

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 non-technical. 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?

```
|— presentation
|   |— version-control-slides-first-draft.pptx
|   |— version-control-slides-first-draft-v2.pptx
|   |— version-control-slides-first-draft-v2 AB review.pptx
|   |— version-control-slides-final.pptx
|   |— version-control-slides-final-FINAL.pptx
```

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?

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>

The screenshot shows the GitHub profile for the DataScienceScotland organization. The header includes the GitHub logo, a search bar, and navigation links for Pull requests, Issues, Codespaces, Marketplace, and Explore. The organization's name, DataScienceScotland, is prominently displayed, along with a description: "This GitHub organisation is for analysts across Scotland's public sector to code and collaborate in the open." It also shows 58 followers and an email address: statistics.enquiries@gov.scot. Below this, a navigation bar highlights the Overview tab, with links to Repositories (91), Projects (1), Packages, Teams (11), People (49), and Settings. The main content area is titled "Pinned" and displays a grid of six pinned repositories. On the right, there are sections for "View as: Public", "Discussions", and "People".

DataScienceScotland
This GitHub organisation is for analysts across Scotland's public sector to code and collaborate in the open.
58 followers | statistics.enquiries@gov.scot

Overview | Repositories 91 | Projects 1 | Packages | Teams 11 | People 49 | Settings

Pinned

- welcome** (Public)
New to GitHub or this organisation? Read this first.
8 stars, 1 fork
- COVID-19-Management-Information-Up-to-11-04-2022** (Public)
The data provide past data for the daily updates published by the Scottish Government
R, 12 stars, 6 forks
- COVID-19_dashboard** (Public)
R, 5 forks
- opendatascot** (Public)
An R package to pull data from statistics.gov.scot into R
R, 42 stars, 6 forks
- shiny_cookies** (Public)
How to add Cookie Control to an R Shiny app
JavaScript, 3 stars, 2 forks
- sgtemplates** (Public)
R markdown templates for data analysts working for the Scottish Government.
CSS, 9 stars, 1 fork

View as: Public
You are viewing the README and pinned repositories as a public user.
You can [create a README file](#) visible to anyone.
[Get started with tasks](#) that most successful organizations complete.

Discussions
Set up discussions to engage with your community!
[Turn on discussions](#)

People

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](https://github.com/alicebyers5)