View Source

Chris H. Wiggins
Associate Professor,
Department of Applied Physics and Applied Mathematics,
Center for Computational Biology and Bioinformatics,
Columbia University, New York NY 10027

November 30, 2009

1 Apologia

There are many voices in the conversation on open science, coming from many different communities. There are some coming from the perspective of the arts or the free culture movement; some considering the legal ramifications and how best to construct a license or contract which successfully balances both the necessary precision and the necessary flexibility; some informed by expertise in the history or sociology of science; and others coming from the perspective of the journals, corporations, or universities which often have a fiduciary responsibility to ensure that the communication of information is coupled to a commercial transaction. My personal background is, I confess, none of those. I am coming from the perspective of a computational scientist and faculty member interested in advancing the field and educating the students while facing a level of opacity and mysticism in scientific publication that is often directly at odds with these goals.

2 Context

With each passing year a wider variety of scientific communities comes to enjoy advances in computational power, in the volume and rich structure of available data, and in improved algorithms. Not long thereafter, members in such diverse communities[1] are likely to face a disconnect: this change in our nature of our research occurs without a change in how we communicate our research. Particularly in the statistical analysis of natural data, but also in large scale computational experiments (e.g., simulations of complex proteins or climate models), the algorithm, the mathematics motivating the algorithm, the data (either from natural or computational experiments), and the scientific results

[1] *Cf.*, *e.g.*, recent calls for change in natural language processing [8], databases [7], economics [6], social science [4], statistics [3], and applied mathematics [10].

are intimately connected. The nature of scientific manuscripts, however, remains essentially that developed in the 18th century. In fact, the situation is even worse than in previous centuries owing to the diversity of practical and commercial factors which have driven the length of scientific manuscripts ever shorter. In many fields the most prestigious publications are in journals requiring that the manuscripts are reduced to no more than six or even four pages, or even a one-page "comment" or "letter." Given that the primary quantitative metric for evaluating the contribution and promotion of an individual is the number of such publications, there is little to no restorative force encouraging scientists to describe their results in greater detail (*cf.* Sec. 4).

The result is that the line between scientific and non-scientific publication has become blurred: a four-page description of data, mathematics, a novel and possibly highly-parameterized algorithm, computational experiments, and results, including figures and requisite citations, may serve as an advertisement for a result, or a poetic motivation for future scientists to perform future research inspired by this description but is certainly less than the minimal body of information necessary to comprehend, reproduce, and verify a falsifiable scientific claim.

Along with the many voices in this discussion come many goals and many foci, including manuscripts or data. In this brief essay, I focus on code[2]. Not executable files, not online applets or web interfaces, but human-authored, human-readable, source code.

3 Scientific publications: desiderata

3.1 reproducibility

One of the desiderata of a scientific result is that it should be able to be reproduced, by anyone, anywhere. Anyone, anywhere, can reproduce the Michelson-Morely experiment, on a mountain in Switzerland or in a coal mine in Pennsylvania, and verify that there is no ether. Conversely, the shift in world energy production from fossil fuels to room-temperature fusion has been impeded by the inability of scientists to reproduce the results claimed nearly twenty years ago by Pons and Fleischman [11].

Why, then, is computational science granted exceptional status? Why are the results of computational experiments, large scale simulations, and statistical analysis believed and lauded, even in cases where no group is able to reproduce the claimed scientific facts?

Computational research should be far and away the most reproducible subtopic in scientific publication. Armed with the WWW, we should be able to distribute the source code, data, and ancillary descriptions necessary for anyone

[2]The reader might not have ever seen source code, or encountered the reality that executing the code does not reveal what the code is actually doing. A toy example is shown in Appendix A.

to re-execute the research, re-generate the figures, and comprehend and extend the work.

3.2 workability

A related but distinct goal is workability [5]: source code is an apparatus which should be usable not only with one dataset and not only with one parameter setting. Sharing source code means sharing the apparatus along with the blueprints for the apparatus with researchers who should be able to reproduce the results not only on the authors' original data and original parameter settings. Part of using the apparatus is breaking it: testing its robustness or fragility; testing its ability to generalize to new, different, larger or smaller datasets; and demonstrating the effect of varying the explicit and implicit parameters which enter into the apparatus's design.

3.3 transparency

Finally, even if the code is never compiled or executed, viewing the source code itself is an irreplaceable tool in understanding what, precisely, was done. As a blueprint, the code may not be easily sight-read[3], but it should be a complete set of instructions for executing the analysis or experiment claimed by the authors.

4 Code is different

4.1 The code is not the manuscript

Code shares with manuscripts the property of being human-readable (having been human-authored), and the actual files may be sometimes longer (or sometimes much shorter) than the manuscript files. However, in computational science, it is no longer true that writing a description, in words, of what was done, or a description of the results of the experiments, is necessarily sufficient to reproduce the results described in a scientific manuscript.

4.2 The code is not the record of experimental results

Viewing the source is not a record of experimental results. The former is the apparatus and the blueprint; the latter is the research journal or log book or, often, the punchline of the scientific publication.

