Development Research in Practice

The DIME Analytics Data Handbook

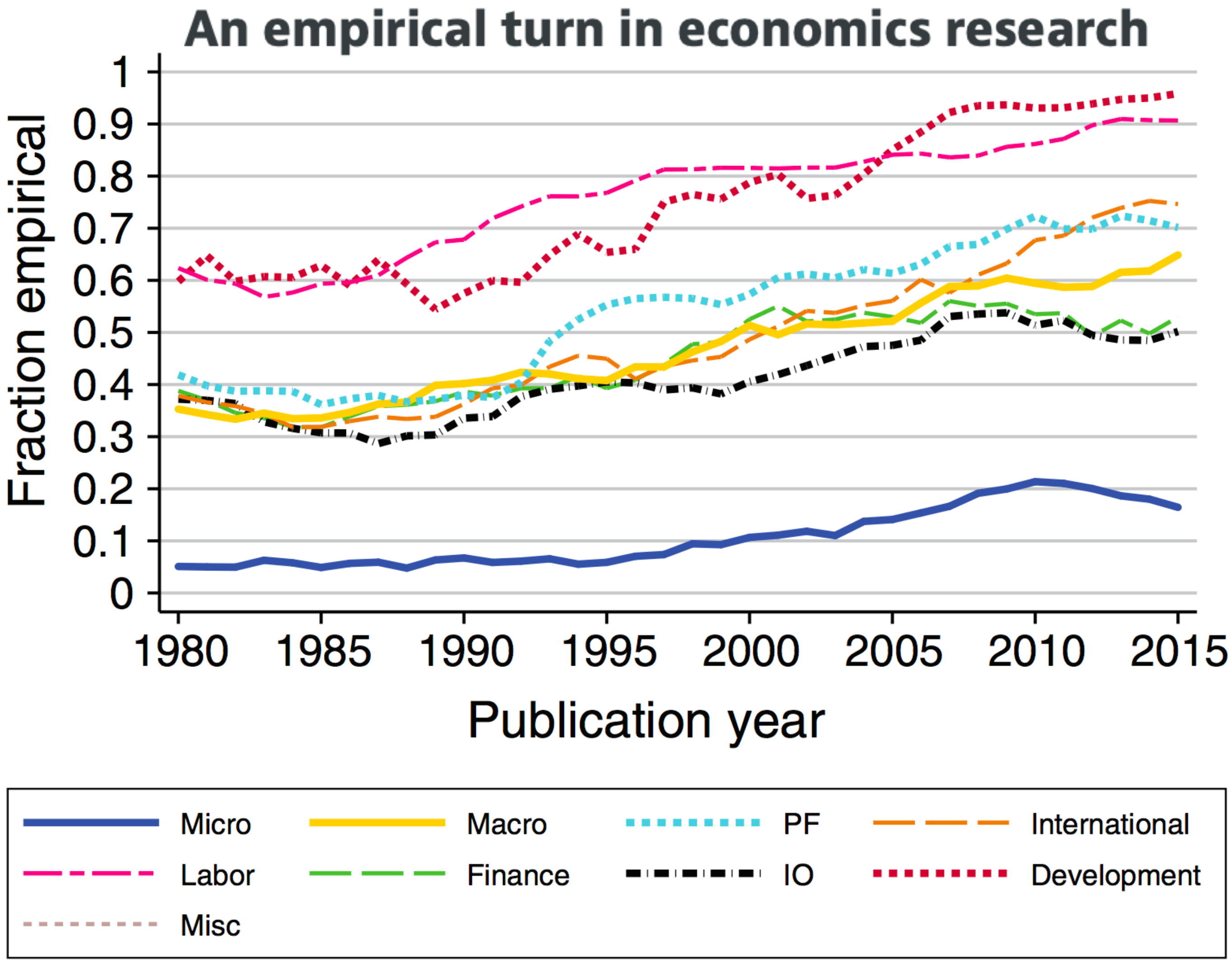
Kristoffer Bjärkefur
Luíza Cardoso de Andrade
Benjamin Daniels
Maria Ruth Jones

WORLD BANK GROUP
DIME analytics
i2i DIME TRANSFORM DEVELOPMENT
UKaid from the British people

Motivation (1)

- Across economics, massive turn towards empirical research
- In development, special focus on “unique” datasets collected by authors

FEATURED CHART
JUNE 26, 2017



https://www.aeaweb.org/research/charts/an-empirical-turn-in-economics-research

AMERICAN ECONOMIC ASSOCIATION

Motivation (2)

- Increasing co-authorship and collaboration
- Development of “laboratory” style organizations with PIs supported by professional staff

Figure 1 Trends in co-authorship in economics

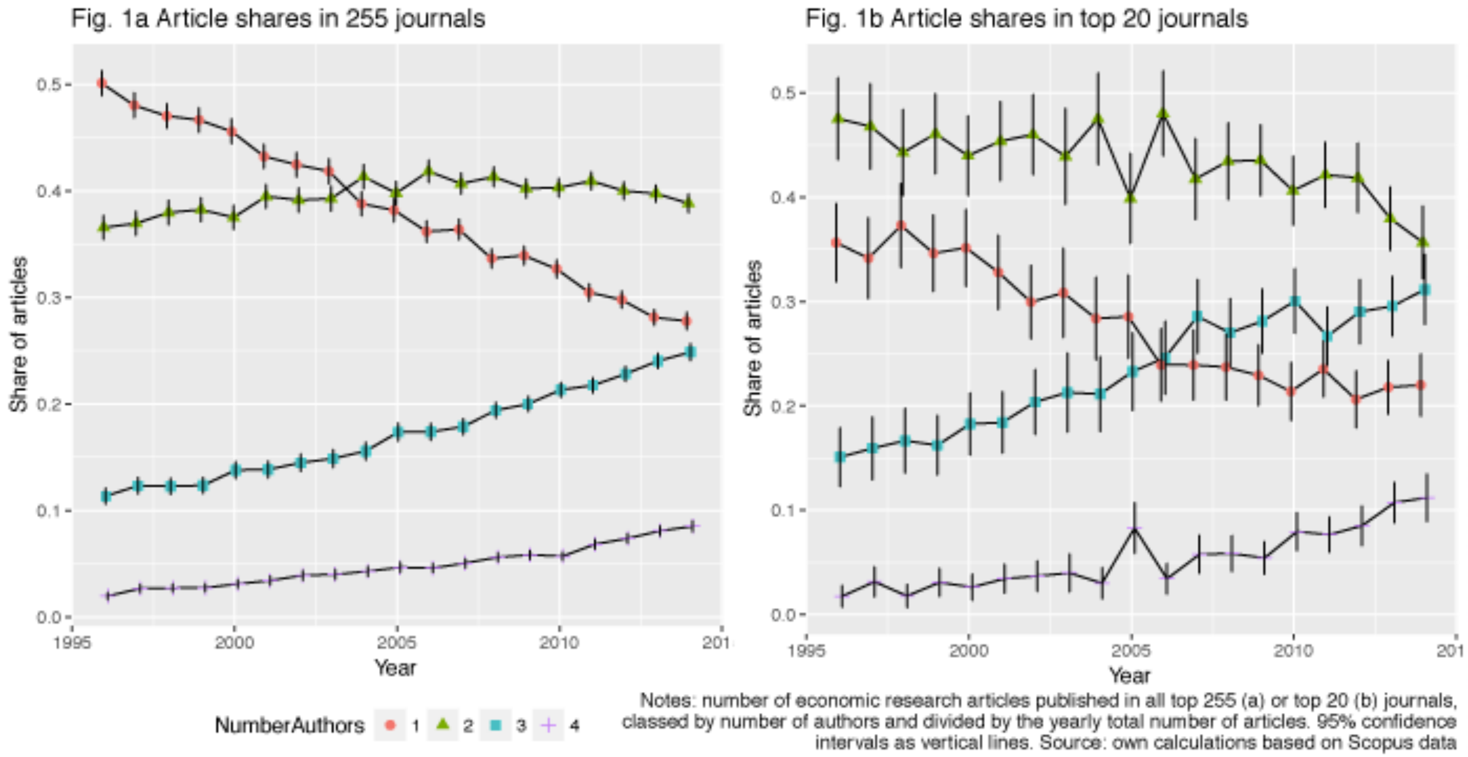


Fig. 1a Article shares in 255 journals

Fig. 1b Article shares in top 20 journals

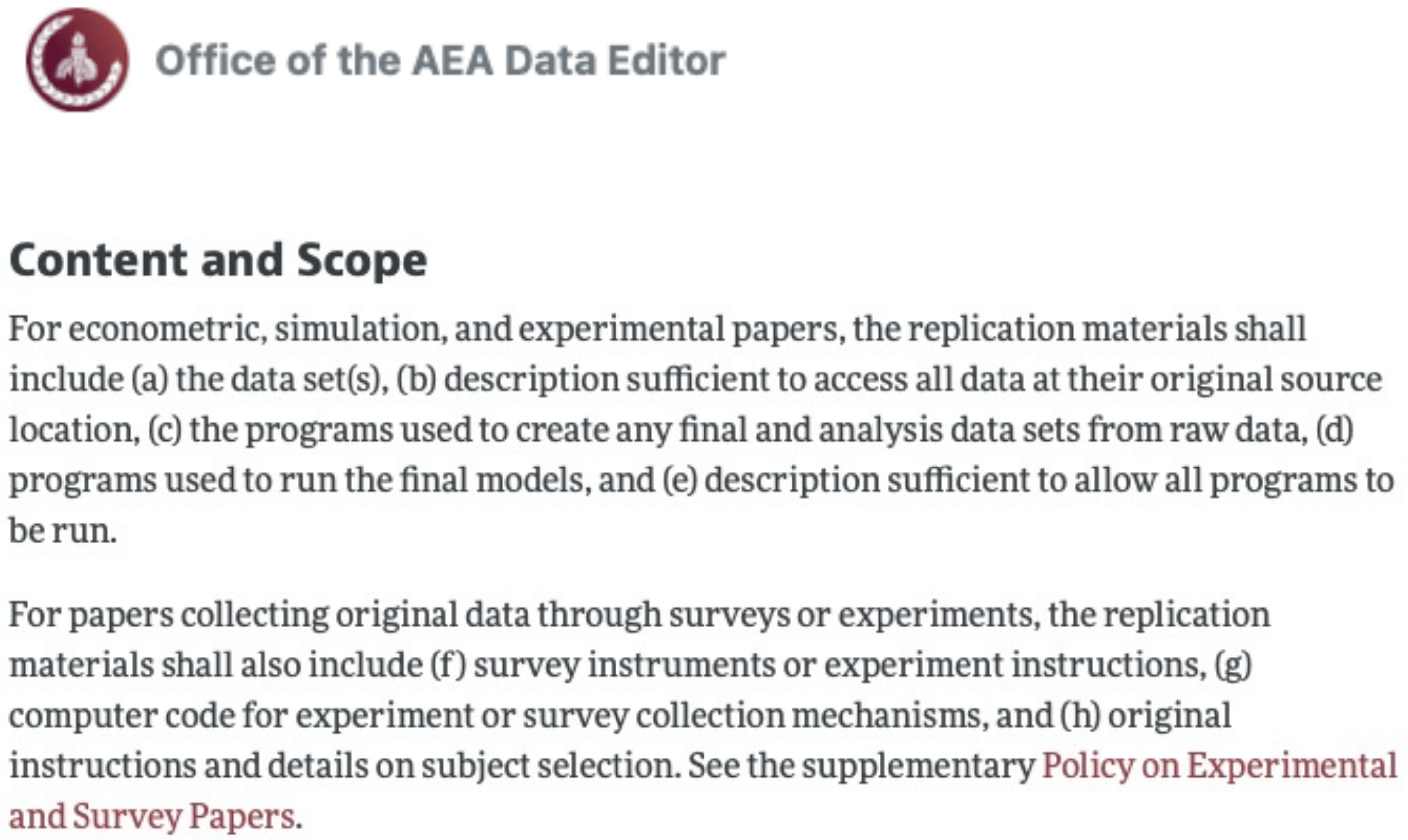
Notes: number of economic research articles published in all top 255 (a) or top 20 (b) journals, classed by number of authors and divided by the yearly total number of articles. 95% confidence intervals as vertical lines. Source: own calculations based on Scopus data

https://voxeu.org/article/growth-multi-authored-journal-articles-economics

VOX EU CEPR

Motivation (3)

- Stricter requirements for publication of code with papers
- Cataloguing of *raw* and *intermediate* data generally expected



Office of the AEA Data Editor

Content and Scope

For econometric, simulation, and experimental papers, the replication materials shall include (a) the data set(s), (b) description sufficient to access all data at their original source location, (c) the programs used to create any final and analysis data sets from raw data, (d) programs used to run the final models, and (e) description sufficient to allow all programs to be run.

