Introduction to the {targets} package

Reproducible Scientific Workflows in R

Ernest Guevarra

2024-02-05

Masons, when they start upon a building,
Are careful to test out the scaffolding;

Make sure that planks won't slip at busy points,
Secure all ladders, tighten bolted joints.

And yet all this comes down when the job's done
Showing off walls of sure and solid stone.

So if, my dear, there sometimes seem to be
Old bridges breaking between you and me

Never fear. We may let the scaffolds fall
Confident that we have built our wall.

- "Scaffolding" by Seamus Heaney, 1939-2013

Outline

- Concepts on scientific workflows
- The `{targets}` package
- Practical session

Concepts on scientific workflows

Concept #1: Reproducibility, reproducibility, reproducibility!

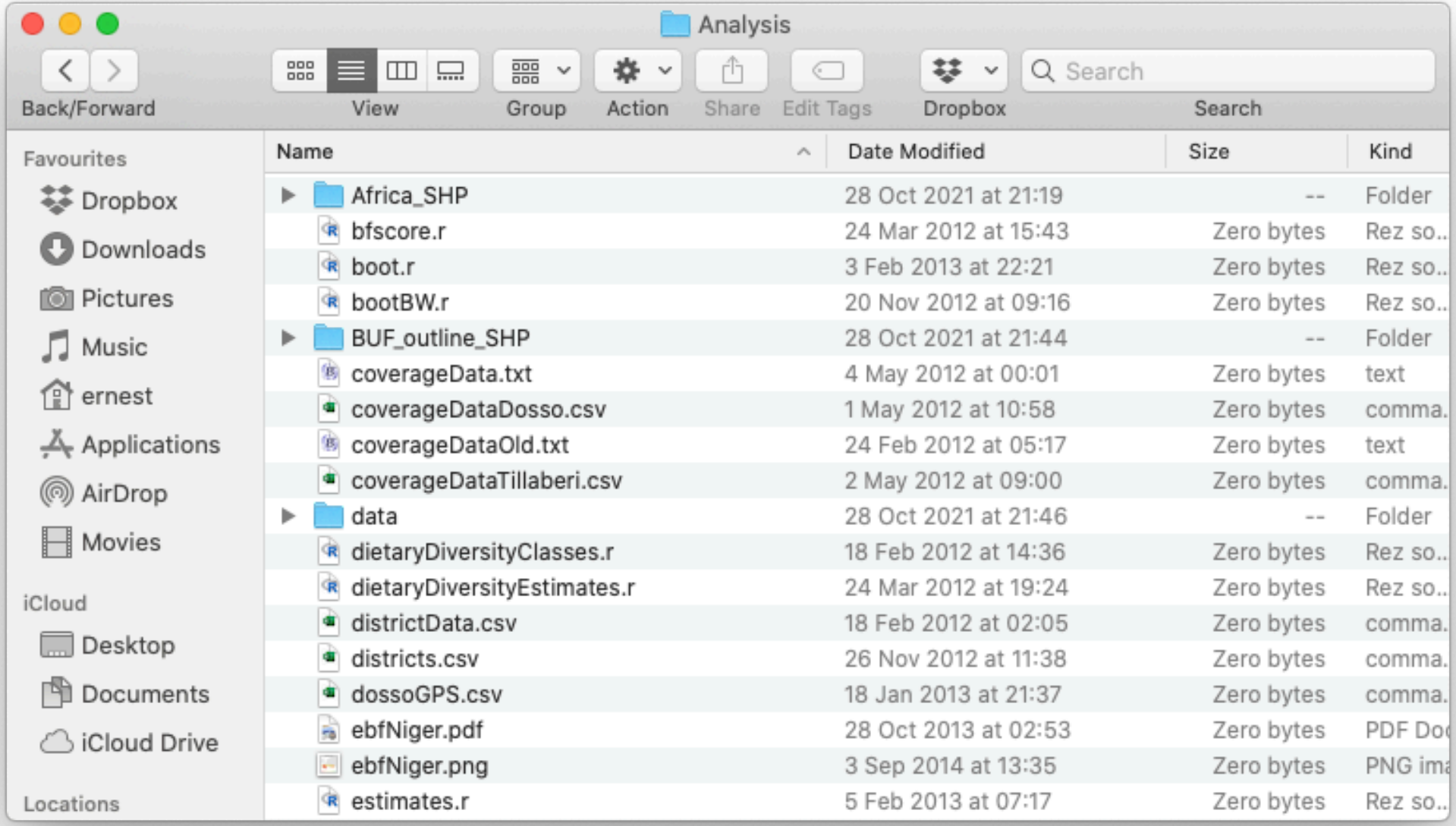
Most important tool for Reproducible Research is the mindset, when starting, taht the end product will be reproducible.

