gesis

Leibniz Institute
for the Social Sciences

Workflows for Reproducible Research with R & Git

Recap & Outlook

Johannes Breuer, Bernd Weiss, & Arnim Bleier

2023-11-17

Leibniz Association

Recap - Day 1

Time	Topic
09:30 - 10:45	**Introduction**
10:45 - 11:00	Coffee Break
11:00 - 12:00	**Computer literacy**
12:00 - 13:00	Lunch Break
13:00 - 15:00	**Git & GitHub - Part 1**
15:00 - 15:15	Coffee Break
15:15 - 16:30	**Git & GitHub - Part 2**
16:30 - 17:00	**Q & A**

Recap - Day 2

Time	Topic
09:00 - 09:30	Recap Day 1
09:30 - 11:00	Dependency management
11:00 - 11:15	Coffee Break
10:15 - 12:00	Build your own Binder
12:00 - 13:00	Lunch Break
13:00 - 14:30	Binder & Notebooks
14:30 - 14:45	Coffee Break
14:45 - 16:00	Saving computational environments
16:00 - 17:00	Recap & Outlook

Other topics in reproducible research

As we said in the introduction, we cannot cover all tools and topics related to reproducible research in this workshop. However, we want to use this session to at least mention and provide pointers to resources for some of these topics. Things we did not cover (TWDNC):

- Reproducible documents/papers
- Data sharing
- Collaboration with `Git` & *GitHub*
- Alternatives to *Docker*
- Alternatives to *Binder*
- Workflow & project templates

TWDNC: Reproducible documents

There are different formats for literate programming that combine text, code, and output, such as:

- `R Markdown` (`.Rmd`)
  - Although not as widely used, there also is the format of `R` Notebooks.
- Jupyter Notebooks (`.ipynb`)
- Quarto (`.qmd`)

Note: Jupyter Notebooks, `Quarto` (depending on the output format), and `R` Notebooks (can) also add interactivity.

TWDNC: Reproducible documents

Some helpful resources for getting started or digging deeper ⛏:

- `R Markdown`
  - our slides from last year's workshop
  - R Markdown: The Definitive Guide by Yihui Xie, J. J. Allaire, & Garrett Grolemund
  - R Markdown Cookbook by Yihui Xie, Christophe Dervieux, & Emily Riederer
- `Quarto`
  - materials from the workshop "Automated Reports & Co with Quarto and Markdown" by David Schoch & Chung-hong Chan
  - Quarto: The Definitive Guide by Mine Çetinkaya-Rundel and Charlotte Wickham

TWDNC: Data sharing

Typically, there are different research products that you can share:

- a publication
- code used for data processing and analysis (= focus of this workshop)
- research data
- other materials (e.g., stimuli, questionnaires, etc.)

TWDNC: Data sharing

Studies have repeatedly shown that "sharing upon request" is not a sustainable solution for data sharing. And we have discussed why *GitHub* is not a good place for sharing research data. However, there are other more suitable options for sharing research data. The best one is the use of dedicated repositories for research data.

The paper by Klein et al. (2018) provides an overview of public repositories that hold psychological data.[1] A good tool for finding suitable repositories is the *Registry of Research Data Repositories*.

[1] However, parts of this overview have inevitably become somewhat outdated since the paper was published.

TWDNC: Data sharing

Some good options for sharing research data (as everything, each with their own pros and cons):

- General purpose (data) repositories, such as the *Open Science Framework* (OSF), *Zenodo*, or *Harvard Dataverse*
- Curated discipline-specific archives, such as the *GESIS Data Archive*, *PsychArchives* by *ZPID*, or *ICPSR*
- Archives for specific types of data, such as *Qualiservice* or *The Qualitative Data Repository*

TWDNC: Collaboration with `Git` & *GitHub*

Typically, you collaborate with others on research projects. `Git` and *GitHub* can also be used for this purpose. However, this requires that all collaborators are willing and able to do so.

For some guidance on how `Git` & *GitHub* can be used for collaboration, you can have a look at our slides on this topic from last year or the slides by Frederik Aust from a similar workshop (taught together with Johannes Breuer).

TWDNC: *GitHub* as a social network 

You can *star* repositories ⭐, *follow* users (or organizations), and have a personalized newsfeed on *GitHub*.

It is also easily possible to contribute to the work of others or have others contribute to your work, e.g., via creating or closing an *issue* or *pull requests*.

TWDNC: Alternatives to *Docker*

- *podman* as an open-source alternative to *Docker*
- using the package manager *Nix* with the `R` package `rix`
  - Bruno Rodrigues wrote a series of blog posts about "Reproducible Data Science with Nix"
  - *note*: only works for Linux and macOS; for Windows you need to use WSL

TWDNC: Alternatives to *Binder*

Note: All of the following alternatives are commercial (but offer limited free use).

- *Code Ocean*
- *Observable*
- *Posit Cloud* (formerly *RStudio Cloud*)

TWDNC: Project setup and templates

In this workshop, we have shown you how to manually set up a reproducible research workflows. However, there are some tools that you can use to automate parts of this process. These can range from very simple to very elaborate solutions.

TWDNC: Project setup and templates

We have already seen and tested the file `create-project.sh` (which is small shell script for initializing a basic project folder structure that can be easily adapted and extended using any text editor). However, there also are several other (more complex) packages and templates that can be used for the creation and maintenance of reproducible research workflows, such as...

