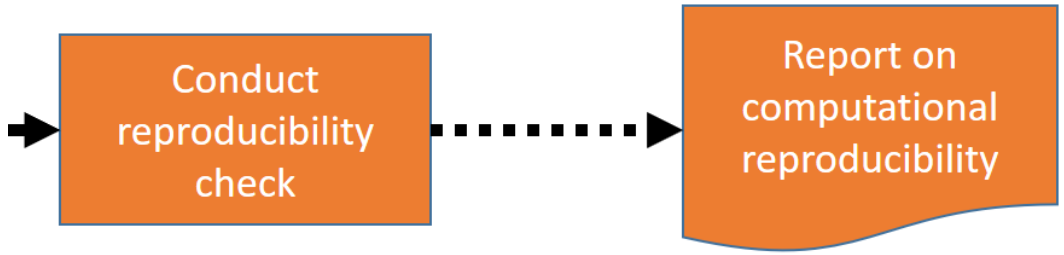
How do you document data provenance when you don't control the process?

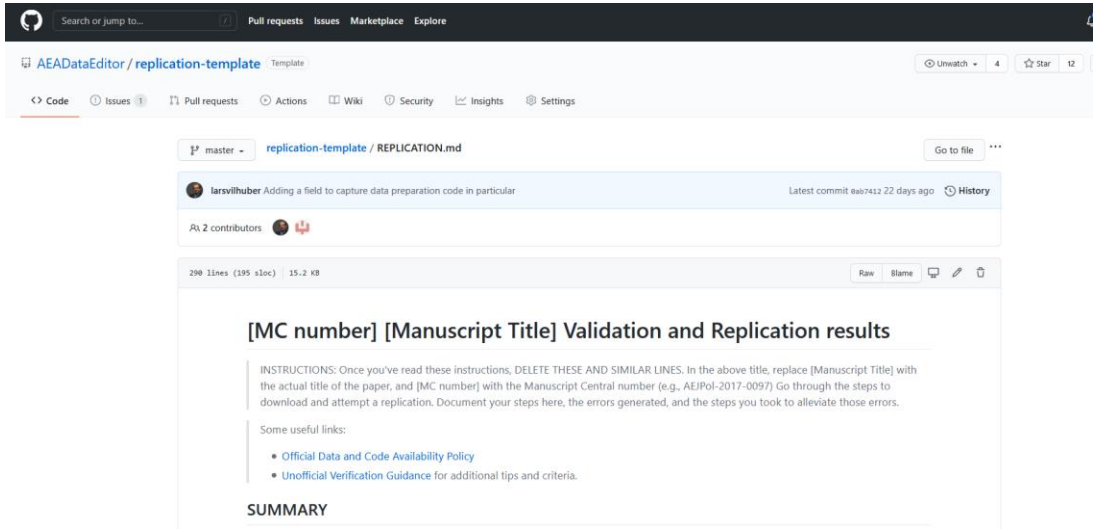

Request for evaluation
Data citation and provenance analysis
Report on provenance and data citations
Can we access all the data?
Conduct reproducibility check
Report on computational reproducibility
Do we know how/ somebody who can?
Contact third party
Conducts reproducibility check
Conduct Code check
Report on potential reproducibility


Request for evaluation
Data citation and provenance analysis
Conduct reproducibility check
Contact third party
Conducts reproducibility check
Conduct Code check
Report on provenance and data citations
Report on computational reproducibility
Report on potential reproducibility

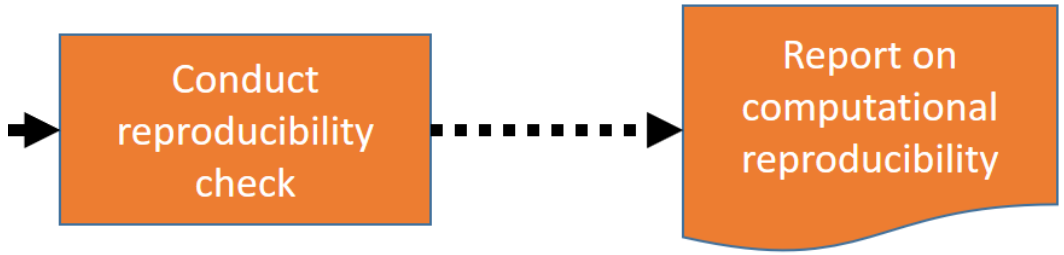

Request for evaluation
Data citation and provenance analysis
Conduct reproducibility check
Contact third party
Conducts reproducibility check
Conduct Code check
Report on provenance and data citations
Report on computational reproducibility
Report on potential reproducibility
Summary indicator of reproducibility

What is the reproducibility check?





What is the reproducibility check?



- Data checks
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

What is the reproducibility check?

- **Data checks**
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS: When data are present, run checks:

- **Can data be read** (using software indicated by author)?
- Is data in **archive-ready formats** (CSV, TXT) or in custom formats (DTA, SAS7BDAT, Rdata)?
- Does the dataset **have variable labels?**
- Run **check for PII**. Apply judgement.

What is the reproducibility check?

- Data checks
- **Code description**
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS:

- **Review the code** (but do not run it yet).
- Identify programs that create "analysis files" ("**data preparation code**").
- Identify **programs that create tables and figures**. Not every deposit will have separate programs for this.
  - Identify all Figure, Table, and any in-text numbers.

What is the reproducibility check?

- Data checks
- Code description
- **Requirements**
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

- Software Requirements
  - Version of software (Stata 15, Matlab R2019b, etc.)
  - Complete list and version of packages!
- Computational Requirements
  - Type, vintage, memory size, speed of computer
  - Disk space!
- Time Requirements
  - Minutes, hours, days, weeks, months?