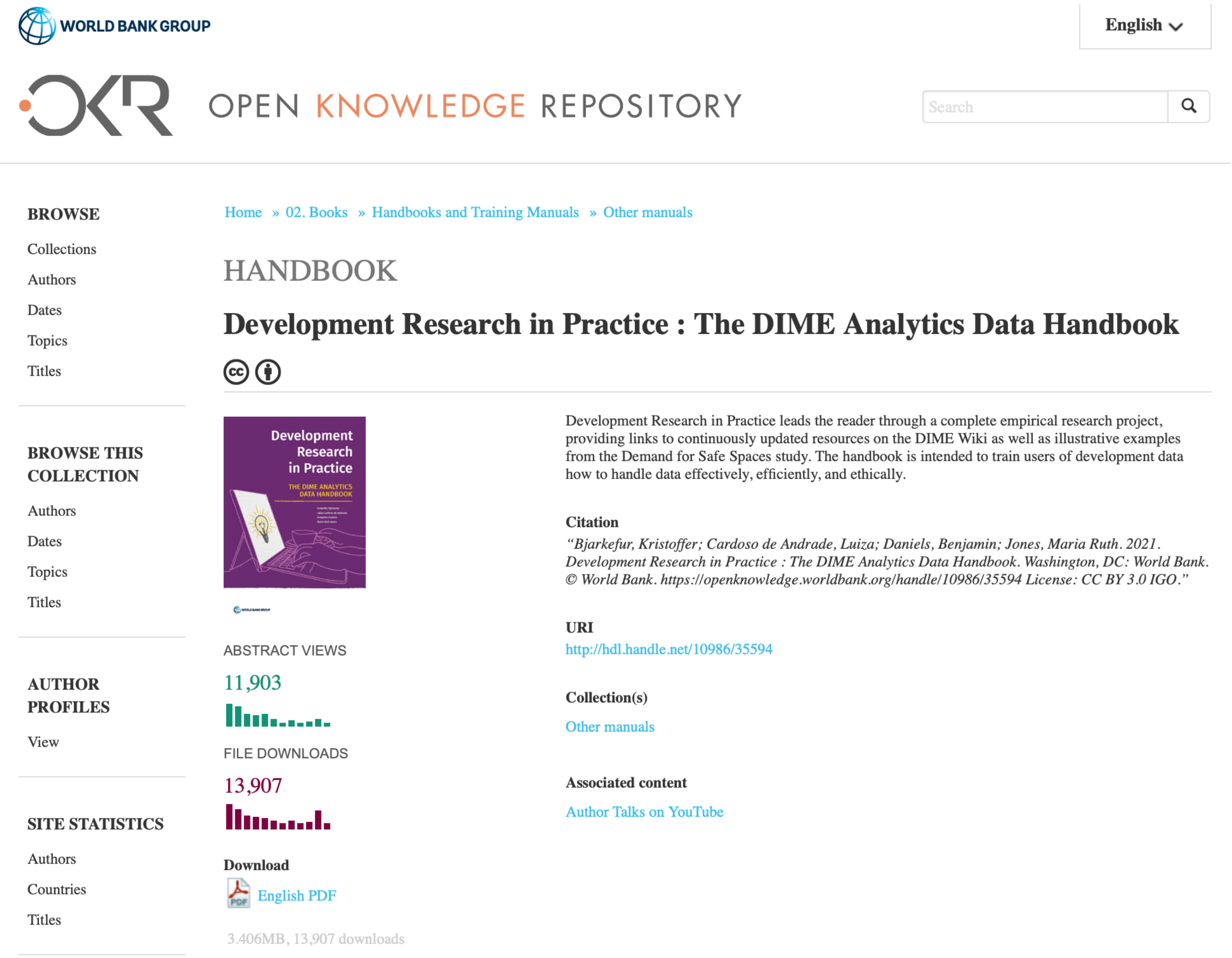
For papers collecting original data through surveys or experiments, the replication materials shall also include (f) survey instruments or experiment instructions, (g) computer code for experiment or survey collection mechanisms, and (h) original instructions and details on subject selection. See the supplementary Policy on Experimental and Survey Papers.

https://www.aeaweb.org/journals/data/data-code-policy

https://aeadataeditor.github.io

Development Research in Practice

- Huge need to professionalize and modernize research responsibilities, especially in collaborative settings
- We provide a comprehensive resource for PI *and* RA/student researchers and staff members



HANDBOOK

Development Research in Practice : The DIME Analytics Data Handbook

Development Research in Practice leads the reader through a complete empirical research project, providing links to continuously updated resources on the DIME Wiki as well as illustrative examples from the Demand for Safe Spaces study. The handbook is intended to train users of development data how to handle data effectively, efficiently, and ethically.

Citation

"Bjarkefur, Kristoffer; Cardoso de Andrade, Luiza; Daniels, Benjamin; Jones, Maria Ruth. 2021. Development Research in Practice : The DIME Analytics Data Handbook. Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/35594 License: CC BY 3.0 IGO."

URI

http://hdl.handle.net/10986/35594

Collection(s)

Other manuals

Associated content

Author Talks on YouTube

ABSTRACT VIEWS

11,903

FILE DOWNLOADS

13,907

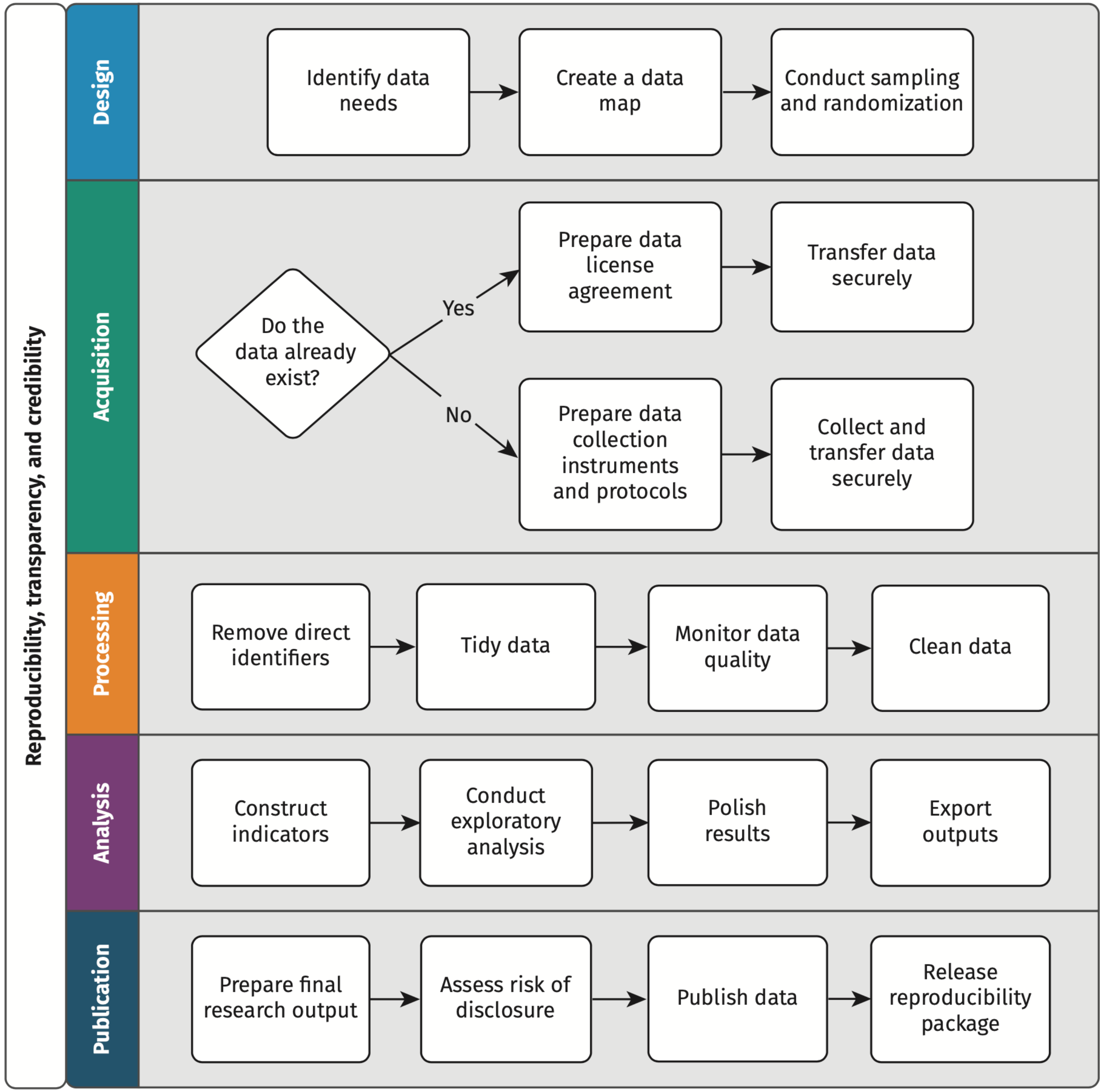
Download

English PDF

3.406MB, 13,907 downloads

Reproducibility, Transparency, and Credibility

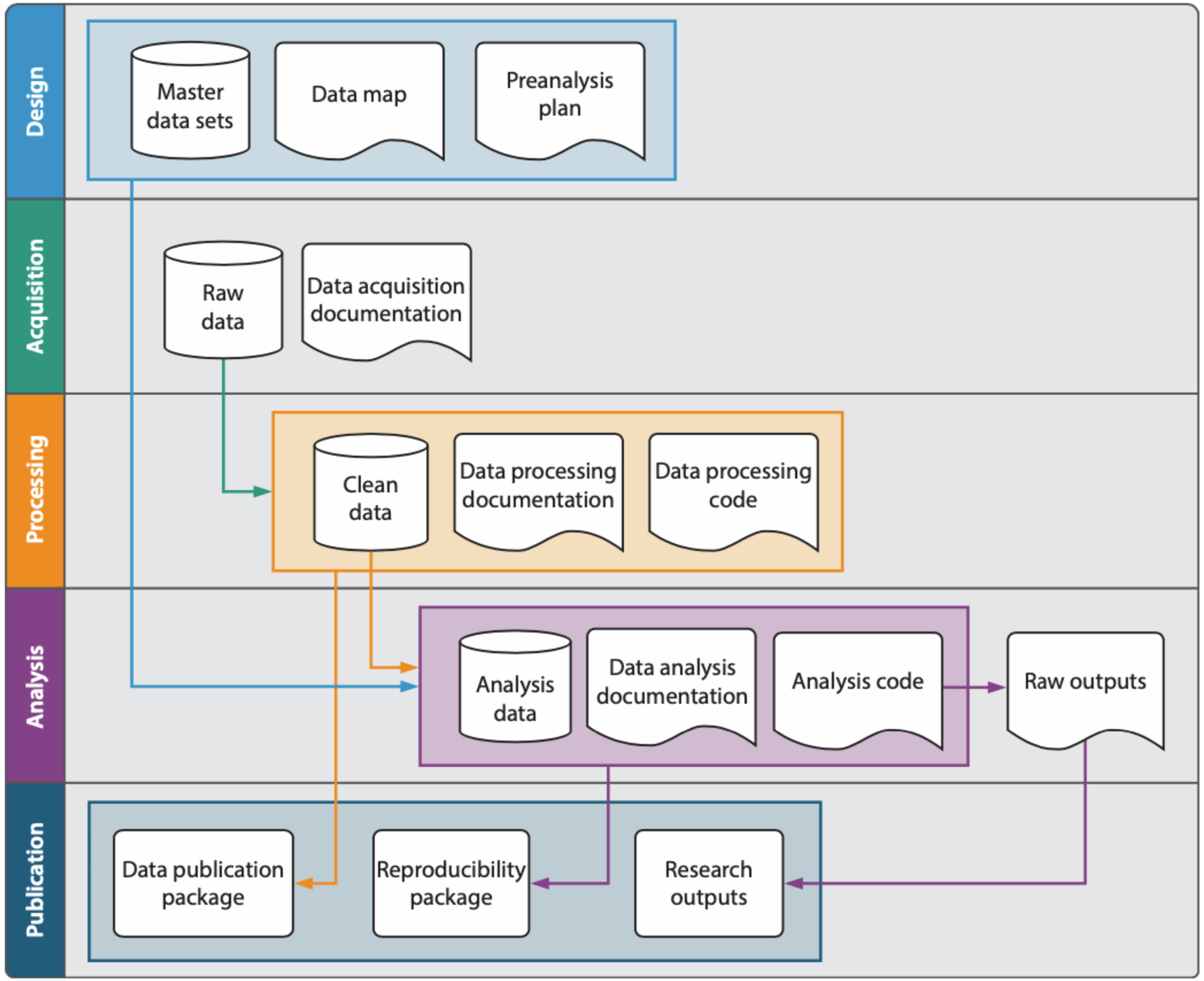
- Modern research should **NOT** be “ticking off boxes”
- No “requirements” or “chores”
- We provide a workflow and framework where research is *easier* to do within and across teams when *done right*


