ILR LDI

#CornellDayofData

DAY OF DATA 2021

Scholarship through Collaboration

Day of Data

Working with Restricted Access / Big Data

Lars Vilhuber and David Wasser

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.

You may, however, find these opinions quite useful.

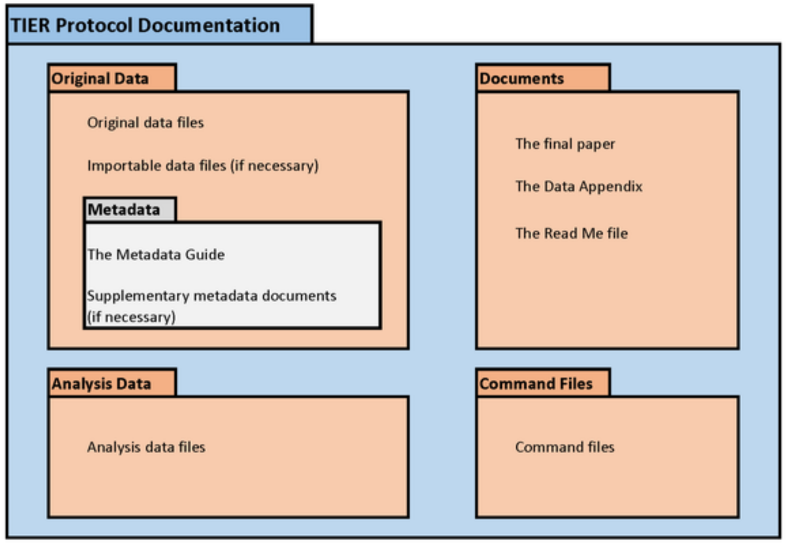
Plan for the workshop

- Generic best practice (2min) [LARS]
- Why that applies to restricted and big data, why is that reproducibility [LARS, DAVID]
- Example: Danish data (restricted) [DAVID]
- Example: FSRDC (restricted + big) [LARS]
- Example: “Large data” (API, “big”) [LARS]

Generic best practice

Basic project setup

- Structure your project
  - Data inputs
  - Data outputs
  - Code
  - Paper/text/etc.
- Version your project (git)
- Track metadata
  - Cite articles you reference
  - Cite data sources you use



TIER Protocol Documentation

Original Data

Original data files

Importable data files (if necessary)

Metadata

The Metadata Guide

Supplementary metadata documents (if necessary)

Documents

The final paper

The Data Appendix

The Read Me file

Analysis Data

Analysis data files

Command Files

Command files

https://www.projecttier.org/tier-protocol/specifications-3-0/

Computational empathy

Computational empathy

- Consider how the next person will (be able to) compute
  - You don’t know what they don’t know
  - Assume some frequent characteristics
    - Empirical background (2-3 yrs undergrad?)
    - Likely to know about frequently used software, but not very specific software
    - Have **none** of your add-on packages/ libraries/ etc. pre-installed
- Don’t force them to do tedious things

More

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

Have “**computational empathy**”

- Consider cross-platform programming 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;
      ```

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

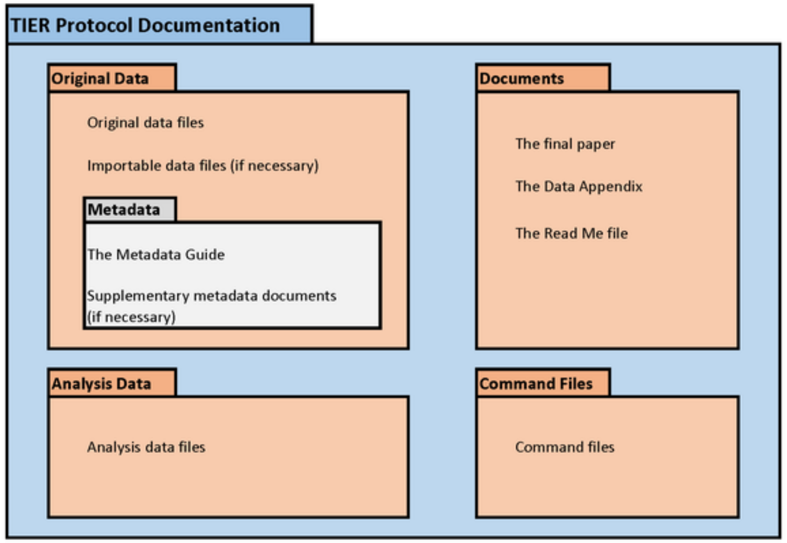
Collaboration when data is restricted or big

What is different?

- Can’t “bring” data to you – have to go to data
- Less control over environment, data stability
- No access to common “open” tools (favorite editor, Stata/R packages)
- Time cost

Basic project setup

- Structure your project
  - Data inputs
  - Data outputs
  - Code
  - Paper/text/etc.
- Version your project (git)
- Track metadata
  - Cite articles you reference
  - Cite data sources you use



TIER Protocol Documentation

Original Data

Original data files

Importable data files (if necessary)

Metadata

The Metadata Guide

Supplementary metadata documents (if necessary)