What is the reproducibility check?

- Data checks
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- **Verbose description** of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS

- Provide details about your process of accessing the code and data.
- DO describe actions that you did as per instructions
- DO describe any other actions you needed to do ("I had to make changes in multiple programs")
- Findings come later

What is the reproducibility check?

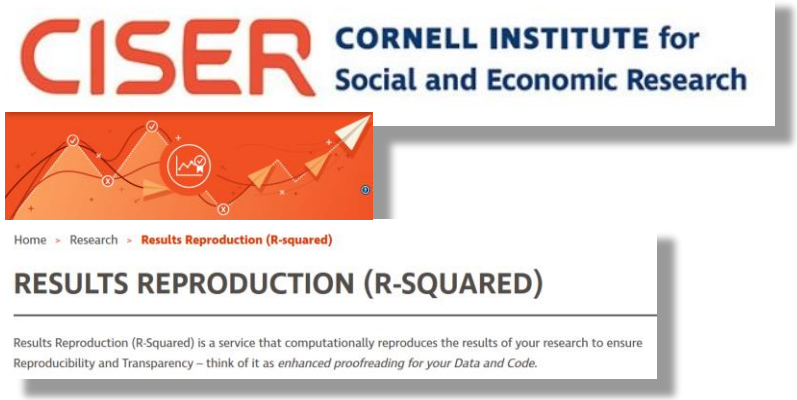
- Data checks
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- **Findings**
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS:

- Describe your findings both positive and negative in some detail, for each **Data Preparation Code, Figure, Table, and any in-text numbers**.
- When errors happen, be as precise as possible.
  - For differences in figures, provide screenshot of manuscript figure, as well as the figure produced by the code you ran.
  - For differences in numbers, provide both the number as reported in the manuscript, as well as the number replicated.

Also work with 3rd parties

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



Also work with 3rd parties

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

Some examples:

- IFAU (Swedish data)
- Fed Board (proprietary data)
- Upjohn Institute (US admin data)
- JCT (US tax data)
- Banco do Portual (LEED)
- IAB (German data)
- *Graduate students* at Wharton, Chicago, Yale, Wisconsin, Cornell

Could this be done in CRDCN?

Yes

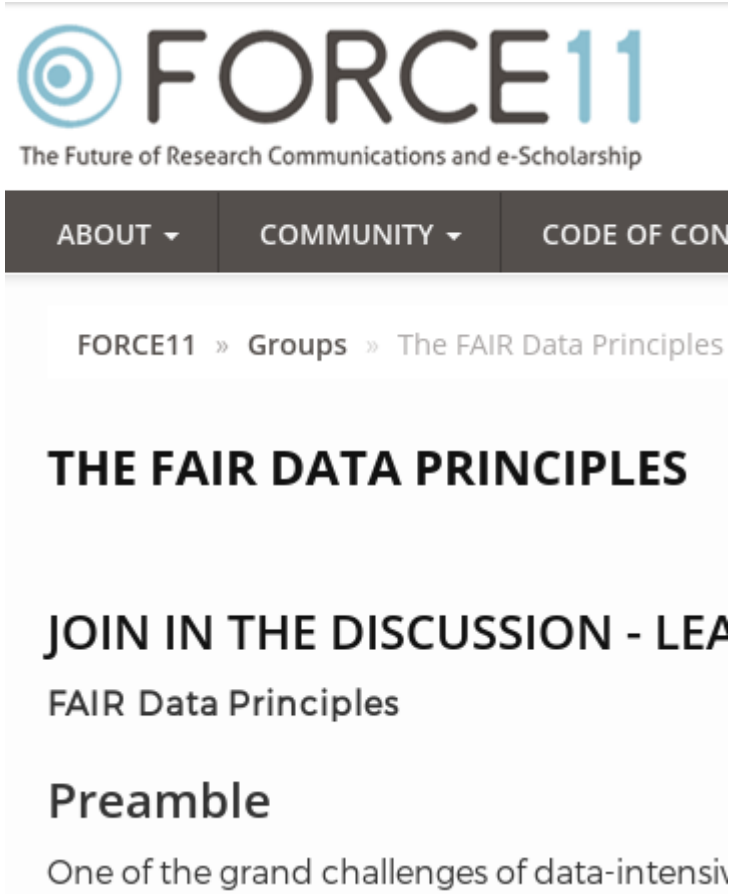
Current efforts at the AEA

- **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: 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

