Open data, open methods, reproducible and sustainable results

By: Dr Brian Ballsun-Stanton, Data Scientist, Research Fellow in the Department of Ancient History and Technical Director of the FAIMS Project, Macquarie University; Ms Kate Carruthers, Chief Data Officer, UNSW Australia and Adjunct Senior Lecturer, UNSW School of Computer Science and Engineering

The views in this letter are those only of the researchers above and do not necessarily represent those of their institutions.

22 August 2016

To the committee producing the National Research Infrastructure Roadmap,

Much of Australian research data is stashed away in desk drawers, faculty file systems, and in mouldering and unmaintained websites. Of the mandated "open data" and data management plan provisions in grant-making bodies, very few researchers see those clauses as anything more than a pro forma obligation. A long history of this lackadaisical approach has contributed to the world-wide replication problem currently being discussed in psychology. (Open Science Collaboration 2015)

We urge this committee, when considering funding and priorities to do more than provide for "write-once, read-never" data management plans. There are multiple facets of the problem: data repositories, methodological repositories, long term repository maintenance and availability plans, and "end of days " plans. There are some larger scale logistical problems which could form the basis of grand projects[1] -- though proposing them is outside the scope of this letter.

Data and methodological repositories

A useful rule to be considered here is "research produced with public funds should (with suitable privacy protections) be fully available to those who paid for it". Research consists of the methodologies to collect data, the data, methods by which the data was analysed, software, and the instrumentation design which allowed the data to be collected and stored. Satisfaction of the term "available" requires more than publishing a few spreadsheets uploaded before a paper is released. For data to be "available" it must be understandable and *extensible*.

Data extensibility requires the entire data lifecycle to be documented and re-usable by other researchers without assistance from the original scholars. Only if the original data set is e-planned in sufficient detail to be extensible should the committee consider the data to be open to reuse.

This committee is in a position to create mandates and requirements of use that will help create a culture of open documentation. One example of an effort in this direction is the "protocols.io" service. Edmunds (2016) notes that "Currently, the most common way of presenting methods in articles is in extremely brief paragraphs as supplemental downloadable PDF files. The result is often incomplete or non-discoverable methodology, which is key for scientists to properly build on

[1] Some random titles for inspiration: "Discovering the discoveries: accession and rediscovery of instutitional research archives," "The Humanities Reproducibility Project," or most anything discussed by Vernor Vinge in his novel *Rainbows End*.

scientific discovery." While parasitologists are currently benefiting from sharing methods on protocols.io, many of the landmark projects discussed last friday in Sydney seem to give no indication of being interested in heading in this direction.

In the interests of reproducibility and open participation in the arts and the sciences, we propose that all projects funded by the roadmap:

1. Are directed to publish complete and original data in suitable repositories with sufficient description such that other facilities with access to tools can *contribute meaningfully* to the data in the dataset;
2. Are directed to publish their methodologies[2] in complete and sharable format, in the paradigm of protocols.io. Of note, a research methodology should encompass both data collection and analysis. This allows peer review to be conducted before the data is available (Findley et al. 2016) and is a way to reduce publication bias and to encourage more researchers to use extant analyses for novel questions (Lakatos 1978);
3. Are encouraged to share tool schematics and other advances in hardware and software instrumentation in the same spirit, much like CERN's Open Hardware Repository (http://www.ohwr.org/); and
4. Are encouraged to publish peer-reviewed data papers: "A data paper is a searchable metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal." ("Data Papers" 2013)

By acting as an advocate for open research, leading the way in the fight against publication bias, and creating a market for (hopefully open access and non-profit) data paper publishers, this committee is in a position to make significant positive impacts on research culture in Australia and the world.

Repository Maintenance and the "End of days"

The second major point we wish to raise is the difficulty of data storage for the long term. It is one thing to require all data and methods be "open." It is another to take steps to ensure that researchers are able to use and access the data in 2,5,10, and 100 years. One successful data repository is the UK's "Archaeology Data Service" (ADS) which owes at least some of its success to regulatory capture[3].

