Appendix

Patrick Charbonneau: Data Sharing Panel

Patrick Charbonneau conducted his undergraduate study at McGill, in Montréal, obtained his Ph.D. in chemical physics from Harvard University in 2006 and then was a Marie-Curie Postdoctoral Fellow at Amolf, in Amsterdam, before joining Duke in 2008, where he is now professor of chemistry and physics.

Practices: How have you integrated data sharing and archiving into your workflows or research practices?

Dr. Charbonneau: When I started at Duke, my archival strategy consisted of waiting until researchers in my group were about to leave before having them save and document their data to a backed up departmental drive. Because that approach was a marked improvement over how I managed research material during my own graduate and postdoc years, I was proud to disclose it in early NSF data management plans (Pasek 2017). The passing years, however, revealed its weaknesses. Out of the blue, I would receive requests for data whose creator had long since left Duke. Handling these requests was stressful because our retroactive and last-minute metadata practices were far from comprehensive. Fortunately, because former students had remained in touch we muddled through, but that was more luck than forethought. Before luck ran out, I sought to implement a model in which (i) former affiliates (or myself) would need not be involved, and (ii) archiving could be handled on a per-publication basis, when everything is fresh. Since 2016, with the help of Duke Libraries, my group has thus developed a protocol by which each publication could cite a DOI-labeled repository containing code, data, and archival-quality metadata. The key to success is having this material at hand by the time the publication is accepted, then promptly passing the data on to the Duke Research Data Repository curators and inserting the repository info in the article at the proofing stage. To make sure this choreography goes smoothly it is now an expectation in my group that in parallel to manuscript preparation, raw data and codes are also getting ready for public release. Interestingly, this routine also provides an additional validation route of the research data before publication.

Perceptions: What are the key challenges or opportunities that everyone in research in your field (or any field) should know about related to data sharing/reproducibility? What do you feel are the most effective ways to engage around the topic of data sharing/reproducibility?

Dr. Charbonneau: Shifting from the old to the new model of data archiving mostly involved overcoming a logistical (or cultural) barrier. There was nothing fundamentally new in our values or intents. Setting a protocol in place, with curatorial oversight, merely kept our practices in line with these ideals. The challenge is therefore to do it for the first time. Afterwards, for students and postdocs this practice becomes the *proper* way to publish research findings.

The key practical question is deciding what to include. All raw data or only (partially) processed data? Should one be concerned as to whether this data might or might not be reused? There’s no right answer to these questions, but setting small goals facilitates getting started, which I feel is the most important consideration. Aiming for the lowest common denominator also broadens disciplinary relevance. Scholars from history to biology to economics and literature publish plots with points, whose coordinates could be deposited. After a few deposition cycles, the scope can grow and specialize, in response to collegial requests or otherwise. It’s also possible to go back and extend the content of an initial deposition. Version control makes that completely transparent.

The most effective—non-coercive—way to encourage the use of repositories is emphasizing that depositing research products comes at essentially no cost and offers key benefits. Research productivity does not suffer because depositing research products is interwoven into the manuscript and figure preparation process. It mostly reorders some of the steps most researchers already take. The benefits, however, are substantial. As mentioned above, it provides an additional guardrail for pre-publication data handling, and it reduces the need to handle individual data requests. It can also improve the long-term impact of a research effort. Who knows what use any given particular dataset could have for future researchers? I at least know that some of my requests for old codes and data have been unfulfilled because of agedness, of either technology or researchers.

Angela Zoss: *Reproducibility Panel*

Angela Zoss holds a Master's degree in Communication from Cornell University and a Ph.D. in Information Science from Indiana University, where she also completed her undergraduate study. She has worked at Duke University since 2012, first as a Data Visualization Coordinator and now as an Assessment and Data Visualization Analyst in the Duke University Libraries.

Practices: How have you integrated reproducibility into your workflows or research practices?

Dr. Zoss: From 2012 to 2018, I was working a full-time job while trying to conduct and complete my doctoral dissertation research. I found it difficult to reserve significant and consistent blocks of time for my research, and my committee members and I would struggle to remember what previous decisions had been made about data collection and analysis. As I had been learning and teaching R during my "day job," I decided to use my dissertation research as an opportunity to improve my proficiency with that tool.

I started by using R to analyze the results of a simple pilot survey. When I realized the many benefits of a fully reproducible analysis workflow, I sought out a framework that would help me organize the various data and script files I was generating. Someone recommended the TIER Protocol[1] to me, and I found it to be a wonderful fit for the kind of analysis work I was doing. Combining R projects and the TIER Protocol, I was able to generate a series of analysis workflow steps that took me all the way from the preparation of components included in my main experiment through the data cleaning of participant responses to the final data modeling and visualizations for my dissertation.

After completing my Ph.D., I have continued to use reproducible workflows in my full-time position as an assessment and data visualization analyst. Any project requiring data cleaning, blending, visualization, and publication is made much simpler by standard reproducibility practices. I have continued to use R for many projects, but I also prioritize other types of reproducible workflows like command-line scripts or tools with exportable steps when R is not appropriate. These practices benefit not only my project partners and me, but also a much broader community. One large project that required gathering and blending data from a variety of sources was recently presented at a national conference. After the presentation, I was immediately approached by someone who thanked me for sharing the files openly on GitHub.

[1] https://www.projecttier.org/tier-protocol/

Perceptions: What are the key challenges or opportunities that everyone in research in your field (or any field) should know about reproducibility? What do you feel are the most effective ways to engage with the topic of reproducibility?

Dr. Zoss: One challenge I see for reproducibility is that exposure to these practices was not centralized in my graduate program, which meant that my understanding of best practices was piecemeal. I was training in a social science discipline where formal statistical analysis and programming for research projects were not widespread practices. I hope in the intervening years these concepts have begun diffusing into all disciplines, but as a new graduate student, finding advice on reproducibility best practices involved haphazard searching and filtering out materials designed either for much more technical work, such as formal software development projects, or for STEM/health fields.

The core principles of reproducibility that I think apply to all fields are as follows: reproducible practices improve research quality, even if the research is never reproduced; reproducibility requires transparency throughout the entire research process; and reproducibility is better served by simple, sustainable documentation than by reliance on complex and ever-changing software. Finding ways to engage those principles, regardless of your discipline or the tools you use for your research, leads to a less stressful, more productive, and higher quality research experience. I highly recommend using a small project early in your research career to explore what reproducibility means for your work, and then building on that with ideas that come from your own experience as well as that of others in similar fields.