

Git and GitHub for Reproducible Data Analysis

Data Standards and Open Data Community of Practice

Alice Byers

Data Innovation Team

27 April 2023

Slides: https://github.com/alicebyers5/presentations/tree/main/2023-04-27_version-control

Aims

- What is version control?
- Introduction to Git and GitHub
- How Git and GitHub can be used for reproducible data analysis
- Note: This aim of this presentation is to introduce concepts and will be nontechnical. If you would like technical support, please get in touch. (Contact details are at the end of the slides.)



What is version control?

Version control is the practice of tracking and managing changes to files.

Does this look familiar?



Git



Git is a free and open source software for version control.

To use:

- Install Git on your computer
- Initiate Git in a project folder (also called a repository)
- Use Git to record any changes you make to files (these records are called 'commits')
- Undo changes and revert to previous version of files
- Branching?

Git



Using Git means you don't need to save multiple copies of the same file to retain older versions. This information is stored by Git.

Commits contain information on:

- who made a change,
- what the change was,
- when the change was made, and
- why the change was made.

GitHub



GitHub is a platform for hosting version controlled code. GitHub is mostly free, but some extra features are available for a fee.

GitHub can be used to:

- Make code publicly available (although repositories can also be private)
- Facilitate code review (via pull requests)
- Manage projects using tools such as issue tracking
- Navigate Git history and view previous versions of files
- View other people's code and collaborate

What is in a code repository?

A code repository will usually contain files for one project and can contain:

- Code (e.g. R scripts)
- Documentation (e.g. README)
- Configuration files
- ...but NOT DATA!

Data can be stored within your code repository, but it should not be tracked by Git or uploaded to GitHub. To ensure data files (and any other files containing sensitive information) are not tracked, a **.gitignore** file can be used to exclude them. Alternatively, data can be stored outside of your code repository.

More information on using Git safely can be found in the Duck Book.

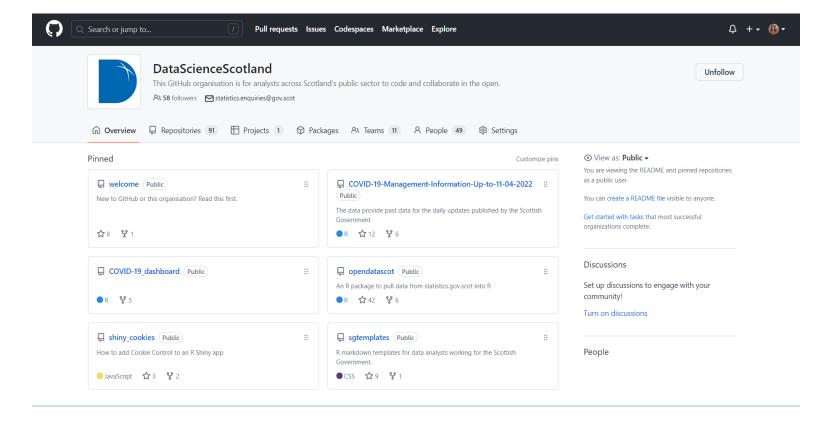
How to use Git and GitHub

- Git can be used without GitHub
- Git can be used via a terminal-like tool called Git Bash, or via R Studio
- GitHub is often used as the main copy of a code repository. Copies (or clones)
 of the repository are taken from GitHub for developers to work on.
- Use Git locally to track changes and regularly 'push' to GitHub
- Use GitHub to facilitate code review and merging of branches



GitHub Organisation

https://github.com/DataScienceScotland



Why use Git and GitHub?

- Preferable to lots of copies of the same file with various names!
- Reproducible Analytical Pipelines (RAP)
 - Reproducible: You can rerun your code as it was at any point in time.
 - Auditable: You have a record of when changes were made and why.
 - Transparent: Your code is publicly available on GitHub



Links and Resources

- Version control Saltire pages (accessible on Scottish Government SCOTS network only)
- Data Science Scotland GitHub Organisation
- Duck Book
 - Version Control
 - Using Git safely
- Government Analysis Function guidance on open sourcing analytical code
- ONS Learning Hub (contact Data.Science.Campus.Faculty@ons.gov.uk to request an account)
 - Command Line Basics
 - Introduction to Git



Contact

Alice Byers RAP Developer Data Innovation Team, Scottish Government

- Email alice.byers@gov.scot
- GitHub alicebyers5