Principles of Transparent Research

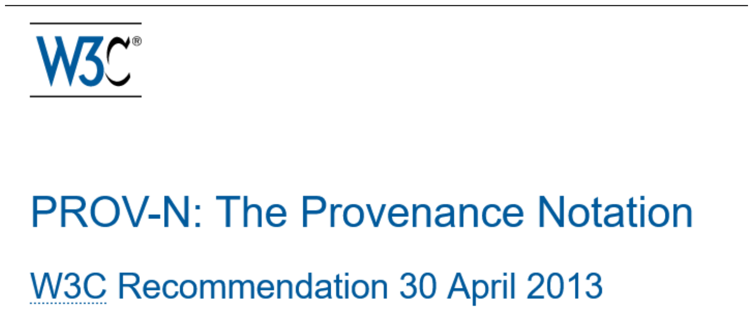
Implementation Challenges

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.

Context



PROV-N: The Provenance Notation

W3C Recommendation 30 April 2013

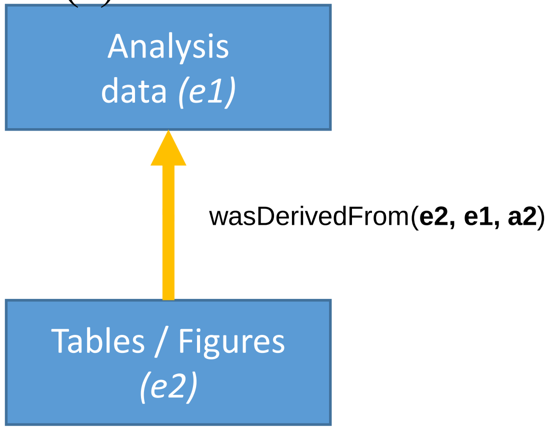
```
wasDerivedFrom(e2, e1, a, g2, u1)
```

Replication packages in Social Sciences

- Started to appear in mid-1980s
- American Economic Review “data availability policy” 2005

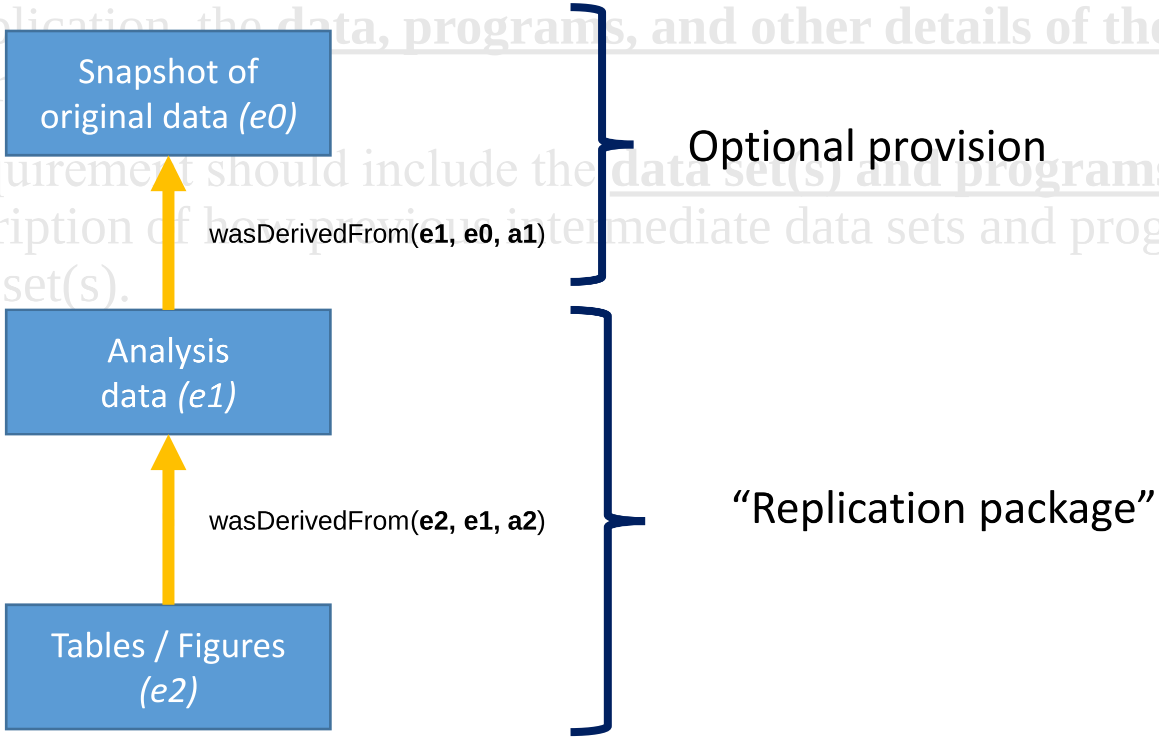
AER policy ca. 2005

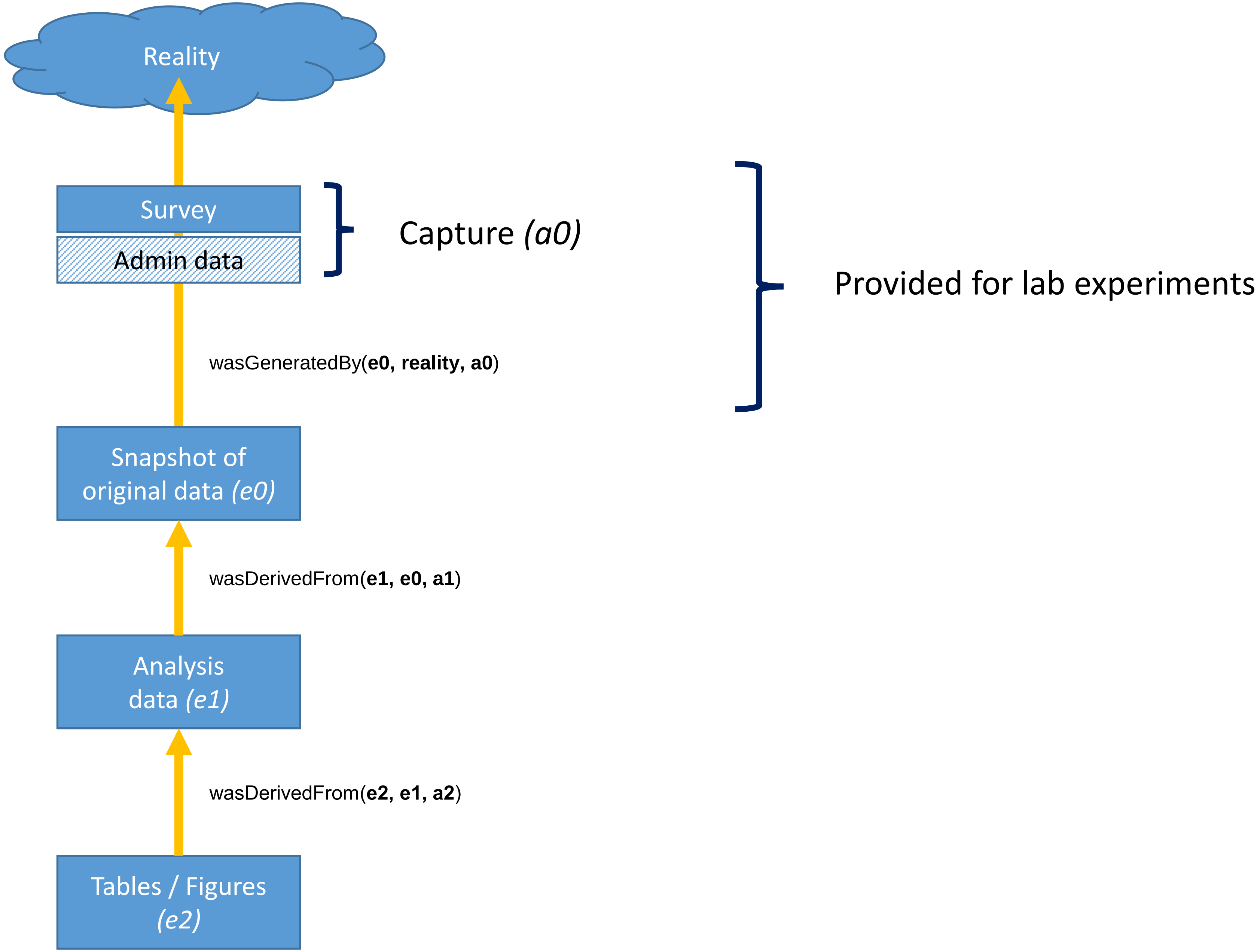
- It is the policy of the AEA to publish papers only if the **data used in the analysis** are clearly and precisely documented and are readily available to any researcher for purposes of replication.
- Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to publication, the **data, programs, and other details of the computations** sufficient to permit replication.
- … the minimum requirement should include the **data set(s) and programs used to run the final models**, plus a description of how previous intermediate data sets and programs were employed to create the final data set(s).



AER policy ca. 2005

- It is the policy of the AEA to publish papers only if the **data used in the analysis** are clearly and precisely documented and are readily available to any researcher for purposes of replication.
- Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to publication, the **data, programs, and other details of the computations** sufficient to permit r
- … the minimum requirement should include the **data set(s) and programs used to run the final models**, plus a description of how previous intermediate data sets and programs were employed to create the final data set(s).


Snapshot of original data *(e0)*
wasDerivedFrom(**e1, e0, a1**)
Analysis data *(e1)*
wasDerivedFrom(**e2, e1, a2**)
Tables / Figures *(e2)*
Optional provision
“Replication package”



Reality

Survey

Admin data

Capture *(a0)*

Provided for lab experiments

wasGeneratedBy(**e0, reality, a0**)

Snapshot of original data *(e0)*

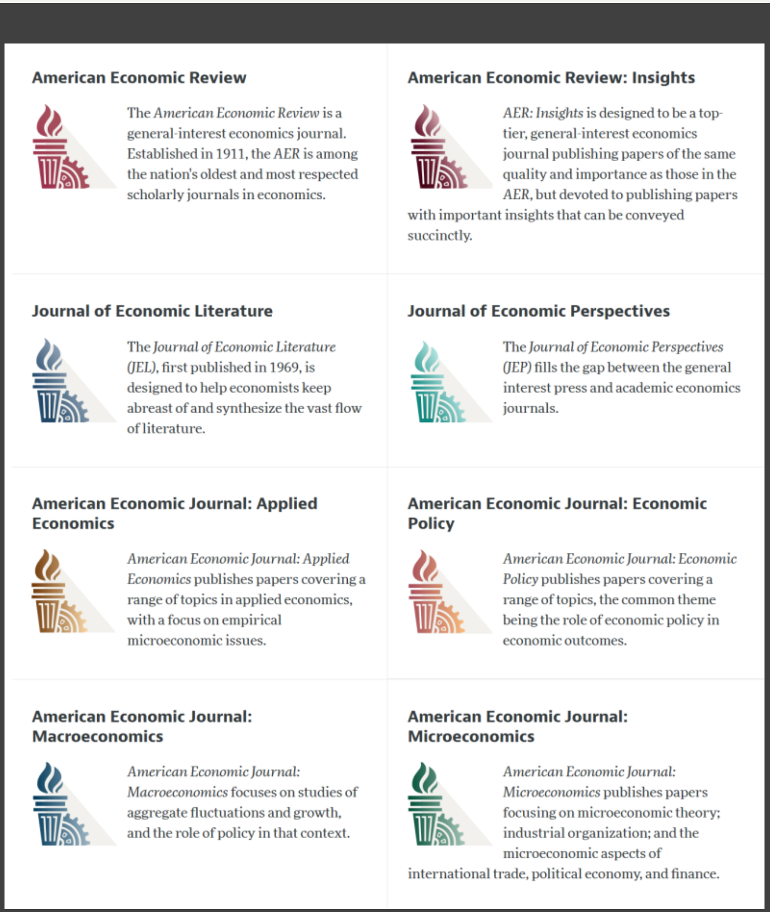
wasDerivedFrom(**e1, e0, a1**)

Analysis data *(e1)*

wasDerivedFrom(**e2, e1, a2**)

Tables / Figures *(e2)*

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.