OPEN ICPSR Find Data Share Data openICPSR Repositories Management

GO 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

```
<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
	application/x-stata	11.9	10/08/2014

OPEN ICPSR Find Data Share Data openICPSR Repositories Management

GO dataeditor@aeapubs.org

Find Data / Imperial Russian Factory Database, 1894-1908

Imperial Russian Factory Database, 1894-1908

Principal Investigator(s): Amanda Gregg, Middlebury College

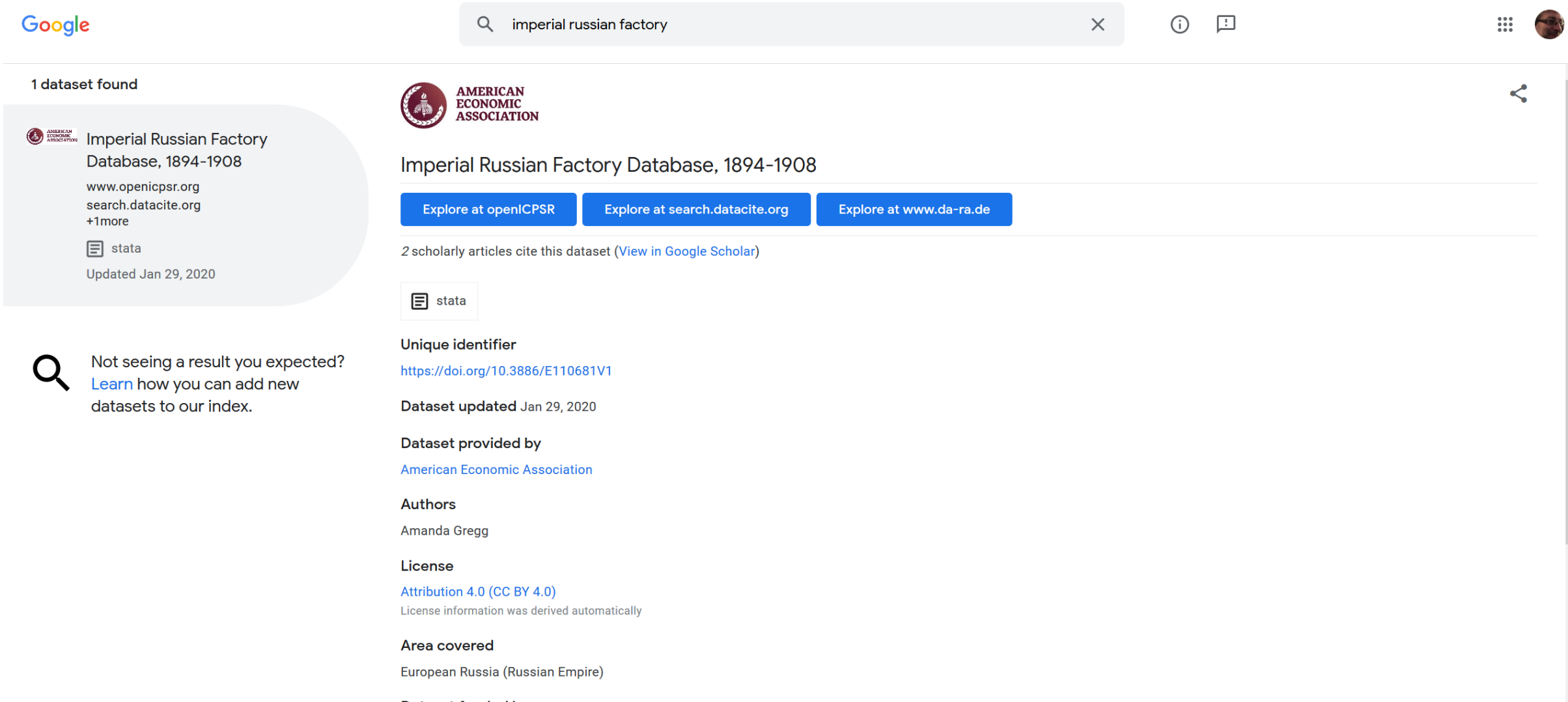
Version: V1

AMERICAN ECONOMIC ASSOCIATION


