BEST PRACTICES FOR COMPUTATIONAL SCIENCE: SOFTWARE INFRASTRUCTURE AND ENVIRONMENTS FOR REPRODUCIBLE AND EXTENSIBLE RESEARCH

Victoria Stodden
Department of Statistics
Columbia University

Sheila Miguez
Department of Statistics
Columbia University

September 6, 2013

Abstract

Scholarly dissemination and communication standards are changing to reflect the increasingly computational nature of scholarly research, primarily to include the sharing of the data and code associated with published results. This paper presents a formalized set of best practice recommendations for computational scientists wishing to disseminate reproducible research, facilitate innovation by enabling data and code re-use, and enable broader communication of the output of digital scientific research. We distinguish two forms of collaboration to motivate choices of software environment for computational scientific research. We also present these Best Practices as a living, evolving, and changing document at http://wiki.stodden.net/Best_Practices.

Keywords: best practices, reproducible research, archiving, data sharing, code sharing, wiki, open science, computational science, scientific method

Introduction

Pervasive digitization is changing the practice of science by enabling massive data collection and storage, giving us the tools to carry out and record analyses of these data, and also by providing a mechanism for communicating these new digital scholarly objects through the Internet. The potential is enormous: the technology is available to permit the open communication not only of the results of scientific investigation, but also the tools required to verify, extend, and understand the knowledge.

There is a movement within the computational science community to adapt communication standards to include the data and code associated with published findings, called the Reproducible Research Movement [1]. The July/August 2012 issue of IEEE Computing in Science and Engineering was focused on Reproducible Research [2] and called for "changing the culture" of scientific research [3]. A Roundtable at Yale Law School in 2009 focused on the issue of reproducibility by bringing together computational scientists from many different disciplines [4,5]. Over the past few years many editorials and commentaries have continued these appeals [6,7,8,9,10]. Without the data and computer codes that underlie scientific discoveries, published findings are all but impossible to verify.

Computational results are frequently of a complexity that makes a complete enumeration of the steps taken to arrive at a result prohibitive in typical scientific publications today [11]. As noted in 2009,

> At conferences and in publications, it's now completely acceptable for a researcher to simply say, "here is what I did, and here are my results." Presenters devote almost no time to explaining why the audience should believe that they found and corrected errors in their computations. The presentation's core isn't about the struggle to root out error — as it would be in mature fields — but is instead a sales pitch: an enthusiastic presentation of ideas and a breezy demo of an implementation. Computational science has nothing like the elaborate mechanisms of formal proof in mathematics or meta-analysis in empirical science. Many users of scientific computing aren't even trying to follow a systematic, rigorous discipline that would in principle allow others to verify the claims they make. How dare we imagine that computational science, as routinely practiced, is reliable! [1]

A necessary response to this crisis is the adoption of the practice of reproducible computational research, in which all details of the computations — the underlying data and the code that generated the results — are made conveniently available to others.

In this document we envision a computational environment that facilitates reproducibility as a digital concept beginning from data and tracing through the computational steps taken to achieve the published results. This distinguishes it from the replication of the experiment from first principles, including for example the regeneration of the raw data and the reimplementation of the data analysis de novo. We introduce the concepts of *vertical collaboration* and *horizontal collaboration*, to distinguish between the act of building on previously published research and that of carrying out joint research at the same point in time. Both are important considerations for modern computational science. We do not try to enumerate optimal environment for all possible research settings, rather we outline use cases with the hope of spurning greater discussion and development in this area of research. Although some best practice documents do exist for digital archivists [12], we know of only one other resource designed to communicate best practices for scientific computing [13]. We encapsulate some of these ideas in a wiki designed to facilitate the development and communication of best practices for computational scientists.

