Data Provenance

Documenting how you got to the data

- Access can be **clearly and precisely documented**
- Is **non-exclusive to the authors**
- Intermediate files preserved

(example taken from Fort, Restud 2016)

- NOTE: for AEA, you are required to provide all programs, but a copy may/should be available within the FSRDC as well.

To reproduce the tables and figures in the paper:

1. *All the results in the paper use confidential microdata from the U.S. Census Bureau. To gain access to the Census microdata, follow the directions here on how to write a proposal for access to the data via a Federal Statistical Research Data Center: https://www.census.gov /ces/rdcresearch/howtoapply.html.*
2. *You must request the following datasets in your proposal:*
   - *Longitudinal Business Database (LBD), 2002 and 2007*
   - *Foreign Trade Database – Import (IMP), 2002 and 2007*
   - *Annual Survey of Manufactures (ASM), including the Computer Network Use Supplement (CNUS), 1999*
   - *[...]*
   - *Annual Survey of Magical Inputs (ASMI), 2002 and 2007*
3. *Reference "Technology and Production Fragmentation: Domestic [illegible] proposal. This will give you access to the programs and input datasets required to reproduce the results. Requesting a search of archives with the articles DOI ("10.1093/restud/rdw057") should yield the same results.*

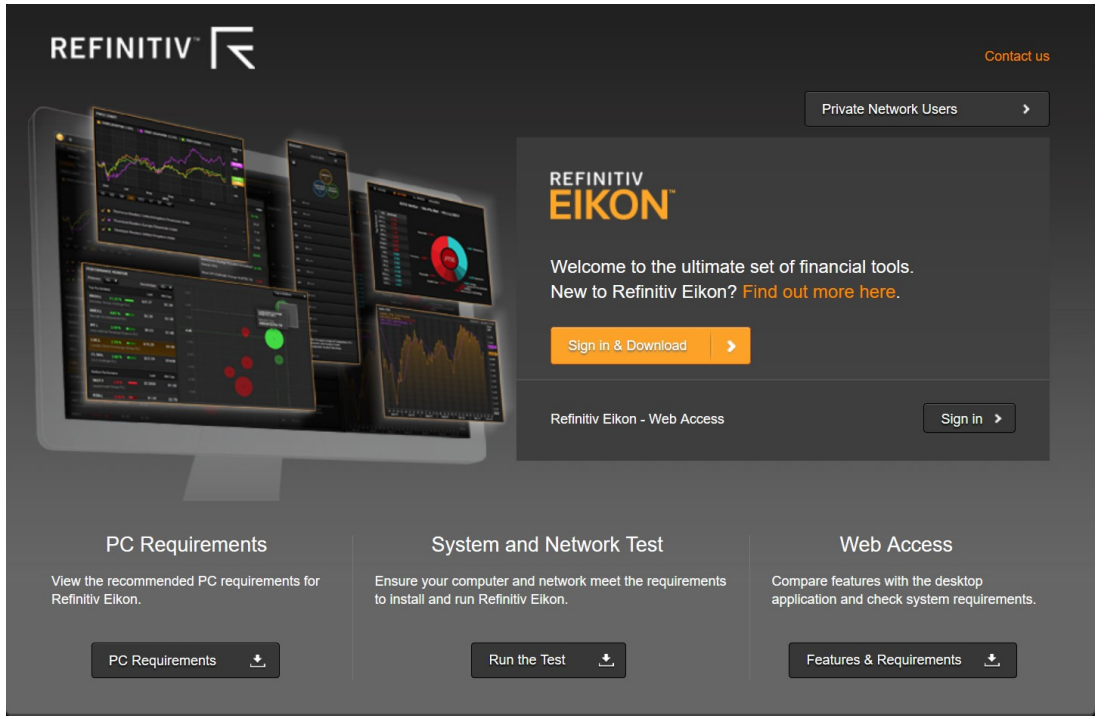
NOTE: Project-related files are available for 10 years as of 2015.

https://social-science-data-editors.github.io/guidance/DCAS_Restricted_data.html#us-census-bureau-and-fsrdc