Reproducibility, transparency, and credibility
Design
Identify data needs
Create a data map
Conduct sampling and randomization
Acquisition
Do the data already exist?
Yes
No
Prepare data license agreement
Transfer data securely
Prepare data collection instruments and protocols
Collect and transfer data securely
Processing
Remove direct identifiers
Tidy data
Monitor data quality
Clean data
Analysis
Construct indicators
Conduct exploratory analysis
Polish results
Export outputs
Publication
Prepare final research output
Assess risk of disclosure
Publish data
Release reproducibility package

Full research lifecycle view

- Suited for individuals as well as large labs
- Goal is always to know what outputs should exist and how to re-create them
- Where to go back and get old info is essential in 7+ year life cycles

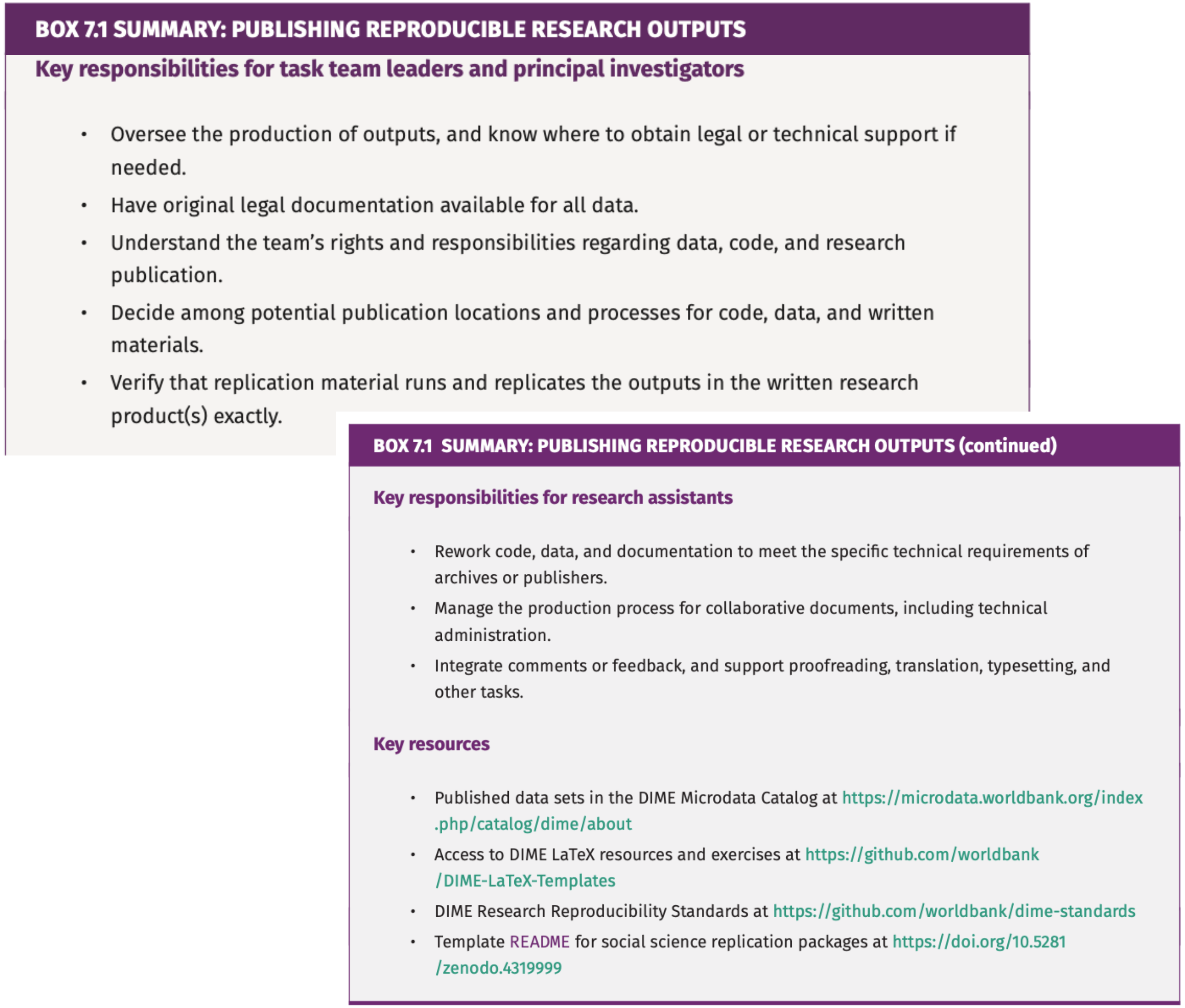
FIGURE 8.1 **Research data work outputs**



Source: DIME (Development Impact Evaluation), World Bank.

Explicit tasks and outputs divided by role

- What works for reproducibility is also required for group collaboration
- By focusing on within-team efficiency, achieve reproducibility "by default"



BOX 7.1 SUMMARY: PUBLISHING REPRODUCIBLE RESEARCH OUTPUTS

Key responsibilities for task team leaders and principal investigators

- Oversee the production of outputs, and know where to obtain legal or technical support if needed.
- Have original legal documentation available for all data.
- Understand the team's rights and responsibilities regarding data, code, and research publication.
- Decide among potential publication locations and processes for code, data, and written materials.
- Verify that replication material runs and replicates the outputs in the written research product(s) exactly.

BOX 7.1 SUMMARY: PUBLISHING REPRODUCIBLE RESEARCH OUTPUTS (continued)

Key responsibilities for research assistants

- Rework code, data, and documentation to meet the specific technical requirements of archives or publishers.
- Manage the production process for collaborative documents, including technical administration.
- Integrate comments or feedback, and support proofreading, translation, typesetting, and other tasks.

Key resources

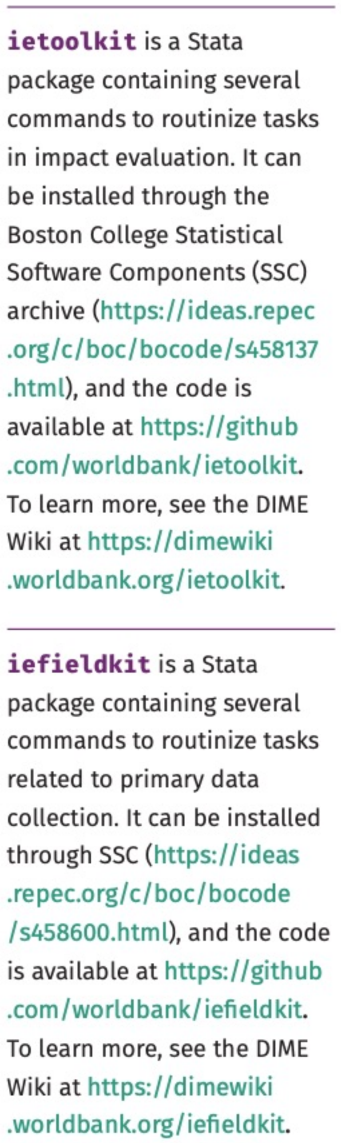
- Published data sets in the DIME Microdata Catalog at https://microdata.worldbank.org/index.php/catalog/dime/about
- Access to DIME LaTeX resources and exercises at https://github.com/worldbank/DIME-LaTeX-Templates
- DIME Research Reproducibility Standards at https://github.com/worldbank/dime-standards
- Template README for social science replication packages at https://doi.org/10.5281/zenodo.4319999

Using Stata for Research

CODE EXAMPLES, STYLE GUIDE, AND GENERAL ADVICE

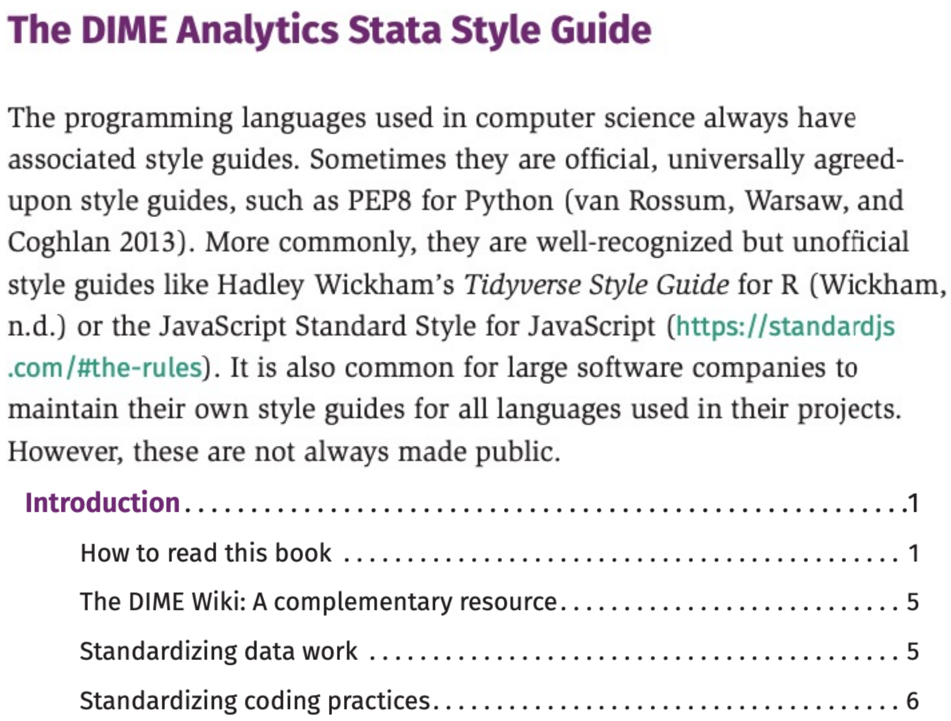
Stata remains dominant in economics work

- But there is little “computer science style” training in basic Stata practices
- We provide basic resources for students with non-CompSci background



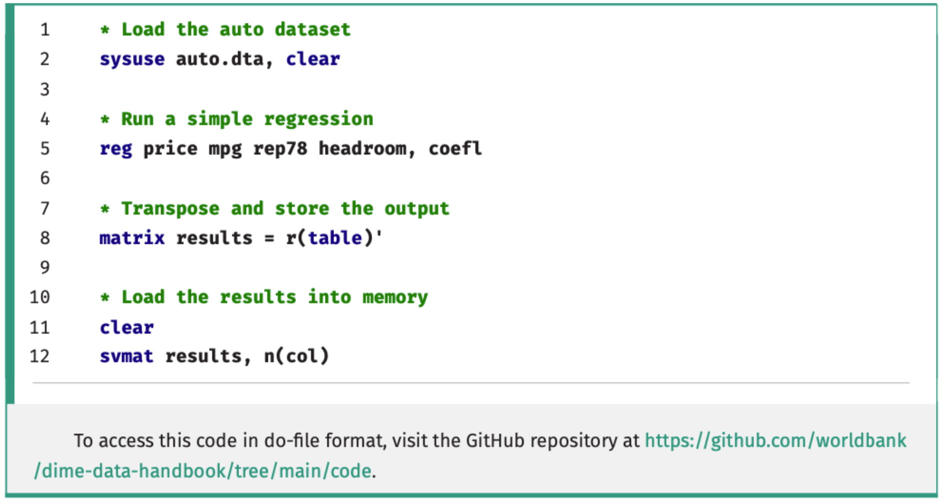
`ietoolkit` is a Stata package containing several commands to routinize tasks in impact evaluation. It can be installed through the Boston College Statistical Software Components (SSC) archive (https://ideas.repec.org/c/boc/bocode/s458137.html), and the code is available at https://github.com/worldbank/ietoolkit. To learn more, see the DIME Wiki at https://dimewiki.worldbank.org/ietoolkit.

