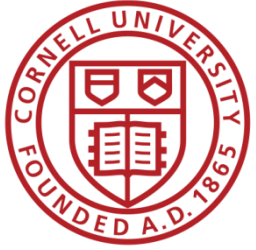


# Data Provenance: Why and How

Lars Vilhuber  
Cornell University  
2020-10-09

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



# On the menu today

- Data provenance:  
what you want is what you should get
- Data citations:  
what you got is not yours



# 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

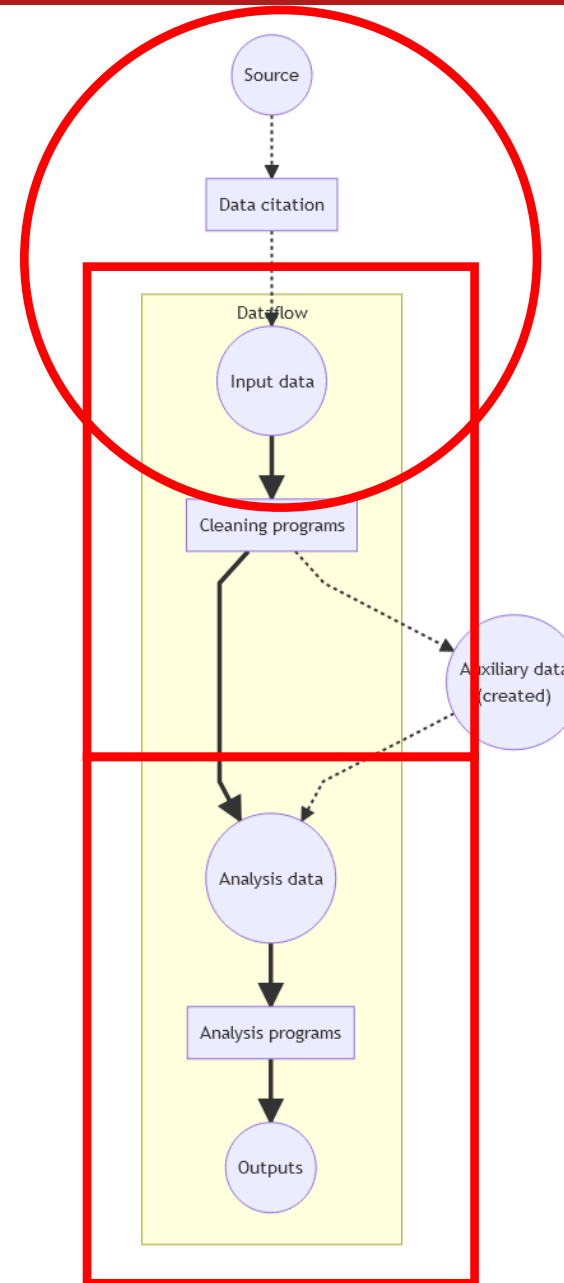
- **Pre-emptively improve code archives**
  - By conducting reproducibility checks when we can
  - By working with groups that conduct reproducibility checks when we cannot
- **Better archives**
  - Greater transparency of the code and data archives
- **Better provenance tracking**
  - Leave code where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

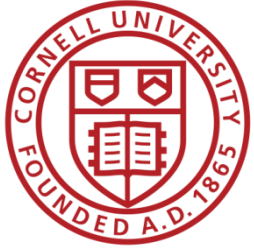
# Data provenance



# Simplified model

- Basic data-oriented flow
- Historically: analysis data
- More recently: input data, too
- But where does it come from? (“provenance”)





# Poor citation practices

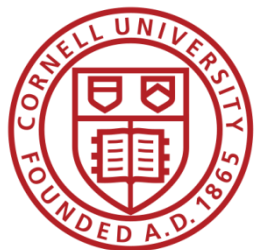
- **Macrodata:**

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

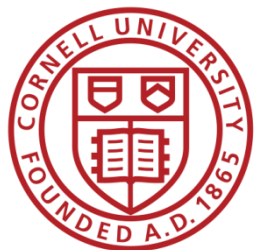
- **Microdata:**

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





An analogy...



