Reproducible Science via Open Source Requirements:

Increasing Impacts of and Public Support for NASA Mission Science

Michael Hirsch, Ph.D.
January 2018

In response to NASA Solicitation for white papers, PIN DEPS-SSB-17-01

Corresponding Author:
Michael Hirsch, Ph.D., mhirsch@bu.edu, (202) 765-4770, Boston University.

The author is speaking as an individual, not as a representative.

Introduction

I have authored or contributed code to over 200 open source projects, including NumPy and SciPy. In the startup/corporate realm, as an electrical and computer engineering professor mentoring hundreds of students, and as a geoscientist, my experience is that the benefits of free open source software (FOSS) are worth the challenges of balancing stakeholder IP rights. For the purposes of this paper, I define FOSS as having an OSI-approved license. The balance of open code incentives, policy and guidance NASA promulgates will set a new bar for federal and international agencies as well as journals and institutions. Besides undiscovered errors, the unseen price of closed source science includes the intangible loss of collaborations due to chance virtual encounters that never occur. Perhaps a year or two before their work is in arXiv and ADS Abstracts, I find the first hints of work from future collaborators in GitHub searches.

Given the widely varying restrictions (especially export control) that NASA-funded effort encounter, incentives and requirements must balance public and community benefits with grantee and vendor rights and policies. Journals with open code policies across diverse disciplines make routine exemptions for export control or IP protection. Newly developed code inherent to the NASA grant/contract should generally be FOSS, with a general exception for legacy proprietary libraries. Consistent with AGU Best Practices, analysis codes must at least generally be made available upon request. The voluntary but encouraged gold standard for code is an independent third-party reviewer (perhaps automated by free services such as Travis-CI) that runs the code replicating the figures and other science output (Benureau, 2018). Making open-source the norm for NASA-funded research will help attract and retain top geospace talent as evinced by the world's largest corporations open-sourcing and sponsoring FOSS projects (e.g. Google Summer of Code).

Author and Institution Impacts of Open Source

Numerous funding agencies require data to be released or at least embargoed for a fixed length of time (say, one year) before public release, since the research uses public funds. This data often effectively remains locked away for broad-based studies since only a limited clique connected to the author is able to obtain the source code needed to read the data. Author and institution intellectual property rights may be retained and robustly protected by the adoption of copyleft licenses such as GNU Affero v3+ (aGPL). The most stringent licenses like aGPL prohibit integration into closed source software, including software in the cloud. These "share-alike" software licenses have found increasing parallel adoption in the form of Creative Commons licenses for journal and conference publications. Non-open geospace software stifles potential new projects and delays novel research due to unavailability of code to read data and replicate plots in the literature/documentation. When original authors change institutions or

retire, experiments costing millions of dollars in public funds are lost to future use with non-open software.

Retroactive Open Source Impacts

Where NASA can compel authors and institutions to make prior NASA-funded code available under open source, this should generally be done. Downsides include code that had private funding from industry and/or defense agencies, that may be more difficult to get everyone to sign on to open sourcing. Finite resources would be required to ensure copyrightable code by all authors has either been excised or signed on to by all parties. This problem may seem insurmountable at first glance, but there is practical precedent by a number of very widely used open source software such as OpenSSH. Given limited resources, some kind of non-monetary NASA award/recognition/commendation that provides "currency" for career promotion could be devised.

The well-publicized NASA Langley FUN3D competition attracted over 1,800 participants to modernize and enhance NASA Fortran CFD code. This is remarkable as the subset of programmers with sufficient expertise in Fortran supercomputing architecture is a tiny fraction of skilled Python or C++ programmers. However, the FUN3D competition was canceled due to the need to verify US citizenship of the unexpectedly large number of participants, which exceeded NASA vetting resources. There is a clear groundswell of support and interest in furthering NASA's science interests by US citizens and global participants, if only NASA-sponsored code licensing would allow them to participate. Moving forward, this mean software planning, particularly for NASA-funded efforts, should include as a high priority segregating publicly disclosable code from proprietary code.

Future Open Source Impacts

There will be exceptions necessary for NASA-funded open code policy. I regularly deal with cases where FOSS is interfaced with proprietary code. In such cases, more permissive licenses such as LGPL or BSD/MIT licenses are often usable. The technical impacts are generally negligible, since best practices for efficient computing lead to segmentation of code in most cases. Where NASA sets policy, OSI-approved licenses should be required to extract maximum value from the policy by maximizing code license compatibility.

With appropriate licensing, algorithms deeply embedded in private code for specialized proprietary tasks may be publicly shared in its general form. These are the practices taken every day by the largest corporations such as Facebook, Netflix, Google, etc. as well as startups laser-focused on building IP as their *raison d'être*. Corporations share novel generalized code even when it could give their competitors some advantage, to help attract top talent when the talent sees excellent work output. Releasing such code allows the corporation or agency to be a

standard bearer, increasing the relevance to the general population, and underlining the impact of an agency on National strength and technological output.