```
<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.&amp;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":["Middlebu
"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>
```

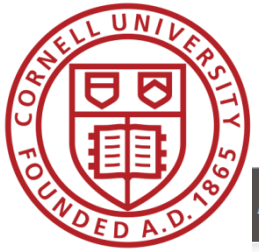

AG_Corp_CleaningandDatabaseCompiler.do		KB	11:02:AM
AG_Corp_Prod_AppendixCode.do	text/x-stata-syntax	42.2 KB	12/09/2019 09:19:AM
AG_Corp_Prod_Code.do	text/x-stata-syntax	26.6 KB	12/12/2019 03:01:AM
AG_Corp_Prod_Database.dta	application/x-stata	11 MB	08/07/2019 08:55:AM

… and findability relies on metadata



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



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].

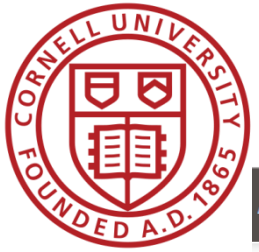
5. Access

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].

DC1

Data Citation Principles

2. Credit and Attribution

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].

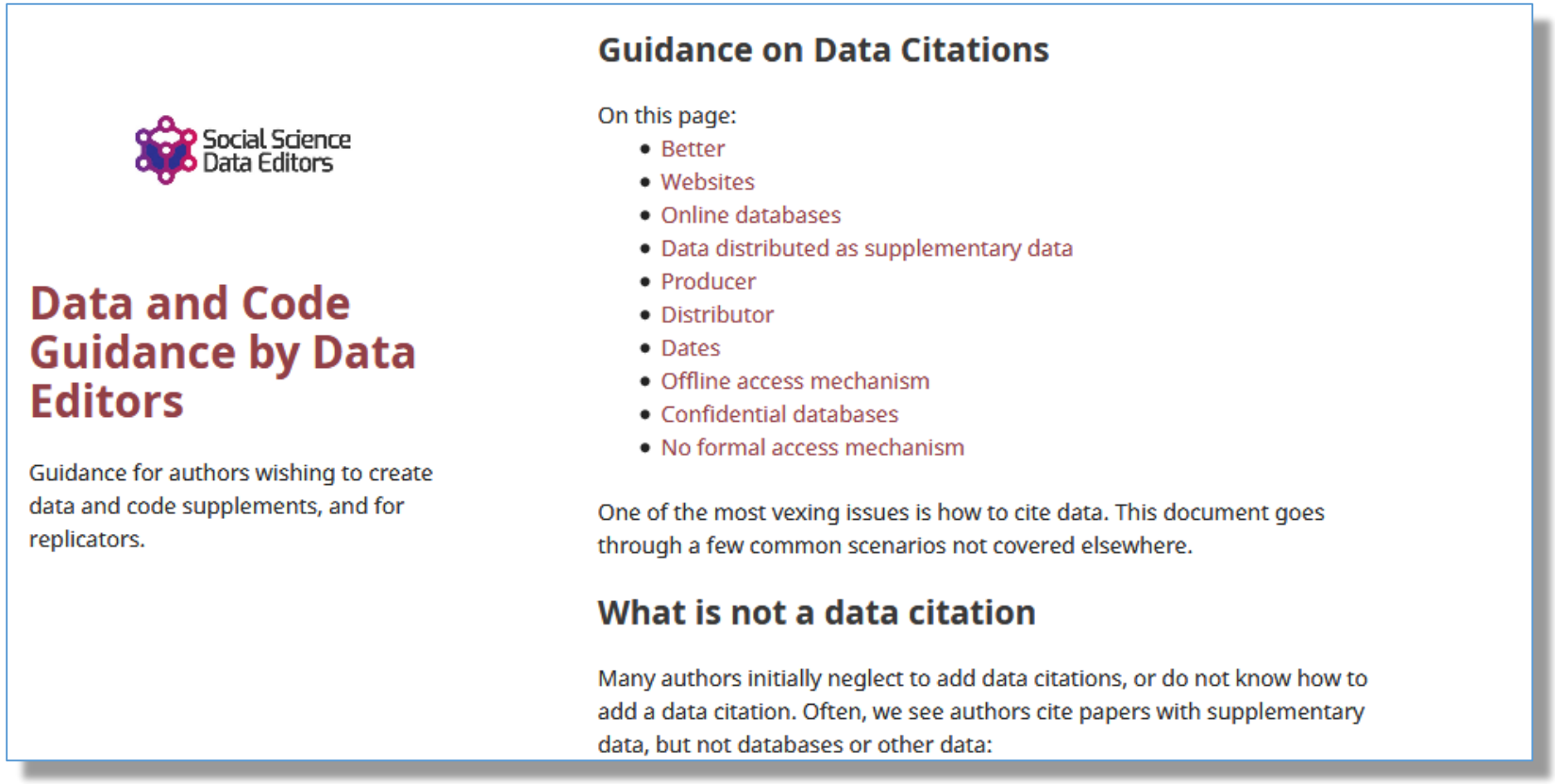
5. Access

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].

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

Some practical tips (based on 1000+ assessments)

Citing restricted-access data