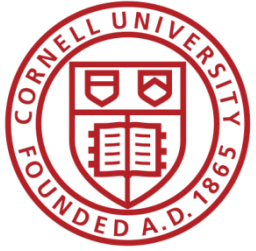
Would you buy a car from this guy?





# Provenance!

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



You don't want end up like this





Translating into the data world



# Would you use this data?

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





Or would you trust this data?





## Provenance!

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

# Metadata!





# Provenance!

- “CDC warns not to ingest Lysol.”
- “It suggests that Clorox is better.”





# Find the source!

Facebook CDC - Home | Facebook

https://www.facebook.com/CDC/

Me W B Passkey CED2AR T AEA-tools Music Admin Search Travel Info Banks DCAP Asana: J Asana: L Jira Search GDS R Table Convert Online -... SocSci Guide

facebook

Email or Phone Password Log In Forgot account?

**cdc.gov/coronavirus**

**CDC** @CDC

Home Posts Videos Photos About Comment Policy Events

Wear a mask in public spaces Stay at least 6 feet from other people Frequently wash your hands

Like Share Suggest Edits ... Learn More Send Message

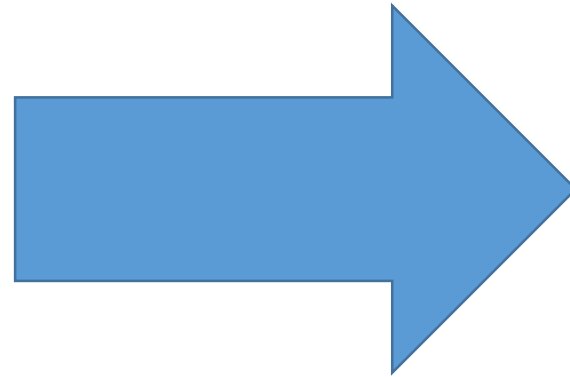
**Posts**

CDC 23 hrs



“It’s a file called stockmarket.xlsx”

```
00000000 cfd0 e011 b1a1 e11a 0000 0000 0000 0000
00000100 0000 0000 0000 0000 003e 0003 fffe 0009
00000200 0006 0000 0000 0000 0000 0000 0004 0000
00000300 008f 0000 0000 0000 1000 0000 fffe ffff
00000400 0000 0000 fffe ffff 0000 0000 008b 0000
00000500 008c 0000 008d 0000 008e 0000 ffff ffff
00000600 ffff ffff ffff ffff ffff ffff ffff ffff
*
00002000 0809 0010 0600 0005 209a 07cd c0c9 0000
00002100 0306 0000 00e1 0002 04b0 00c1 0002 0000
00002200 00e2 0000 005c 0070 0001 4c00 2020 2020
00002300 2020 2020 2020 2020 2020 2020 2020 2020
*
00002900 2020 2020 2020 2020 0042 0002 04b0 0161
00002a00 0002 0000 013d 0002 0001 009c 0002 000e
00002b00 0019 0002 0000 0012 0002 0000 0013 0002
00002c00 0000 01af 0002 0000 01bc 0002 0000 003d
00002d00 0012 0000 000f 3f1b 27f6 0038 0000 0000
00002e00 0001 0258 0040 0002 0000 008d 0002 0000
00002f00 0022 0002 0000 000e 0002 0001 01b7 0002
```