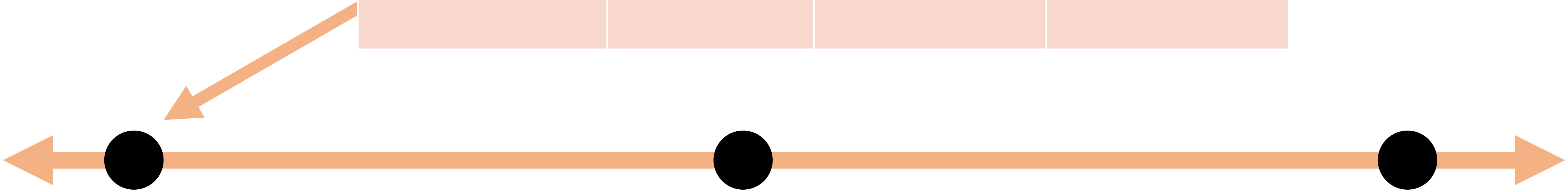
American Economic Journal: Microeconomics

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

Replication continuum

https://doi.org/10.17226/25303

Same data	Same code	Same methods	Same context



Reproducibility

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

Replicability

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

Generalizability

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

Historical challenges

Poor citation practices

Provenance!

- **Macrodata:**

“We use data downloaded from
the Bureau of Economic Analysis…”

- **Microdata**:

“… this paper uses data from
the Current Population Survey…”

Poor coding practices

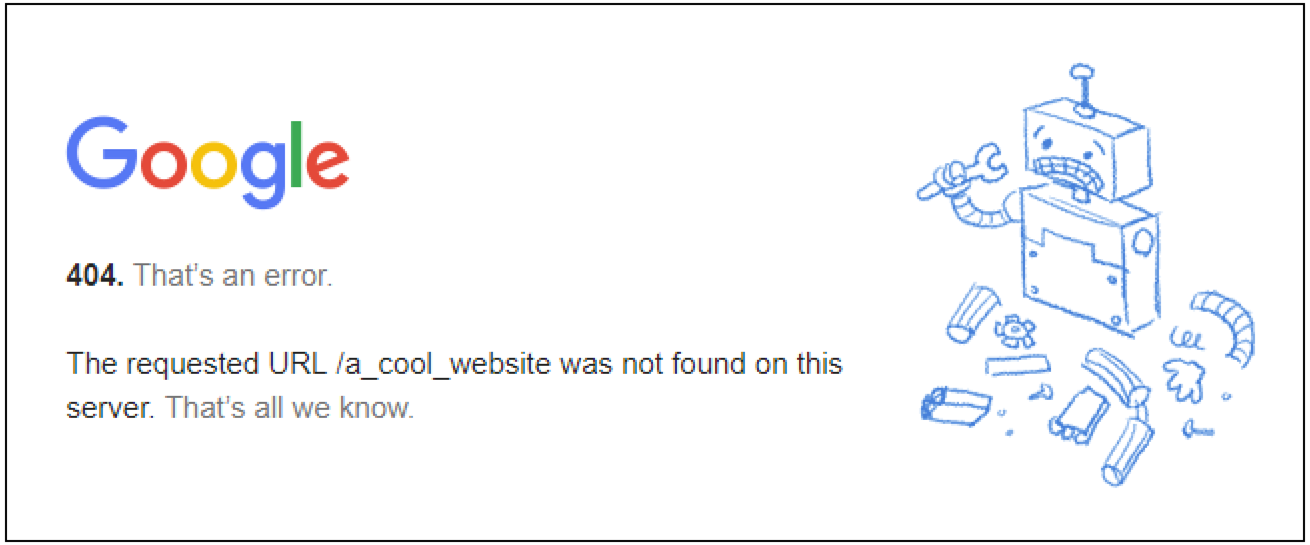
- **Manual/non-automation**

Code produces no meaningful output

- **Lack of robustness**:

Bugs in the code

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.

AEA Data & Code Availability Policy (2019)

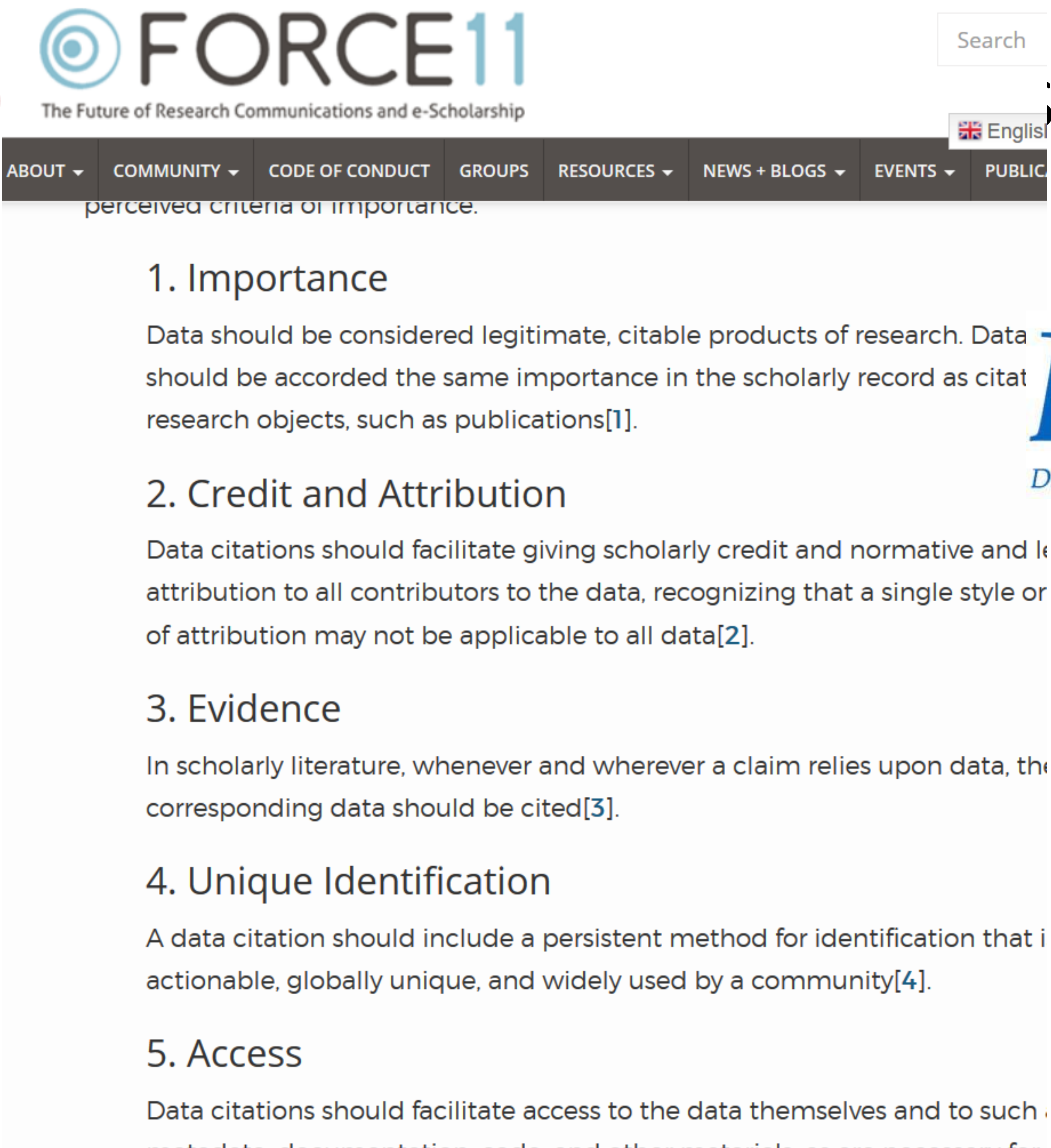
- It is the policy of the American Economic Association to publish papers only if the data used in the analysis are **clearly and precisely** documented and **access to the data and code is clearly and precisely documented and is non-exclusive to the authors.**
- Authors of accepted papers that contain 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