"Well, I can't download the data, so I can't cite it."

AEA "Data 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.**

Every manuscript is checked

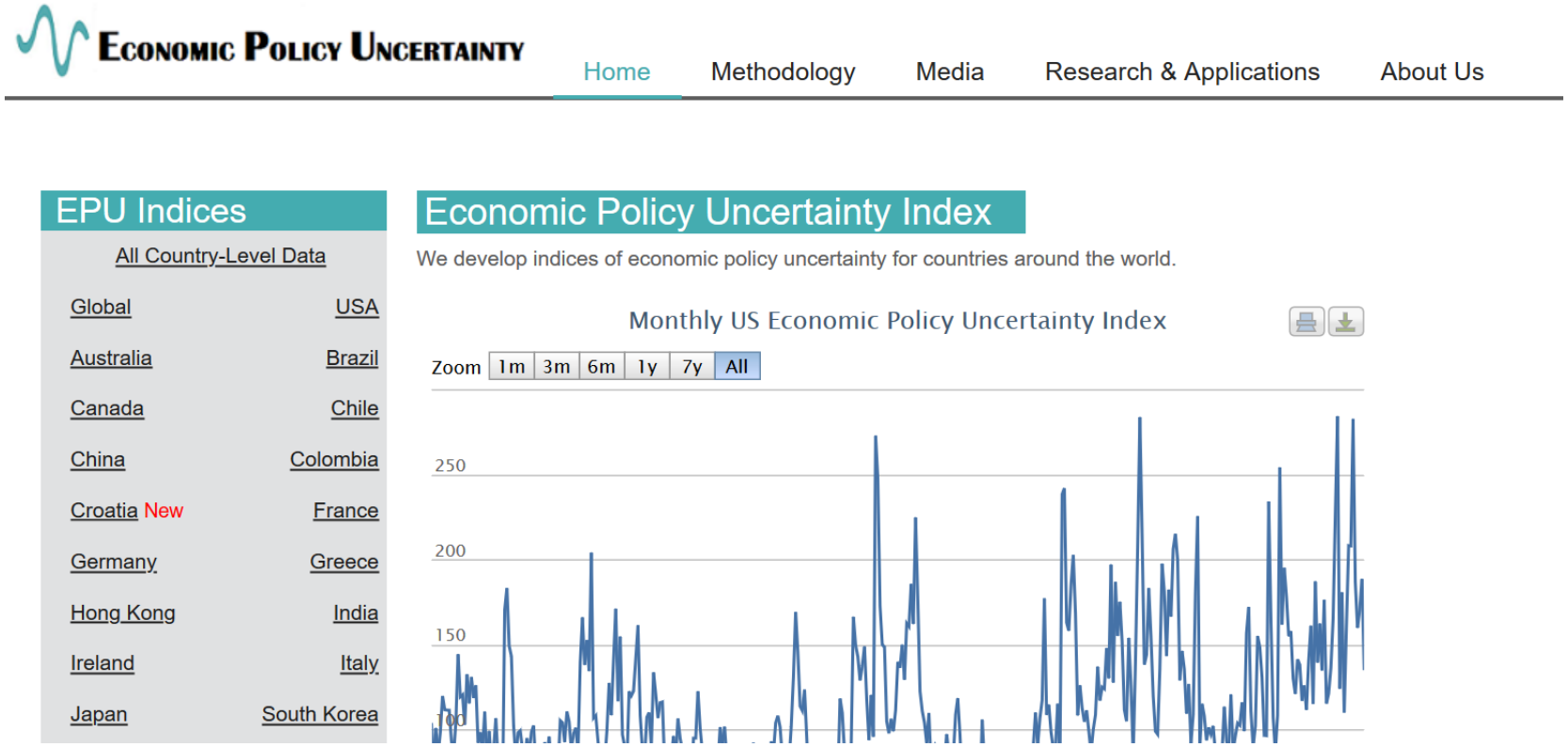
- What datasets are used
- Are they cited?
- Is there additional information access?
  - → URL leads to exact data?
  - → URL leads to application procedure?
  - → other access procedure is described?

Le labyrinthe des juges, par Guillaume Beck

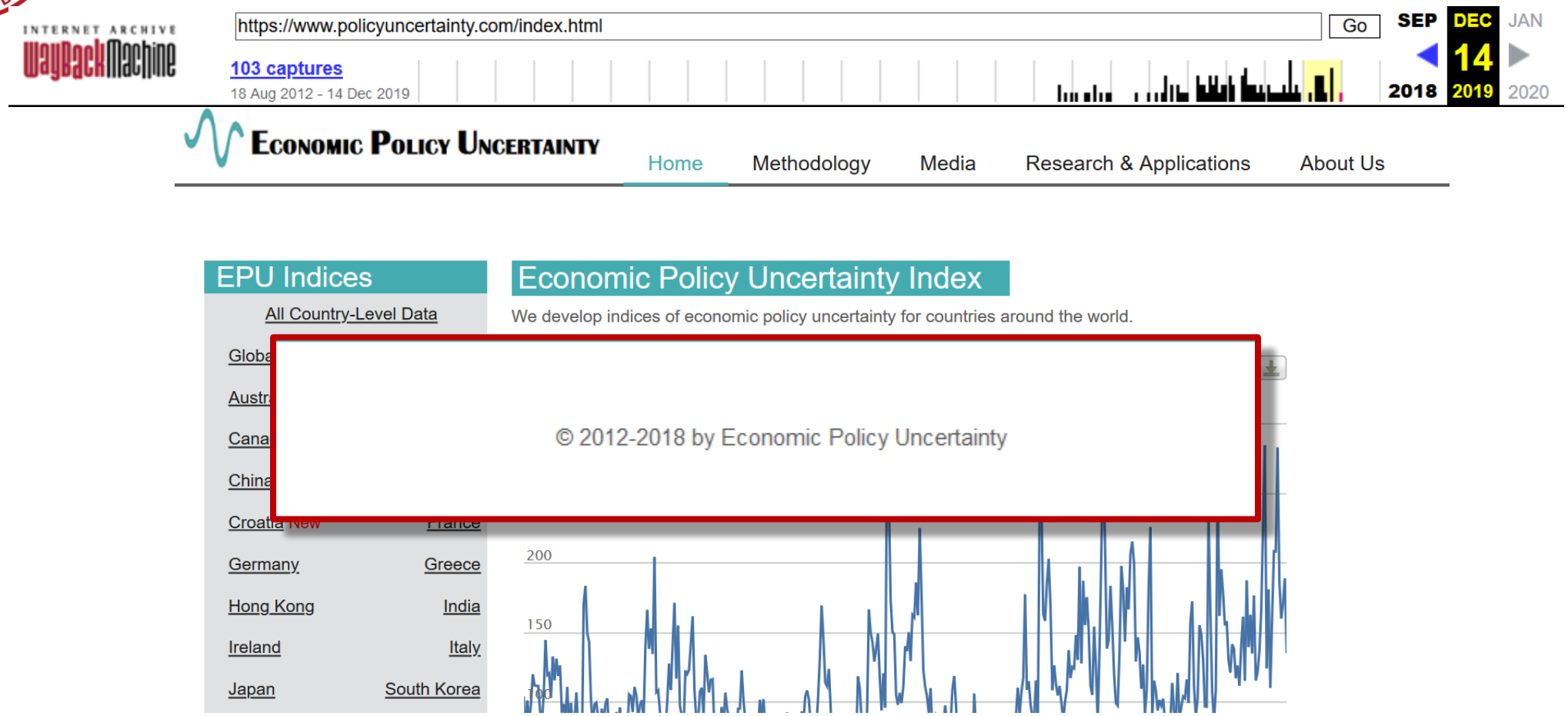
Every manuscript is checked

- What datasets are used
- Are they cited?
- Is there additional information on access?
- Is there license/ data use information?
  - → **Should** the author provide the data?
  - → Is the author **allowed** to provide data?

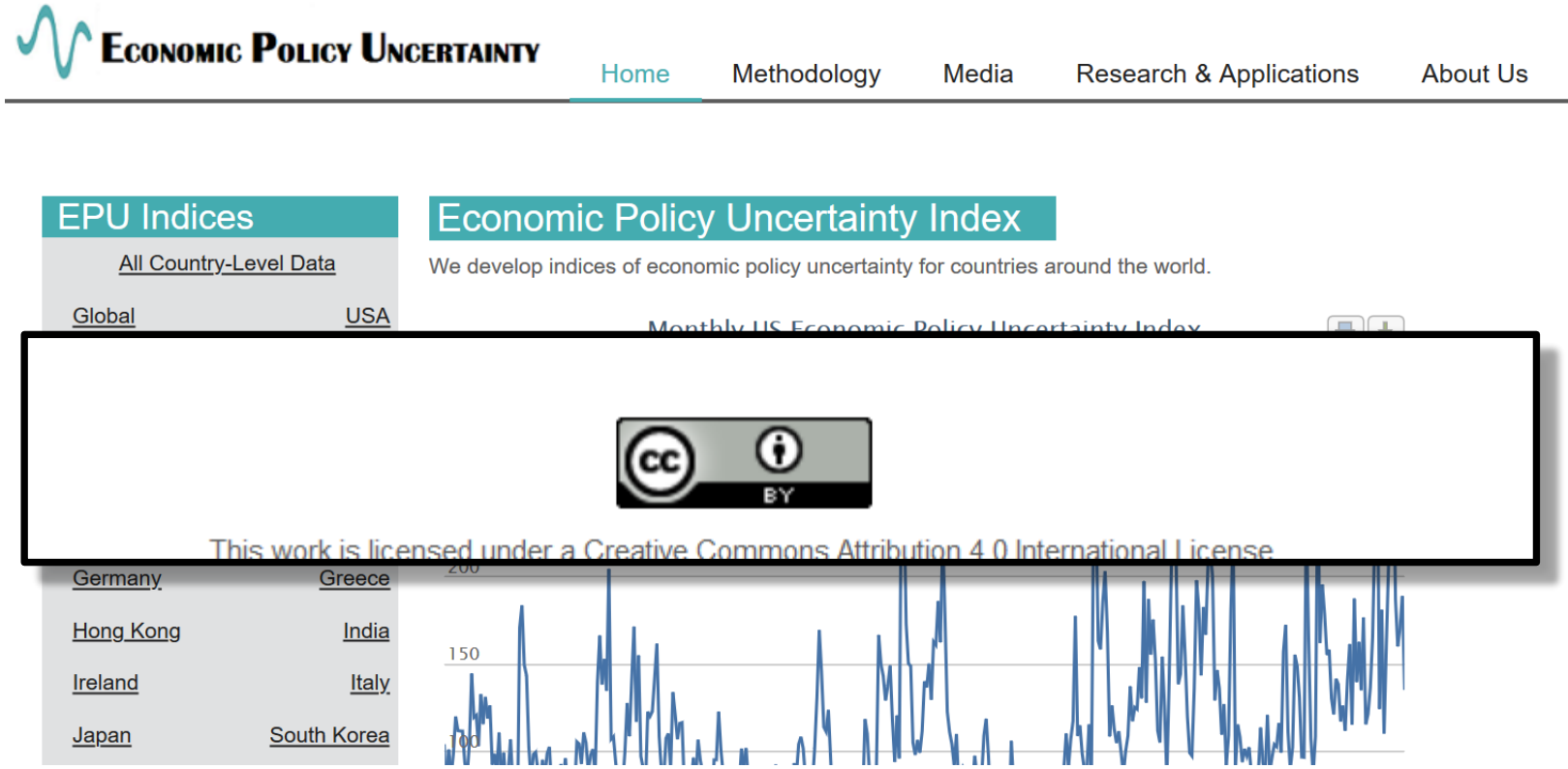
Example 2: Academic data publisher



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

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!

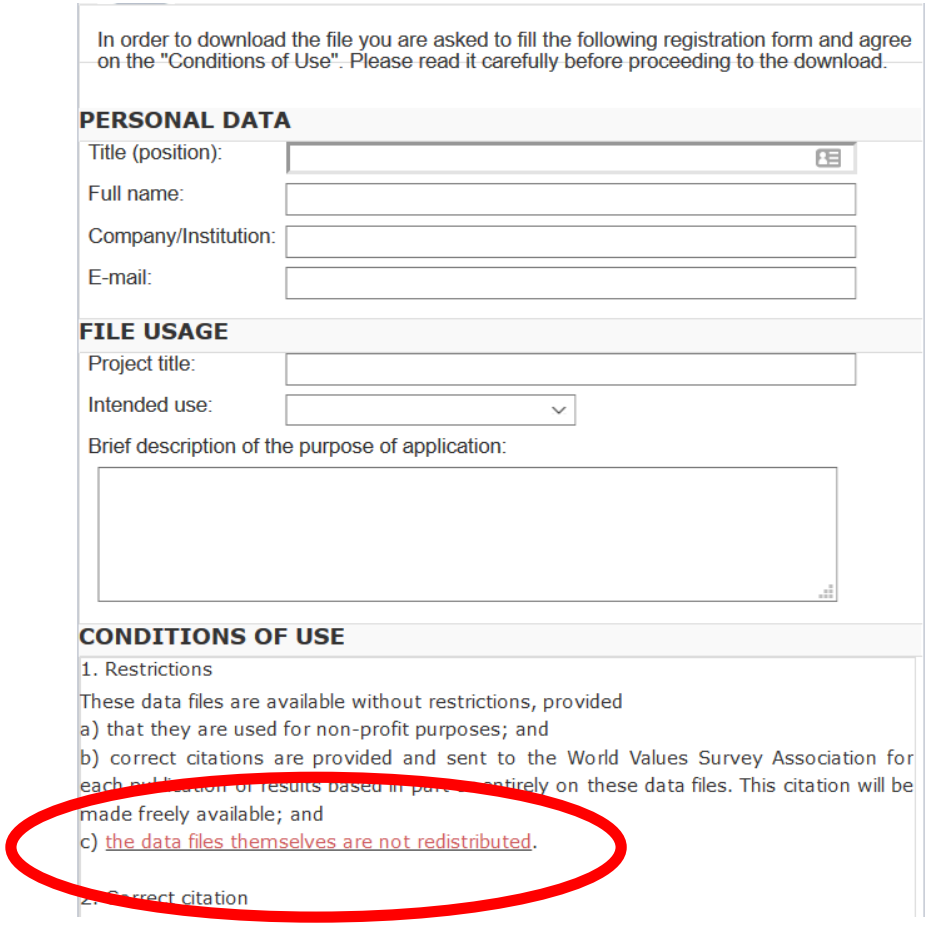
- If the URL does not work, we make a note.
- If the site requires registration, we try it out.
  - How long?
  - Any requirements?