Developing Best Practices

A computational scientist today is being inundated with new software tools to help with research [14], new requirements for publication [15], and evolving standards as his or her field responds to the changing nature and increasing quantity of available data [16]. Because of the speed of the changes occurring in scientific research, we chose to implement the best practice recommendations given in this paper as a wiki, available at http://wiki.stodden.net/Best_Practices_for_Researchers_Publishing_Computational_Results. The hope is that parties with specialized or more complete knowledge will be able to add their expertise to the best practices document, to create a maximally useful document.

What follows is a series of principles for producing really reproducible computational science, and examples of implementations. We take as a starting point a National Academies of Science 2003

report, “Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences” [17] stating:

> **Principle 1. (Chapter 3) Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims.**
> This is a *quid pro quo*—in exchange for the credit and acknowledgement that come with publishing in a peer-reviewed journal, authors are expected to provide the information essential to their published findings. (p. 5)
> **Principle 2. (Chapter 3) If central or integral information cannot be included in the publication for practical reasons (for example, because a dataset is too large), it should be made freely (without restriction on its use for research purposes and at no cost) and readily accessible through other means (for example, on-line). Moreover, when necessary to enable further research, integral information should be made available in a form that enables it to be manipulated, analyzed, and combined with other scientific data.** (p. 5)
> **Principle 3. (Chapter 3) If publicly accessible repositories for data have been agreed on by a community of researchers and are in general use, the relevant data should be deposited in one of these repositories by the time of publication.** (p. 6)

Here Principle 1 calls for the dissemination of data, software, and all information necessary for a researcher to “verify or replicate the claims” made in the publication. Principles 2 and 3 provide guidance for the implementation of Principle 1. We adapt and expand these ideas into a series of Best Practice principles for computational scientists generally, and include these on the associated wiki. This is not meant to be a comprehensive list of all possible best practices in all circumstances, but a first cut at a generic case with the means provided through the wiki for modification and improvement.

Best Practice Principles for Computational Science

1. **Permissions and Licensing.** This document assumes you have the legal right to make the data and code publicly available. Usually this means you either generated the data and code yourself, or you have permission from others who did so. Best Practices indicate making the data and code maximally available and open for re-use. One way to make this legally possible is through the use of open licensing and the Reproducible Research Standard [18].

1. **Pre-publication Best Practices.** Provenance, workflow tracking, and publishing environments: The preceding sections of this report focus on best practices after publication of the computational article. If one had been using a version control system such as GitHub.com or BitBucket.org throughout the project, making the code available at the time of publication would be simplified. Setting up an issue tracker and/or an agile wall can make communication more efficient.

2. **Data must be available and accessible.** Availability and accessibility can be broken down into three sub-discussions.

i. **Version Control for Data**: If you did not generate or collect the data yourself, provide a link and citation to the source of each dataset you incorporated, including which *version* of the data you used (if the data source does not provide version information, provide the exact time and date you accessed the data).

ii. **Raw Data Availability**: Results should be reproduced from the earliest digital data in the experiment, whether that is raw data coming from instruments or observations, or data as accessed from a secondary source. It defeats the purpose to supply a "cleaned" version of the data if it is impossible to access the methodology of the cleaning, for example. The goal is that all data manipulations be made transparent, beginning with the initial version of the data with which the researcher started working.

iii. **Store the Data Externally**: In the simplest case, there are no external data files, for example in some simulations. In the most complex case, data are massive, distributed, and possibly updated in real time. The intermediary cases involve data files that can be readily downloaded and accessed by the user. Going roughly from the simplest to the most challenging cases:

- **Simulated Data**: In the case of simulated data, sharing the code that generated the data is enough if the code executes reasonably quickly. When a simulation takes an extended amount of time to regenerate the simulated data, pre-calculated data should be provided along with the code used to generate them.
- **"Small" Static Data**: We classify small data as datasets less than 2GB in size, but this is a relative term that may change depending on your system, download speeds, and repository storage capacity. If you are able to store your datasets at your institution and link to them from your institutional webpage, that is a good step. It will help your citation count, help others find your data, and help verification of your work. But it is *insufficient*. You must make your datasets available at an external repository dedicated to providing access to scientific datasets in perpetuity.
- **Large "Static" Data**: Datasets greater than 2GB encounter a number of problems smaller datasets do not. The first is access, since uploading and download very large files is very time-consuming if not prohibitive. If you created the dataset yourself, you may have to make a one-time upload (either over the internet or shipping disks via snail mail). If you did not create the dataset yourself, it is likely sufficient to cite the version of the third party data that you accessed and provide the computer code you used to manipulate the data.