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

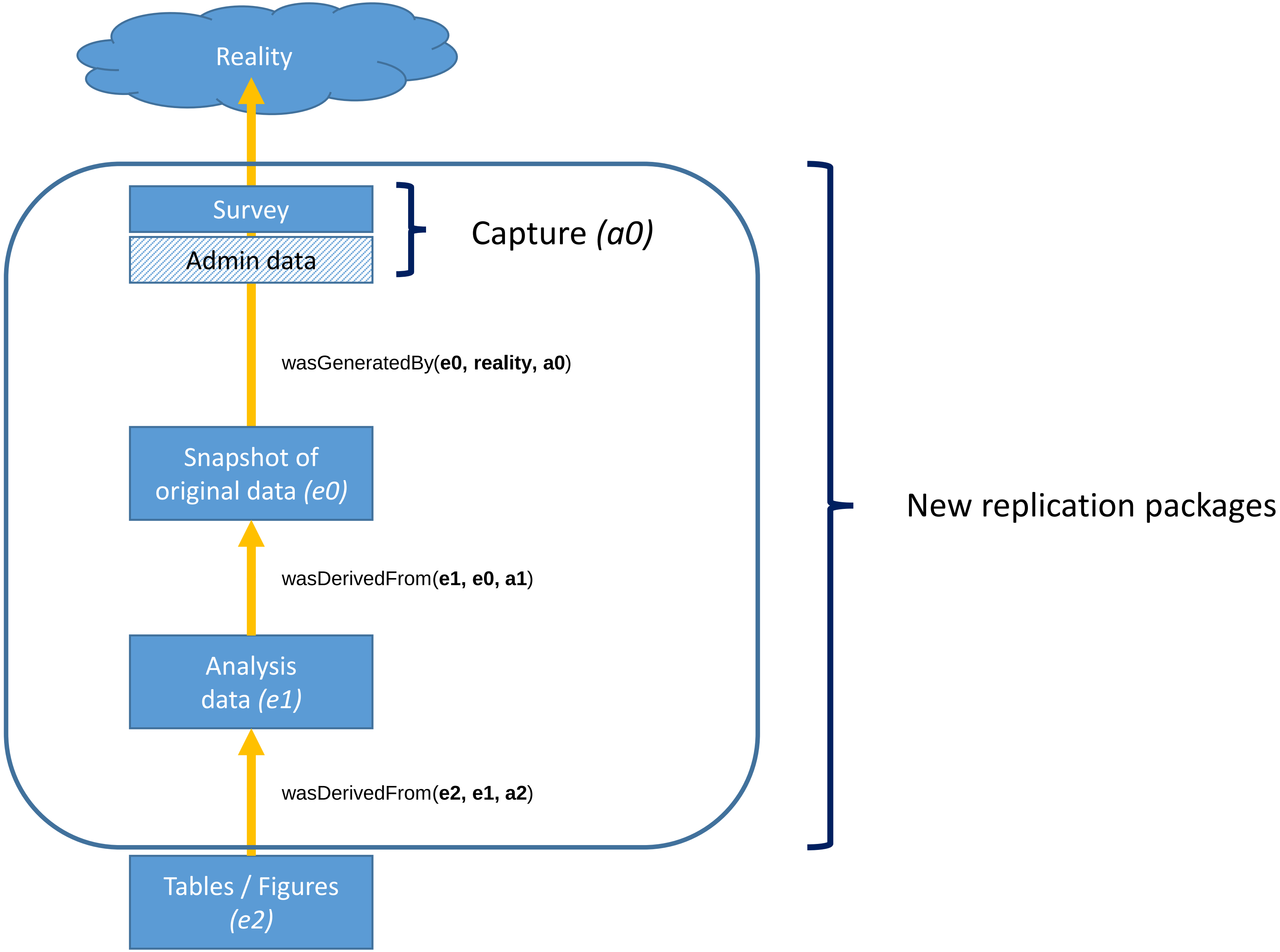
A data citation should include a persistent method for identification that i actionable, globally unique, and widely used by a community[4].

5. Access

Data citations should facilitate access to the data themselves and to such metadata, documentation, code, and other materials as are necessary for

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].



Reality

Survey

Admin data

Capture *(a0)*

wasGeneratedBy(**e0, reality, a0**)

Snapshot of original data *(e0)*

wasDerivedFrom(**e1, e0, a1**)

Analysis data *(e1)*

wasDerivedFrom(**e2, e1, a2**)

Tables / Figures *(e2)*

New replication packages

Remaining challenges

Poor citation practices

- **Macrodata:**

> “We use data downloaded from
> the Bureau of Economic Analysis…”

- **Microdata**:

> “… this paper uses data from
> the Current Population Survey…”

Poor coding practices

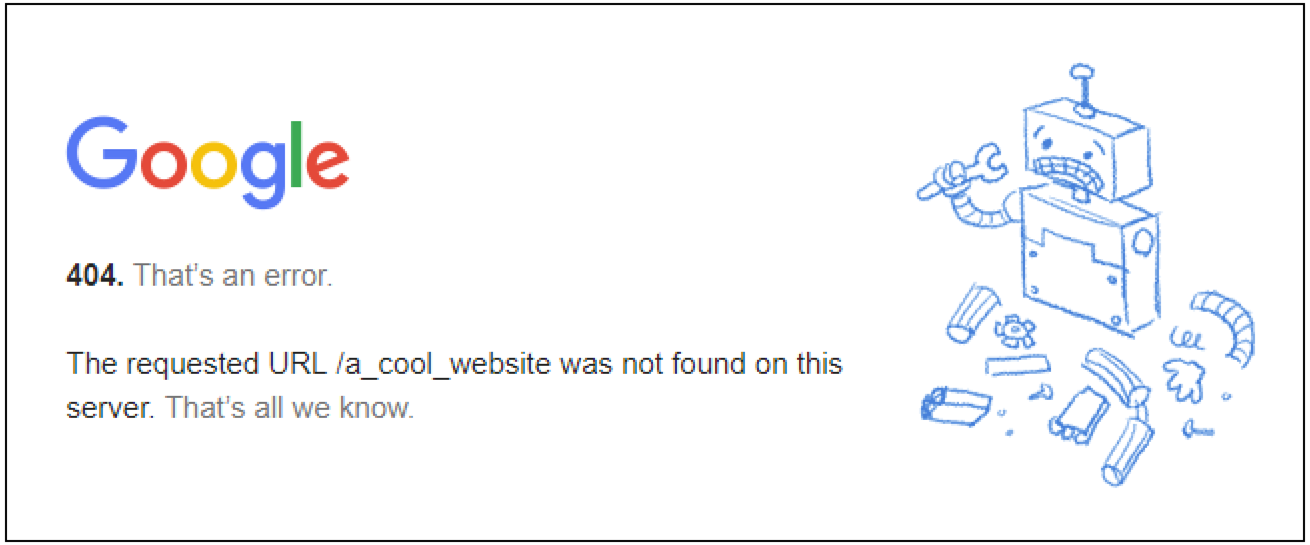
- **Manual/non-automation**

Code produces no meaningful output

- **Lack of robustness**:

Bugs in the code

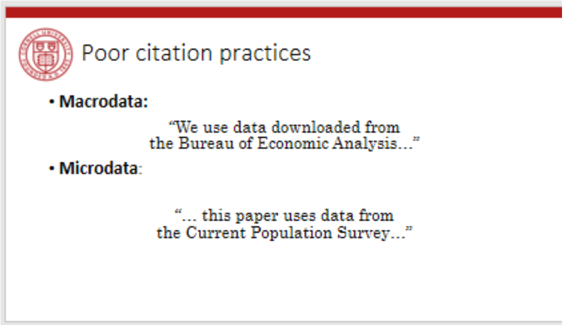
Failure to curate



404. That's an error.

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

Challenges remaining



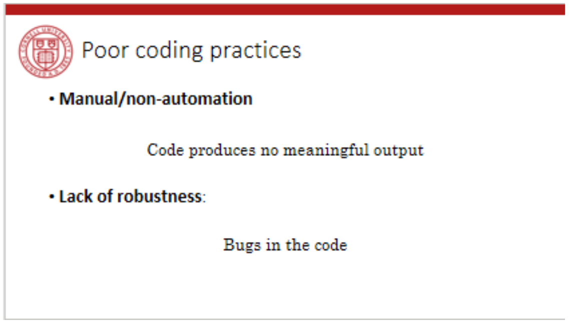
Poor citation practices

- **Macrodata:**

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

- **Microdata**:

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



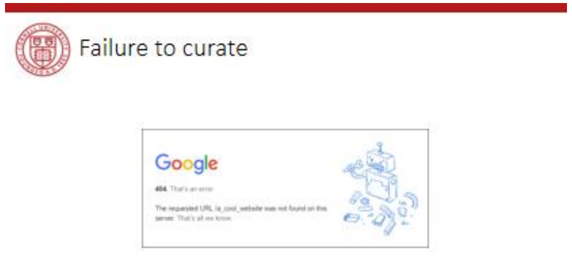
Poor coding practices

- **Manual/non-automation**

  Code produces no meaningful output

- **Lack of robustness**:

  Bugs in the code



Failure to curate

404. That's an error.

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

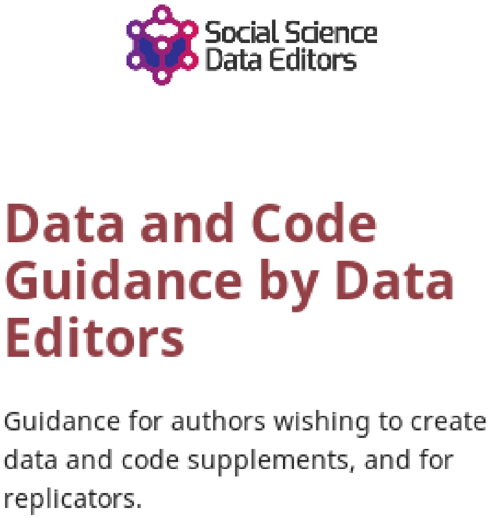
- Required now
- Verified by Data Editor
- Not always in manuscript
- Challenging!

- Still an issue
- Self-check needed
- "Computational empathy" required

- Solved for code
- Fallback available for data
- Lots of issues with data without redistribution rights (confidential, otherwise)

Actions

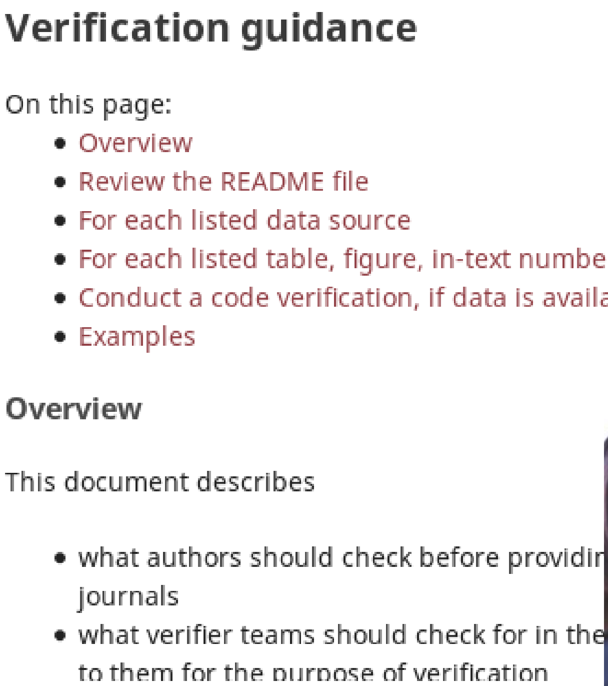
Action: Reproducibility Check



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 providin journals
- what verifier teams should check for in the to them for the purpose of verification