4.3 Code is not data

Code shares with data the important property that submission of code and data can be required before publication, such that enforcement of data and code sharing is not based on a promise by the author to cooperate with future

[3] In fact, the community of C and Perl users has long-elevated writing difficult to interpret source code to a competitive art [1].

requests [9], but rather a precondition before a journal decrees the publication to be a complete, transparent, reproducible scientific communication. However, data themselves do not constitute the complete statement of what was created by the computational scientist. Data files are often much much larger and present technical challenges not faced by code. Code is naturally dynamic, benefits from the editing and refinement of a community (and, consequently, a variety of software engineering solutions exist for the sharing, modification, community-wide editing and version control of code).

5 A taxonomy of excuses (and replies)

5.1 Blaming commercial interest

E: I can't share the code because I wrote the paper with someone from industry.

R: Then you should advertise it somewhere other than a scientific journal. A scientific communication is not a commercial.

5.2 Blaming the student

E: Only the student had the necessary meta-codes to execute everything, and only she/he knew the parameters.

R: Then take the time to reverse-engineer your own research so that the reviewers, readers, and future students don't have to. Moreover, in the future, don't declare a student 'graduated' without making the code reproducible. This is part of communicating the computational research as a publication.

5.3 Blaming the code

E: My code's not pretty.

R: Beauty isn't the goal here. Your code, on average, is as pretty as that of the average researcher in your field.

5.4 Bald-face Charlatanism

E: I'm not sure it's right.

R: Why are you publishing it? Do you have students? Think of the children!

6 Example cases, possibly to be emulated

6.1 Protein DataBank (PDB)

The single example case I think most worthy of emulation is that of the Protein Databank (PDB). PDB was created in 1971 and has thus has had 38 years to struggle to become a standard within the structural biology community as well as to work out the difficult relationship between the journals, the author-scientists, and the resource itself. It is part of a world-wide effort, funded by a variety of agencies, with main hubs in the US, Japan, and Europe.

With the rise of the WWW PDB usage became more intimately connected with publication: first with the understanding that data were to be available within months or a year of publication, then, owing to the coordinated decisions of the editors of Nature, Science, Cell, and PNAS, the promise of eventual availability was replaced by the far more simple and effective precondition for publication[2]. This has in turn enabled an entire field of statistical studies and molecular dynamics based on the structural data, a field impossible without access to the data as part of each publication.

6.2 Conferences

SIGMOD 2008, 2009 have aimed explicitly at reproducibility and workability [5]. In this case, the goals were completely voluntary, yet still met with resistance.

7 What is to be done? Who is to do it?

7.1 Community

1. Establish 3rd party sites, with transparent version control (cf., e.g., code.google.com or sourceforge.net). Funding must be by multiple sources (e.g., PDB) in case one blows up, bails out, or re-evaluates its mission. Establish unique IDs (e.g., doi) for code and data.

2. Use them. Version control and mirroring are solved problems (unlike, e.g., the technical challenges associated with uploading, downloading, and sharing massive image and fMRI datasets). Work with database experts to ensure that data are searchable. Do not tolerate datasets or code which are available upon request or hosted on the websites of the journal or (worse) the author.

7.2 Funding agencies

1. Before funding, demand that proposals include not only a plan for code sharing but a third-party URL which reviewers and program officers can check to verify that the plan was enacted as promised before grants are renewed or future grants are issued.

2. When evaluating renewals or continued funding, check whether the code- and data- sharing plans were executed as proposed.

3. Help fund the 3rd party sites, in coordination with the other funding agencies (as is the case with PDB).

7.3 Journals

Do not publish a computational publication without the code and the data. Use 3rd party sites and provide unique identifiers (e.g., doi) for the code and data. Gary King has, in fact, already written for you a statement of policy: http://gking.harvard.edu/repl.shtml . You need only copy and execute.

References

[1] cf http://en.wikipedia.org/wiki/Obfuscated_Perl_Contest and http://en.wikipedia.org/wiki/The_International_Obfuscated_C_Code_Contest.

[2] The gatekeepers. 5(3):165–6, Mar 1998. http://www.nature.com/nsmb/wilma/v5n3.892130820.html.

[3] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[4] G. King. Replication, replication. *PS: Political Science and Politics*, 28(3):444–452, 1995.

[5] I. Manolescu, L. Afanasiev, A. Arion, J. Dittrich, S. Manegold, N. Polyzotis, K. Schnaitter, P. Senellart, S. Zoupanos, D. Shasha, et al. The Repeatability Experiment of SIGMOD 2008. *SIGMOD Record*, 37(1):39, 2008.

[6] BD McCullough. Got replicability. *The Journal of Money, Banking and Credit Archive. Econ Journal Watch*, 4(3):326–337, 2007.

[7] R.V. Nehme. Black Hole in Database Research.

[8] T. Pedersen. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470, 2008.

[9] CJ Savage and AJ Vickers. Empirical study of data sharing by authors publishing in PLoS journals. *PloS one*, 4(9):e7078, 2009.

[10] V. Stodden. Enabling Reproducible Research: Open Licensing for Scientific Innovation.

[11] G. Taubes. Bad science: the short life and weird times of cold fusion. 1993.

A Example

```
wiggins@tonkotsu{~}108: source foo.sh
hello world
wiggins@tonkotsu{~}109: source bar.sh
hello world
wiggins@tonkotsu{~}110: cat foo.sh
echo "hello world"
wiggins@tonkotsu{~}111: cat bar.sh
setenv str 'hello world'
echo $str
wiggins@tonkotsu{~}112:
```