While advocating regulatory capture for field-specific data repositories is outside the scope of the committee, broadening the scope (and funding) of state libraries to allow and encourage capture *all products of their state's researchers*, will allow them to serve as archives of last resort. Beyond that, requiring that applicants address how their data will be available in 5-20 years will help to refine and create opportunities worldwide for dataset repositories.

One item when assessing landmark proposals should be long term plans for data curation. In our experience, severe security issues with any major repository crop up around twice a year. Applicants should be aware of these issues and budget system administration accordingly[4].

[2] Documentation of the *choices* made during methodology creation is also important. The factors which influence model making are of huge import for any predictive model. (Hourdin et al. 2016)

[3] See point #5 "Meet governmental requirements" http://archaeologydataservice.ac.uk/advice/WhyDeposit#section-WhyDeposit-WhyDepositData

[4] We found that it took 1 system administrator 1 day a week (on average) to keep up the load of the FAIMS repository. System administrators scale fairly well, but any price point below the tens of thousands per year should be treated with skepticism. An alternative to requiring funding is to require participating institutions to sign a "service level agreement" for data availability for the expected lifetime of their data.

Another innovation from the ADS is an "end of days fund" (Richards 2012): a portion of all receipts they collect is allocated towards this fund which allows for the complete, professional, and orderly migration of all of their data to a national library or other major archival repository. Thus, even if the repository dies, the data is preserved for future researchers.

Role of Government

There is a role for Government. It is not necessarily in the doing regarding data, instead it is largely in providing a legislative and regulatory framework. However, there are several key initiatives that Government could and should consider.

Government open data repositories are insufficient. Universities should also be funded to provide data repositories that are linked into a national collaborative network, that would provide an open data network for researchers. The data must have relevant metadata stored alongside and the metadata needs to be stored by the National Library in a searchable form.

A service that could be offered by Government agencies, such as the NSW Data Analytics Centre, on a fee for service basis is de-identification as a service. Numerous researchers have no idea of how to effectively de-identify data and a service such as this would provide enormous benefit.

Additionally, the role of repositories for the code that is associated with the generation of data is critical for reproducibility. It is a critical resource and one which ought to be considered. Possibly the Australian National Data Service (ANDS) could offer this as a service along similar lines as the *Cite My Data* service for Digital Object Identifiers (DOIs).

Conclusion

In conclusion, we commend these suggestions to you. We recognise the strength and commitment of the Australian research community. We believe that the suggestions outlined above will support them in the creation of open data, using open methods, with reproducible and sustainable results for decades to come.

Sincerely,

Brian Ballsun-Stanton

Kate Carruthers

Citations

“Data Papers.” 2013. *GBIF.ORG*. August 19. http://www.gbif.org/publishing-data/data-papers.

Edmunds, Scott. 2016. “Reproducible Research Resources for Research(ing) Parasites.” *GigaBlog*. June 3. http://blogs.biomedcentral.com/gigablog/2016/06/03/reproducible-research-resources-researching-parasites/.

Findley, Michael G., Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. 2016. “Can Results-Free Review Reduce Publication Bias? The Results and Implications of a Pilot Study.” *Comparative Political Studies*, July. doi:10.1177/0010414016655539.

Hourdin, Frederic, Thorsten Mauritsen, Andrew Gettelman, Jean-Christophe Golaz, Venkatramani Balaji, Qingyun Duan, Doris Folini, et al. 2016. “The Art and Science of Climate Model Tuning.” *Bulletin of the American Meteorological Society* 0 (0): null.

Lakatos, I. 1978. “Science and Pseudoscience.” presented at the BBC Radio Talk. http://www.lse.ac.uk/philosophy/department-history/science-and-pseudoscience-overview-and-transcript/.

Open Science Collaboration. 2015. “PSYCHOLOGY. Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716.

Richards, Julian. 2012. “Funding the ADS.” presented at the FAIMS Stocktaking Workshop, Sydney. https://www.fedarch.org/workshop%20presentations/sustainability-strategies/.