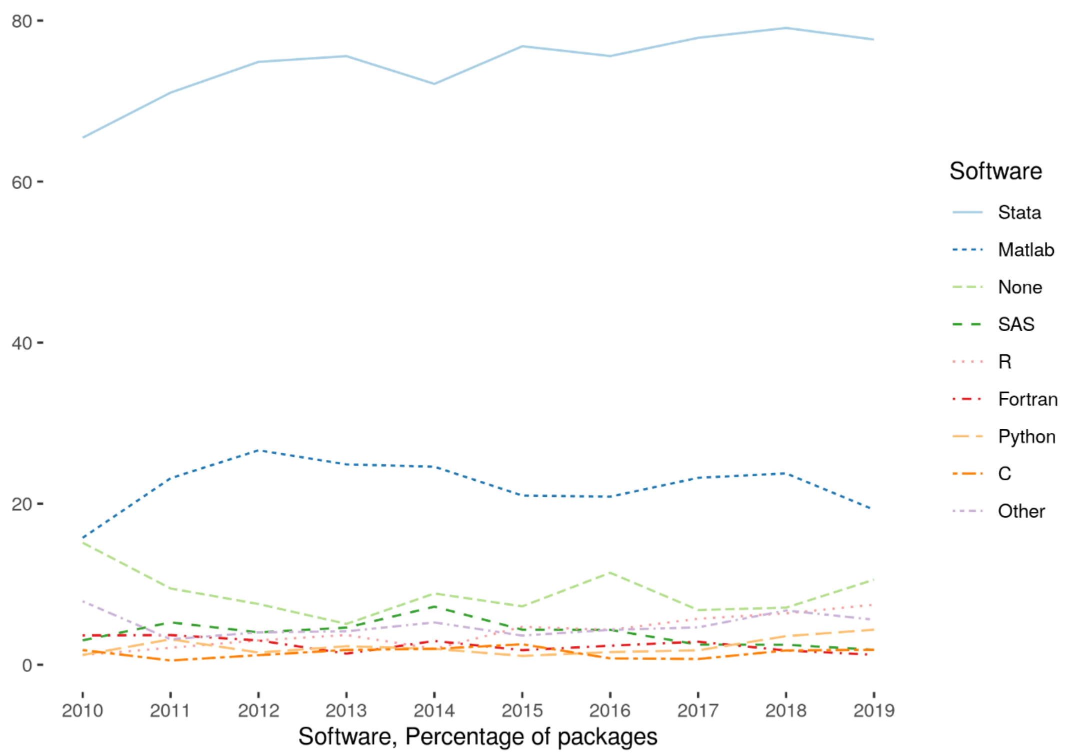
Stats on reproduced articles

Between Dec 1, 2019, and Nov 30, 2020, the AEA Data Editor team conducted

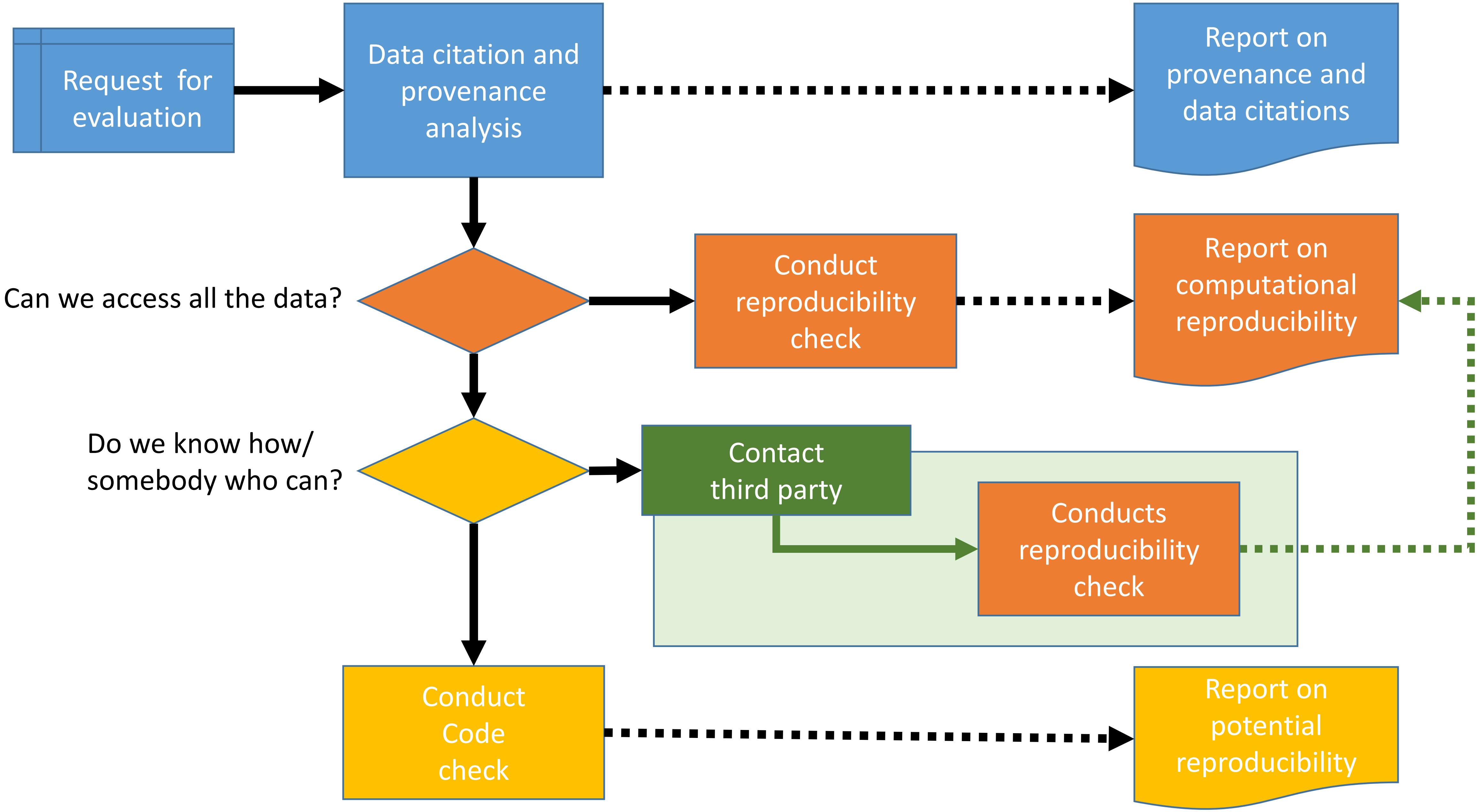
- **~832 assessments**
- ~**509 manuscripts** have been "accepted"

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%**)



How this works



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

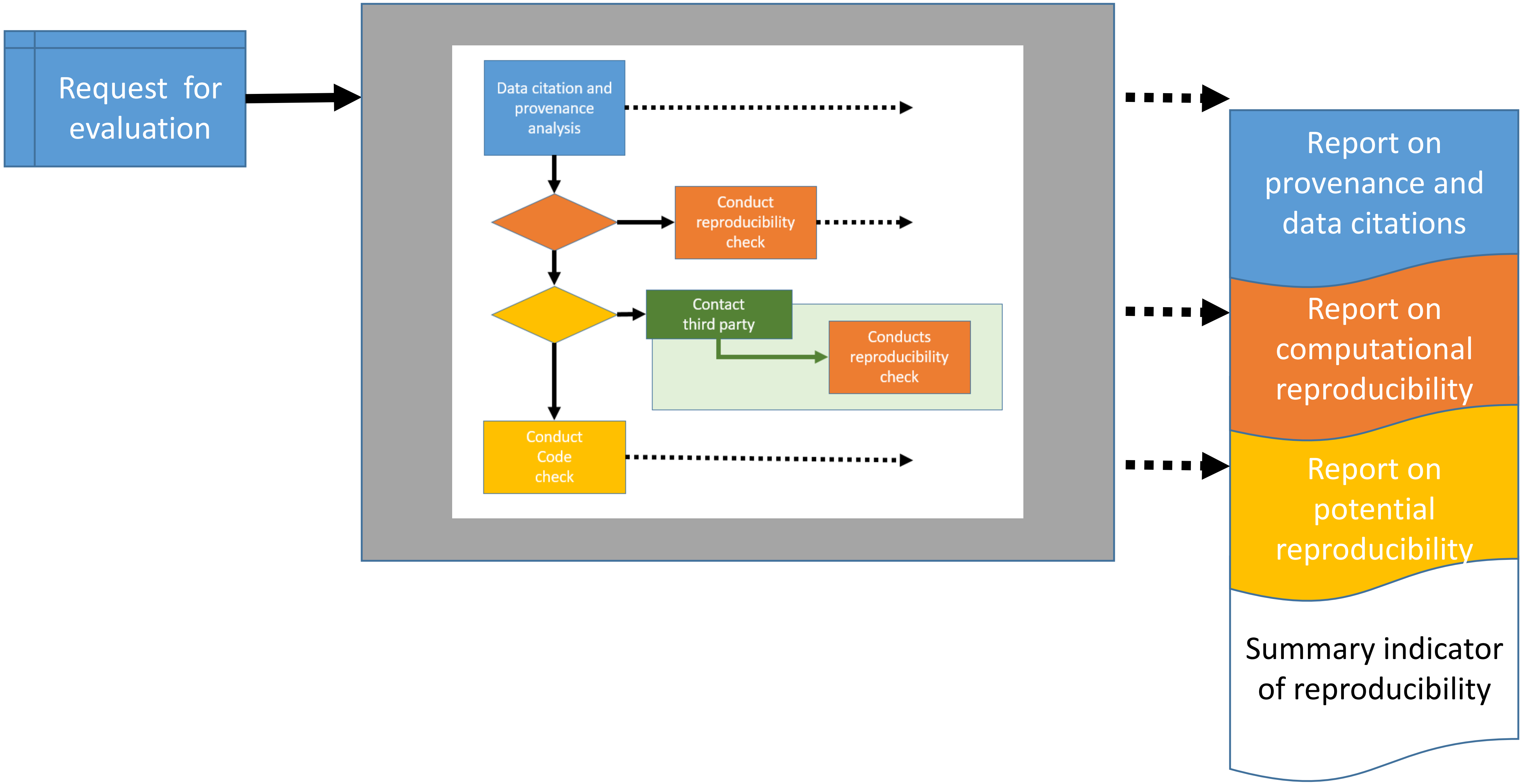
Do we know how/ somebody who can?

Contact third party

Conducts reproducibility check

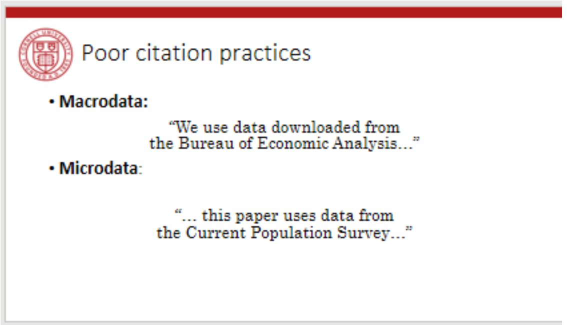
Conduct Code check

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
Summary indicator of reproducibility

Challenge: Where’s the Data?