- `template` by Frederik Aust & Marius Barth
- `WORCS` - *Workflow for Open Reproducible Code in Science*
- `workflowr`
- `starter` - a toolkit for starting new projects
- `rrtools` - Tools for Writing Reproducible Research in R
- `targets` - Function-oriented Make-like declarative workflows for R

start your lab also provides an `R` Project Template.

Other resources on reproducible research with R

- Building reproducible analytical pipelines with R by Bruno Rodrigues
- Blog post An overview of what's out there for reproducibility with R by Bruno Rodrigues
- Chapter on "Computational Reproducibility" in the book/course "Improving Your Statistical Inferences" by Daniel Lakens
- BERD Course Booklet: Make Your Research Reproducible by Heidi Seibold
- Guide for Reproducible Research by *The Turing Way*

Final note: Showing appreciation 

The creation and maintenance of FOSS takes a lot of time and this is rarely recognized as much as it should be. One thing we can do to change this is to at least give credit where credit is due and cite the tools and resources that we use.

Final note: Showing appreciation 👏

```
citation("rang")
```

```
## To cite rang in publications use:
##
##   Chan C, Schoch D (2023). "rang: Reconstructing reproducible R computational environments
##   doi:10.1371/journal.pone.0286761 <https://doi.org/10.1371/journal.pone.0286761>,
##   <https://github.com/gesistsa/rang>.
##
## Ein BibTeX-Eintrag für LaTeX-Benutzer ist
##
##   @Article{,
##     title = {rang: Reconstructing reproducible R computational environments},
##     journal = {PLOS ONE},
##     author = {Chung-hong Chan and David Schoch},
##     url = {https://github.com/gesistsa/rang},
##     year = {2023},
##     doi = {10.1371/journal.pone.0286761},
##   }
```

FOSS is boss 😎

"Open source is a hard requirement for reproducibility" (Bruno Rodrigues)

Looking back

You created a *GitHub* repository containing materials for a fully reproducible research pipeline! 🎉

If you created a public *GitHub* repository: Head over to http://starlogs.net/ and paste the URL of the repository to recap your heroic journey into the universe of reproducible research! 🌌

The path to reproducibility

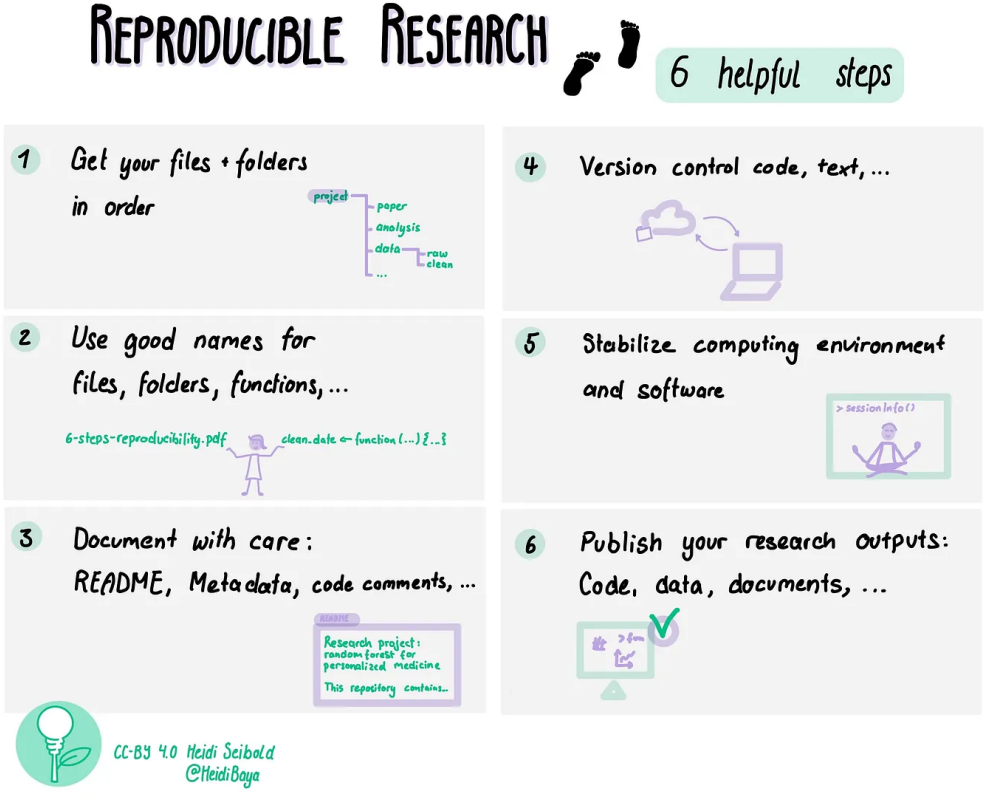

Illustration by Heidi Seibold from her newsletter/*Substack* blog post "Document with care: README, Metadata, code comments, … to make both others and future you happy"

Any other questions or things you want to say/discuss?

Looking forward

We hope that we could get you started or help you with with making your research (more) reproducible. Of course, as always, there is much more to explore and learn. The only way to really get familiar with the tools and workflows is if you use them for your own research.

And remember that making your research (more) reproducible is an incremental process. Every step towards reproducibility is an improvement 👣. You don't have to (and probably should not) leap to a full `R` + `Git` + `Docker` workflow all at once.

Keep calm and stay reproducible! 😊

Thank you very much for participating in this workshop!

We hope that you learned something and also had some fun (at least a little bit...)