RESEARCH PRIORITIES

Shining Light into Black Boxes

Funders, publishers, and research institutions must act to ensure that research computer code is made widely available.

A. Morin,[1] J. Urban,[2] P. D. Adams,[3] I. Foster,[4] A. Sali,[5] D. Baker,[6] P. Sliz[1]*

The publication and open exchange of knowledge and material form the backbone of scientific progress and reproducibility and are obligatory for publicly funded research. Despite increasing reliance on computing in every domain of scientific endeavor, the computer source code critical to understanding and evaluating computer programs is commonly withheld, effectively rendering these programs "black boxes" in the research work flow. Exempting from basic publication and disclosure standards such a ubiquitous category of research tool carries substantial negative consequences. Eliminating this disparity will require concerted policy action by funding agencies and journal publishers, as well as changes in the way research institutions receiving public funds manage their intellectual property (IP).

Policy actions to eliminate source code withholding in research computation

Institutional support	Publicly funded research institutions and university TTOs must remove organizational impediments to OSS licensing of computer code and embrace a wider variety of methods for exploiting and sharing their intellectual property. Creating a "standard set" of open software licensing tools within and across institutions that includes established OSS licenses would be an important step toward that goal.
Funding policy	Public funding and policy-setting agencies must explicitly and clearly state their strong preference for open dissemination, sharing, and publication of scientist-created software and source code. Although not an absolute requirement in recognition of the enormous diversity of research receiving public funds, the burden of justifying proprietary research products would be left to the applicant.
Publishing requirement	Scientific journal publishers must enact editorial policies requiring, as a condition of publication, that researchers make available new computer source code generated in the course of the research and necessary to reproduce the published research findings. Policies in place at journals already meeting this requirement (*16*–*18*, *36*) could provide guidance for wider implementation.

Disparity Without a Cause

In publicly funded research outside of computational science, the creation and dissemination of new tools, techniques, and methods requires detailed publication and disclosure of information necessary to satisfy peer review, experimental reproduction, and the ability to build upon another's work. Research tools created using public funds, such as animal models or cell lines, even those intended for commercialization, must fulfill disclosure and publication requirements (*1*).

Disclosure practices among scientist-programmers often do not meet these standards. Computer programs created in the course of research can range from single-command line scripts to multigigabyte code repositories. Many scientist-created programs are ad hoc efforts never intended for distribution or release, but all can be equally critical to research outcomes. Although it is typical to publish general conceptual and functional descriptions of new, major pieces of scientist-created software, it is not uncommon to withhold the program source code and instead release only the binary (executable) version of a program. Source code is the human readable form of a programming language and contains the complete set of instructions for how a computer processes input data. In the absence of source code, the inner workings of a program cannot be examined, adapted, or modified.

The consequences of relying on these black boxes in research computation can be far-reaching. Common implementation errors in programs, such as failing to convert units correctly or assigning missing values as zero, can be difficult to detect without access to source code (*2*). Recent retractions, resignations, and canceled clinical drug trials at Duke University involved unreleased and unreproducible code (*3*). Calls for greater focus on reproducibility in scientific research have mounted in recent years (*4*, *5*), and the inability to reproduce many published computational results or to perform credible peer review in the absence of program source code has contributed to a perceived "credibility crisis" for research computation (*6*, *7*). Source code withholding causes duplication of efforts by preventing sharing and reuse of validated computer code (*8*) and is incompatible with the stated goals of science funding agencies and policy advisory bodies (*9*).

How and why this unique disparity in disclosure practices persists within research computation is complex and goes beyond simple protectionism. Contributing factors may include the informal means by which most scientist-programmers attain their programming skills (*10*, *11*). It is not uncommon for self-taught programmers to be insecure about publishing "ugly" code: programs that work but do not conform to accepted best practices, are inefficient, or are aesthetically lacking (*12*). Lack of awareness and education around issues of code dissemination among scientist-programmers may also contribute. Among the small number of programming courses geared toward scientists, issues of code publishing or software licensing are seldom addressed.

Systems of attribution and citation, frequently relied on as metrics for career evaluation and achievement, which have evolved to accommodate publication of traditional scientific methods and techniques, may not adequately assure authorship credit when source code is adapted by other researchers. Tendencies toward traditional IP protection regimes at institutional technology transfer offices (TTOs) can result in proprietary licensing and distribution schemes that discourage release of source code (*13*).