`iefieldkit` is a Stata package containing several commands to routinize tasks related to primary data collection. It can be installed through SSC (https://ideas.repec.org/c/boc/bocode/s458600.html), and the code is available at https://github.com/worldbank/iefieldkit. To learn more, see the DIME Wiki at https://dimewiki.worldbank.org/iefieldkit.



The DIME Analytics Stata Style Guide

The programming languages used in computer science always have associated style guides. Sometimes they are official, universally agreed-upon style guides, such as PEP8 for Python (van Rossum, Warsaw, and Coghlan 2013). More commonly, they are well-recognized but unofficial style guides like Hadley Wickham’s *Tidyverse Style Guide* for R (Wickham, n.d.) or the JavaScript Standard Style for JavaScript (https://standardjs.com/#the-rules). It is also common for large software companies to maintain their own style guides for all languages used in their projects. However, these are not always made public.

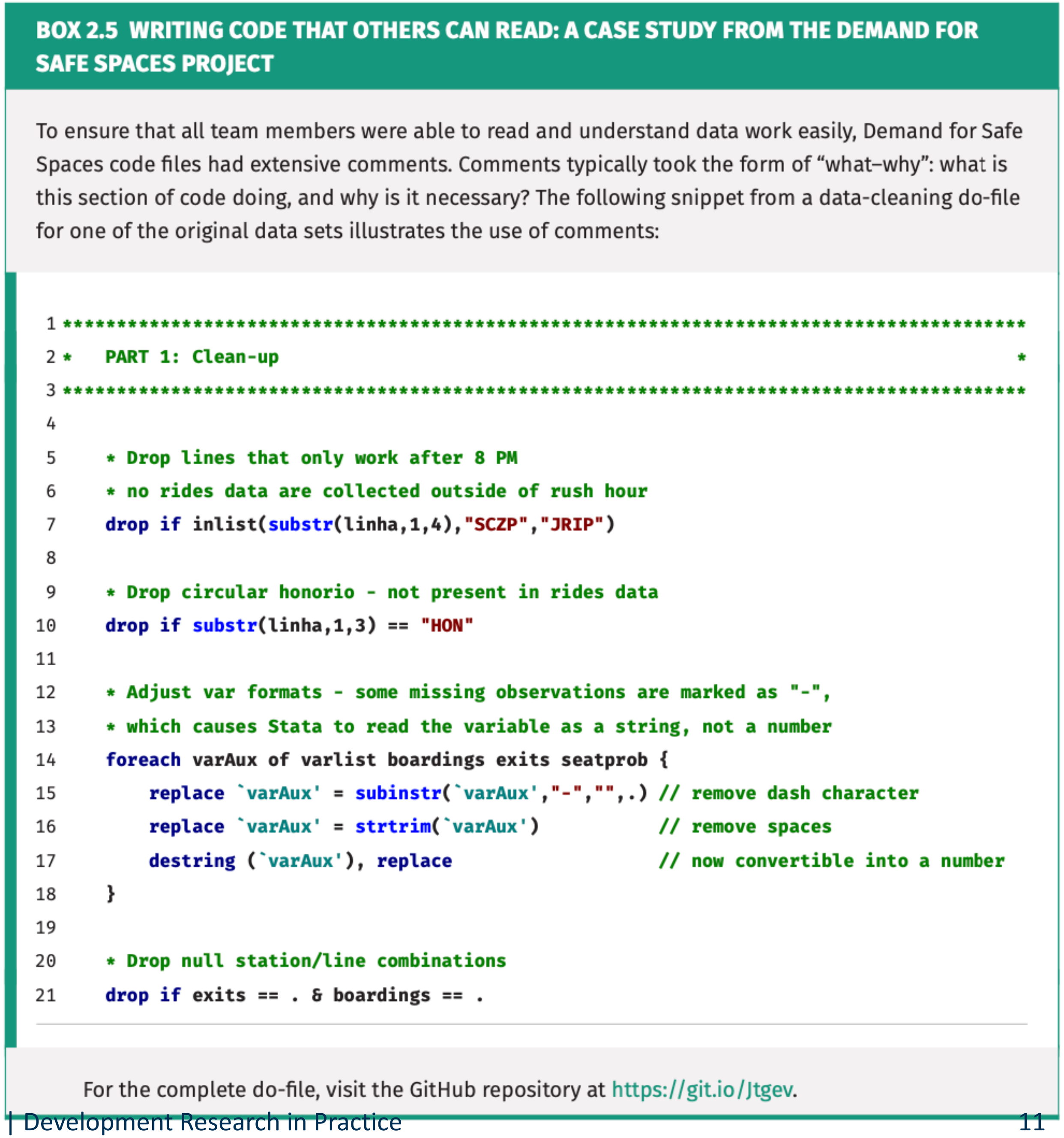


```
* Load the auto dataset
sysuse auto.dta, clear

* Run a simple regression
reg price mpg rep78 headroom, coefl

* Transpose and store the output
matrix results = r(table)'

* Load the results into memory
clear
svmat results, n(col)
```

To access this code in do-file format, visit the GitHub repository at https://github.com/worldbank/dime-data-handbook/tree/main/code.

Goal: Accessible common coding standards

- Every task is illustrated with real Stata code from a published project
- See: https://bit.ly/drip-rio-demo



BOX 2.5 WRITING CODE THAT OTHERS CAN READ: A CASE STUDY FROM THE DEMAND FOR SAFE SPACES PROJECT

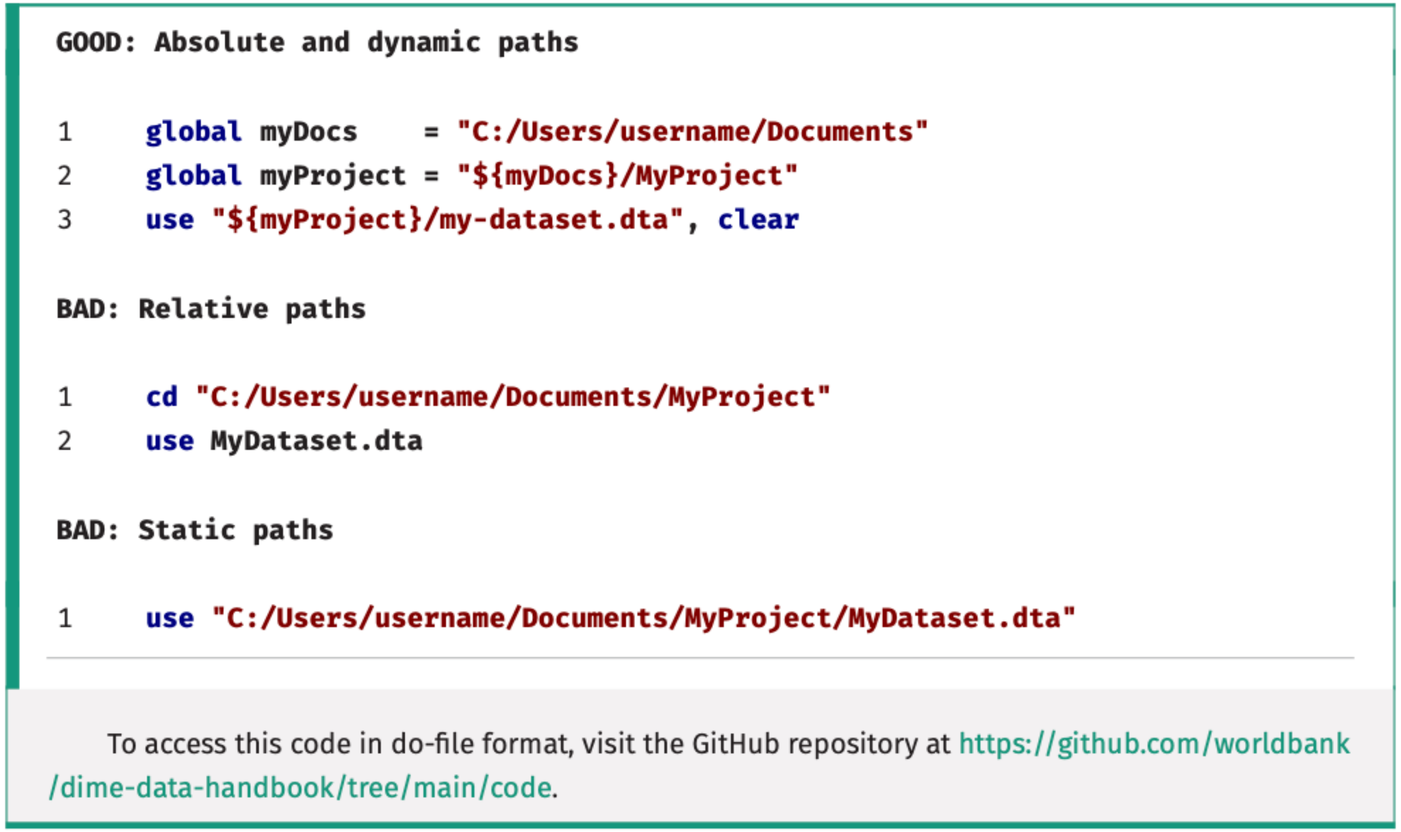
To ensure that all team members were able to read and understand data work easily, Demand for Safe Spaces code files had extensive comments. Comments typically took the form of "what–why": what is this section of code doing, and why is it necessary? The following snippet from a data-cleaning do-file for one of the original data sets illustrates the use of comments:

```
*****************************************************************************************
*   PART 1: Clean-up                                                                    *
*****************************************************************************************

    * Drop lines that only work after 8 PM
    * no rides data are collected outside of rush hour
    drop if inlist(substr(linha,1,4),"SCZP","JRIP")

    * Drop circular honorio - not present in rides data
    drop if substr(linha,1,3) == "HON"

    * Adjust var formats - some missing observations are marked as "-",
    * which causes Stata to read the variable as a string, not a number
    foreach varAux of varlist boardings exits seatprob {
        replace `varAux' = subinstr(`varAux',"-","",.) // remove dash character
        replace `varAux' = strtrim(`varAux')           // remove spaces
        destring (`varAux'), replace                   // now convertible into a number
    }

    * Drop null station/line combinations
    drop if exits == . & boardings == .
```

For the complete do-file, visit the GitHub repository at https://git.io/Jtgev.