Documents

The final paper

The Data Appendix

The Read Me file

Analysis Data

Analysis data files

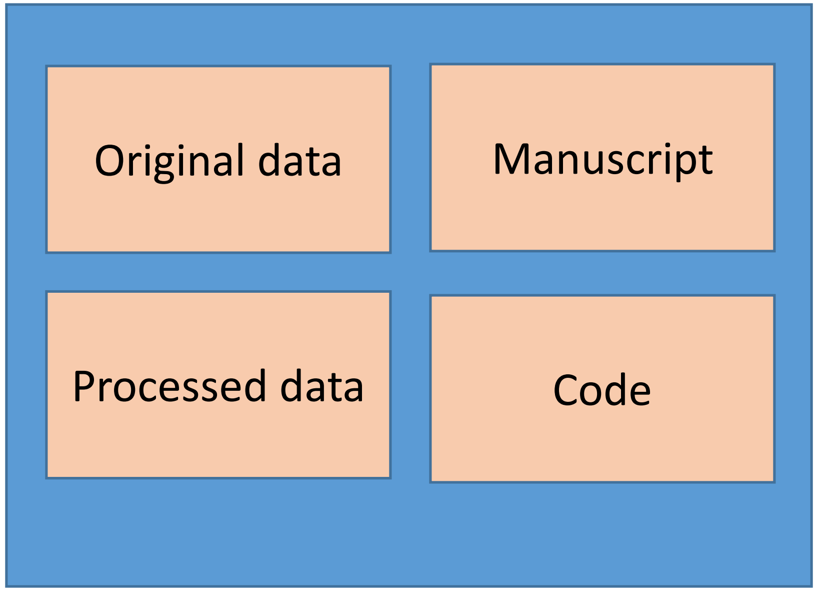
Command Files

Command files

https://www.projecttier.org/tier-protocol/specifications-3-0/

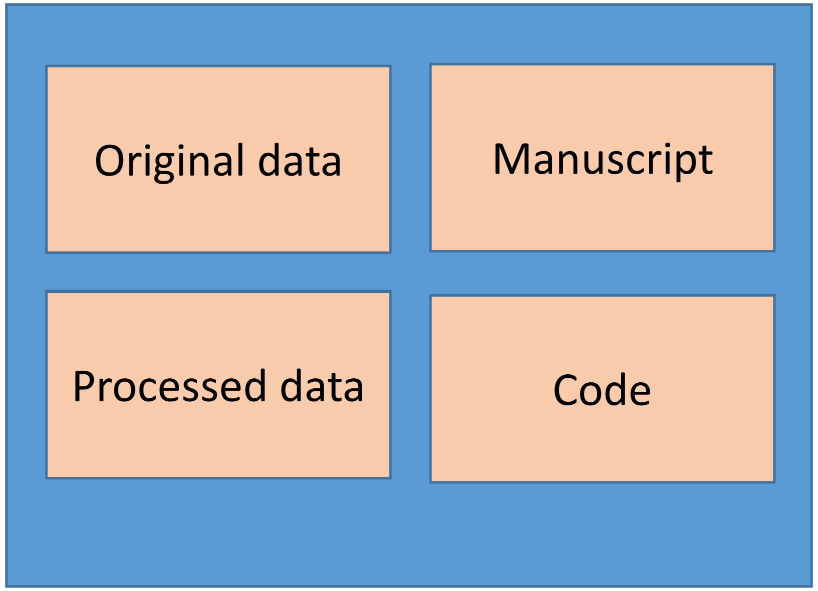
Basic project setup

- Structure your project
  - Data inputs
  - Data outputs
  - Code
  - Paper/text/etc.
- Version your project (git)
- Track metadata
  - Cite articles you reference
  - Cite data sources you use



