git: A powerful tool to facilitate greater reproducibility and transparency in science.

Karthik Ram, Ph.D.
Environmental Science, Policy, and Management.
University of California, Berkeley.
Berkeley, CA 94720. USA.
karthik.ram@berkeley.edu

Abstract

Reproducibility is the hallmark of good science. Maintaining a high degree of transparency in scientific publications is essential not just for gaining trust and credibility within the scientific community but also to facilitate development of novel science. Sharing data and computer code associated with publications is becoming increasingly common, motivated partly in response to data deposition requirements from journals and mandates from funders. Despite this increase in transparency, it is still difficult to reproduce or build upon the findings of most scientific publications without access to a complete workflow.

Version control systems (VCS), which have long been used to maintain code repositories in the software industry, are now finding new applications in science. One such open-source VCS, **git**, provides a robust framework that allows scientists to track every component (data, code, figures, and text) of a research endeavor from start to finish, with the ability to revert any file or an entire project back to any stage in its development. Since the system is decentralized, every copy of a repository not only contains all the data and code but also the entire history of changes along with detailed notes documenting each of those decisions. The power of a git repository can be further extended by linking it to a git hosting service (e.g. GitHub) which makes it easy for multiple collaborators to work asynchronously and merge their changes as needed. A git based workflow is particularly suited for science because it is designed to protect against data loss, quickly retrace errors, and allow multiple ideas and methods to co-exist in parallel.

In this paper I review how git benefits science and why more scientists should make git-based workflows an integral part of their research. I also provide several real world use-cases demonstrating the idea that sharing git repositories can foster collaborations, increase accountability, track contributions, in addition to lowering barriers to data reuse and supporting novel synthesis.