- 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?

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

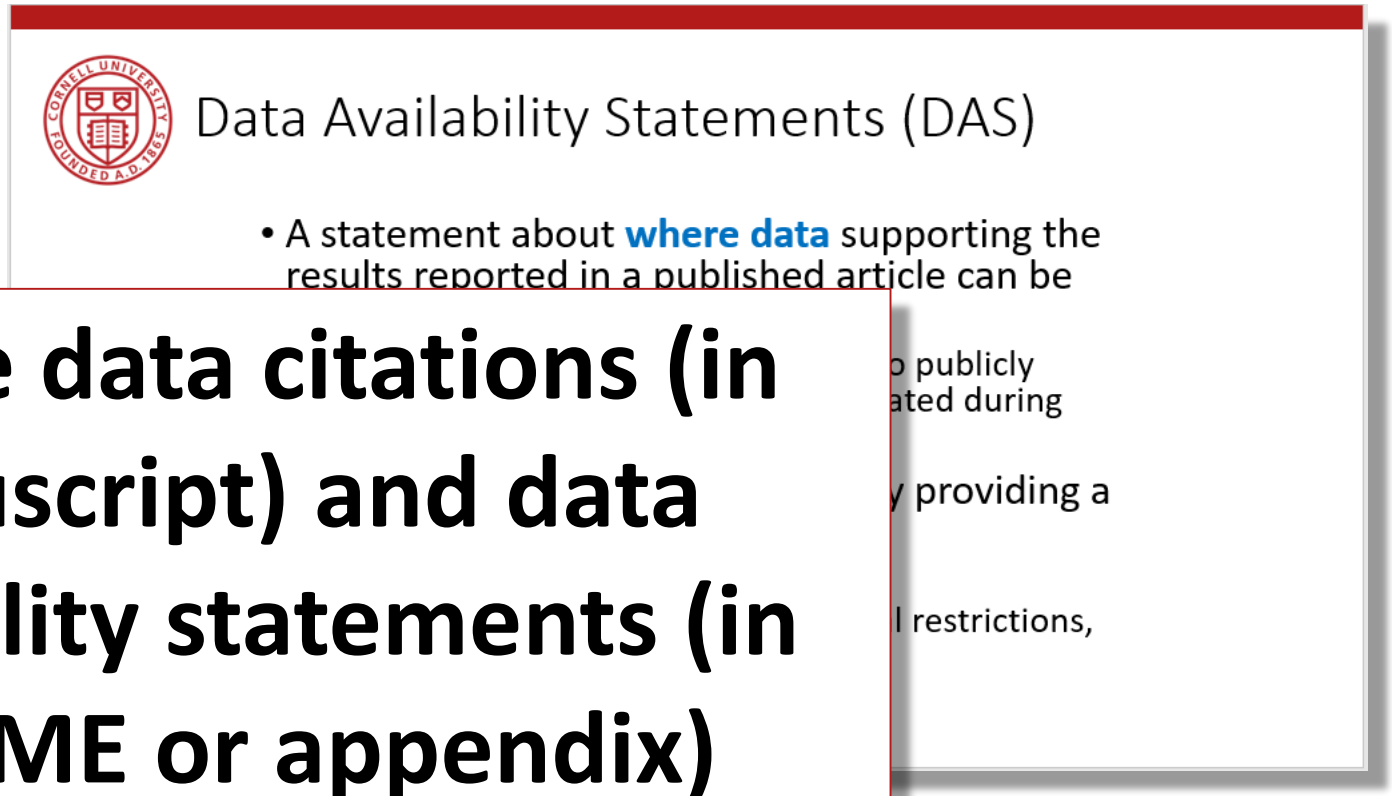
These data files are available without restrictions, provided
a) that they are used for non-profit purposes; and
b) correct citations are provided and sent to the World Values Survey Association for each publication of results based in part or entirely on these data files. This citation will be made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

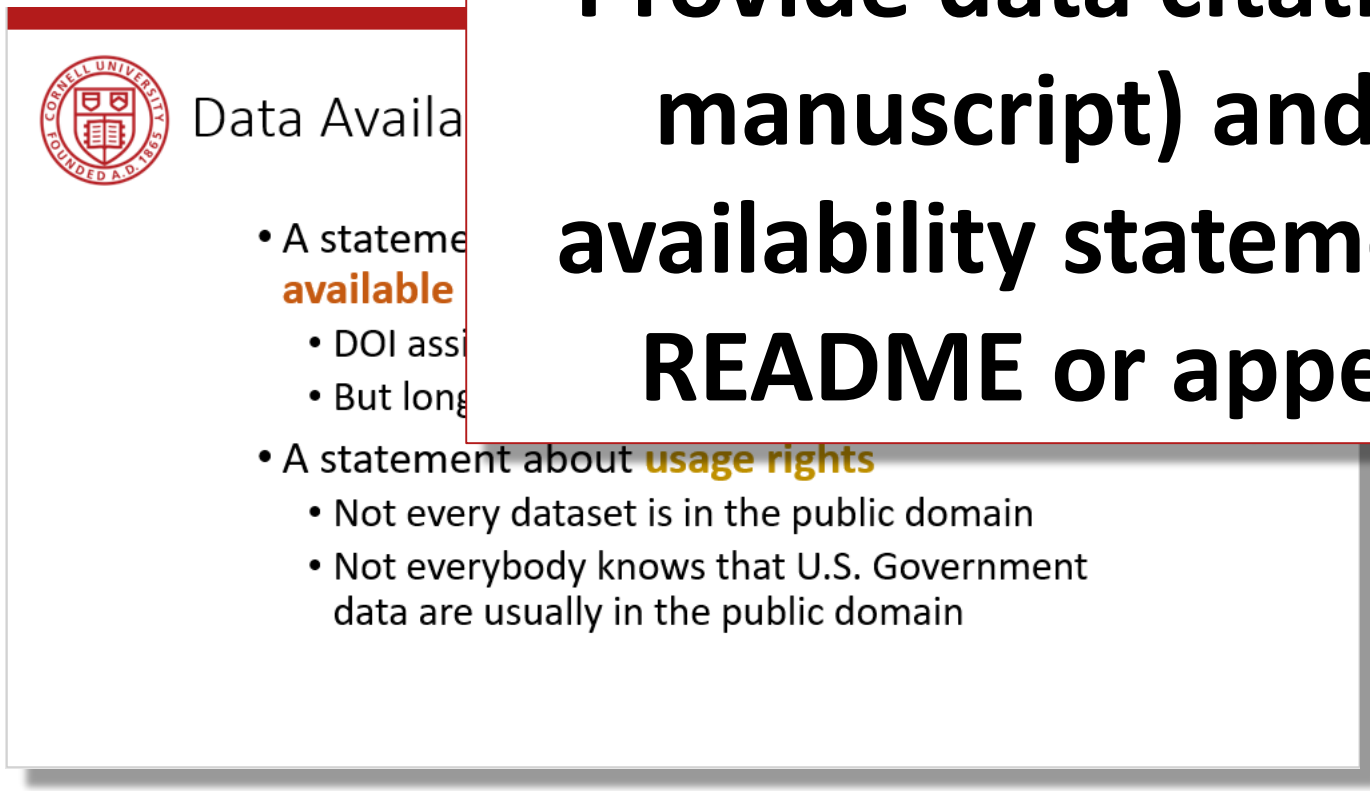
2. Correct citation

- 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?





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

Take-away

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!

Coding for Reproducibility

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!

Some tips from the "frequently gotten wrong" bin

- Set the project directory **ONCE** in code, or **NEVER** (Stata, R, Python)
- Use **placeholders** (globals, libnames, etc.) for common locations ($CONFDATA, $TABLES, $CODE) (Stata, R, Python, SAS)
- **Write out all tables, figures**, and in-text numbers into separate files

If you need to manually modify the code to obtain a series of tables/figures/columns, you're doing something wrong:

- Use **functions, ado files, programs, macros, subroutines**
- Use **loops, parameters, parameter files** to call those subroutines

Some tips from the “frequently gotten wrong” bin

Cleanly **separate**

- Confidential data and public use data
  - You are going to have to provide copies of the public use data without compromising confidentiality
- Confidential parameters and the rest of the code
  - Reduces need to redact programs

- Use **placeholders** (globals, libnames, etc.) for common locations ($CONFDATA, $TABLES, $CODE) (Stata, R, Python, SAS)

Some tips from the “frequently gotten wrong” bin

Have “**computational empathy**”

- Consider cross-platform programmng practices
- Consider that the replicator can learn from the process
  - They probably don’t have the same knowledge
- Consider that the replicator might not have the same modules/packages/etc.

- Path and filenames:
  - Stata: always use forward slashes, even on Windows