The DIME Analytics Stata Style guide

- Online at: https://worldbank.github.io/dime-data-handbook/coding.html



GOOD: Absolute and dynamic paths

```
global myDocs    = "C:/Users/username/Documents"
global myProject = "${myDocs}/MyProject"
use "${myProject}/my-dataset.dta", clear
```

BAD: Relative paths

```
cd "C:/Users/username/Documents/MyProject"
use MyDataset.dta
```

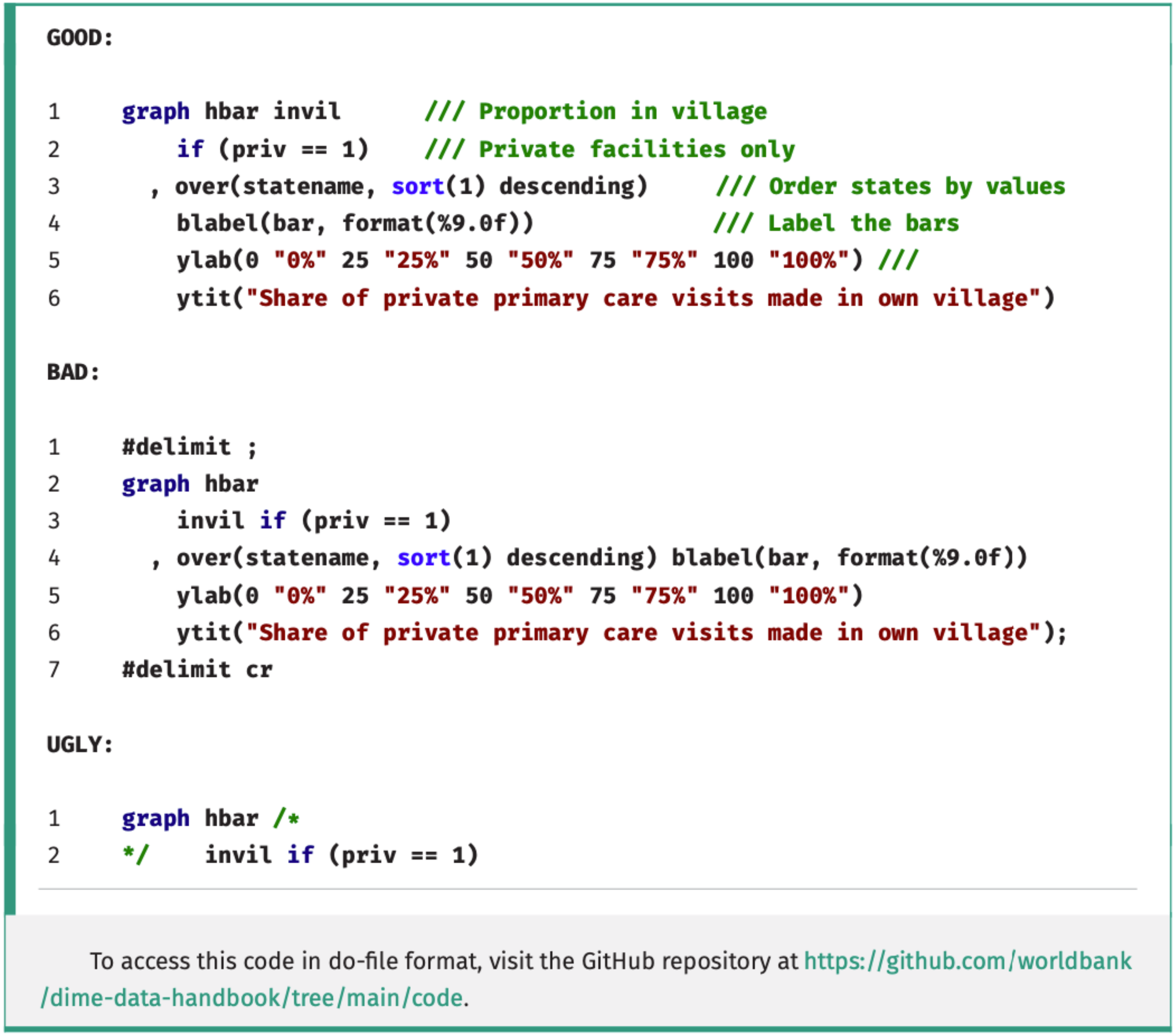
BAD: Static paths

```
use "C:/Users/username/Documents/MyProject/MyDataset.dta"
```

To access this code in do-file format, visit the GitHub repository at https://github.com/worldbank/dime-data-handbook/tree/main/code.

The DIME Analytics Stata Style guide

- Highly opinionated, for the sake of consistency!
- We do not want people to “get creative” with code style when they don’t have to be



GOOD:

```
graph hbar invil      /// Proportion in village
    if (priv == 1)    /// Private facilities only
  , over(statename, sort(1) descending)    /// Order states by values
    blabel(bar, format(%9.0f))             /// Label the bars
    ylab(0 "0%" 25 "25%" 50 "50%" 75 "75%" 100 "100%") ///
    ytit("Share of private primary care visits made in own village")
```

BAD:

```
#delimit ;
graph hbar
    invil if (priv == 1)
  , over(statename, sort(1) descending) blabel(bar, format(%9.0f))
    ylab(0 "0%" 25 "25%" 50 "50%" 75 "75%" 100 "100%")
    ytit("Share of private primary care visits made in own village");
#delimit cr
```

UGLY:

```
graph hbar /*
*/    invil if (priv == 1)
```

To access this code in do-file format, visit the GitHub repository at https://github.com/worldbank/dime-data-handbook/tree/main/code.

Goal: Automate common Stata workflows

- We provide recommended templates for folder organization in all projects
- [ietoolkit] programs like [iefolder] automatically manage this structure

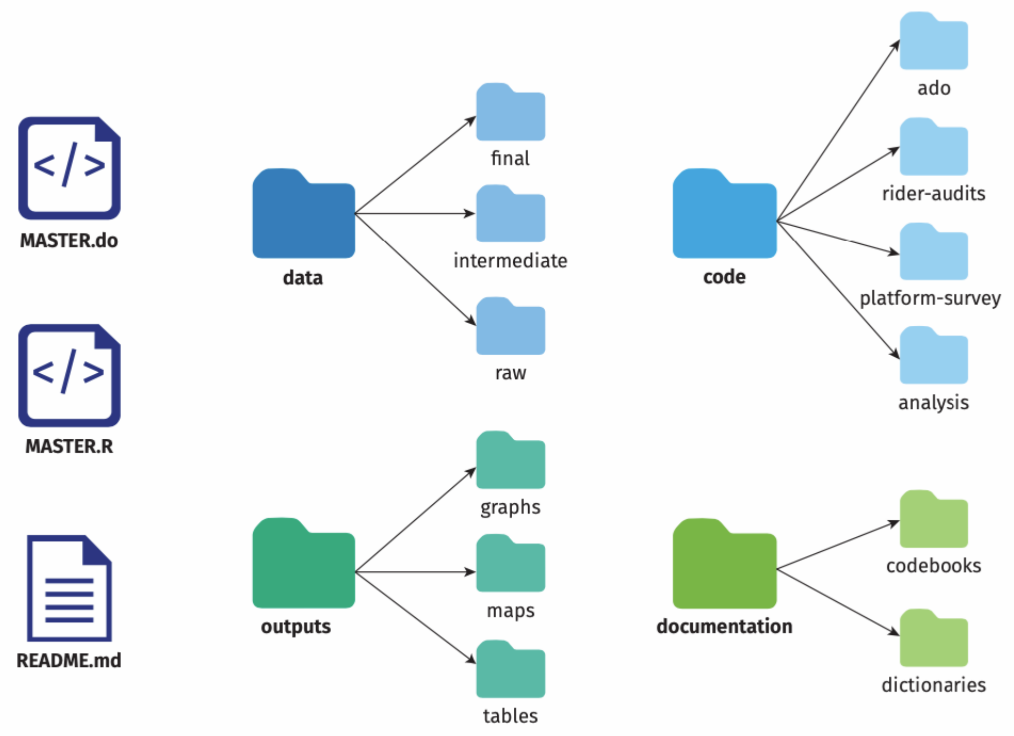
`iefolder` is a Stata command to set up a standardized folder structure for all research team members. It is part of the `ietoolkit` package. For more details, see the DIME Wiki at https://dimewiki.worldbank.org/iefolder.

The **DataWork folder** is the root of DIME's recommended folder structure. For more details, see the DIME Wiki at https://dimewiki.worldbank.org/DataWork_Folder.

Chapter 2

Setting the stage for effective and efficient collaboration

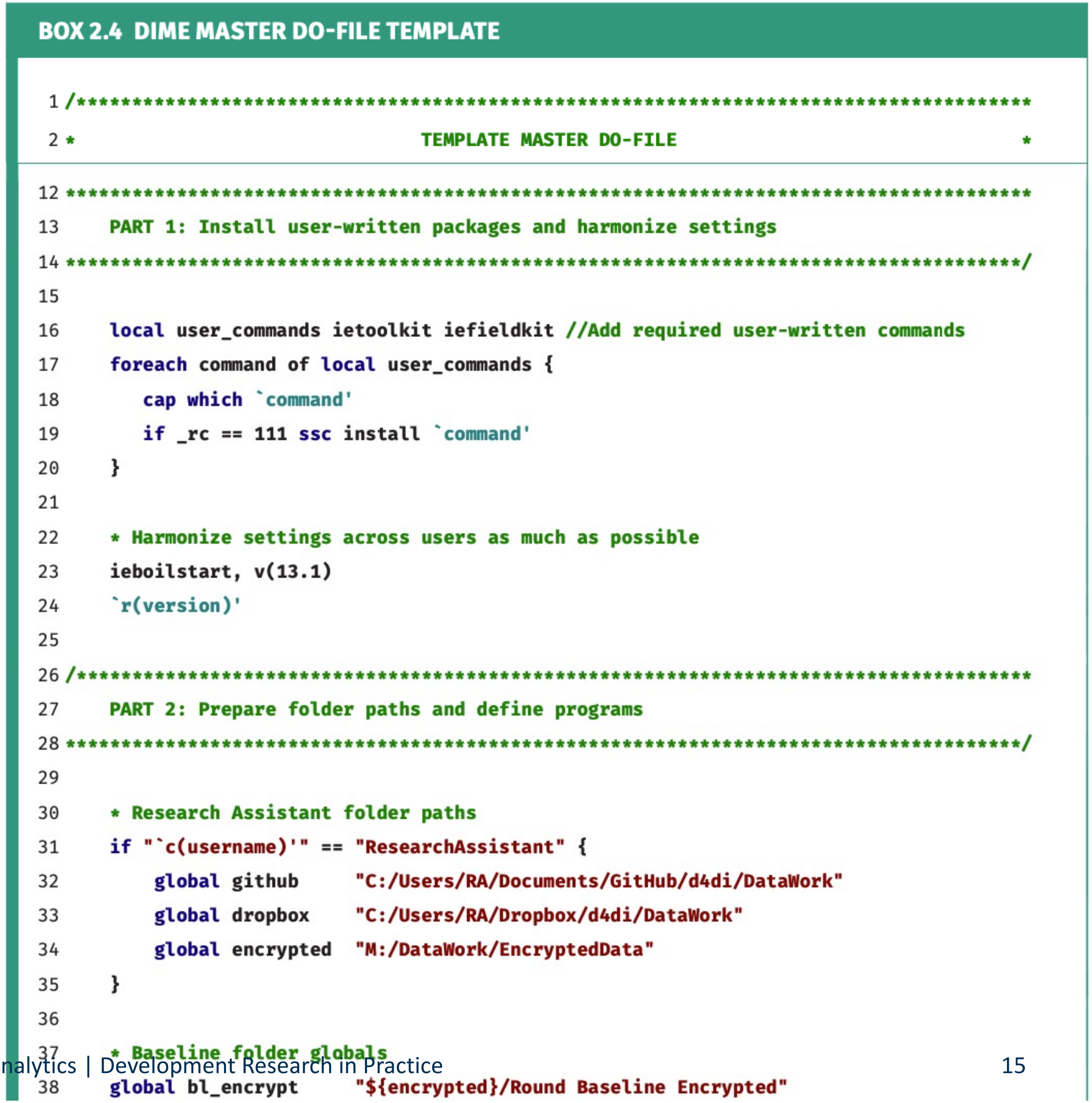
FIGURE B2.3.1 **Folder structure of the Demand for Safe Spaces data work**



Source: DIME (Development Impact Evaluation), World Bank.