Keith Baggerly, via @kwbroman tweet

Concept #2: Organisation

File organization and naming are powerful weapons against chaos.

@JennyBryan


Analysis
Back/Forward
View
Group
Action
Share
Edit Tags
Dropbox
Search
Search
Favourites
Dropbox
Downloads
Pictures
Music
ernest
Applications
AirDrop
Movies
iCloud
Desktop
Documents
iCloud Drive
Locations
Name
Date Modified
Size
Kind
Africa_SHP 28 Oct 2021 at 21:19 -- Folder
bfscore.r 24 Mar 2012 at 15:43 Zero bytes Rez so..
boot.r 3 Feb 2013 at 22:21 Zero bytes Rez so..
bootBW.r 20 Nov 2012 at 09:16 Zero bytes Rez so..
BUF_outline_SHP 28 Oct 2021 at 21:44 -- Folder
coverageData.txt 4 May 2012 at 00:01 Zero bytes text
coverageDataDosso.csv 1 May 2012 at 10:58 Zero bytes comma.
coverageDataOld.txt 24 Feb 2012 at 05:17 Zero bytes text
coverageDataTillaberi.csv 2 May 2012 at 09:00 Zero bytes comma.
data 28 Oct 2021 at 21:46 -- Folder
dietaryDiversityClasses.r 18 Feb 2012 at 14:36 Zero bytes Rez so..
dietaryDiversityEstimates.r 24 Mar 2012 at 19:24 Zero bytes Rez so..
districtData.csv 18 Feb 2012 at 02:05 Zero bytes comma.
districts.csv 26 Nov 2012 at 11:38 Zero bytes comma.
dossoGPS.csv 18 Jan 2013 at 21:37 Zero bytes comma.
ebfNiger.pdf 28 Oct 2013 at 02:53 Zero bytes PDF Doc
ebfNiger.png 3 Sep 2014 at 13:35 Zero bytes PNG ima
estimates.r 5 Feb 2013 at 07:17 Zero bytes Rez so..

Concept #3: DRY - Don't repeat yourself

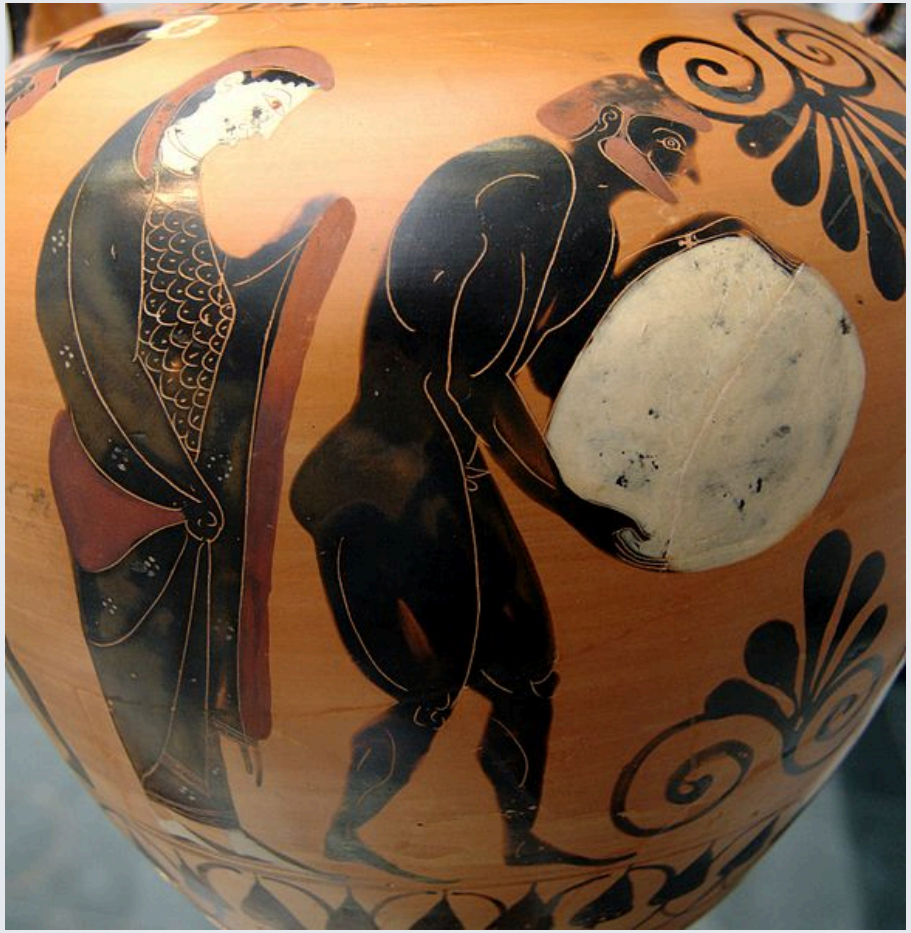
Don't repeat yourself. It's not only repetitive, it's redundant, and people have heard it before.

