Software Practices for improved collaboration among space scientists

Asti Bhatt, SRI International; Ryan McGranaghan, NASA JPL;
Tomoko Matsuo, U. of Colorado; Yolanda Gil, U. of Southern California

Open data and code are required to evolve traditional methods and enable scientific discovery to keep pace with progressions in the digital age. To actualize better data and code practices we must promote larger, trans-disciplinary networks (cutting across science and computer science, among others). Evidence shows that the growth of open access publishing has been driven in part by clear and specific mandates on the part of funding agencies [State of Open Data, 2017 https://doi.org/10.6084/m9.figshare.5481187.v1]. In this white paper, we propose recommendations for NASA space science to adopt to improve robustness and collaboration in coding practices.

PROPOSED RECOMMENDATION #S1: Software used in a paper, particularly models, should be:

1. *available in a shared community repository*, so anyone can access it
2. *have a license*, so anyone can understand the conditions for use and extension of the software
3. *have an associated digital object identifier (DOI) or persistent URL (PURL)* for the version used in the paper, so that the software is available permanently and URL rot in papers is avoided
4. *cited properly in the paper in the references section*, so readers can identify the software version unequivocally and software creators can get credit for their work

When this recommendation cannot be met, a brief explanation should be included in the paper that may include the results from use of the software.

PROPOSED RECOMMENDATION #S2: When using or downloading a model or software from a repository, a user should be given the license and a DOI or PURL for the version used as well as a preferred citation that they can use in their papers. Model repositories that serve the CEDAR community should consider serving models in this manner.

PROPOSED RECOMMENDATION #S3: A software registry for geospace science software developed by various stakeholders should be created such as in partnership with Ontosoft. Such software registries create a common place for the domain scientists to search for existing software solution and make their own software searchable. The software registry isn't a replacement for software hosting service.

PROPOSED RECOMMENDATION #S4: For open source software solutions using non-proprietary languages, a web-based platform should be implemented that can access data repositories and software to reproduce results, such as from a paper or a proposal with appropriate data and software IDs. This solution is recommended in the light of the recent push for reproducible science.

Permanent Unique Identifiers for Software
A separate DOI should be assigned to meaningful versions of the software, such as a version used for a paper. GitHub offers an option to obtain a DOI for a software version, which is done by storing that version permanently in the Zenodo data repository. Any software can be uploaded manually to community data repositories such as Zenodo, figshare, and Dataverse. PURLS can be assigned by anyone to any software version that has a URL on the Web, using a trusted service such as w3id.org.

Community Software Repositories
Many general software repositories can be used by scientists in any domain, and as such they are available to the CEDAR community. These repositories will often inquire about choosing a license, and specifying a descriptive name and authors for the software. These repositories include GitHub and BitBucket, to name a few. General data repositories accept software as an entry, and as with any dataset they always offer DOIs, licenses, and citations.

Geospace Community Model and Software Repositories
Model repositories that serve the geospace community should adopt mechanisms for assigning DOIs or PURLs to software versions that they run for users. The management of PURLs or DOIs can be complex, and organizations such as FORCE11, the Research Data Alliance, and ESIP have working groups with extensive and detailed recommendations in this respect.

Ancillary software used by the geospace community is currently maintained in personal computers and in some cases in general software repositories. A software registry for geospace community would help make these software available to the larger community subject to the developer's interest.

Licenses for Software
Recommended licenses for software are the standard licenses from the Open Source Initiative, preferably Apache v2 or MIT (unlimited reuse as long as there is attribution) but other more restrictive licenses are available.

Resources for the specific objectives of the committee

Below we provide reference material pertaining to each of the objectives that the committee has been tasked with.

Review and describe examples of code / modeling policies developed by research teams and communities in the NASA-supported disciplines of Earth Science and Applications from Space, the Space Sciences, and other research communities, as appropriate;

- Software Sustainability Institute Case Studies (https://www.software.ac.uk/index.php/resources/case-studies)

Develop a set of lessons learned from these established approaches-- paying particular attention to issues such as, but not limited to, proprietary, export control, code/model maintenance, and documentation considerations;

- *Reproducible Research: Tools and Strategies for Scientific Computing* workshop materials (http://www.stodden.net/AMP2011/)
- Particularly, *Barriers to Data and Code Sharing in Computation Science: A survey of the Neural Information Processing Systems Community (NIPS)* (https://web.stanford.edu/~vcs/talks/VictoriaStoddenNESS2010.pdf)

Define and describe options for policies on open codes and open models for research supported by NASA Science Mission Directorate (SMD) and assess the pros and cons of these options from the perspective of the research community and the interests of NASA; and

- NASA Earth Science Data System Open Code Policy (https://earthdata.nasa.gov/earth-science-data-systems-program/policies/esds-open-source-policy)

Recommend a set of best practices for NASA to consider should SMD decide to adopt an open code / open model policy for research supported by the agency. The committee may also choose to present alternate sets of best practices rather than just one recommended set.

- Here we refer the committee to the recommendations provided on page 1 of this white paper. These recommendations have been put together after a pilot project by NSF's EarthCube initiative, called the Integrated Geoscience Observatory was implemented for 2 years to create an online software platform and a software registry for geospace scientists to share their open source, Python-based software using.

Summary

This white paper makes recommendations for NASA's open code policy for adoption in space sciences. These recommendations are put together with the goal of increasing transparency and improving collaboration among space scientists.