https://www.projecttier.org/tier-protocol/specifications-3-0/

Basic project setup



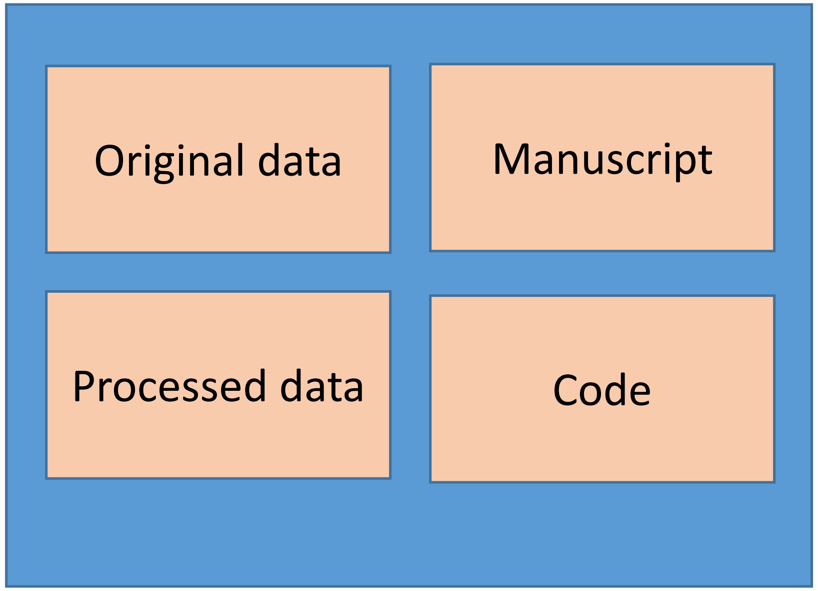
Original data

Manuscript

Processed data

Code

Basic project setup



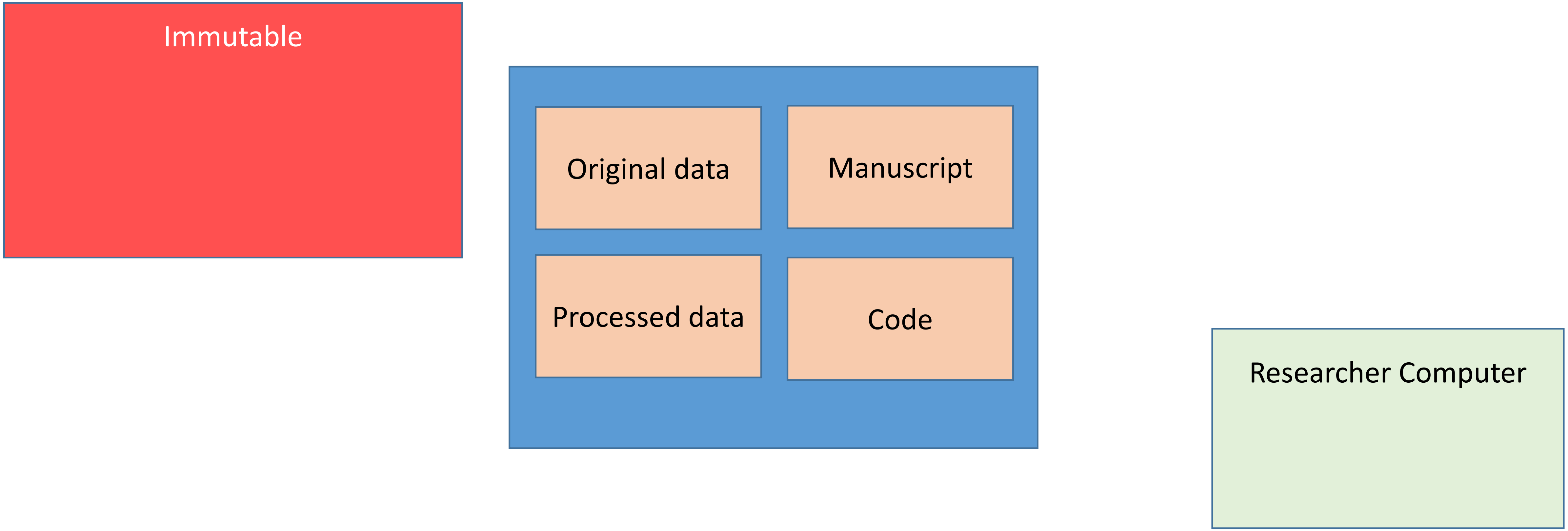
Original data

Manuscript

Processed data

Code

Basic project setup: restricted data



Immutable

Original data

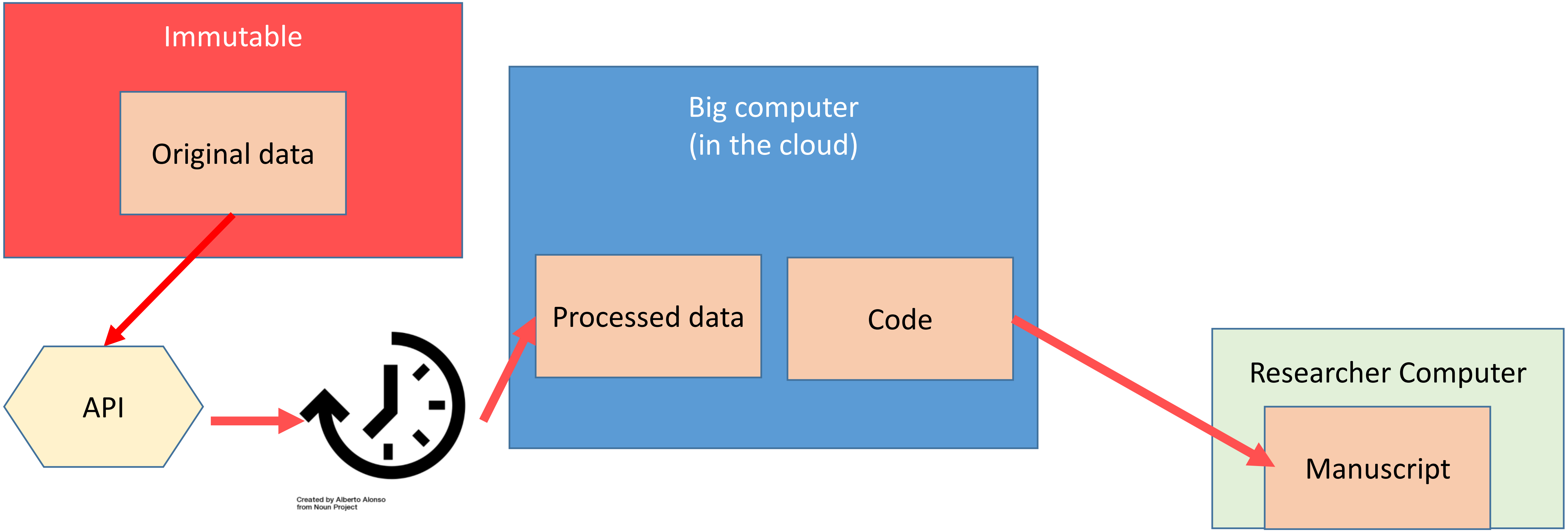
Manuscript

Processed data

Code

Researcher Computer

Basic project setup: big data



Created by Alberto Alonso from Noun Project

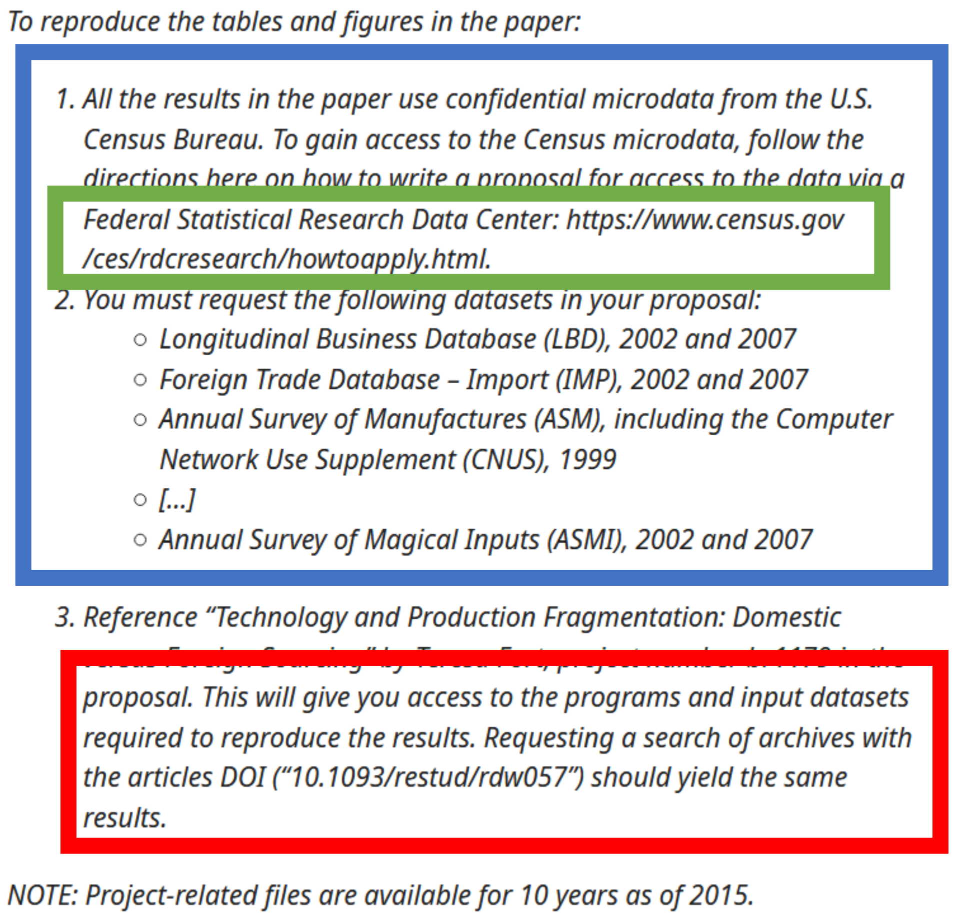
Data Provenance

