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

Karthik Ram
Environmental Science, Policy, and Management.
University of California, Berkeley.
Berkeley, CA 94720. USA.
karthik.ram@gmail.com

Abstract

A hallmark of good scientific endeavours is that the research is described in adequate detail that it can be accurately reproduced, verified, and extended to support novel research questions.

Science can only be reproduced when all the components of a publication, such as the underlying data and code, in addition to other methods, are shared alongside the paper.

To take this one step further, additional decisions made during the analysis phase, such as excluding certain data, are often lost since there is often not enough room to describe such valuable bits of information.

One potential way to avoid such problems is to use a version control system such as git.

Not only does git track versions of manuscripts, code, and data, a decentralized system such as git carried the entire history of changes from the start of the project to the final publication with every copy.

This allows others to revisit changes and decisions made at any stage of the effort, the ability to branch off from an earlier point in the effort, or explore a new direction before certain key decisions were made during the evolution of the project.

The power of git can be further extended by linking a repository to a remote repo hosting service such as github. Unlike centralized VCS, anyone is free to work on their existing copy without having to check out a current version from the central repo. This allows multiple authors to work asynchronosly and merge changes as needed and as posssible. Although git can merge most changes seamlessly, in cases where it cannot automatically merged, git leaves placeholders for any author to go in and resolve it manually. Doing so, along with meaninful commit messages makes it a valuable way to document author contributions and assign credit at a fine grained level.

Even with the manuscript writing process alone, it allows for easy collaboration.

Science is often irreproducible. We don't even plan how we manage our data. But we really need to manage not just data but also code and the way we write our manuscripts. git arose in the software industry as a distributed way to manage code. However, it is really well suited for science. It is robust, not prone to a single point of failure, favors collaboration, and easily allows anyone to branch off from the most current iteration of the project or from any point along the way. Also it allows scientists to explore multiple directions and yet keep everything manageable and accountable. git can also automatically sync to a remote repository hosting service such as GitHub or BitBucket, providing additional fail safes against data loss. Since these services have a large userbase, it becomes extremely easy to add collaborators, and receive contributions from other viewers. In this manuscript, I outline some of the reasons why more scientists, particularly bioinformaticians need to incorporate git into their regular academic workflow. I also include a few examples of how a successful collaboration could work.

1. The hallmark of good science is that it is reproducible. One way to ensure reproducibility is to include all of the underlying data, and code used in the analysis alongside publications in the form of supplements or by depositing them in persistent repositories. While such efforts despite being rare are extremely valuable, they are still likely to lose valuable provenance such as key decisions made at any point from the data acquisition and analysis to manuscript development and final submission.

2. git is an open source distributed version control system (DVCS) that was originally developed to manage versions of the linux kernel code. The system provides a powerful way to track a version an entire research effort from start to finish. Unlike previous iterations of version control systems, git is particularly powerful because it is distributed allowing users to work asynchronously, and each copy contains the complete history of changes. Although git, and web-based git hosting services such as GitHub, have been popular in the software development community, this power has only recently been leveraged by the academic community and in a limited way.

3. Hosted git services allow several researchers to work asynchronously and synchronize their efforts through a central server. Users can work asynchronously and merge their copies as necessary. While most merges are handled automatically, conflicts need to be resolved manually by an author.

development on a project, a user can commit changes up to that point with a human-friendly message. This allows anyone to return the project to that state or pursue a different direction along a different branch. Thus, key decisions that may serious affect the trajectory of a paper can be quickly retraced (to either improve upon, pursue a different direction, or troubleshoot errors).

A git managed research folder also allows members within the lab to reuse methods and easily determine why certain decisions were made during the course of the project.

Much like a lab notebook, a git history, tracks the evolution of a research efforts with waypoints and notes that allow both the original researcher (and anyone else down the line) to retrace steps, pursue a different direction, or build something new from any point forward.

Unlike track changes in Word documents, git can keep track of contributions from unlimited authors maintaining an audit trail.

It also allows for soliciting feedback.

4. In this review, I outline reasons why more scientists should incorporate a **git** based workflow into their research. I also provide several use-cases for why a git based project can improve overall transparency and reproducibility lacking in most scientific endeavors.