Challenges remaining



Poor citation practices

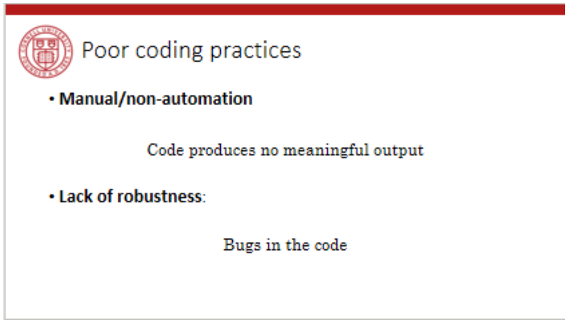
- **Macrodata:**

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

- **Microdata**:

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

- Required now
- Verified by Data Editor
- Not always in manuscript
- Challenging!



Poor coding practices

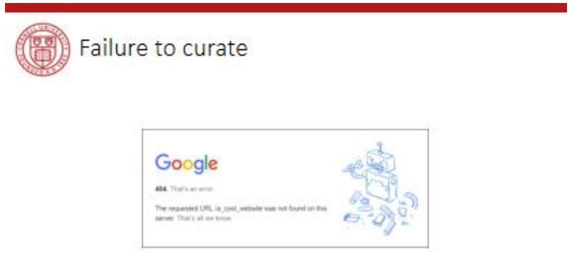
- **Manual/non-automation**

  Code produces no meaningful output

- **Lack of robustness**:

  Bugs in the code

- Still an issue
- Self-check needed
- "Computational empathy" required



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.

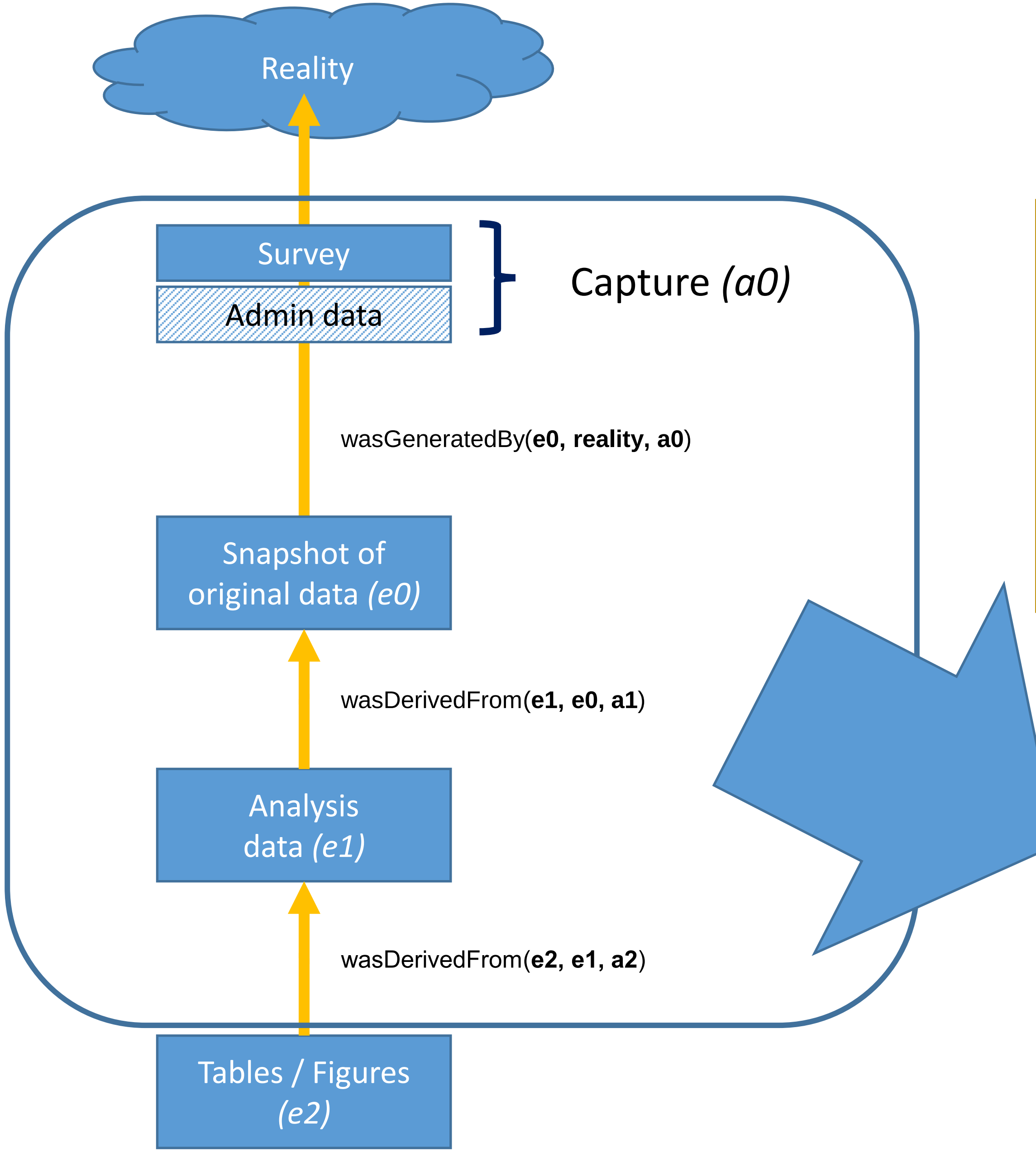
- Solved for code
- Fallback available for data
- Lots of issues with data without redistribution rights (confidential, otherwise)

Three pieces

Where did the author get the data?

Where can others get the same data?

Persistence of the data



Reality

Survey

Admin data

Capture *(a0)*

wasGeneratedBy(**e0, reality, a0**)

Snapshot of original data *(e0)*

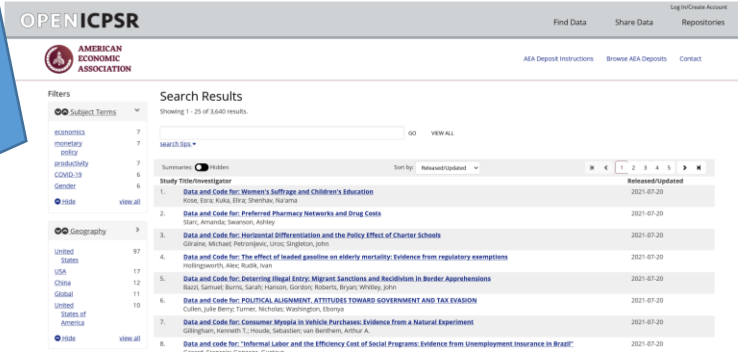
wasDerivedFrom(**e1, e0, a1**)

Analysis data *(e1)*

wasDerivedFrom(**e2, e1, a2**)

Tables / Figures *(e2)*

Easy case:
Author-conducted survey



OPENICPSR

Log In/Create Account

Find Data Share Data Repositories

AMERICAN ECONOMIC ASSOCIATION

AEA Deposit Instructions Browse AEA Deposits Contact

Filters

Subject Terms

economics	7
monetary policy	7
productivity	7
COVID-19	6
Gender	6

Hide view all

Geography

United States	97
USA	17
China	12
Global	11
United States of America	10

Hide view all

Search Results

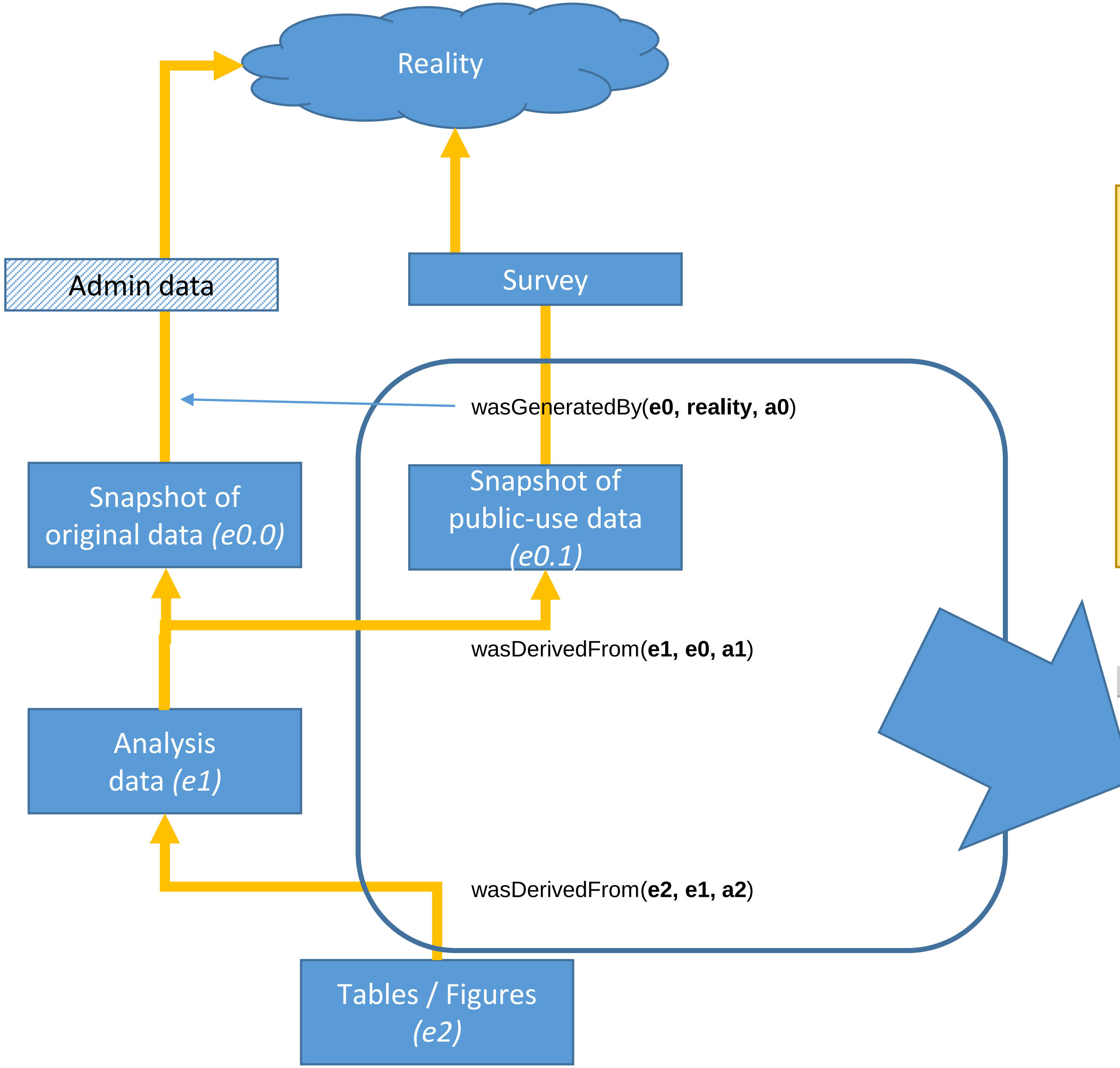
Showing 1 - 25 of 3,640 results.