Goal: Common organization, comment styles

- We also provide guidance for organization of do-files and run-files
- Built from Day 1 for reproducibility, including user-written command management and user-switching

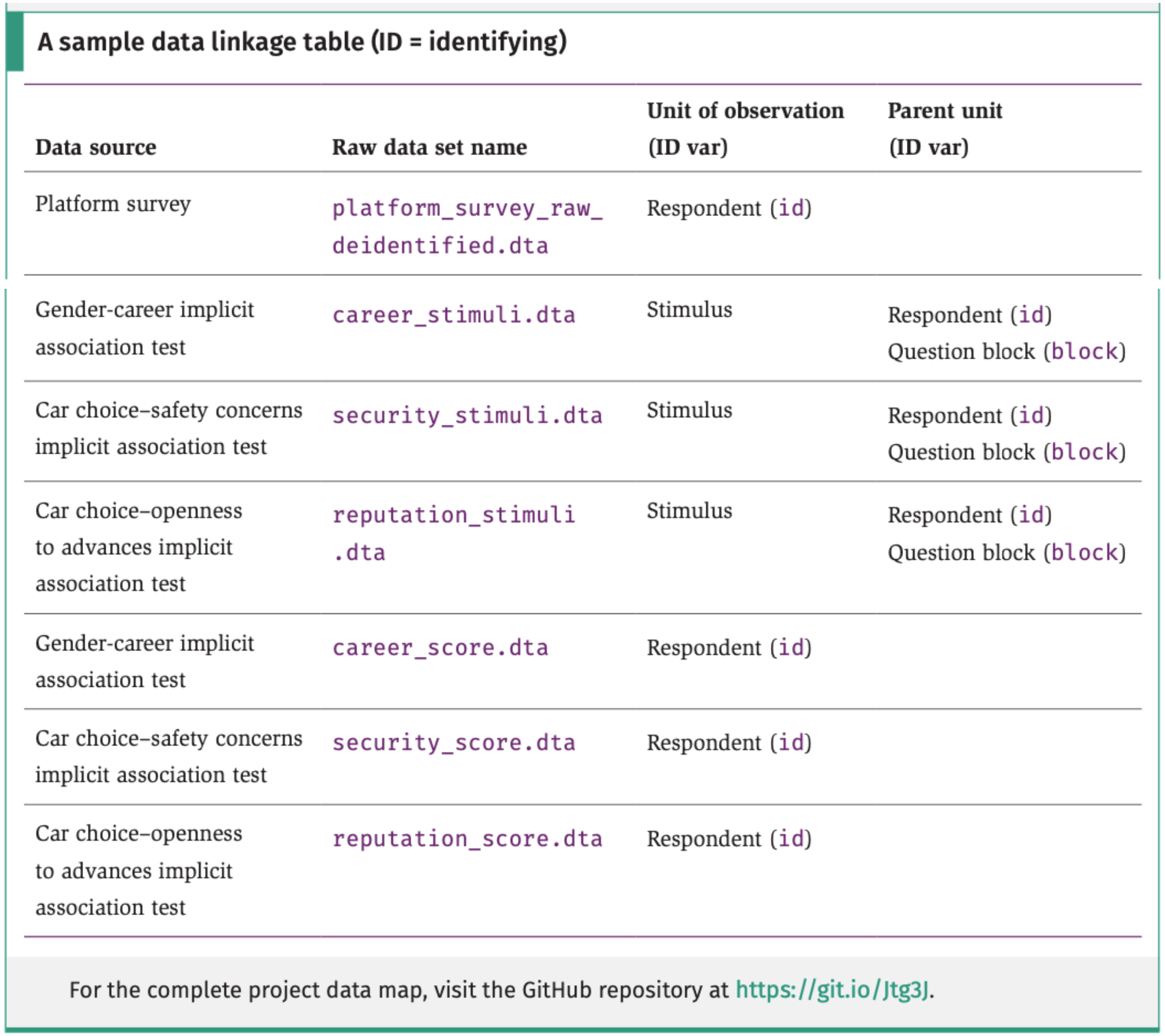


BOX 2.4 DIME MASTER DO-FILE TEMPLATE

```
/*******************************************************************************
*                           TEMPLATE MASTER DO-FILE                           *
********************************************************************************
    PART 1: Install user-written packages and harmonize settings
*******************************************************************************/

    local user_commands ietoolkit iefieldkit //Add required user-written commands
    foreach command of local user_commands {
       cap which `command'
       if _rc == 111 ssc install `command'
    }

    * Harmonize settings across users as much as possible
    ieboilstart, v(13.1)
    `r(version)'

/*******************************************************************************
    PART 2: Prepare folder paths and define programs
*******************************************************************************/

    * Research Assistant folder paths
    if "`c(username)'" == "ResearchAssistant" {
        global github      "C:/Users/RA/Documents/GitHub/d4di/DataWork"
        global dropbox     "C:/Users/RA/Dropbox/d4di/DataWork"
        global encrypted   "M:/DataWork/EncryptedData"
    }

    * Baseline folder globals
    global bl_encrypt      "${encrypted}/Round Baseline Encrypted"
```

Goal: Linking documentation to Stata code

- Data linkage table lists location and description of all data sets
- Specific reference to identifying variables (which satisfy [isid] checks) and [merge] keys



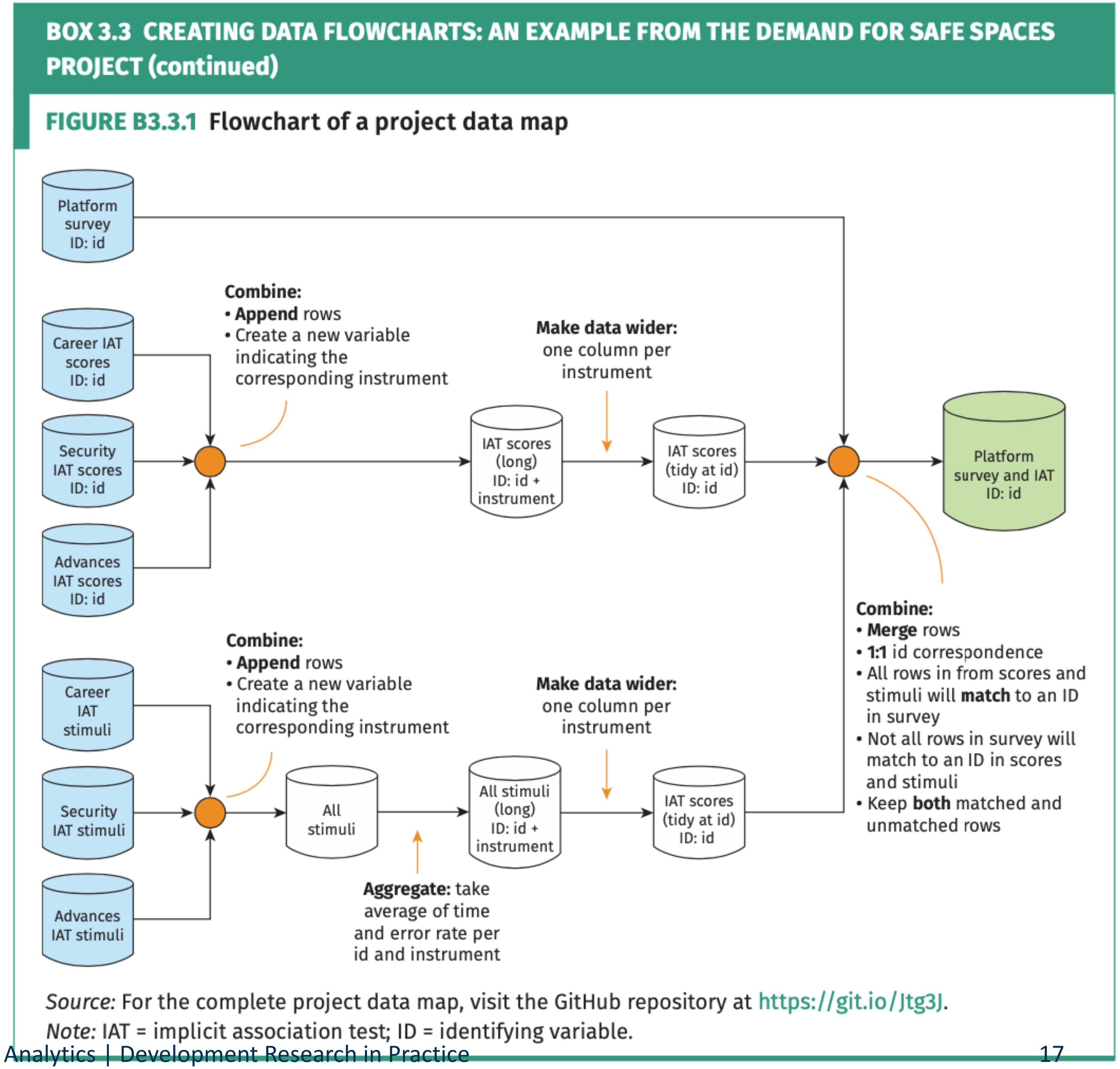
A sample data linkage table (ID = identifying)

Data source	Raw data set name	Unit of observation (ID var)	Parent unit (ID var)
Platform survey	`platform_survey_raw_deidentified.dta`	Respondent (`id`)	
Gender-career implicit association test	`career_stimuli.dta`	Stimulus	Respondent (`id`) Question block (`block`)
Car choice–safety concerns implicit association test	`security_stimuli.dta`	Stimulus	Respondent (`id`) Question block (`block`)
Car choice–openness to advances implicit association test	`reputation_stimuli.dta`	Stimulus	Respondent (`id`) Question block (`block`)
Gender-career implicit association test	`career_score.dta`	Respondent (`id`)	
Car choice–safety concerns implicit association test	`security_score.dta`	Respondent (`id`)	
Car choice–openness to advances implicit association test	`reputation_score.dta`	Respondent (`id`)	

For the complete project data map, visit the GitHub repository at https://git.io/Jtg3J.

Goal: Linking documentation to Stata code

- Data flow chart shows how these are supposed to be combined
- Specific reference to commands like [append], [collapse], [merge x:x], [reshape]



BOX 3.3 CREATING DATA FLOWCHARTS: AN EXAMPLE FROM THE DEMAND FOR SAFE SPACES PROJECT (continued)

FIGURE B3.3.1 **Flowchart of a project data map**

Platform survey
ID: id

Career IAT scores
ID: id

Security IAT scores
ID: id

Advances IAT scores
ID: id

Combine:
- **Append** rows
- Create a new variable indicating the corresponding instrument

IAT scores (long)
ID: id + instrument

Make data wider:
one column per instrument

IAT scores (tidy at id)
ID: id

Platform survey and IAT
ID: id

Career IAT stimuli

Security IAT stimuli

Advances IAT stimuli

Combine:
- **Append** rows
- Create a new variable indicating the corresponding instrument

All stimuli

Aggregate: take average of time and error rate per id and instrument

All stimuli (long)
ID: id + instrument

Make data wider:
one column per instrument

IAT scores (tidy at id)
ID: id

Combine:
- **Merge** rows
- **1:1** id correspondence
- All rows in from scores and stimuli will **match** to an ID in survey
- Not all rows in survey will match to an ID in scores and stimuli
- Keep **both** matched and unmatched rows

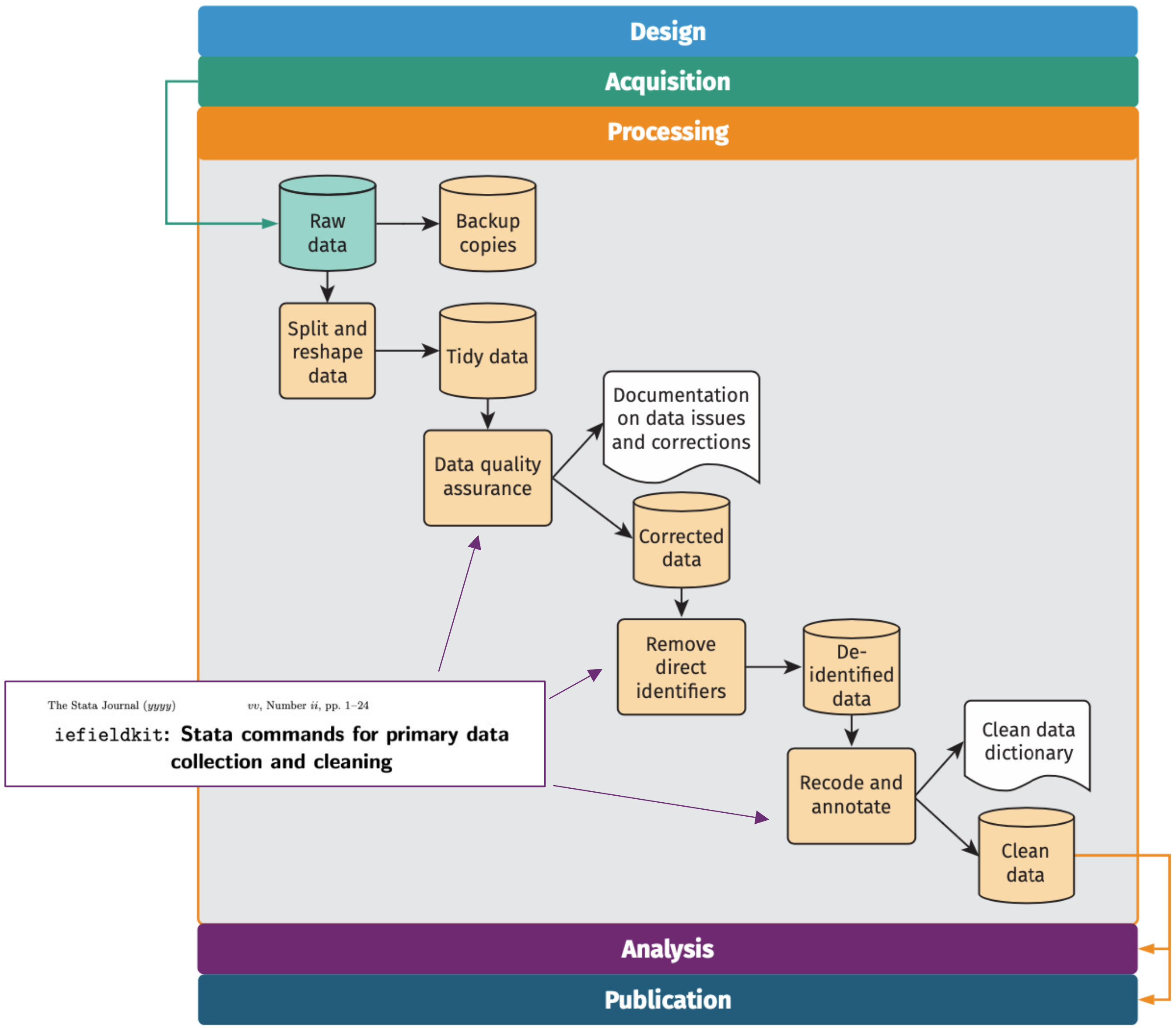
Source: For the complete project data map, visit the GitHub repository at https://git.io/Jtg3J.

