Tips for Sharing Data and Code

Lisa DeBruine

2022-03-17

Contents

If you're sharing data or code with a paper, here are a few tips to make sure your resources are as useful as possible. People have variable experience with data and code sharing, so this document provides tips with different levels of complexity and links for further in-depth tutorials. Do as much as you have time and expertise for, and build on your skills in future projects.

> Computational reproducibility leads to more transparent and accurate research. ... fear of a crisis and focus on perfection should not prevent curation that may be 'good enough.' (Sawchuk and Khair 2021)

Data

Save it in an accessible format

- Use tab-separated value (.tsv) or comma-separated value (.csv) files
- Use UTF-8 (or UTF-16) encoding to avoid problems in an international context (e.g., so characters like ü or é aren't mangled)

Excel is less preferable because of the proprietary format and its tendency to mangle anything that resembles a date. SPSS and other proprietary formats are also not ideal, but data in a proprietary format is better than no data.

Include a codebook

- Beginner: include a text file with each column name and an explanation
- Intermediate: You can include a data dictionary with further info like data type (string, numeric) or possible ranges/values (Buchanan et al. 2021)
- Intermediate: Use `faux::codebook()` to make a machine-readable codebook in PsychDS format
- Intermediate: Use the R codebook to include a report with detailed metadata (Horstmann, Arslan, and Greiff 2020)

Ethical Sharing

- Check that you are not sharing any identifiable data (without clear consent), such as names, student ID numbers, postcodes, IP addresses, or uniquely identifying combinations of demographic variables.
- Add a license so others know how they can use the data. he most common licenses for data are:
  - CC-0: Waives all rights and releases work to public domain
  - CC-BY: By Attribution, which permits sharing and reuse of the material, for any purpose, as long as the original authors are credited
  - CC-BY-SA: By Attribution, with a Share-Alike clause which means that anyone sharing or modifying the original work must release it under the same license

Make it findable

- Use a persistent archive to host your data, like the OSF, figshare, or zenodo. These platforms are free and can give your data a DOI.
- Include the citation info in a README
- Remember to make the data accessible for reviewers before submission. The OSF allows you to create a blinded review-only link.
- Make the data accessible to the public before publication.
- Make sure the paper contains the correct links to the data before publication.

Sharing Code

Make it reproducible

- Include all the external files (e.g., data files) needed to run it
- Use relative paths so that it can run on any computer
- Set a seed if your code uses simulations or random number generation

Make it clear

- Include a README that explains how to run the code
- Assume that the audience has varying technical expertise and doesn't necessarily know the conventions of the language you're using
- Indicate which parts of your code produce any figure, table, or value in your manuscript
  - Beginner: include comments in your code like `# produces Figure 3.1`
  - Intermediate: Use code the generate the text of the results section so you only have one thing to copy and paste into the manuscript
  - Advanced: Use code to generate the entire manuscript (e.g., using the papaja package)

Indicate software versions

- Beginner: Include a text file with the info, e.g., devtools::session_info() in R, or requirements.txt in python
- Intermediate: Or use dependency management, like renv or packrat in R
- Advanced: Or use containers like Binder, Docker or CodeOcean for full reproducibility

Confirm reproducibility

- Beginner: Access your shared materials from a new computer and run the code
- Intermediate: Or ask a colleague to try running your code on their machine
- Intermediate: Or set up more formal code review (Vable, Diehl, and Glymour 2021) in your group
- Advanced: Or use a service like CodeCheck

NB: If your code takes a very long time to run, such as when you have extremely large datasets or are running simulations, you can include smaller test datasets or run a smaller number of replications and include code at the top of the script to toggle real and test data or low and high numbers of reps.

Finally, make your code findable using the same tips from the data section above. GitHub is a common place to share code, but doesn't create a DOI, so if you use github, consider archiving a snapshot of your repository on zenodo.

Further Resources

Thank to everyone who responded to my tweet about this topic. Many of the tips and links are from their comments.

- Tools for Reproducible Research course by Karl Broman
- How to name a file: Interoperability considerations

References

AlNoamany, Yasmin, and John A Borghi. 2018. "Towards Computational Reproducibility: Researcher Perspectives on the Use and Sharing of Software." *PeerJ Computer Science* 4: e163. https://doi.org/10.7717/peerj-cs.163.

Buchanan, Erin M, Sarah E Crain, Ari L Cunningham, Hannah R Johnson, Hannah Stash, Marietta Papadatou-Pastou, Peder M Isager, Rickard Carlsson, and Balazs Aczel. 2021. "Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set." *Advances in Methods and Practices in Psychological Science* 4 (1): 2515245920928007. https://doi.org/10.1177/2515245920928007.

Horstmann, Kai T, Ruben C Arslan, and Samuel Greiff. 2020. "Generating Codebooks to Ensure the Independent Use of Research Data: Some Guidelines." https://doi.org/10.1027/1015-5759/a000620.

Sawchuk, Sandra L, and Shahira Khair. 2021. "Computational Reproducibility: A Practical Framework for Data Curators." *Journal of eScience Librarianship* 10 (3): 7. https://doi.org/10.7191/jeslib.2021.1206.

Vable, Anusha M, Scott F Diehl, and M Maria Glymour. 2021. "Code Review as a Simple Trick to Enhance Reproducibility, Accelerate Learning, and Improve the Quality of Your Team's Research." *American Journal of Epidemiology* 190 (10): 2172–77. https://doi.org/10.1093/aje/kwab092.