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 versus Foreign Sourcing" by Teresa Fort, project number br1170 in the 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.*

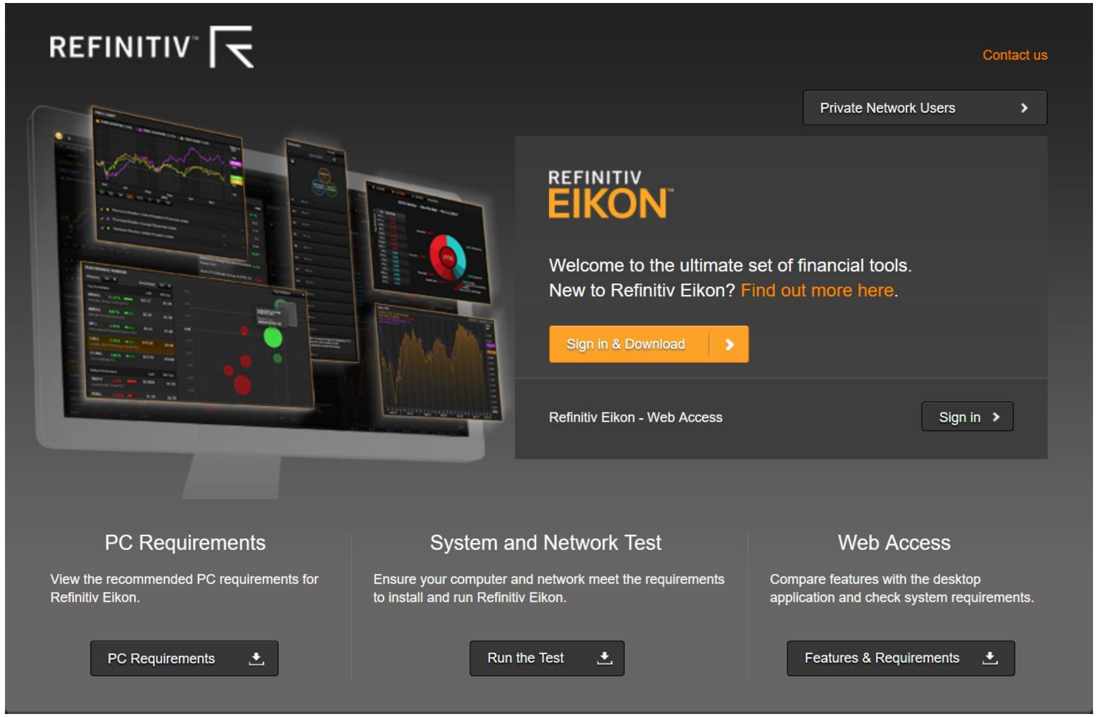
NOTE: Project-related files are available for 10 years as of 2015.

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

…



Welcome to the ultimate set of financial tools.
New to Refinitiv Eikon? Find out more here.

PC Requirements

View the recommended PC requirements for Refinitiv Eikon.

System and Network Test

Ensure your computer and network meet the requirements to install and run Refinitiv Eikon.