Note: IAT = implicit association test; ID = identifying variable.

Goal: Self-documenting data processing

- Creating documentation (codebooks) for data
- Removing and documenting duplicate observations
- Rapidly de-identifying and re-coding data
- Spreadsheets are both code and documentation

FIGURE 5.1 Data-cleaning tasks and outputs



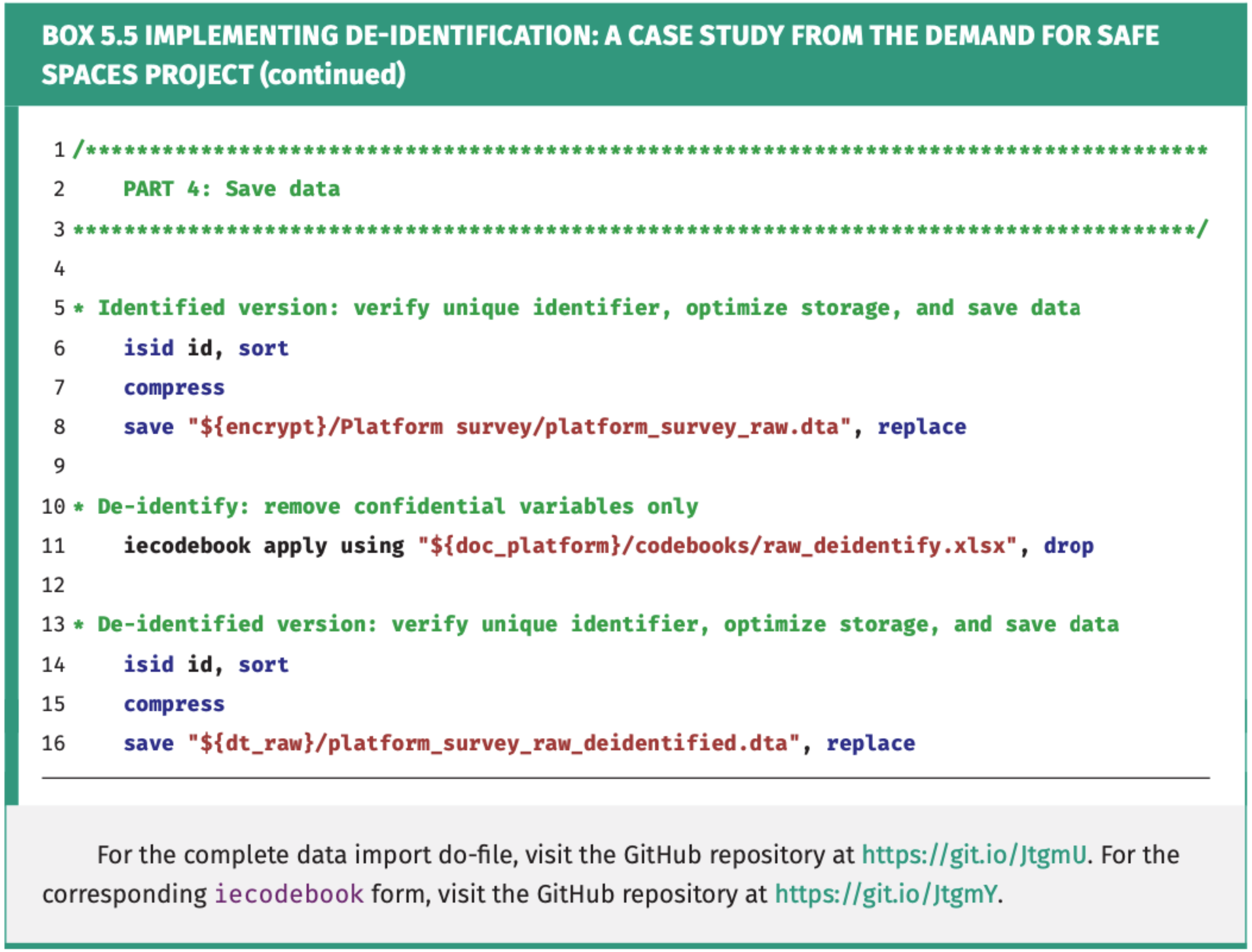
The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–24

`iefieldkit`: Stata commands for primary data collection and cleaning

Source: DIME (Development Impact Evaluation), World Bank.

Goal: Basic data hygiene is made simple

- Reducing the amount of coding expertise necessary
- Making essential tasks like security and de-duplication automatic
- Reducing the amount of *coding* necessary



BOX 5.5 IMPLEMENTING DE-IDENTIFICATION: A CASE STUDY FROM THE DEMAND FOR SAFE SPACES PROJECT (continued)

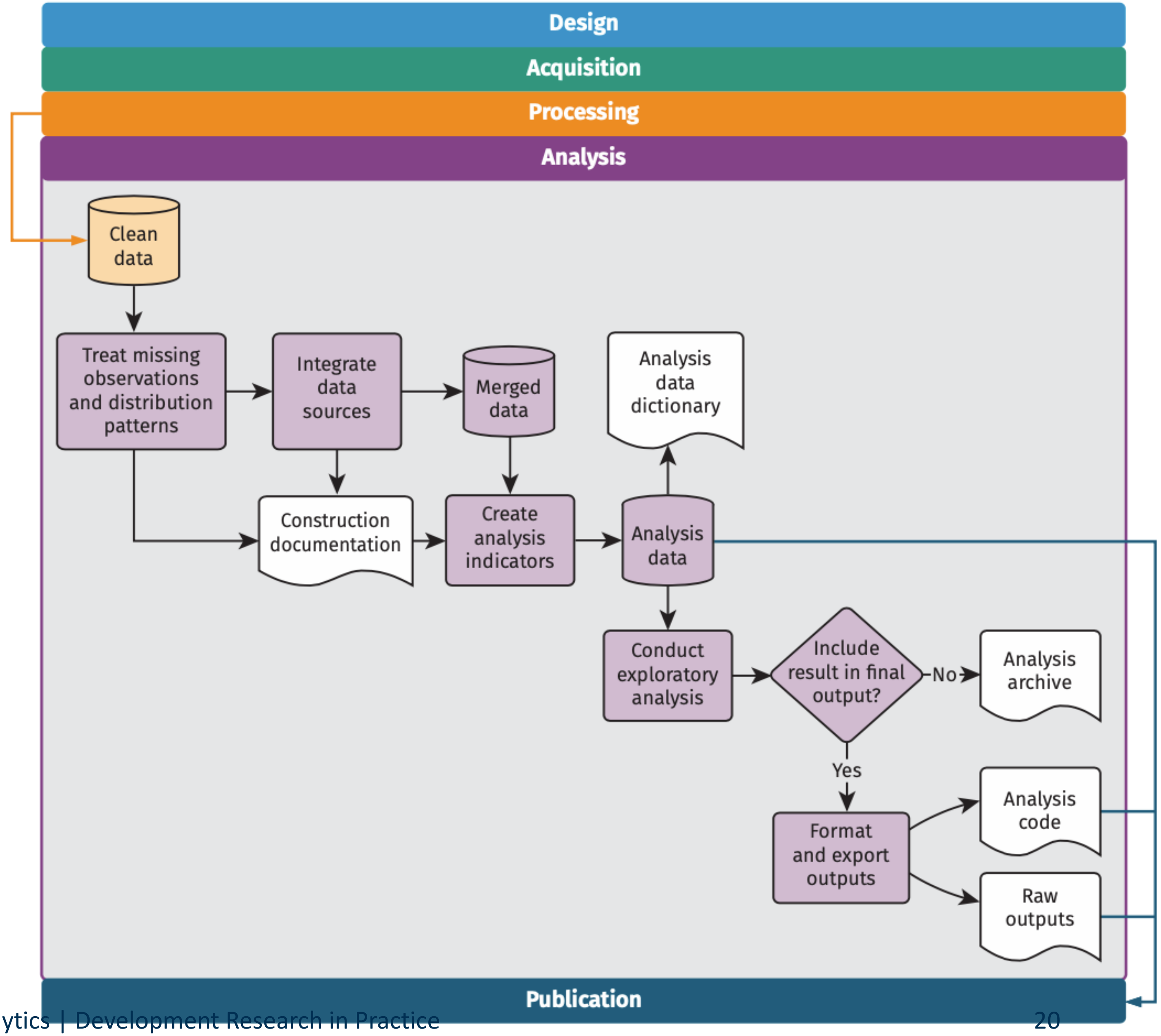
```
/****************************************************************************************
    PART 4: Save data
****************************************************************************************/

* Identified version: verify unique identifier, optimize storage, and save data
    isid id, sort
    compress
    save "${encrypt}/Platform survey/platform_survey_raw.dta", replace

* De-identify: remove confidential variables only
    iecodebook apply using "${doc_platform}/codebooks/raw_deidentify.xlsx", drop

* De-identified version: verify unique identifier, optimize storage, and save data
    isid id, sort
    compress
    save "${dt_raw}/platform_survey_raw_deidentified.dta", replace
```

For the complete data import do-file, visit the GitHub repository at https://git.io/JtgmU. For the corresponding `iecodebook` form, visit the GitHub repository at https://git.io/JtgmY.