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



# “It’s a file called SP500.xlsx”

SP500

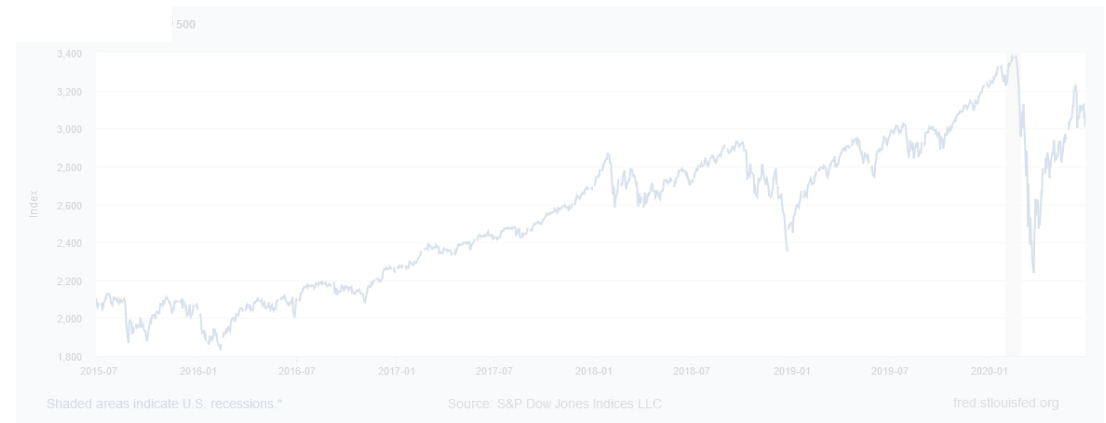
S&P 500, Index, Daily,  
Not Seasonally Adjusted

Frequency: Daily, Close

observation\_date

SP500

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





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

SP500

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

Frequency: Daily, Close

observation\_date

SP500

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



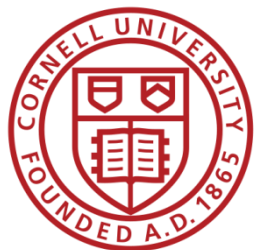


“It’s a file called SP500.xlsx, downloaded from FRED.”

|                         |   |
|-------------------------|---|
| SP500                   | S&P 500, Index, Daily,<br>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.





We're all set then.



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

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

Frequency: Daily, Close  
observation\_date

SP500

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

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







# Data Availability Statements



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

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

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

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.



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



# Data Citation



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

Attributes the file to the proper source

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

|            |         |
|------------|---------|
| 2015-07-08 | 2101.49 |
| 2015-07-09 | 2057.64 |
| 2015-07-10 | 2063.11 |
| 2015-07-13 | 2075.12 |
| 2015-07-14 | 2076.78 |
| 2015-07-15 | 2076.78 |
| 2015-07-16 | 2068.76 |
| 2015-07-17 | 2081.34 |
| 2015-07-20 | 2046.68 |
|            | 2051.31 |
|            | 2076.62 |
|            | 2099.60 |
|            | 2108.95 |
|            | 2107.40 |
|            | 2124.29 |
|            | 2126.64 |
|            | 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.





The Future of Research Communications and e-Scholarship

 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 citable research objects, such as publications[1].

*DC<sup>1</sup>*  
*Data Citation Principles*

## 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].

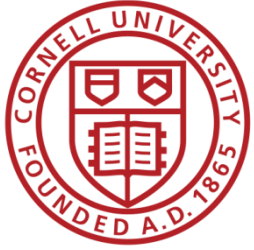
## 4. Unique Identification

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 metadata, documentation, code, and other materials as are necessary for

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

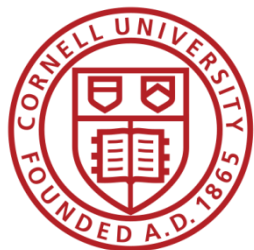


Data Citations are Hard!

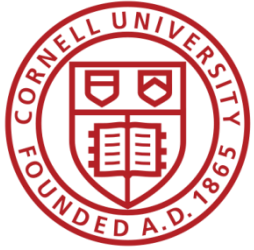


At the AEA, every manuscript is checked

- 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?



# Data Rights



# Example 2: Academic data publisher



**ECONOMIC POLICY UNCERTAINTY**

[Home](#)

[Methodology](#)

[Media](#)

[Research & Applications](#)

[About Us](#)

## EPU Indices

[All Country-Level Data](#)

|  |                             |
|--|-----------------------------|
| <a href="#">Global</a>                   | <a href="#">USA</a>         |
| <a href="#">Australia</a>                | <a href="#">Brazil</a>      |
| <a href="#">Canada</a>                   | <a href="#">Chile</a>       |
| <a href="#">China</a>                    | <a href="#">Colombia</a>    |
| <a href="#">Croatia</a> <span>New</span> | <a href="#">France</a>      |
| <a href="#">Germany</a>                  | <a href="#">Greece</a>      |
| <a href="#">Hong Kong</a>                | <a href="#">India</a>       |
| <a href="#">Ireland</a>                  | <a href="#">Italy</a>       |
| <a href="#">Japan</a>                    | <a href="#">South Korea</a> |