GO VIEW ALL

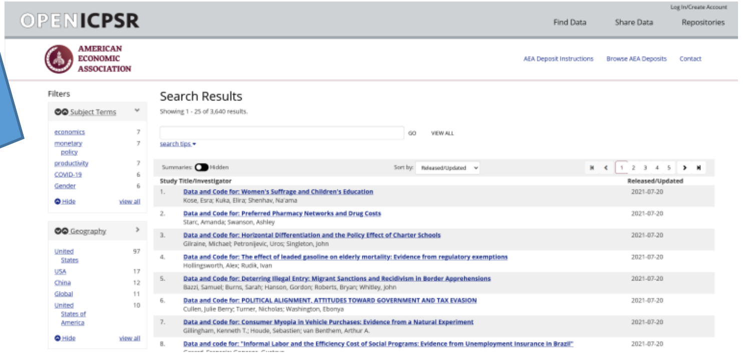
search tips ▾

Summaries: Hidden Sort by: Released/Updated 1 2 3 4 5

	Study Title/Investigator	Released/Updated
1.	Data and Code for: Women's Suffrage and Children's Education Kose, Esra; Kuka, Elira; Shenhav, Na'ama	2021-07-20
2.	Data and Code for: Preferred Pharmacy Networks and Drug Costs Starc, Amanda; Swanson, Ashley	2021-07-20
3.	Data and Code for: Horizontal Differentiation and the Policy Effect of Charter Schools Gilraine, Michael; Petronijevic, Uros; Singleton, John	2021-07-20
4.	Data and Code for: The effect of leaded gasoline on elderly mortality: Evidence from regulatory exemptions Hollingsworth, Alex; Rudik, Ivan	2021-07-20
5.	Data and Code for: Deterring Illegal Entry: Migrant Sanctions and Recidivism in Border Apprehensions Bazzi, Samuel; Burns, Sarah; Hanson, Gordon; Roberts, Bryan; Whitley, John	2021-07-20
6.	Data and Code for: POLITICAL ALIGNMENT, ATTITUDES TOWARD GOVERNMENT AND TAX EVASION Cullen, Julie Berry; Turner, Nicholas; Washington, Ebonya	2021-07-20
7.	Data and Code for: Consumer Myopia in Vehicle Purchases: Evidence from a Natural Experiment Gillingham, Kenneth T.; Houde, Sebastien; van Benthem, Arthur A.	2021-07-20
8.	Data and code for: "Informal Labor and the Efficiency Cost of Social Programs: Evidence from Unemployment Insurance in Brazil"	2021-07-20


Reality
Admin data
Survey
wasGeneratedBy(**e0, reality, a0**)
Snapshot of original data *(e0.0)*
Snapshot of public-use data *(e0.1)*
wasDerivedFrom(**e1, e0, a1**)
Analysis data *(e1)*
wasDerivedFrom(**e2, e1, a2**)
Tables / Figures *(e2)*

Difficult case:
Confidential data with
auxiliary public-use data



Filters

Subject Terms

economics	7
monetary policy	7
productivity	7
COVID-19	6
Gender	6

Geography

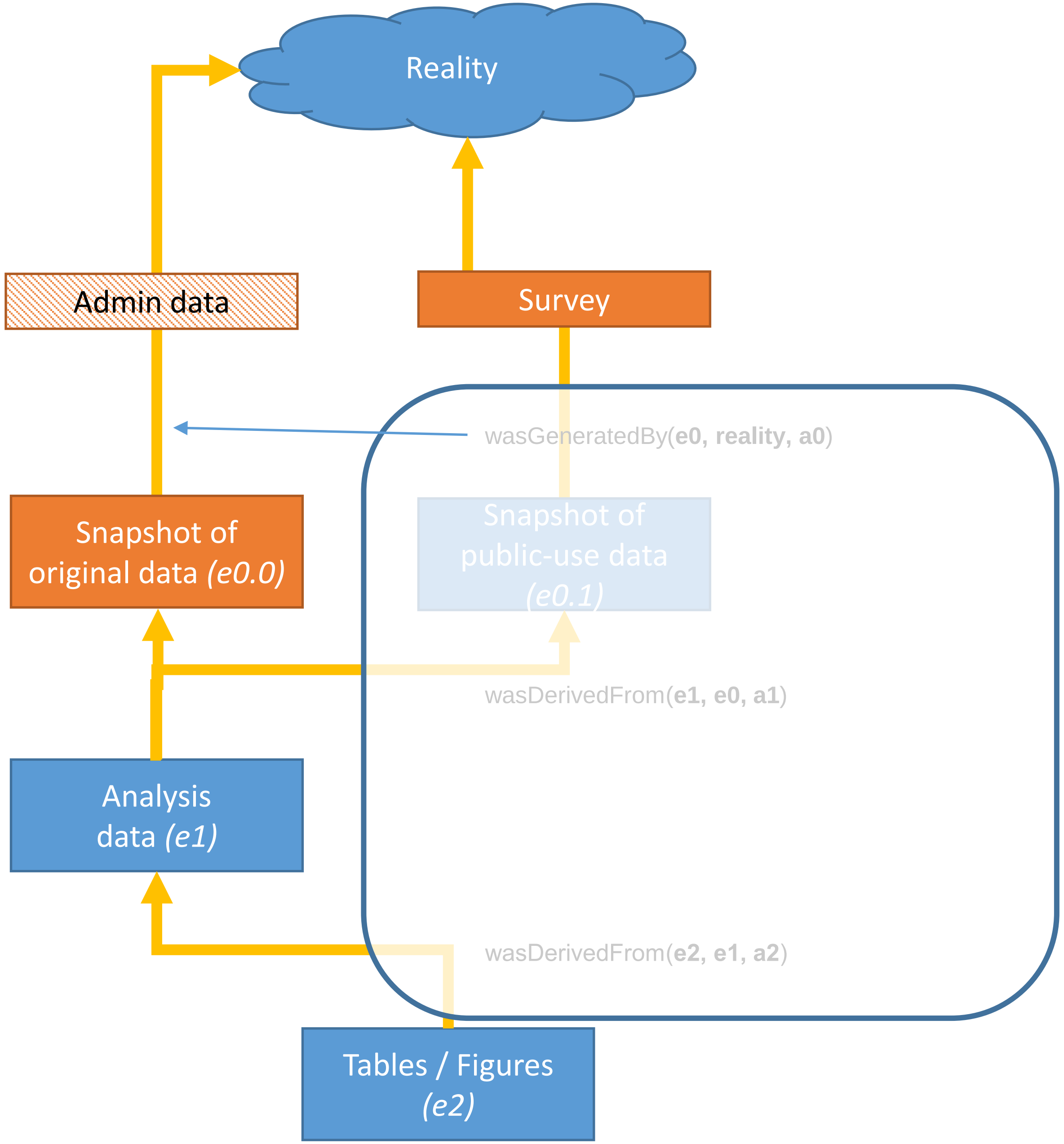
United States	97
USA	17
China	12
Global	11
United States of America	10

Search Results

Showing 1 - 25 of 3,640 results.

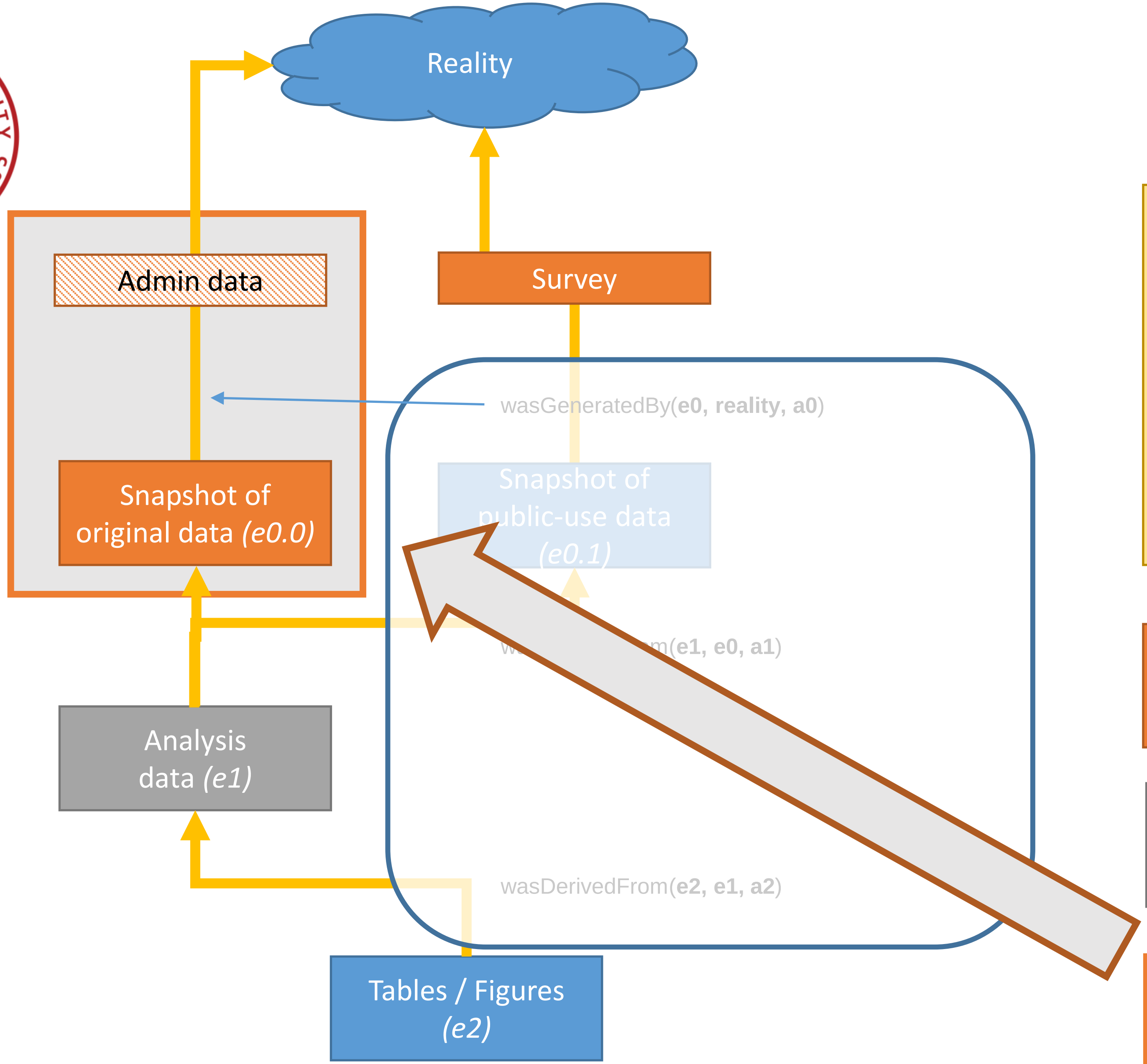
Summaries: Hidden Sort by: Released/Updated 1 2 3 4 5