Public-funding and policy-setting agencies have yet to enumerate clear, comprehensive, and universal policies promoting the publishing and dissemination of computer source code. Some specific funding initiatives evaluate applicants, in part, on software sharing and dissemination plans

[1]Harvard Medical School, Boston, MA 02115, USA. [2]School of Law, University of California, Berkeley, Berkeley, CA 94720, USA. [3]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [4]Argonne National Laboratory and University of Chicago, Argonne, IL 60439, USA. [5]University of California, San Francisco, San Francisco, CA 94158, USA. [6]University of Washington and Howard Hughes Medical Institute, Seattle, WA 98195, USA.

*Author for correspondence. E-mail: piotr_sliz@hms.harvard.edu

Published by AAAS

[e.g., (*14*)]. Such grants are typically for, or specifically include, large software development projects, however, and thus fail to address the large majority of scientist-created code.

Most significant may be the absence of a universal disclosure requirement by the gatekeepers of scientific publishing. Of the 20 most-cited journals in 2010 from all fields of science (*15*), only three (*16*–*18*) (including *Science*) have editorial policies requiring availability of computer source code upon publication. This stands in stark contrast to near-universal agreement among the 20 on policies regarding availability of data and other enabling materials.

Mechanisms of Code Dissemination

Source code can be made available through a variety of mechanisms. Posting code for download on laboratory Web sites, deposition in public code repositories, or making use of publisher facilities for supplemental materials are just a few existing options (*6*). Because of the complexity and unique characteristics of computer source code, however, preserving the systems of attribution and citation that have evolved to accommodate traditional channels of scientific publishing (e.g., data sets, journal articles, and lecture materials) requires additional measures. Fortunately, a variety of software licensing tools exist to help scientist-programmers retain the benefits of authorship, as well as protect IP rights, when disseminating their code.

Beyond allowing others to inspect and understand the inner workings of a computer program, open source software (OSS) licenses encourage the free adoption, reuse, and adaptation of computer source code while also assuring the attribution and citations customary in scientific research. For the scientist-programmer, disseminating software under an OSS license can be a simple method for enabling community participation in development, use, and adoption of a program and can lead to enhanced influence, reputation, and increased rates of citation for the author (*19*). Numerous types of OSS licenses exist to meet the diverse needs of academic environments, many of which were developed by and for academics working at research institutions [e.g., Berkeley Software Distribution (*20*), MIT (*21*), and Educational Community License (*22*)]. OSS licenses are also fully compatible with commercialization of scientist-created software (*23*) and Bayh-Dole requirements that allow the patenting of inventions created using public funds (*24*).

Although OSS licensing options are well suited to the open access and dissemination goals of publically funded research (*25*–*28*), not every research software development effort will find OSS licensing an appropriate vehicle for source code dissemination. Many large, publicly funded research software development projects (*29*, *30*) have found a mixture of standard OSS and custom licensing to be appropriate means of achieving source code disclosure while also generating commercial licensing revenue.

Eliminating the Disclosure Disparity

As reliance on scientist-created software grows across scientific fields, the common practice of source code withholding carries significant costs, yields few benefits, and is inconsistent with accepted norms in other scientific domains. Changing this practice will require concrete and unambiguous policy action (see the table). Less definitive disclosure policies are unlikely to achieve desired results. For example, a recent article (*31*) makes a persuasive case for the necessity of source code release in reproducing scientific results, but fails to lay out efficacious policy recommendations likely to achieve significant and timely change in withholding practices.

Calls for change in disclosure practices from within the scientific community are not new. Similar actions, initiated by the research community and with the cooperation of publishers, have proven successful in the past. In the late 1980s, a group of structural biologists petitioned journal editors to help end then-common data-withholding practices by making the deposition of protein structure data into public databases a condition of publication (*32*). As a result, today, the Worldwide Protein Data Bank (*33*) is a vital enabling resource for the biomedical research community that has helped fuel the emergence of multiple fields.

More recently, the field of genomics underwent a community-driven consensus process on data publication and availability. The resulting "Bermuda principles" (*34*) state that data should be publicly released prior to publication, within 24 hours of generation. Similar principles have since been adopted by other publicly funded 'omics initiatives, including structural genomics (*35*).