  Large data is very likely to come with its own infrastructure. It may already reside in a domain-specific repository designed for access, such as the Sloan Digital Sky Survey, the National Institute for Health's caBIG data sharing portal or its Genome-wide Association Studies, EarthCube, or a number of others. Each of these have or are developing policies on data re-integration to permit uploading of data that has undergone changes, or it may be possible to link to / cite the version of the dataset(s) you used in your research and provide code that replicates the manipulations you carried out on that snapshot of the data.

  Infrastructure for large data is becoming available for researchers beyond these groups of domain-specific data repositories. Both Globus Online and HUBzero provide different types of computational environments for non-domain-specific scientific research and their own methods for data availability. Both are geared toward cloud computation, as is the

National Science Foundation's XSEDE scientific computing environment. TheDataHub.org is an entirely open source data repository.

- **Streaming Data**: These data seem like the most challenging case but are actually likely to fall into one of the above categories. Published results must be obtained on some amount of fixed data, and this particular dataset can be readily shared as above. In these cases it is likely scientifically relevant to validate models on future streams of data, but that is left to the domain of new, potentially publishable research that will share its data when published.

Some domain specific dissemination platforms include the Machine Learning Open Source Software (MLOSS) (both software and data), The Stanford Microarray Database, and the Protein Data Bank (PDB).

There are exceptions to this principle, including confidential data and proprietary data. Workarounds should be attempted and may exist for confidential data [19] and proprietary data [20].

3. **Code and methods must be available and accessible.**

- **Version Control for Code / Making the Code Available Externally**: These goals can be accomplished together by using a version control system with a public facing option, for example Github.com or BitBucket.org. There are many advantages to using version control for the code you and your collaborators write during a project, and releasing the code to the wider world using version control is important. Doing so permits others to know precisely which version of the code generated what results, allows others to make modifications and feed them back into the system without disrupting the original code, and perhaps most importantly permits a community to develop around the research questions, complete with mature functionality for bug tracking and fixes, new code developments, centralized code dissemination, and collaboration. Here is an example of scientific code associated with a published paper, available on GitHub.com, and a second example with the code available on BitBucket.org. This paper (http://arxiv.org/abs/1201.3035) has its code available on Github.com at https://github.com/ketch/RK-opt and this paper (http://arxiv.org/abs/1111.6583) has its code available on BitBucket.org at https://bitbucket.org/ahmadia/pyclaw-sisc-rr. There is no effective size limit on the code bases that could use either of these sites.
- RunMyCode.org and the DataVerse Network will both host code. RunMyCode.org permits readers to run the code and verify results in the cloud through their website, but neither has version control functionality.
- Some collaborative scientific software projects host their own website to disseminate code. These tend to be projects with larger codebases and have a large number of users and developers. Examples include the DANSE project at CalTech for developing software for state of the art neutron scattering experiments, and the DUNE toolbox for solving partial differential equations.
- **"Really Big" Codebases**: These codebases are likely already embedded in version control systems, but are of a complexity that makes visual interpretation of the code next to impossible. Reproducibility in this case requires testing the functionality of the code, to ensure whether it is

operating as the researchers expect. Common software testing methods can be applied to these codebases, such as unit tests, integration tests, and regression tests [13]. Such testing is not broadly implemented in shared scientific software, but it is standard practice in the open source software community.