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

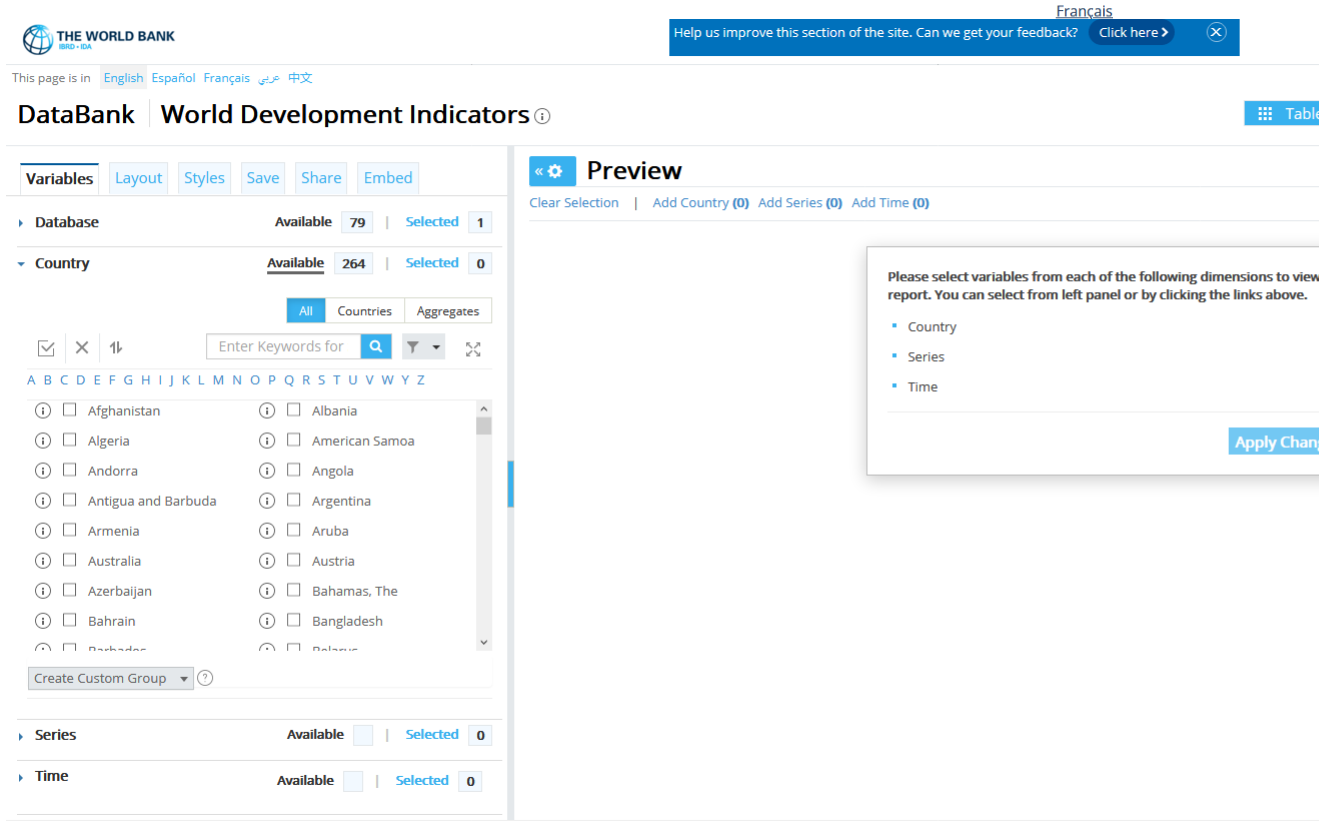
...



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?

- 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!

Example: Danish administrative data

- Access can be **clearly and precisely documented**
- Is **non-exclusive to the authors**

(example taken from Fadlon and Nielsen, AEJ:Applied 2021)

The information used in the analysis combines several Danish administrative registers (as described in the paper). The data use is subject to the European Union's General Data Protection Regulation(GDPR) per new Danish regulations from May 2018. The data are physically stored on computers at Statistics Denmark and, due to security considerations, the data may not be transferred to computers outside Statistics Denmark. Researchers interested in obtaining access to the register data employed in this paper are required to submit a written application to gain approval from Statistics Denmark. The application must include a detailed description of the proposed project, its purpose, and its social contribution, as well as a description of the required datasets, variables, and analysis population. Applications can be submitted by researchers who are affiliated with Danish institutions accepted by Statistics Denmark, or by researchers outside of Denmark who collaborate with researchers affiliated with these institutions.

Health Data. To identify fatal and severe non-fatal health events we use two complementary datasets. Our first dataset is the *Death Registry* (Statistics Denmark 2020b), which includes deceased individuals' date of death. Our second dataset is the *National Patient Registry*

ization r

Statistics Denmark (2020a). Befolkningen (BEF, Population Demographics, 1985-2011 [database]. Danmarks Statistiks Forskningsservice, accessed 2014.

Statistics Denmark (2020b). Dœde i Danmark (DOD, Deaths in Denmark, 1980-2013 [database]. Danmarks Statistiks Forskningsservice, accessed 2014.

Statistics Denmark (2020c). Hustande og familier (FAIN, Households and Families, 1980-2007 [database]. Danmarks Statistiks Forskningsservice, accessed 2014.

https://social-science-data-editors.github.io/guidance/Requested_information_dcas.html#example-for-government-registers

http://www.dst.dk/extranet/forskningvariabellister/Oversigt%20over%20registre.html

Example 4: German Restricted-access

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

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

Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13
BHP 7517 v1	10.5164/IAB.BHP7517.de.en.v1	2018-12-12
BHP 7516 v1	10.5164/IAB.BHP7516.de.en.v1	2018-04-11

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

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

Data Versions

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

Data Version	DOI (Link to Description of Data Version)	Availability (yyyy-mm-dd)
BHP 7518 v1 (current)	10.5164/IAB.BHP7518.de.en.v1	2020-01-13

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

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.

Exercise 4

Exercise 4: Constructing data citations

- Read the Data Citation Guidance at https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html

- Consider the following description:
  "We conducted our experiment with the tax authority of Uruguay. There are 120,123 firms registered in the agency's database, whom we contacted by mail."

Element of a (data) citation

Construct a data citation using the key:

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

Constructed Citation:

Exercise 4: Constructing data citations

- Add this data citation to the README in your repository, and commit to your Github

- Now do the SAME for the data used in the repository

Thank you