Study Title/Investigator	Released/Updated
1. Data and Code for: Women's Suffrage and Children's Education Kose, Esra; Kuka, Elira; Shenhav, Na'ama	2021-07-20
2. Data and Code for: Preferred Pharmacy Networks and Drug Costs Starc, Amanda; Swanson, Ashley	2021-07-20
3. Data and Code for: Horizontal Differentiation and the Policy Effect of Charter Schools Gilraine, Michael; Petronijevic, Uros; Singleton, John	2021-07-20
4. Data and Code for: The effect of leaded gasoline on elderly mortality: Evidence from regulatory exemptions Hollingsworth, Alex; Rudik, Ivan	2021-07-20
5. Data and Code for: Deterring Illegal Entry: Migrant Sanctions and Recidivism in Border Apprehensions Bazzi, Samuel; Burns, Sarah; Hanson, Gordon; Roberts, Bryan; Whitley, John	2021-07-20
6. Data and Code for: POLITICAL ALIGNMENT, ATTITUDES TOWARD GOVERNMENT AND TAX EVASION Cullen, Julie Berry; Turner, Nicholas; Washington, Ebonya	2021-07-20
7. Data and Code for: Consumer Myopia in Vehicle Purchases: Evidence from a Natural Experiment Gillingham, Kenneth T.; Houde, Sebastien; van Benthem, Arthur A.	2021-07-20
8. Data and code for: "Informal Labor and the Efficiency Cost of Social Programs: Evidence from Unemployment Insurance in Brazil"	2021-07-20


Reality
Admin data
Survey
wasGeneratedBy(e0, reality, a0)
Snapshot of original data (e0.0)
Snapshot of public-use data (e0.1)
wasDerivedFrom(e1, e0, a1)
Analysis data (e1)
wasDerivedFrom(e2, e1, a2)
Tables / Figures (e2)

Difficult case:
Confidential data with auxiliary public-use data

Author does not control data

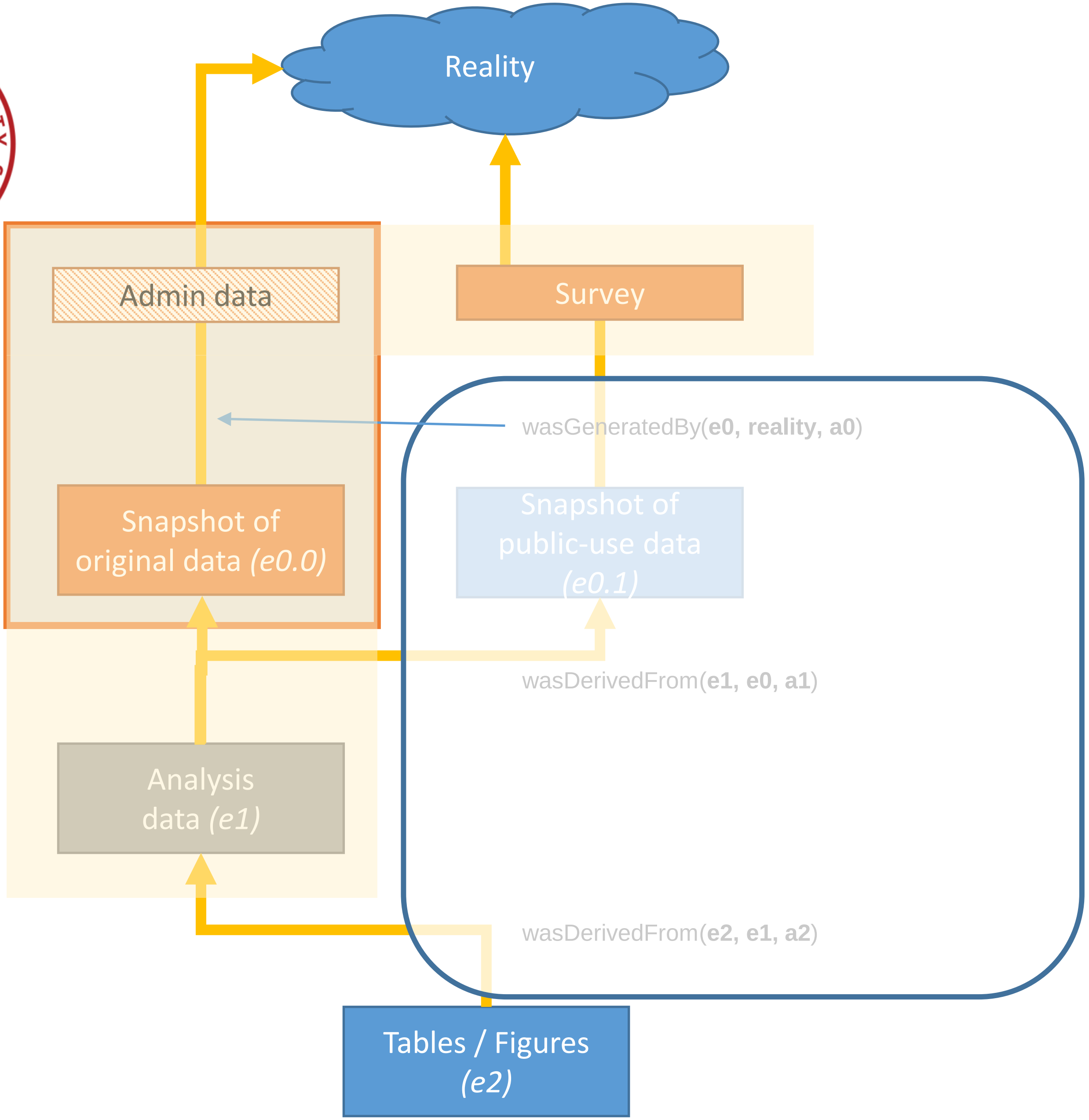

Reality
Admin data
Survey
wasGeneratedBy(e0, reality, a0)
Snapshot of original data (e0.0)
Snapshot of public-use data (e0.1)
w
m(e1, e0, a1)
Analysis data (e1)
wasDerivedFrom(e2, e1, a2)
Tables / Figures (e2)

Difficult case:
Confidential data with
auxiliary public-use data

Author does not control data

Author may not be able to preserve

Data source not well defined



Reality

Admin data

Survey

wasGeneratedBy(**e0, reality, a0**)

Snapshot of original data *(e0.0)*

Snapshot of public-use data *(e0.1)*

wasDerivedFrom(**e1, e0, a1**)

Analysis data *(e1)*

wasDerivedFrom(**e2, e1, a2**)

Tables / Figures *(e2)*

Difficult case:
Confidential data with auxiliary public-use data

Author does not control data

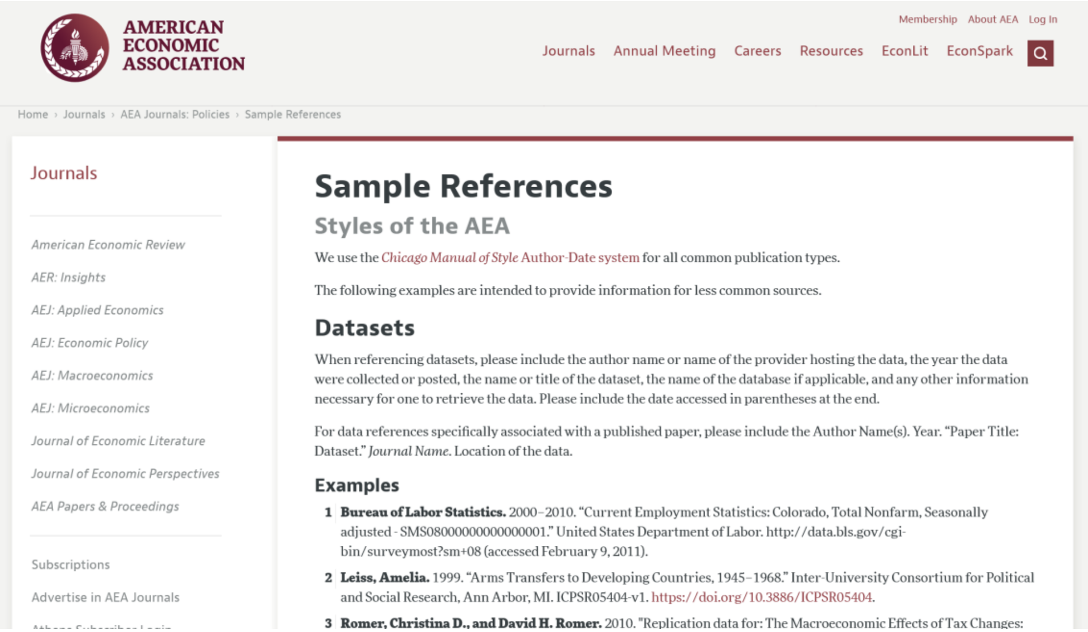
Author may not be able to preserve

Data source not well defined

Data has no PID (or even a webpage)

Actions: Provide guidance

Direct guidance to authors



AMERICAN ECONOMIC ASSOCIATION

Membership About AEA Log In

Journals Annual Meeting Careers Resources EconLit EconSpark

Home › Journals › AEA Journals: Policies › Sample References

Journals

American Economic Review

AER: Insights

AEJ: Applied Economics

AEJ: Economic Policy

AEJ: Macroeconomics

AEJ: Microeconomics

Journal of Economic Literature

Journal of Economic Perspectives

AEA Papers & Proceedings

Subscriptions

Advertise in AEA Journals

Sample References

Styles of the AEA

We use the *Chicago Manual of Style* Author-Date system for all common publication types.

The following examples are intended to provide information for less common sources.

Datasets

When referencing datasets, please include the author name or name of the provider hosting the data, the year the data were collected or posted, the name or title of the dataset, the name of the database if applicable, and any other information necessary for one to retrieve the data. Please include the date accessed in parentheses at the end.

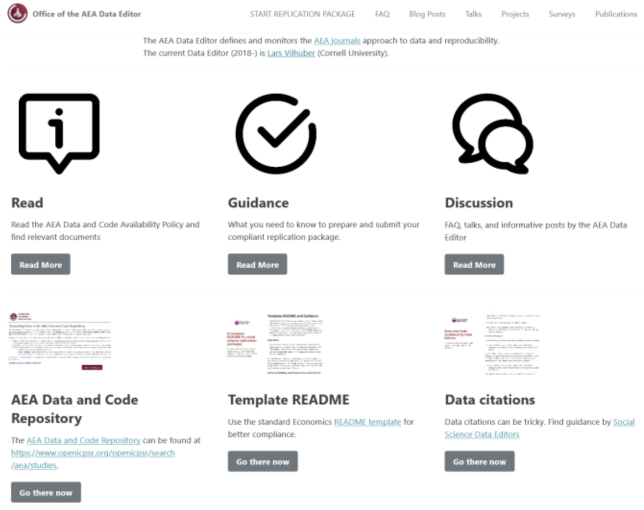
For data references specifically associated with a published paper, please include the Author Name(s). Year. "Paper Title: Dataset." *Journal Name*. Location of the data.