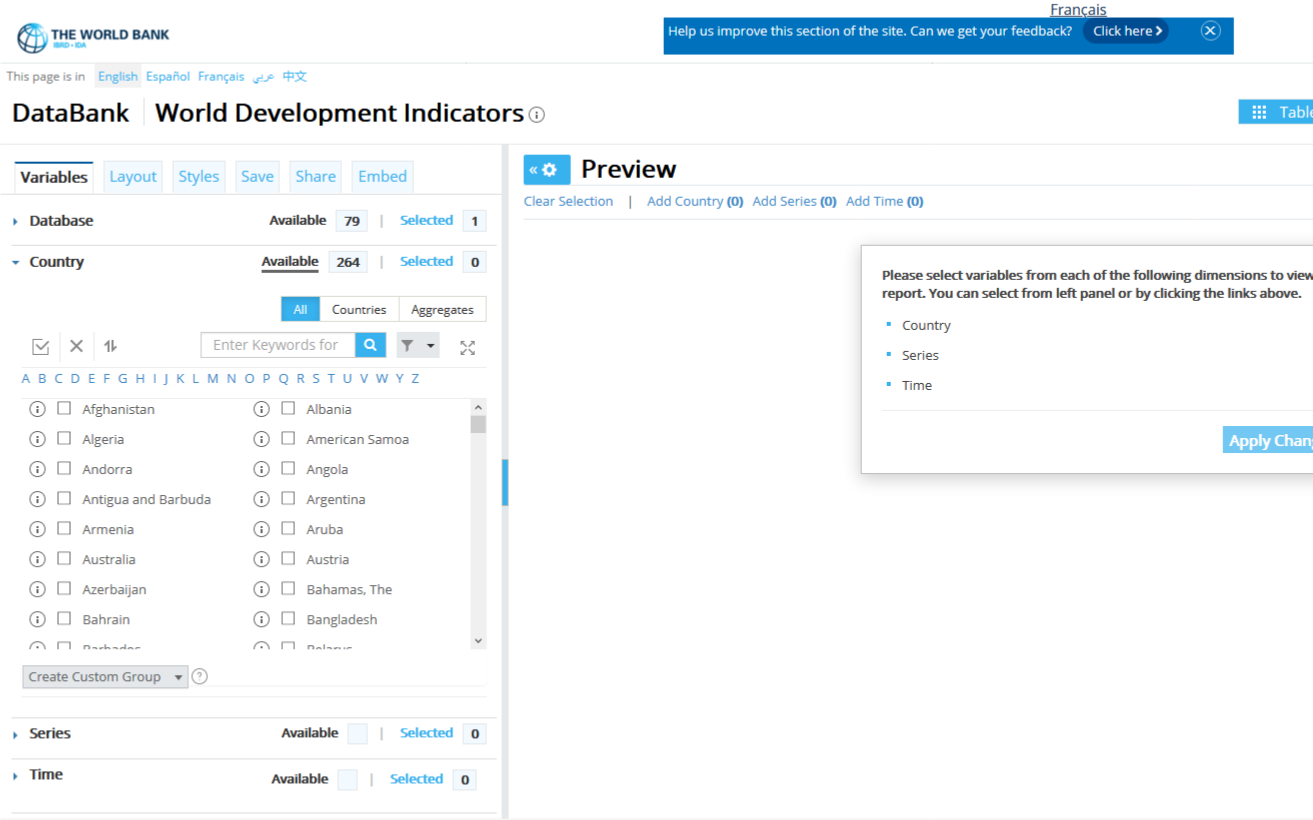
Web Access

Compare features with the desktop application and check system requirements.

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.)

…


Français
THE WORLD BANK
Click here ›
This page is in English Español Français عربي 中文
DataBank
World Development Indicators
Tabl
Variables
Layout
Styles
Save
Share
Embed
Preview
Clear Selection | Add Country (0) Add Series (0) Add Time (0)
All
Countries
Aggregates
Enter Keywords for
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Create Custom Group
Apply Chan

Help us improve this section of the site. Can we get your feedback?

	Available		Selected
Database	79	|	1
Country	264	|	0
Series		|	0
Time		|	0

☐ Afghanistan ☐ Albania
☐ Algeria ☐ American Samoa
☐ Andorra ☐ Angola
☐ Antigua and Barbuda ☐ Argentina
☐ Armenia ☐ Aruba
☐ Australia ☐ Austria
☐ Azerbaijan ☐ Bahamas, The
☐ Bahrain ☐ Bangladesh

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

- Country
- Series
- Time

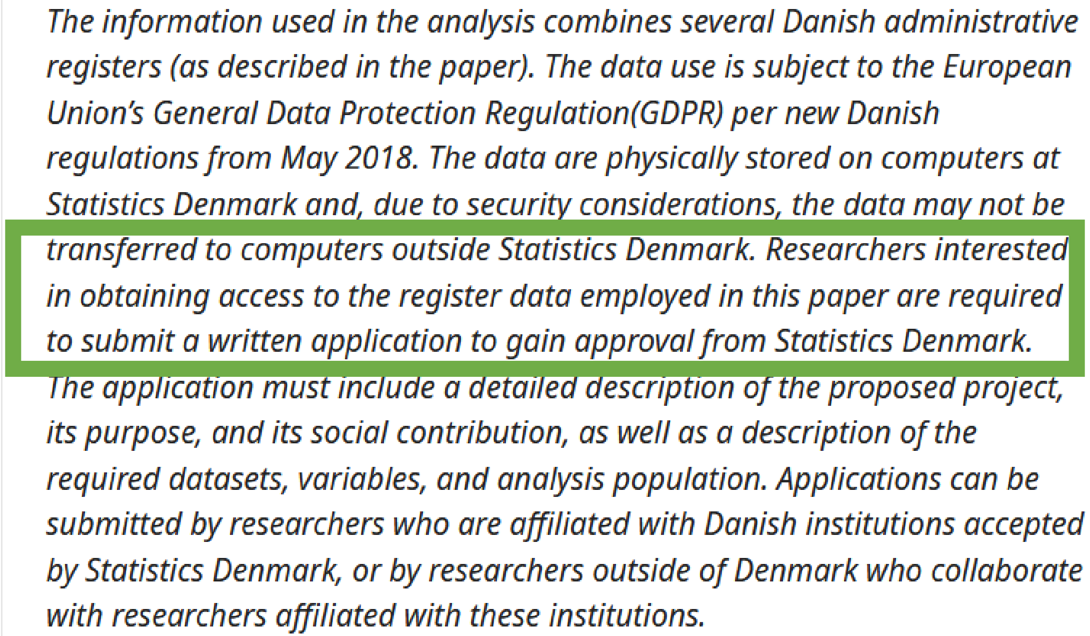
How do you document data provenance?

- 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!

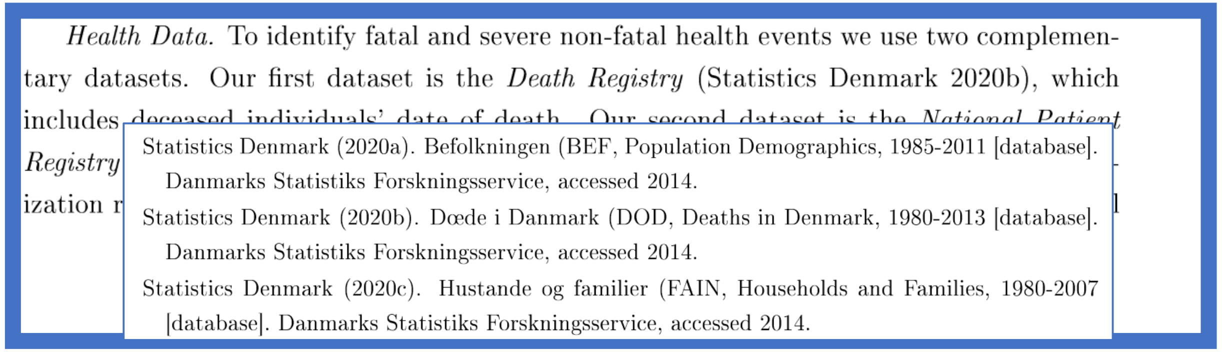
Example: Danish administrative data

- 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.