Enforce and encourage standard archival references

NASA should enforce with policy and encourage with non-policy measures the use of standard archival references such as DOI or NIH-sponsored Research Research Identifier (RRID) (Chawla, 2015) as appropriate to the particular science artifact. AGU journals open data requirement requires making data and code used to make figures available on request. Public code archives whether reviewed as by IEEE, published as by JOSS or simply made available in any form as by Zenodo are far more effective in maintaining data/code utility. Data and software have useful lifetimes of decades, particularly when doing systems science important for current and future policy decisions affecting the global economy. These requirements are not unduly burdensome, particularly when enacted at the time of data request (for data existing before the NASA policy change). Public, free data archives such as Zenodo are suitable for an unlimited number of 50 GB datasets. Zenodo is accessible via a web browser or Python API.

Addressing Open Source Fears

Fear of being "scooped" for their next journal article or grant proposal leads some researchers to keep their code private or for a small group of collaborators only, even after refereed publication or grant completion. The level of knowledge necessary to "copy" results is still substantial, particularly where complex hardware is involved. Industry-leaders such as Ettus Research, a National Instruments company have succeeded with open source *and* open hardware, using Commercial Off the Shelf (COTS) components.

GNU Radio (GPL v3+ license) is a popular software package for software defined radio (SDR) equipment. Ettus Research licenses the firmware and drivers necessary to use their hardware such as the popular USRP SDR under GPLv3. If a customer develops a confidential application, they can pay for an proprietary license from Ettus Research.

Non-policy incentives to open source

NASA should encourage authors to open source existing and future code in the following ways:

1. Intern/coop study on citations and author metric improvements for papers with FOSS
2. Work with journal editors to provide a certification e.g. badges when paper authors have FOSS or open data referenced by DOI, RRID or similar standard archival reference.
3. Encourage journal editors to create special commendation and credit to the authors and non-anonymous reviewer, where a third party reviewer (perhaps automated) replicates the figures using FOSS and open data. This "super" reviewed paper has not only passed

the standard rigor of blind review, but also the scrutiny of the software reviewer. This would help avoid repeating mistakes of non-replicable research (Ziemann, 2016)

NASA's non-policy thrust to incentivize replicability will be felt in higher public and policymaker confidence in NASA-funded studies and mission research output. Gold standard archival reproducible code and data is replicable into free cloud providers such as Google Cloud Always Free Tier, Microsoft Azure Free and Heroku Free, among others.

Publisher open data/code requirement is increasing. Discussions at conferences indicate growing community support for requiring code used to generate publication plots from data. Cloud-hosted data storage, offline and online processing are increasingly adopted by projects such as NCAR CHORDS, Earthcube and MIT Haystack Mahali among many others.

Difficulties in Preserving Archival Code

To this day, code is disclosed by certain agencies as Appendices in Technical Reports. Even with a PDF of the appendix, copying and pasting the code is error-prone and poorly indexable. Perhaps the code is in an institutional repository, until the neglected server gets hacked. These and other non-centralized efforts have in my experience been inadequate for discoverability and preservation of archival code. NASA's PDS has proven to be generally robust with a heterogeneous assortment of data. NASA does not need to create a new service for code. However, a master database of DOI, RRID, etc. may be useful to be sponsored or encouraged by NASA so that in the event of impending future repository difficulty, the code could be copied and the digital identifier pointed to the new URI.

Conclusions

Open code should be favored and supported by NASA in policy and non-policy actions. The benefit to NASA's public image will be immediately realized as a standard-bearer in reproducible science, setting a new paradigm in all sciences. NASA should encourage and enforce use of RRID for software and other archival artifacts essential to reproducible science. The challenge is in finding the right balance between creator rights and the public rights, since they ultimately paid for the code development.

Appendix 1

Speaking for themselves as individuals, the undersigned generally support this white paper.

1. Joshua Katz, jk369@njit.edu
2. Guy Grubbs, guygrubbs@gmail.com, Catholic University of America
3. Matthew Zettergren, zettergm@erau.edu
4. Neel Pandeya, neel.pandeya@ettus.com
5. Asti Bhatt, asti.bhatt@sri.com, SRI International
6. Ashton Reimer, ashton.reimer@sri.com, SRI International

References

D. Chawla (2015). Researchers argue for standard format to cite lab resources. Nature, doi:10.1038/nature.2015.17652

M. Ziemann, Y. Eren and A. El-Osta (2016). Gene name errors are widespread in the scientific literature. Genome Biology, doi: 10.1186/s13059-016-1044-7

F. Benureau, N. Rougier (2018). Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions. Frontiers in Neuroinformatics, doi: 10.3389/fninf.2017.00069