```
use “$data/path/data.dta”
```

  - R: use “`file.path()`”

```
x <-
read(file.path(data,”data.dta”)
```

  - SAS: use `filename` and `libname` to abstract

```
data DATALIB.step1;
set CONFLIB.slid_1996;
```

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

- 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
- ….

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

How to prepare the replication package

- README
- Now ask an RA/ colleague/

AEA Data and Code Guidance

AMERICAN ECONOMIC ASSOCIATION

Guidance for autho
data and code sup
replicators.

Steps for the Third-party Replicator

- Download the author's replication archive(s) from the designated URL (public, or privately shared)
- Ensure access to any confidential files that are described in the replication archive's README
  - The replicator should consider whether a third-party person not
    ironment could reasonably
    olely on the instructions in the
- Follow the checklist to conduct the reproducibility exercise, relying exclusively on the README for instructions and guidance.
- Write a report
- Send the report to the AEA Data Editor
- Report any interactions with the author in the course of conducting the reproducibility exercise (help, assistance, clarifications)

That's our Protocol!

Tension between access and reproducibility

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

In a nutshell

- **40%** use restricted-access data
- **25%** use public-use data and are mostly or completely reproducible
- **25%** use public-use data and are only partially reproducible
- **10%** fail to yield useful results

Assume we can access the data

Sometimes we cannot

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

Sometimes we can:

- In the past 6 months, 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

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/

Thank you!

https://doi.org/10.5281/zenodo.3981458

Some resources

- https://social-science-data-editors.github.io/guidance/
- https://aeadataeditor.github.io/aea-de-guidance/
  - **template README**
  - discussion of licensing
  - data citation guidance
- German example:
  - Establishment History Panel (BHP) DOI: 10.5164/IAB.BHP7516.de.en.v1
- French verification service “cascad” within French RDC CASD
  - https://www.casd.eu/en/le-centre-dacces-securise-aux-donnees-casd/certification-de-resultats-cascad-casd/

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.