4. **Citation**. If you use data you did not collect from scratch, or code you did not write, however little, cite it. Include the **source**, include the **author**, and include the **date and time** you accessed the data or code you used. Best practices indicate including a unique identifier in your citation, preferably a SHA-1 hash such as The DataVerse Network's UNF (see http://thedata.org/book/unf). GitHub.com also uses SHA-1 hashing to identify and track code. Having a unique identifier is important since it provides a check that the data and code are what they are thought to be and it gives a way of versioning and establishing provenance.

   Help people cite the data and code you release. Include a suggested citation such as:

   ```
   Stodden V, Guo P, Ma Z (2013) "Toward Reproducible Computational
   Research: An Empirical Analysis of Data and Code Policy Adoption
   by Journals." PLoS ONE 8(6): e67111. doi:10.1371/journal.pone.
   0067111
   ```

   Citation Standards: There are several entities working to establish citation standards for data, for example ORCID (http://about.orcid.org/faq), DataCite is another (http://www.datacite.org) and EZID (http://n2t.net/ezid), but we are not aware of any specifically for scientific code. Assigning a Digital Object Identifier is excellent and helps establish provenance and citation. At the moment there are no @data or @code fields for BibTex entries. In the meantime best practices would suggest using the @misc field to create a citation for data or code. Here is the BibTex format for @misc:

   ```
    @MISC{citation_key,
        required_fields [, optional_fields] }
   Required fields: none
   Optional fields: author, title, howpublished, month, year, note,
   key
   ```

   Here is an example of a code citation, as suggested by http://www.ict.swin.edu.au/research/projects/helix/helix-cite.html:

   Rajesh Vasa, Markus Lumpe and Allan Jones, Helix - Software Evolution Data Set, http://www.ict.swin.edu.au/research/projects/helix, Swinburne University of Technology, 2010.

   Here is the corresponding BibTex code:

   ```
   @MISC{Helix10a,
      title = {{Helix - Software Evolution Data Set}},
   ```

```
    author={Rajesh Vasa, Markus Lumpe, and Allan Jones}
    howpublished = {\url{http://http://www.ict.swin.edu.au/
research/projects/helix}},
    year = {2010},
    key = {Helix10a},
    url = {http://www.ict.swin.edu.au/research/projects/helix}
   }
```

Note that you may need `\usepackage{url}` to execute.

Code and Plagiarism: Unlike other authors' text, code can and should be re-used exactly as written , but all use should be cited, exactly as is standard practice for research articles. Terms of use accompanying the software should be respected.

Influences from Sources External to the Research Process. There are often specific data and code sharing guidelines associated with funded research. For example,

- National Institutes of Health Funded Research: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
- National Science Foundation Funded Research: NSF Data Management Plan, Jan. 2011. http://www.nsf.gov/bfa/dias/policy/dmp.jsp
- Wellcome Trust Funded Research: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/
- University of Minnesota "Funding Agency and Data Management Guidelines."
- Stanford University Office of Technology Licensing: http://otl.stanford.edu/about/resources/about_resources.html

Conclusion

This article presents a seed attempt at a conversation around best practices for publishing computational scientists, through the traditional medium of the published paper and a community-editable wiki. This can be seen as a response to the Reproducible Research movement, through which computational scientists have been moving toward research practices that include making the data and code underlying a published results conveniently available. Open questions remain. This effort has several lofty goals, including clarifying best practices for computational scientists, establishing community standards, providing a central discussion point for evolving best practices, accelerating discoveries by facilitating reproducible computational science and data and code re-use, and supporting the transfer of the technology underlying scientific results. This paper takes a first step toward reaching these goals by creating a starting point for a set of Best Practices for a computational scientist publishing today, and in years to come.