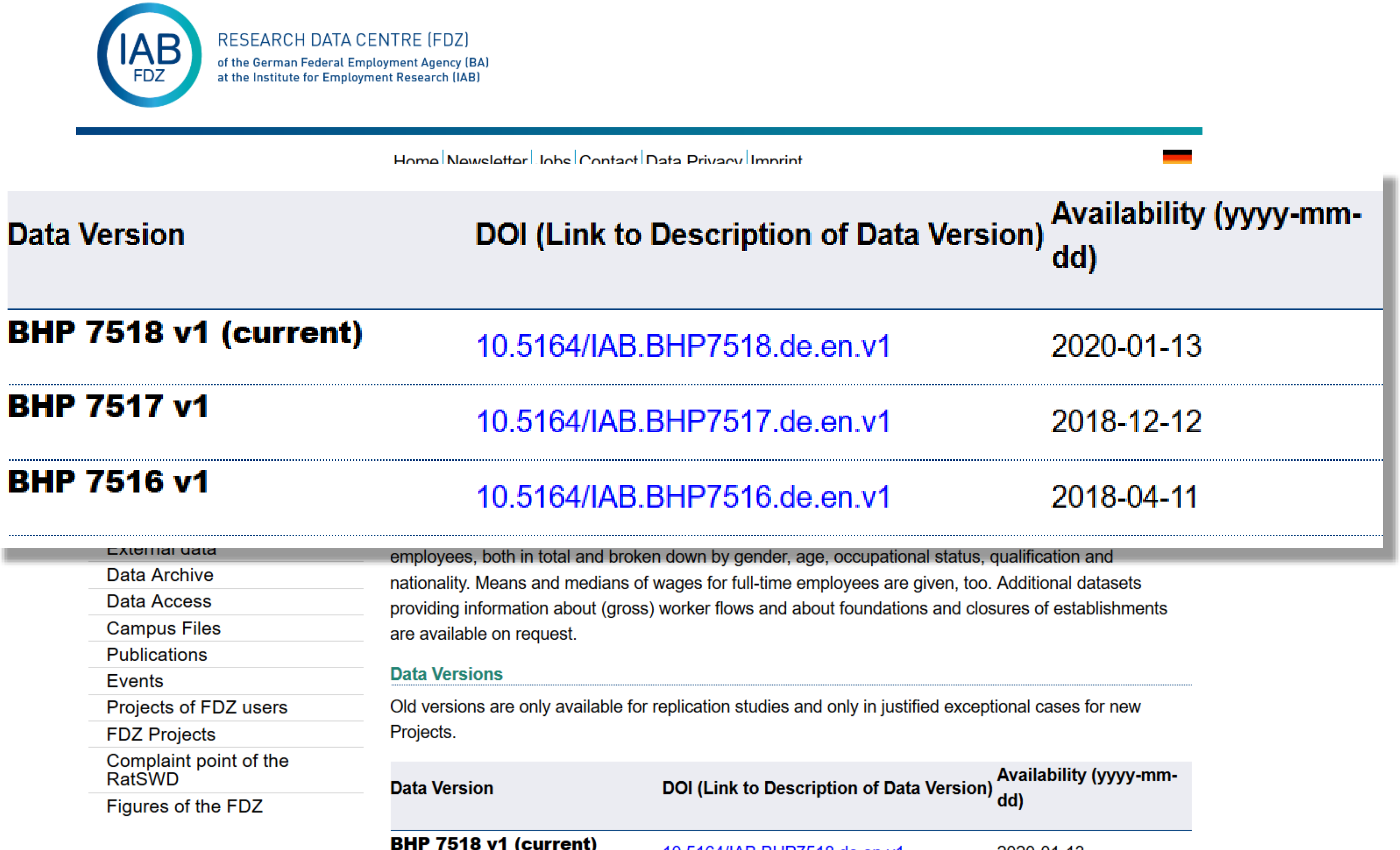
Statistics Denmark (2020b). Døde i Danmark (DOD, Deaths in Denmark, 1980-2013 [database]. Danmarks Statistiks Forskningsservice, accessed 2014.

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)