The parallels between past and current debates over data withholding and the agreed solutions in favor of disclosure and publication are striking. Requiring that source code be made available upon publication would also be expected to yield substantial benefits—including improved code quality, reduced errors, increased reproducibility, and greater efficiency through code reuse and sharing. Achieving this would bring disclosure and publication requirements for computer codes in line with other types of scientific data and materials.

References and Notes

1. D. Weitz, *High Tech L. J.* (1993); http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/berktech8&section=13.
2. G. Wilson, *Am. Sci.* **94**(1), 5 (2006).
3. R. S. Tuma, *Oncology Times*, 7 January 2011 [blog]; http://journals.lww.com/oncology-times/blog/FRESHSCIENCEforClinicians/pages/post.aspx?PostID=10.
4. J. P. Mesirov, *Science* **327**, 415 (2010).
5. J. P. A. Ioannidis, M. J. Khoury, *Science* **334**, 1230 (2011).
6. R. D. Peng, *Science* **334**, 1226 (2011).
7. V. Stodden, *Comput. Sci. Eng.* **11**, 35 (2009).
8. V. Stodden, MIT Sloan School Working Paper 4773-10 (2010); http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1550193.
9. NSF, *National Science Foundation*; www.nsf.gov/pubs/2011/oise11003/.
10. S. M. Baxter *et al.*, *PLOS Comput. Biol.* **2**, e87 (2006).
11. Z. Merali, *Nature* **467**, 775 (2010).
12. N. Barnes, *Nature* **467**, 753 (2010).
13. R. Litan *et al.*, *Yale Law Econ. Res. Paper no. 426*; 10.2139/ssrn.1757982.10.2139/ssrn.1757982
14. PAR-10-266, http://grants.nih.gov/grants/guide/pa-files/PAR-10-266.html.
15. Thomson Reuters, *2010 Journal Citation Reports*, Science Edition (Thomson Reuters, Philadelphia, 2012).
16. B. Hanson, A. Sugden, B. Alberts, *Science* **331**, 649 (2011).
17. Editors, *J. Biol. Chem.*; www.jbc.org/site/misc/edpolicy.xhtml.
18. Editors, *Proc. Natl. Acad. Sci. USA*; www.pnas.org/site/misc/iforc.shtml.
19. H. A. Piwowar, R. S. Day, D. B. Fridsma, *PLoS ONE* **2**, e308 (2007).
20. The BSD License, www.opensource.org/licenses/bsd-license.php.
21. The MIT License, www.opensource.org/licenses/MIT.
22. D. Greenstein, B. Wheeler, *Open Source Collaboration in Higher Education* (Indiana Univ., Indianapolis, IN, 2007).
23. F. Hecker, *IEEE Softw.* **16**, 45 (1999).
24. B. N. Sampat, *Nature* **468**, 755 (2010).
25. 32006R1906–Regulation (EC) Official Journal L 391, 30/12/2006 P. 0001i; http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R1906:EN:HTML.
26. National Science Foundation, www.nsf.gov/pubs/policy-docs/pappguide/nsf11001/aag_6.jsp#VID4.
27. National Institutes of Health, *Fed. Regist.* **64**(246), 72090 (1999).
28. T. R. Cech *et al.*, *Plant Physiol.* **132**, 19 (2003).
29. Phenix, www.phenix-online.org/.
30. RosettaCommons, www.rosettacommons.org/.
31. D. C. Ince *et al.*, *Nature* **482**, 485 (2012).
32. F. M. Richards *et al.*, http://hkl.hms.harvard.edu/si/disclosure_letter.pdf.
33. H. Berman, K. Henrick, H. Nakamura, *Nat. Struct. Biol.* **10**, 980 (2003).
34. E. Birney *et al.*, *Nature* **461**, 168 (2009).
35. A. Edwards, *Nat. Struct. Mol. Biol.* **15**, 116 (2008).
36. Editors, *PLoS Comput. Biol.* (2008); www.ploscompbiol.org/static/policies.action.

37. We thank K. Keating and P. Suber for discussions, and S. K. Burley for providing the document in *(32)*.

10.1126/science.1218263