References

[1] Donoho, D., A. Maleki, I. Ur Rahman, M. Shahram, V. Stodden, "Reproducible Research in Computational Harmonic Analysis," Computing in Science and Engineering, Vol. 11, Issue 1, 2009, p. 8-18. http://www.computer.org/csdl/mags/cs/2009/01/mcs2009010008-abs.html

[2] "Reproducible Research," Special Issue, Computing in Science and Engineering, Vol. 14, Issue 4, 2012, p. 11-56. http://www.computer.org/portal/web/csdl/abs/mags/cs/2012/04/mcs201204toc.htm

[3] Stodden, V., I Mitchell, R. LeVeque, "Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture," Computing in Science and Engineering, Vol. 14, Issue 4, p. 13-17, 2012. http://www.computer.org/csdl/mags/cs/2012/04/mcs2012040013-abs.html

[4] See Yale Roundtable on Reproducible Research, 2009. http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/

[5] Yale Roundtable Participants, "Reproducible Research," Vol. 12, Issue 5, 2010. http://www.computer.org/portal/web/csdl/doi/10.1109/MCSE.2010.113

[6] "Devil in the details," Nature editorial, Vol. 470, p. 305-306, Feb. 17, 2011. http://www.nature.com/nature/journal/v470/n7334/full/470305b.html

[7] Hanson, B., A. Sugden, B. Alberts, "Making Data Maximally Available," Science Magazine editorial, Vol. 331, Issue 6018, Feb. 11, 2011. http://www.sciencemag.org/content/331/6018/649.short

[8] Merali, Z., "Computational Science:… Error," Nature, Vol. 467, Oct. 13, 2010. http://www.nature.com/news/2010/101013/full/467775a.html

[9] Barnes, N., "Publish your computer code: it is good enough," Nature, Vol. 467, Oct. 13, 2010. http://www.nature.com/news/2010/101013/full/467753a.html

[10] Morin, A, J. Urban, P. Adams, I. Foster, A. Sali, D. Baker, P. Sliz. "Shining Light into Black Boxes," Science Magazine, Vol. 336, p. 159-160, 2012.

[11] Hunold, S. and J. Larsson Träff, "On the State and Importance of Reproducible Experimental Research in Parallel Computing," 2013. http://arxiv.org/abs/1308.3648

[12] The University Library of the University of Illinois at Urbana-Champaign, "Best Practices for Creating Digital Collections." http://www.library.illinois.edu/dcc/bestpractices/contents.html

[13] Wilson, G. et al. "Best Pratices for Scientific Computing," Nov 29, 2012. http://arxiv.org/abs/1210.0530

[14] Stodden, V., “Reproducible Research: Tools and Strategies for Scientific Computing,” Editor’s Introduction, Computing in Science and Engineering, Vol. 14, Issue 4, p. 11-12, 2012. http://dx.doi.org/10.1109/MCSE.2012.82

[15] Stodden, V., P. Guo, Z., Ma, “Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals,” PLoS ONE 8(6), 2013. doi:10.1371/journal.pone.0067111 http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0067111

[16] Lazer, D., A. Pentland, L, Adamic, S. Aral, A. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, “Computational Social Science,” Science Magazine, Vol. 323, Issue 5915, p. 721-723, 2009. http://www.sciencemag.org/content/323/5915/721.summary

[17] Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council, Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences, National Academies Press, Washington, D.C., U. S. A., 2003. http://www.nap.edu/openbook.php?record_id=10613&page=27

[18] Stodden, V., "Enabling Reproducible Research: Open Licensing For Scientific Innovation,” International Journal of Communications Law and Policyhttp://ijclp.net/old_website/article.php?doc=1&issue=13_2009, Issue 13, 2009.

[19] Dwork, C., “Differential Privacy,” in Automata, Languages and Programming - ICALP, Springer, 2006. http://research.microsoft.com/en-us/projects/databaseprivacy/

[20] Stodden, V., “Innovation and Growth through Open Access to Scientific Research: Three Ideas for High-Impact Rule Changes,” in Rules for Growth: Promoting Innovation and Growth Through Legal Reform, edited by The Kauffman Task Force on Law, Innovation, and Growth. February, 2011.