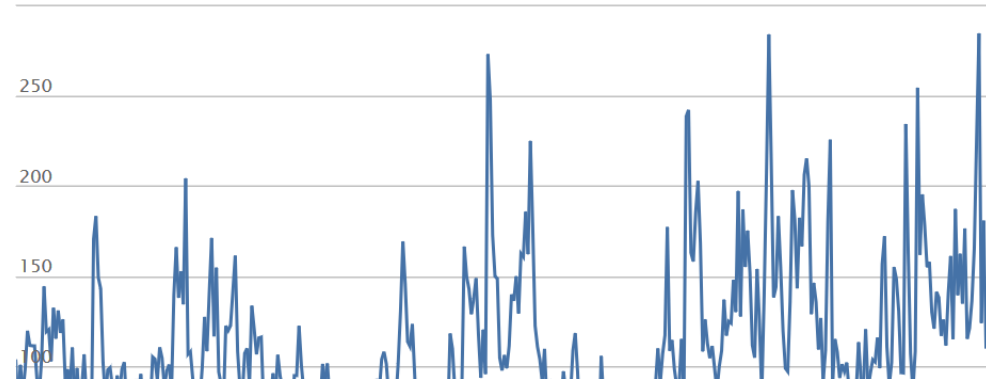
## Economic Policy Uncertainty Index

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

### Monthly US Economic Policy Uncertainty Index



Zoom [1m](#) [3m](#) [6m](#) [1y](#) [7y](#) [All](#)





# Example 2: Academic data publisher

INTERNET ARCHIVE  
WayBack Machine

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

Go

SEP DEC JAN

[103 captures](#)

18 Aug 2012 - 14 Dec 2019

2018 2019 2020



ECONOMIC POLICY UNCERTAINTY

[Home](#)

[Methodology](#)

[Media](#)

[Research & Applications](#)

[About Us](#)

## EPU Indices

[All Country-Level Data](#)

[Global](#)

[Australia](#)

[Canada](#)

[China](#)

[Croatia](#)

[Germany](#)

[Hong Kong](#)

[Ireland](#)

[Japan](#)

[France](#)

[Greece](#)

[India](#)

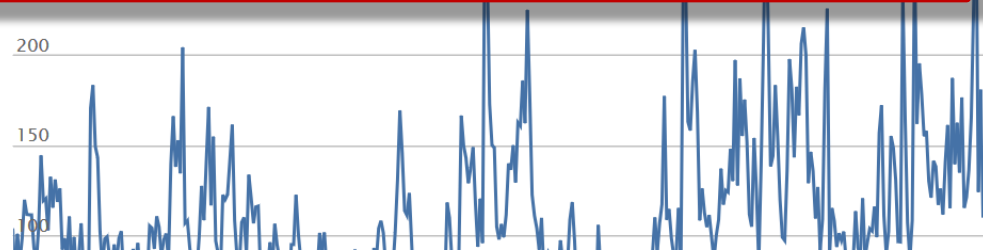
[Italy](#)

[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!





# 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.



# Rights around data

## Usage rights

- 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!)

## Redistribution rights

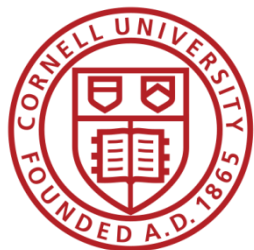
- If you created the data, you decide.
- If you got it from somewhere else:

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

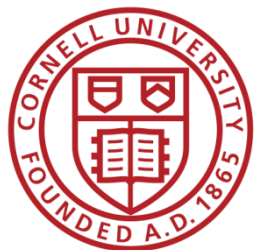


## Citing restricted-access data

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



Wrong



# These are archives



Source: <https://chiddicksfamilytree.com/2017/11/08/use-it-or-lose-it/>



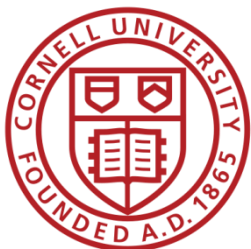
Source: <https://www.archives.gov/publications/prologue/2008/spring/frc.html>



# They are citable

- Typescript of short story Brothers and Sisters by Budge Wilson, 2000, **MS-2-650.2013-070, Box 3, Folder 9**, **Budge Wilson fonds, Dalhousie University Archives**, Halifax, Nova Scotia, Canada.