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

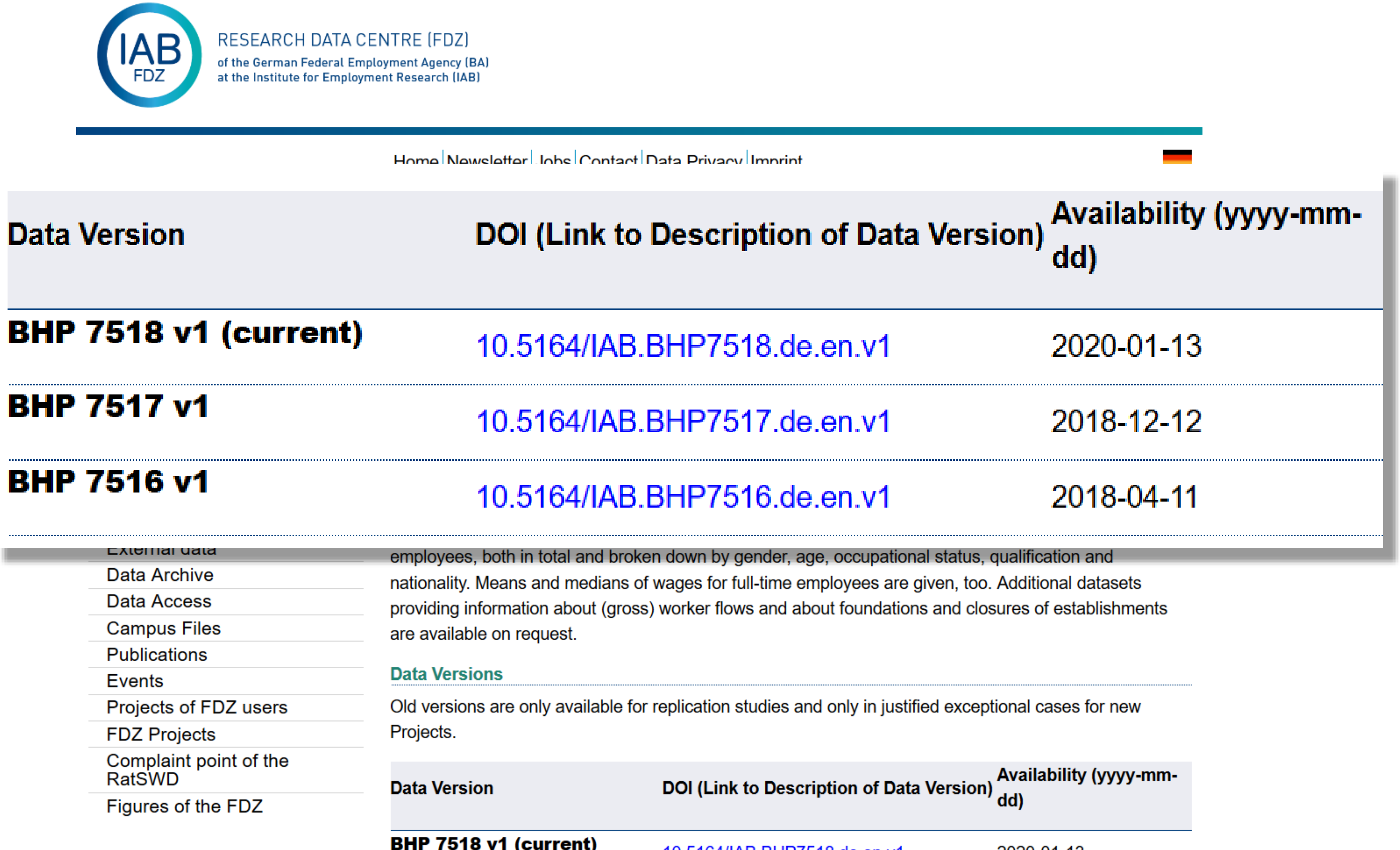
employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are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

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

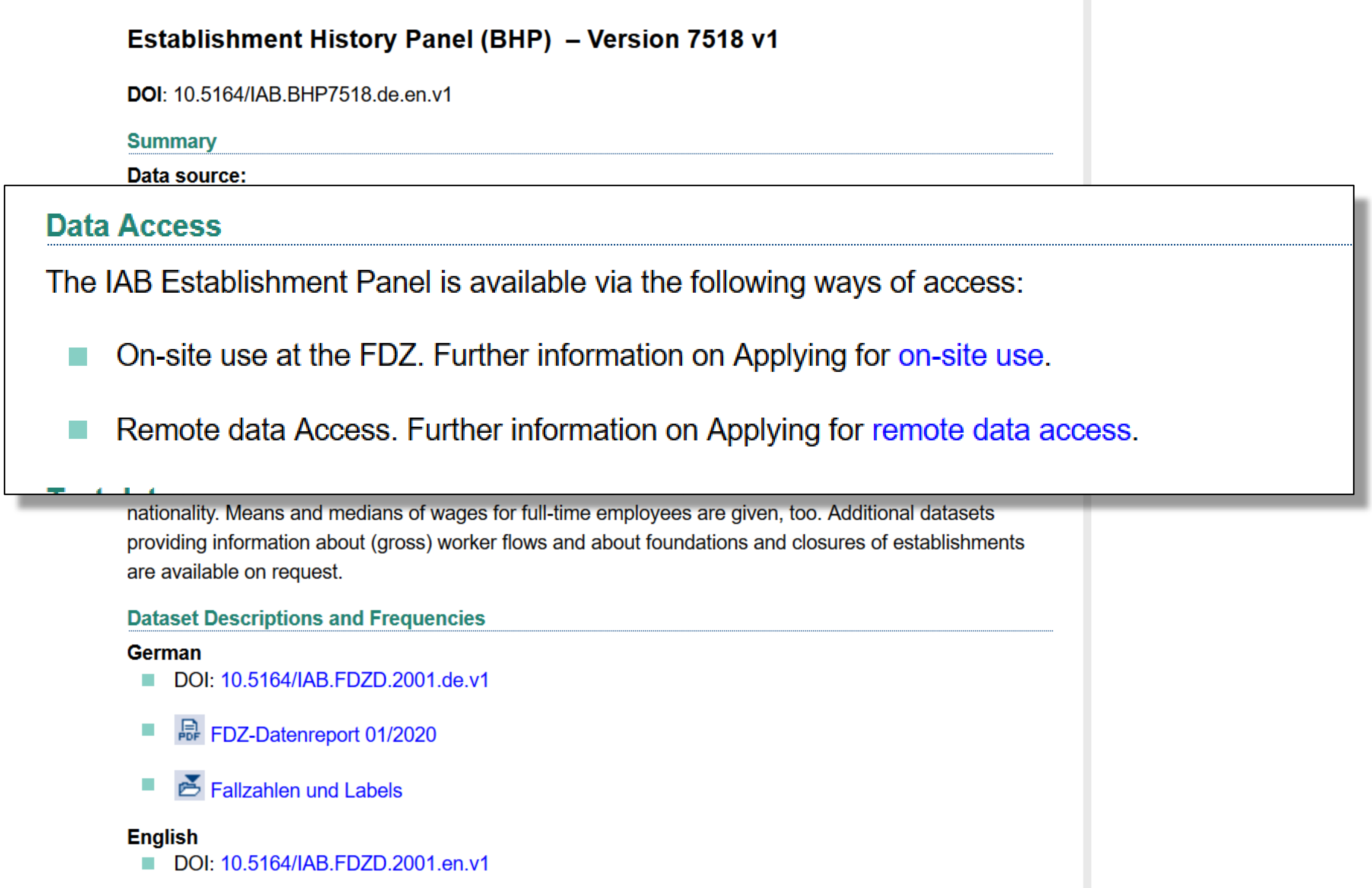
employees, both in total and broken down by gender, age, occupational status, qualification and nationality. Means and medians of wages for full-time employees are given, too. Additional datasets providing information about (gross) worker flows and about foundations and closures of establishments are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