Lemony Snicket

"You should consider writing a function whenever you've copied and pasted a block of code more than twice (i.e. you now have three copies of the same code)"

Challenges with scientific workflows

Sisyphean loop

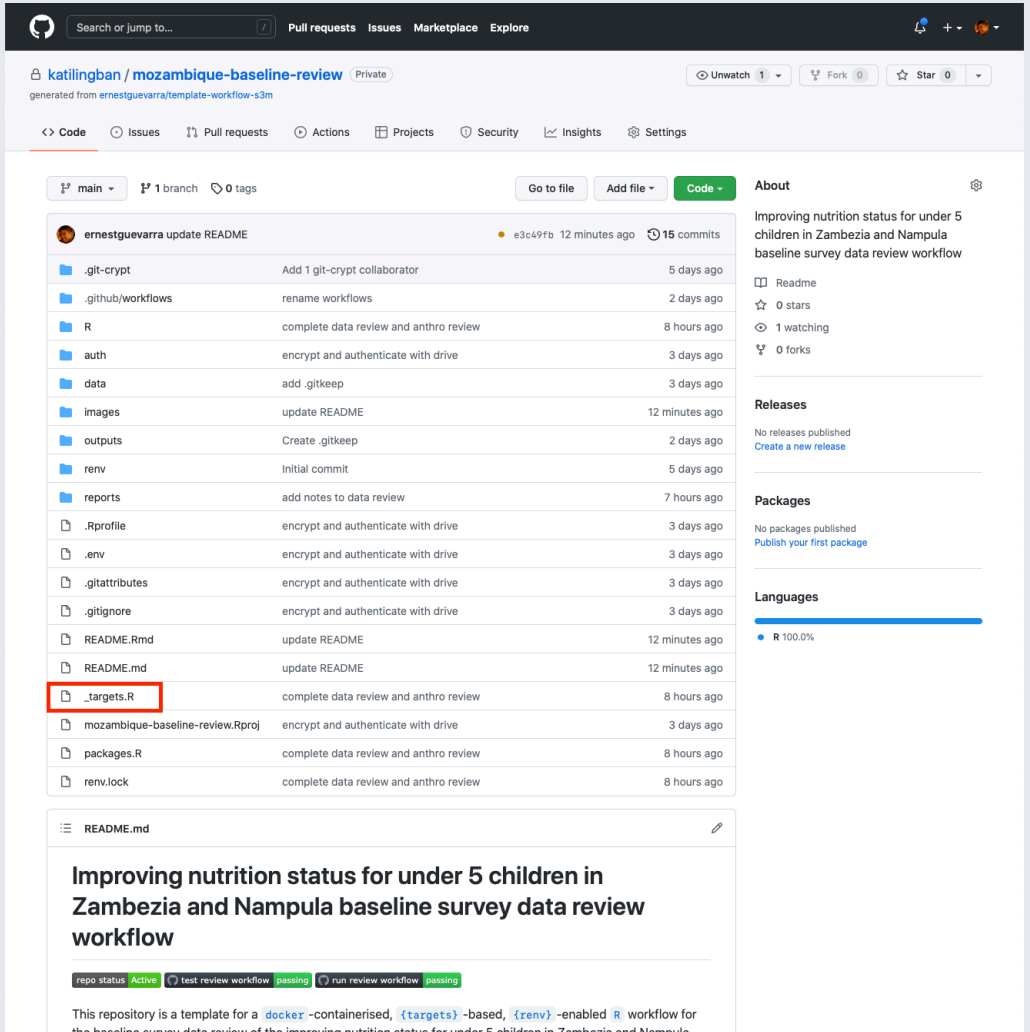


1. Launch the code.
2. Wait while it runs.
3. Discover an issue.
4. Restart from scratch.

The {targets} package

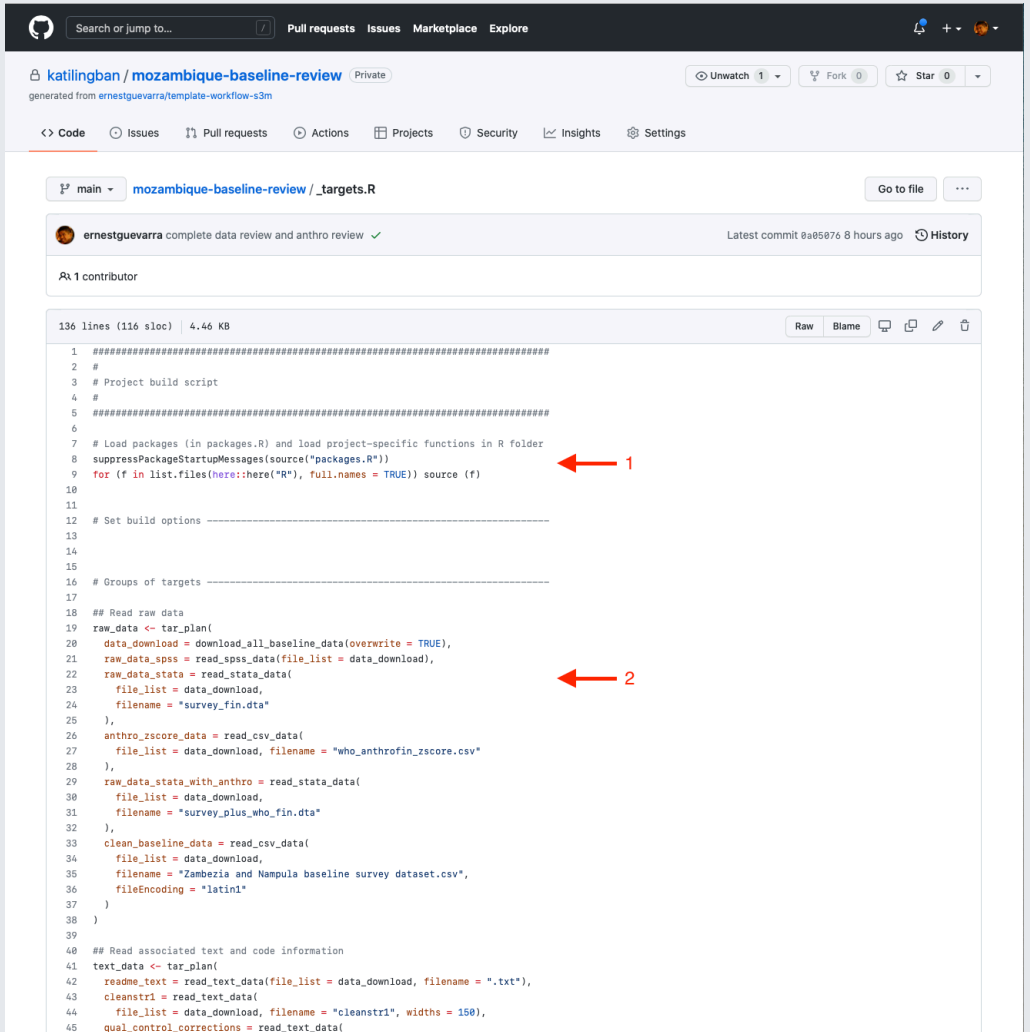
- a pipeline toolkit for Statistics and data science in R
- maintain a reproducible workflow without repeating yourself
- learns how your workflow fits together
- skips costly runtime for tasks that are already up-to-date
- runs only the necessary computation
- supports implicit parallel computing
- abstracts files as R objects
- shows tangible evidence that the results match the underlying code and data

{targets} file organisation



- this is a typical file structure with user-defined components of any project-oriented workflow
- the `_targets.R` file, however, is special and specific to a `{targets}` workflow - it is the target script file
- the target script file should be in the project's root directory