# 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) |
|------------------------------|--|---------------------------|
| <b>BHP 7518 v1 (current)</b> | <a href="#">10.5164/IAB.BHP7518.de.en.v1</a> | 2020-01-13                |
| <b>BHP 7517 v1</b>           | <a href="#">10.5164/IAB.BHP7517.de.en.v1</a> | 2018-12-12                |
| <b>BHP 7516 v1</b>           | <a href="#">10.5164/IAB.BHP7516.de.en.v1</a> | 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

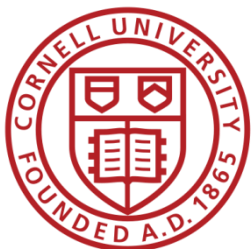
Old versions are only available for replication studies and only in justified exceptional cases for new Projects.

| 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) |
|------------------------------|--|---------------------------|
| <b>BHP 7518 v1 (current)</b> | <a href="#">10.5164/IAB.BHP7518.de.en.v1</a> | 2020-01-13                |
| <b>BHP 7517 v1</b>           | <a href="#">10.5164/IAB.BHP7517.de.en.v1</a> | 2018-12-12                |
| <b>BHP 7516 v1</b>           | <a href="#">10.5164/IAB.BHP7516.de.en.v1</a> | 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](#)



# Element of a (data) citation

- CMOS notes (16<sup>th</sup> ed., 14.1)

“Regardless of the convention being followed, **source citations** must always provide sufficient information either to **lead readers directly to the sources** consulted or, for materials that may ***not be readily available***, to enable readers to ***positively identify them, regardless of whether the sources are published or unpublished or in printed or electronic form.***”

ICPSR notes that a citation should include the following items:

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



# 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.



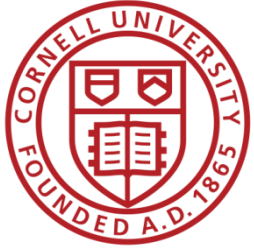
# Element of a (data) citation

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.



# 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](http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp).

Madrid: JD Systems Institute.

- Is that in the README / Paper/ Appendix?





And we

In order to download the file you are as  
on the "Conditions of Use". Please read

#### PERSONAL DATA

Title (position):

Full name:

Company/Institution:

E-mail:

#### FILE USAGE

Project title:

Intended use:

Brief description of the purpose of applic

#### 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

#### 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

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



## Data Availability Statements (DAS)

- A statement about **where data** supporting the results reported in a published article can be

to publicly  
ated during

y providing a

I restrictions,

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



## Data Availa

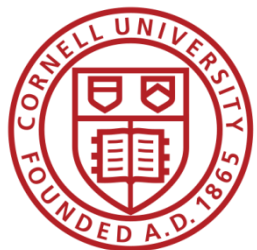
- A statement about **available**
  - DOI assi
  - But long
- A statement about **usage rights**
  - Not every dataset is in the public domain
  - Not everybody knows that U.S. Government data are usually in the public domain

# 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!



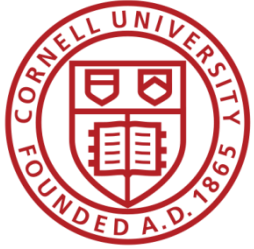
# Social science “guild”



[https://  
social-science  
-data-editors.  
github.io/  
guidance/](https://social-science-data-editors.github.io/guidance/)

# Thank you!

<https://doi.org/10.5281/zenodo.4073995>



# 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](https://doi.org/10.5164/IAB.BHP7516.de.en.v1)