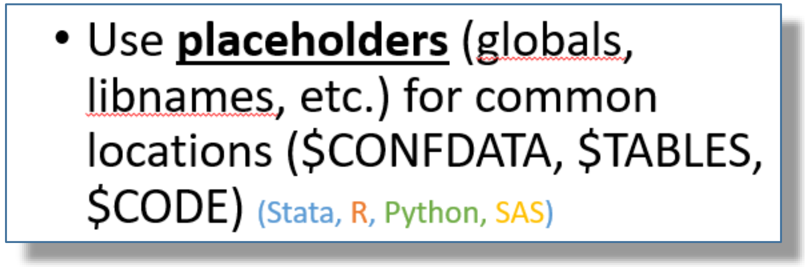
- DOI: 10.5164/IAB.FDZD.2001.en.v1

Coding for Reproducibility

Some tips from the “frequently gotten wrong” bin (restricted-access version)

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

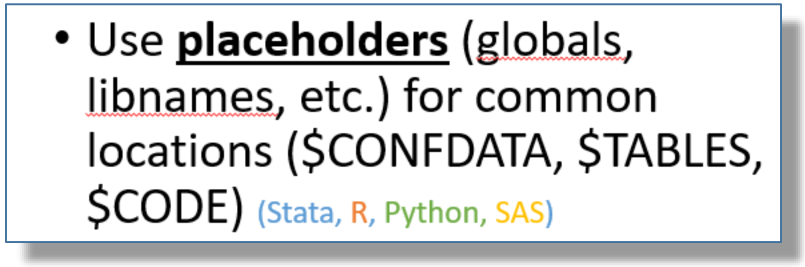


- Use **placeholders** (globals, libnames, etc.) for common locations ($CONFDATA, $TABLES, $CODE) (Stata, R, Python, SAS)

Some tips from the “frequently gotten wrong” bin (big data version)

Cleanly **separate** and **preserve**

- Data acquisition code
  - Or instructions, needs to be re-executable
- Confidential parameters and the rest of the code
  - Reduces need to redact programs
  - API keys and the like
- Intermediate data extracts
  - When its impossible to exactly re-extract data
  - When data extract takes a long time



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

Example: Danish Restricted Data

What happens in the wild?

- **Danish admin data**
  - Raw extracts in project folder
  - Code and extracts live on server
  - Paper lives off server
- **Server not connected to the internet**
  - Administrators install packages
  - Packages are part of your code; newer versions could break older code
- **What does collaboration look like?**

Dynamic Collaboration

- **I was added to a project in the middle of development**
  - No overlap with person previously responsible for code and data
  - They also could no longer access the server
  - This is collaboration, just not at the same time
- **Problem: inflation adjustment was hardcoded, but we needed more years**
  - Source was not documented
  - I was able to find the published numbers online (off server)
  - Updated the hardcoding, but also added URL and short description
  - Not a perfect solution: URL could change in the future!

Solutions for dynamic collaboration

- **Document how code pieces work together from the start**
  - Ideally this follows from how the code is structured
  - Use a README from the start
- **Document package versions**
- **Dynamic collaboration might be with your future self!**

Example: FSRDC (LEHD Data)

What happens in the wild?

- **LEHD data**
  - Raw data in **shared** location
  - Code and extracts live on server
  - Paper lives off server
- **Server not connected to the internet**
  - Administrators install packages
- **Big data: LEHD (**1990-2015**) is**
  - Obs: 4.4 billion
  - Persons: 265 million
  - Firms: 23 million
  - Jobs: 1.7 Billion

This does not fit into Stata on your laptop

Dynamic Collaboration

- **You are working with another person**
  - It takes days to extract the code
  - Need to coordinate on space
  - You may be limited on space
- **You cannot share the data outside of your space**
  - Solution: Share a location
  - Solution: Keep track of which programs have been run, and how long they took
    - Keep log files
    - Program in checks for intermediate files – do not reprocess if present

Checking for intermediate files

R:

```
if ( ! file.exists(file.path(intermediate,”step1.Rdata”)) {
    prepare_file(in=“step0”,out=“step1”,outpath=intermediate)
} else {
   message(“File exists, skipping processing”)
}
```

Checking for intermediate files

Stata:

```
capture confirm file “$intermediate/step1.dta"
  if _rc!=0 {
    process_file step0 step2 $intermediate
}  else {
    display " File exists, skipping processing "
  }
```

Safe Collaboration

- **Census rules discourage frequent removal of results and code**
  - Earlier released results might impede later release of results
  - Code needs to be vetted every time for confidentiality
- **Solution: safe programming**
  - Use samples for early release, late releases that are different (no overlap)
  - Include code that demonstrates safe release
  - Use placeholders for confidential values (for instance, minimum cell size)
  - Store them in external files

Effective Programming

- **Datasets are huge**
  - Bugs in later code may show up many days/weeks later
  - You cannot run code interactively
- **Solution: build in tests**
  - Use samples for debugging
  - Be aware of what you are sampling: *people, jobs, businesses, counties, places, time periods*!
  - Include information about the sampling in output and naming of files

Safe and effective programming

Stata:

```
include “$confcode/confidential_config.do”
// contains parameters that should not be released
use “$confdata/step1.dta”, clear
if “$prelim” == “yes” {
    keep if sample_id == $confsampleid
}
process_stuff
gen flagkeep  ( cellsize > $mincell )
save “$intermediate/step2-conf.dta”, replace // will not be released!
drop if flagkeep==0
label data “Created $rundate - releasable”
save “$outputs/step2-releasable.dta”, replace
```

Solutions for dynamic collaboration

- **Document how code pieces work together from the start**
  - Ideally this follows from how the code is structured
  - Use a README from the start
- **Document package versions**
- **Document critical steps of the process and include automation**
  - Once you are ready to release final product, you may need to run through the whole code again!

Example: Big data via extract

API + confidentiality

- https://labordynamicsinstitute.github.io/day-of-data-2021/safe-and-efficient.html

Thank you