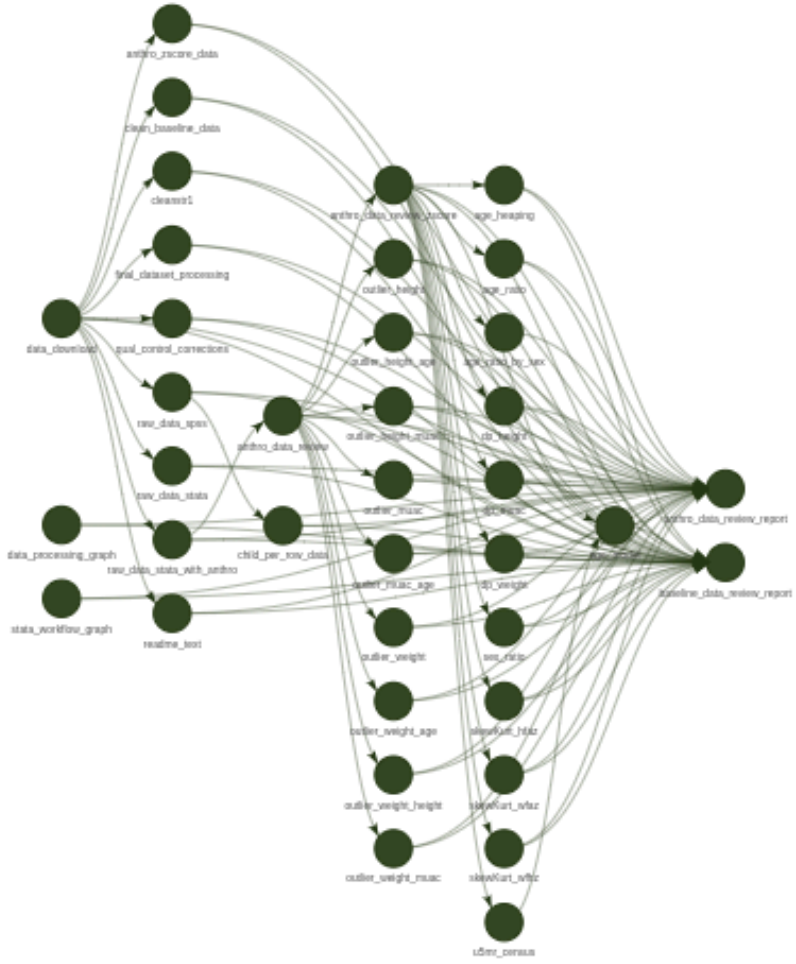
{targets} script file




```
##############################################################################
#
# Project build script
#
##############################################################################

# Load packages (in packages.R) and load project-specific functions in R folder
suppressPackageStartupMessages(source("packages.R"))
for (f in list.files(here::here("R"), full.names = TRUE)) source (f)


# Set build options ------------------------------------------------------------



# Groups of targets ------------------------------------------------------------

## Read raw data
raw_data <- tar_plan(
  data_download = download_all_baseline_data(overwrite = TRUE),
  raw_data_spss = read_spss_data(file_list = data_download),
  raw_data_stata = read_stata_data(
    file_list = data_download,
    filename = "survey_fin.dta"
  ),
  anthro_zscore_data = read_csv_data(
    file_list = data_download, filename = "who_anthrofin_zscore.csv"
  ),
  raw_data_stata_with_anthro = read_stata_data(
    file_list = data_download,
    filename = "survey_plus_who_fin.dta"
  ),
  clean_baseline_data = read_csv_data(
    file_list = data_download,
    filename = "Zambezia and Nampula baseline survey dataset.csv",
    fileEncoding = "latin1"
  )
)

## Read associated text and code information
text_data <- tar_plan(
  readme_text = read_text_data(file_list = data_download, filename = ".txt"),
  cleanstr1 = read_text_data(
    file_list = data_download, filename = "cleanstr1", widths = 150),
  qual_control_corrections = read_text_data(
```

- Load the packages required (1)
- Load custom functions (1)
- Define individual targets - intermediate step of the workflow (2)
- End with a list of targets objects

{targets} workflow



Questions?

Practical session

We will all continue to go through Exercise #1 in the Practical R for Epidemiologists book

Questions?

Thank you!

Slides can be viewed at https://oxford-ihtm.io/open-reproducible-science/session8.html

PDF version of slides can be downloaded at https://oxford-ihtm.io/open-reproducible-science/pdf/session8-reproducible-scientific-workflows.pdf

R scripts for slides available here