Examples

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).

2 **Leiss, Amelia.** 1999. "Arms Transfers to Developing Countries, 1945–1968." Inter-University Consortium for Political and Social Research, Ann Arbor, MI. ICPSR05404-v1. https://doi.org/10.3886/ICPSR05404.

3 **Romer, Christina D., and David H. Romer.** 2010. "Replication data for: The Macroeconomic Effects of Tax Changes:



Office of the AEA Data Editor

START REPLICATION PACKAGE FAQ Blog Posts Talks Projects Surveys Publications

The AEA Data Editor defines and monitors the AEA journals approach to data and reproducibility.
The current Data Editor (2018-) is Lars Vilhuber (Cornell University).

Read

Read the AEA Data and Code Availability Policy and find relevant documents

Read More

Guidance

What you need to know to prepare and submit your compliant replication package.

Read More

Discussion

FAQ, talks, and informative posts by the AEA Data Editor

Read More

AEA Data and Code Repository

The AEA Data and Code Repository can be found at https://www.openicpsr.org/openicpsr/search/aea/studies.

Go there now

Template README

Use the standard Economics README template for better compliance.

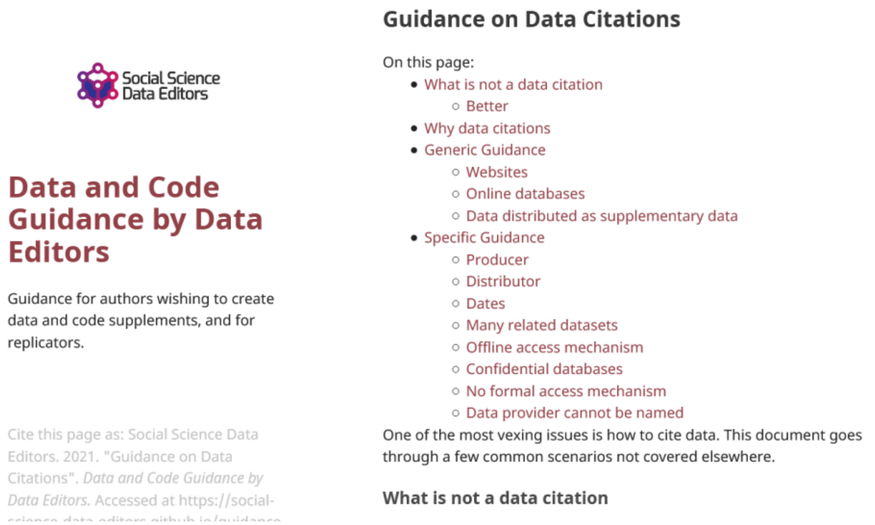
Go there now

Data citations

Data citations can be tricky. Find guidance by Social Science Data Editors

Go there now

Enhanced guidance for authors



Social Science Data Editors

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. 2021. "Guidance on Data Citations". *Data and Code Guidance by Data Editors*. Accessed at https://social-

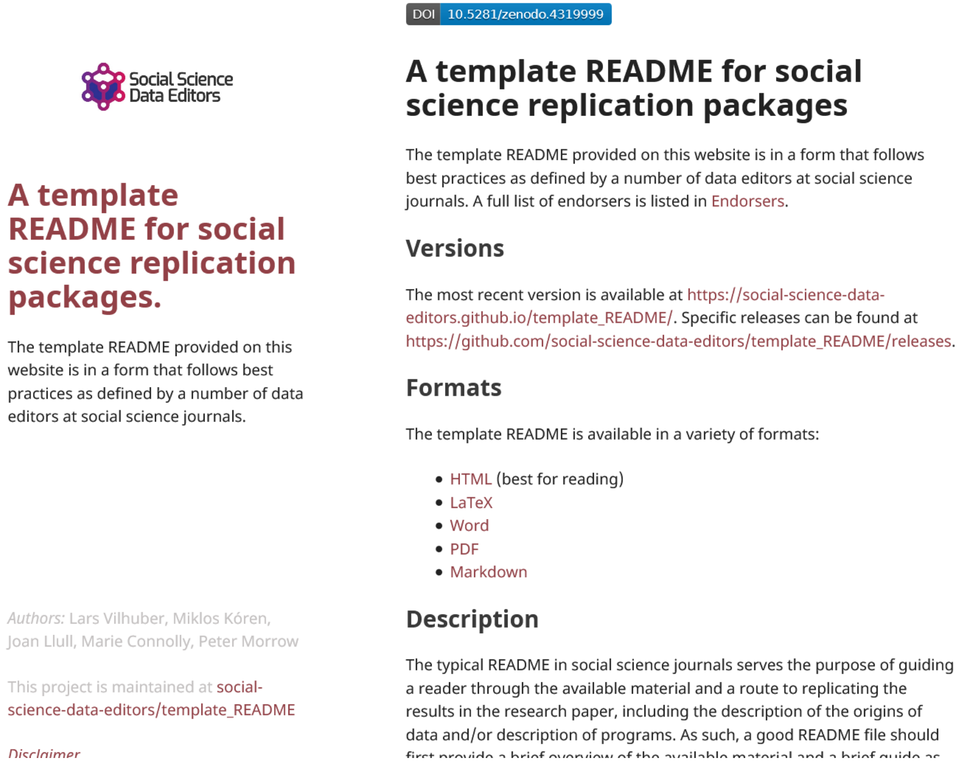
Guidance on Data Citations

On this page:

- What is not a data citation
    - Better
- Why data citations
- Generic Guidance
    - Websites
    - Online databases
    - Data distributed as supplementary data
- Specific Guidance
    - Producer
    - Distributor
    - Dates
    - Many related datasets
    - Offline access mechanism
    - Confidential databases
    - No formal access mechanism
    - Data provider cannot be named

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



Social Science Data Editors

A template README for social science replication packages.

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

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

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

Disclaimer

DOI 10.5281/zenodo.4319999

A template README for social science replication packages

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

Versions

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

Formats

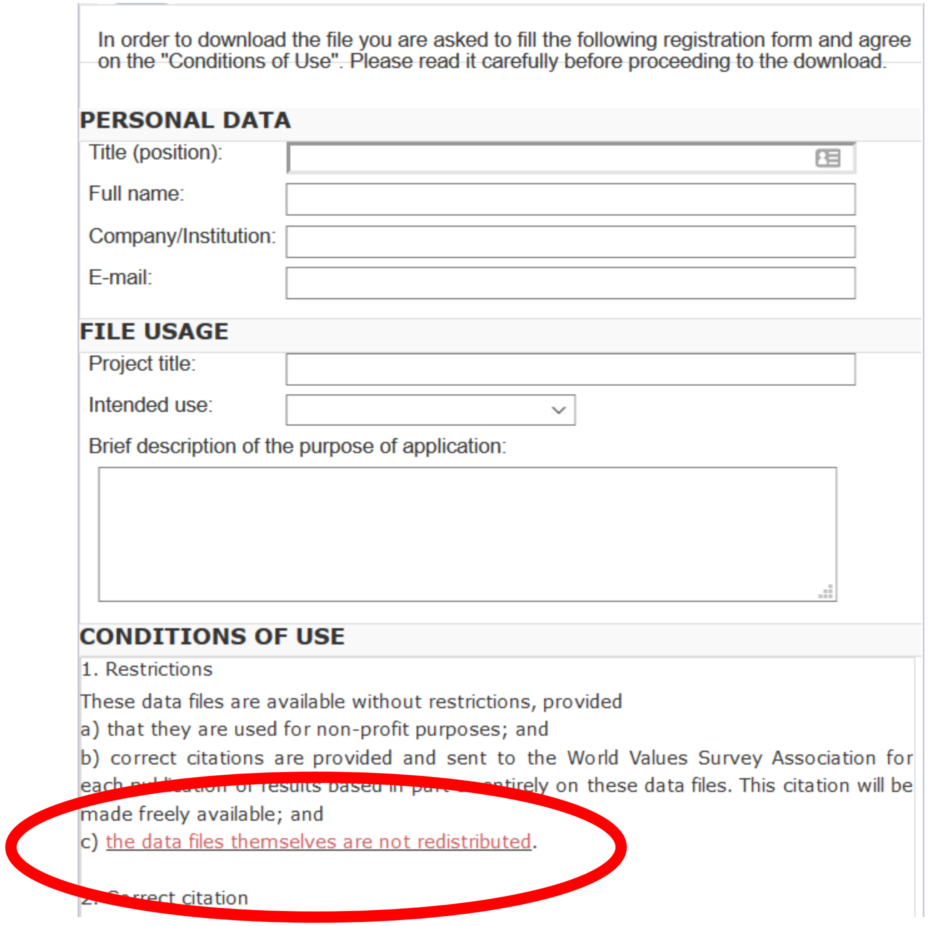
The template README is available in a variety of formats:

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

Description

The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper, including the description of the origins of data and/or description of programs. As such, a good README file should

Edit and checks



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?

Talking to Data Providers

- Federal statistical agencies
- Private data providers
- Individual research groups that produce data

The role for journals

Data citations

Journals should uniformly

monitor

and

enforce

proper data citations!

Goal: Transportability

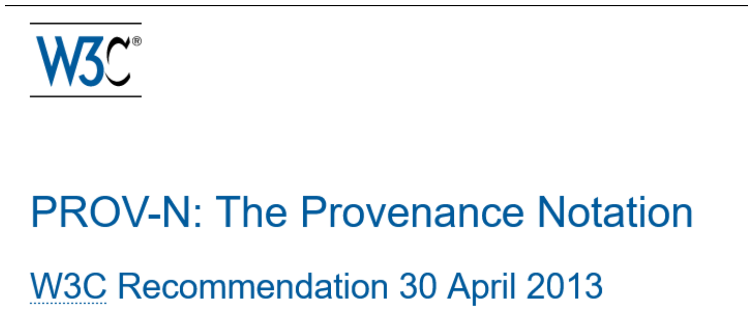
Any standards, tools, methods: must be transportable across journals (no custom solutions)

Social science “guild”

Social Science Data Editors

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

Concluding remarks



PROV-N: The Provenance Notation

W3C Recommendation 30 April 2013

```
wasDerivedFrom(e2, e1, a, g2, u1)
```

Implicitly

- Goal to formally describe
  - Machine readable!
  - PIDs
  - Machine actionable?

In practice

- Goal to formally describe
  - Machine readable!
  - PIDs
  - Machine actionable?

- Very few PIDs
- Data citations a challenge
- Click-through licenses
- Lack of licenses
- Lack of API
- Lack of persistent query extracts
- …

Thank you