Goal: Outputs are modular and organized

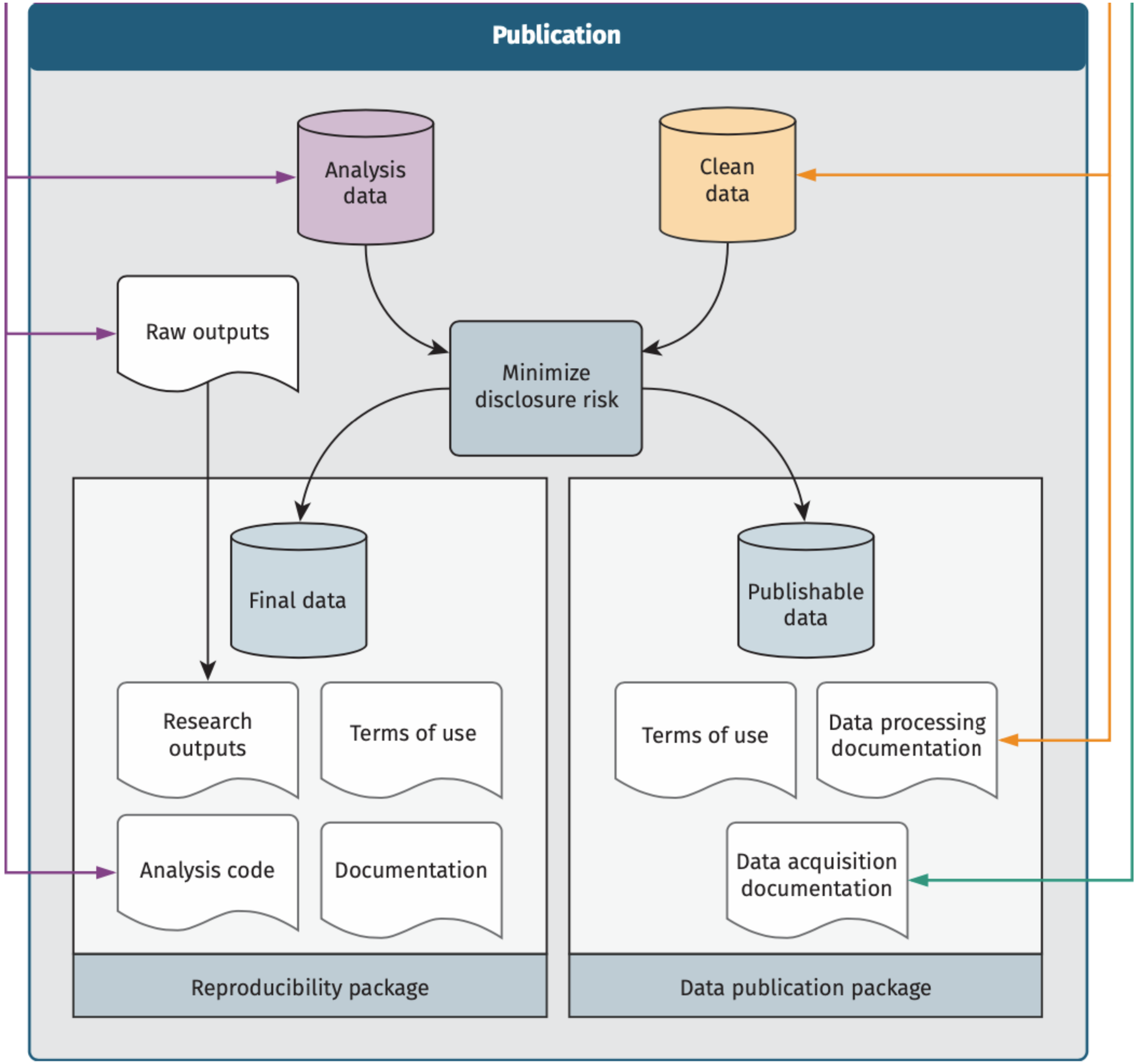
- Basic instructions for exporting figures and tables
- Instructions to organizing and archiving exploratory and final outputs

FIGURE 6.1 Data analysis tasks and outputs


Design
Acquisition
Processing
Analysis
Clean data
Treat missing observations and distribution patterns
Integrate data sources
Merged data
Analysis data dictionary
Construction documentation
Create analysis indicators
Analysis data
Conduct exploratory analysis
Include result in final output?
No
Analysis archive
Yes
Format and export outputs
Analysis code
Raw outputs
Publication

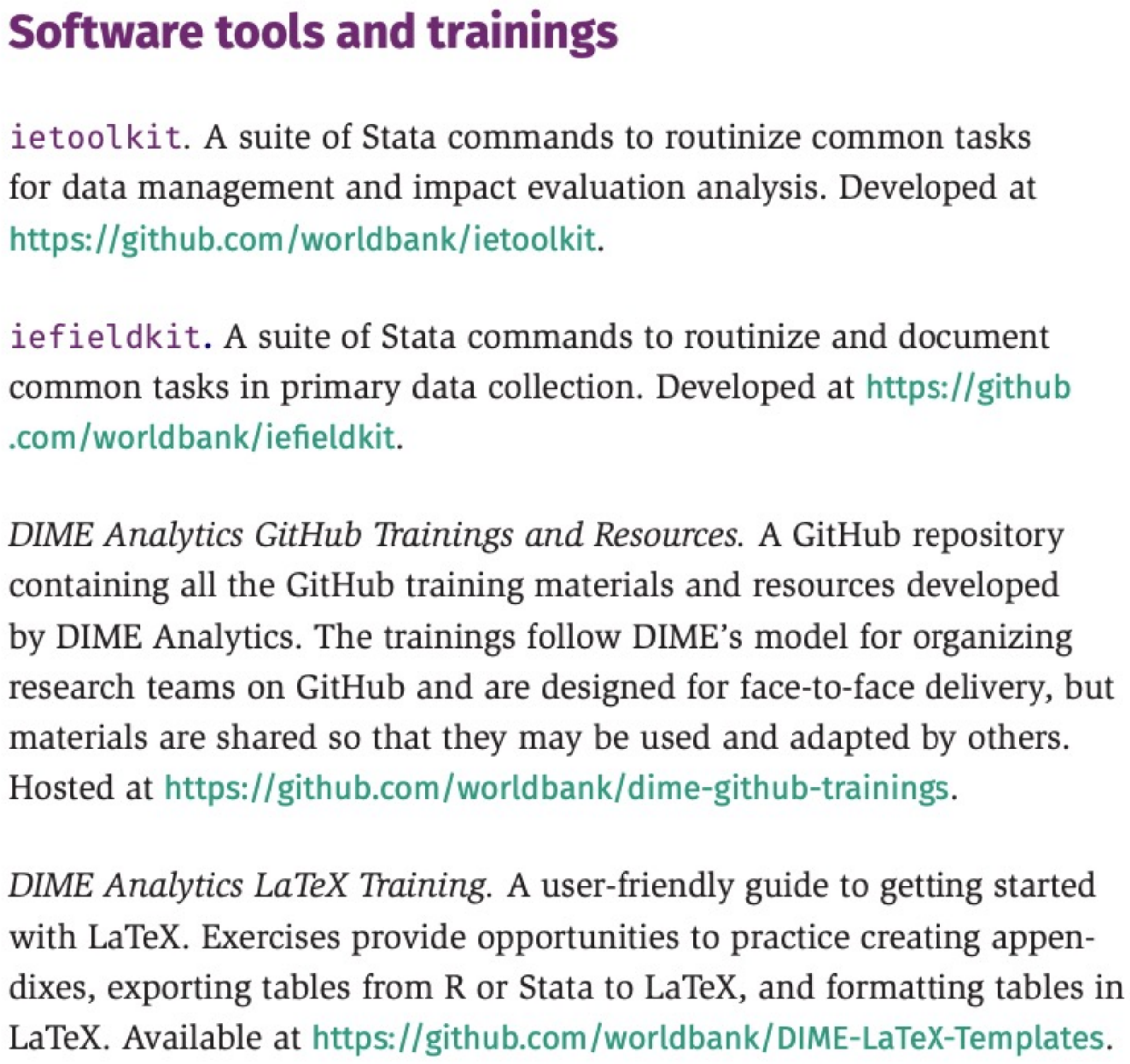
Goal: All steps build toward reproducibility

- Everything that is needed for release packages is already ready
- Highly modular, easily reproducible workflow based on collaboration



Source: DIME (Development Impact Evaluation), World Bank.

Lots more where this came from...



Software tools and trainings

`ietoolkit`. A suite of Stata commands to routinize common tasks for data management and impact evaluation analysis. Developed at https://github.com/worldbank/ietoolkit.

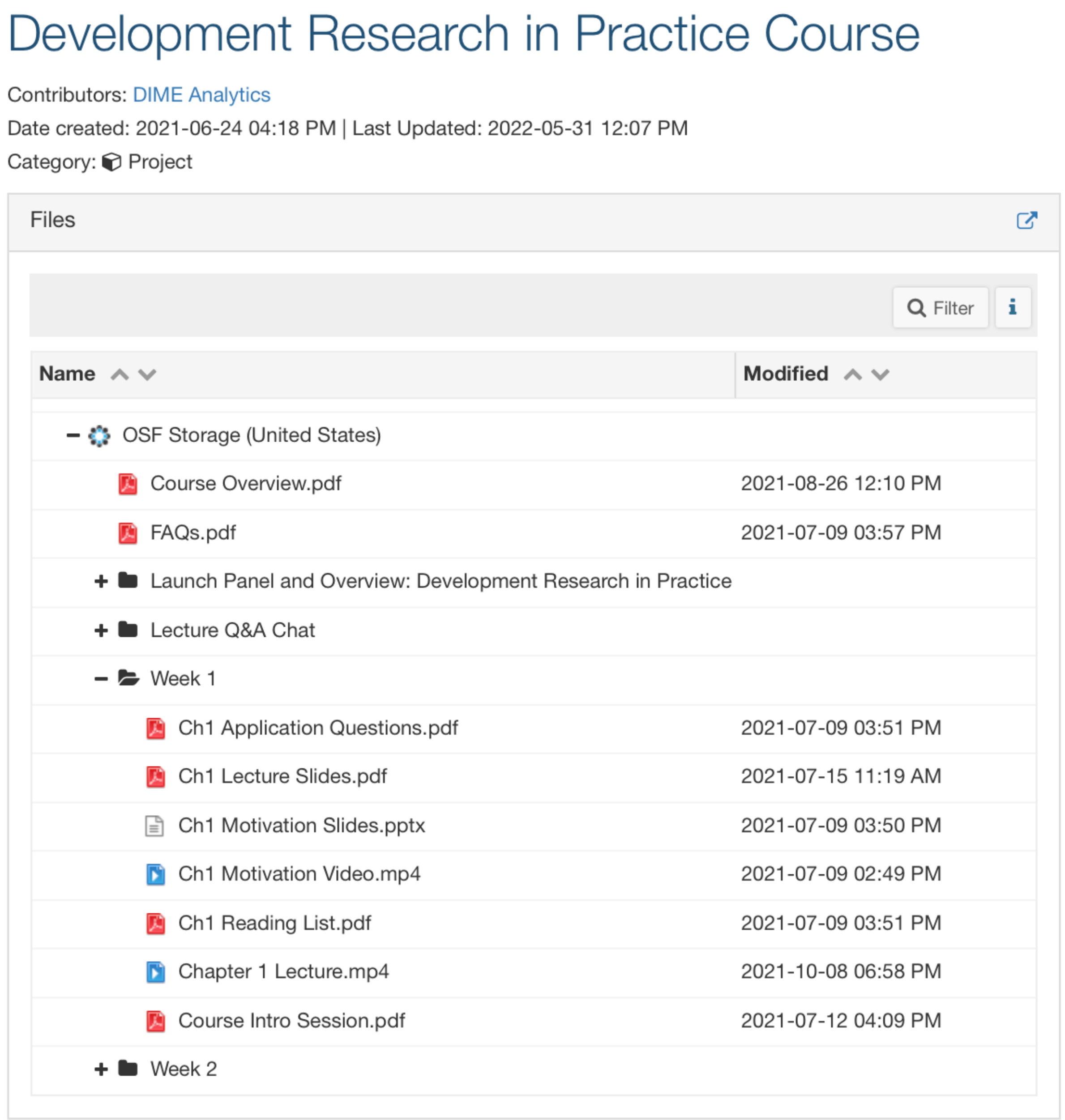
`iefieldkit`. A suite of Stata commands to routinize and document common tasks in primary data collection. Developed at https://github.com/worldbank/iefieldkit.

DIME Analytics GitHub Trainings and Resources. A GitHub repository containing all the GitHub training materials and resources developed by DIME Analytics. The trainings follow DIME's model for organizing research teams on GitHub and are designed for face-to-face delivery, but materials are shared so that they may be used and adapted by others. Hosted at https://github.com/worldbank/dime-github-trainings.

DIME Analytics LaTeX Training. A user-friendly guide to getting started with LaTeX. Exercises provide opportunities to practice creating appendixes, exporting tables from R or Stata to LaTeX, and formatting tables in LaTeX. Available at https://github.com/worldbank/DIME-LaTeX-Templates.

Full self-paced DRiP course

- https://osf.io/6fsz3



Development Research in Practice Course

Contributors: DIME Analytics
Date created: 2021-06-24 04:18 PM | Last Updated: 2022-05-31 12:07 PM
Category: Project

Files

Name	Modified
− OSF Storage (United States)	
Course Overview.pdf	2021-08-26 12:10 PM
FAQs.pdf	2021-07-09 03:57 PM
+ Launch Panel and Overview: Development Research in Practice	
+ Lecture Q&A Chat	
− Week 1	
Ch1 Application Questions.pdf	2021-07-09 03:51 PM
Ch1 Lecture Slides.pdf	2021-07-15 11:19 AM
Ch1 Motivation Slides.pptx	2021-07-09 03:50 PM
Ch1 Motivation Video.mp4	2021-07-09 02:49 PM
Ch1 Reading List.pdf	2021-07-09 03:51 PM
Chapter 1 Lecture.mp4	2021-10-08 06:58 PM
Course Intro Session.pdf	2021-07-12 04:09 PM
+ Week 2	

Thank you!