Version control and git

Bio724D: Spring 2024

2024-03-25

What is version control?

Version control refers to the practice of, and tools for, enabling and tracking complex changes to textual documents (code, ordinary text, markup documents, etc.) over time.

Informal version control

Most of you have likely practiced some sort of informal version at some point in your academic careers.

Typical Scenario:

- You created a document for a paper you were writing: `paper_draft`
- For several iterations of this document, you're the only editor: `paper_draft` evolves
- The document gets to a state where its ready to share with collaborators: `paper_draft_02` (versioning)
- You email your document to collaborators (distribution)
- You get back suggested edits, `paper_draft_02_PMMedits`, `paper_draft_02_GWedits` (multiple working copies)
- You combine the suggested edits into a new draft `paper_draft_03` (merging)

Informal version control, cont.

- You want to experiment with an alternate framing of the discussion so you create a parallel version: `paper_draft_03_alt` (branching)
- You realize that you had a nice paragraph back in version 1, that was deleted in version 2, and you want re-integrate it into your current version (retrieve a prior version and merge text)
- You combine some some parts of your parallel branch back into the main document, `paper_draft_04` (branch merge)
- You reformat the document based on the journal requirements `paper_draft_05_current_bio`
- …etc…

What are the problems that version control must deal with?

- What text was added or removed between versions?
- Simultaneous editing
- Merging text changes from collaborators
- Different branches
- Recovering text that has been deleted or changed
- Figuring out who and when text was changed

Version Control Systems

In the context of computer programming, version control systems (VCS) have been in use for more than 50 years

- Today the most dominant VCS is git, which was first released in 2005

What is git?

- A version control system developed by Linux Torvalds (the creator of the Linux kernel) to support development of Linux
- Distributed version control – full codebase mirrored in many different systems
- Currently the most popular VCS in wide use

What is GitHub?

- Github is a commercial service that provides hosting for git repositories (other popular ones include GitLab, Bitbucket)
- Free hosting of public and private repositories
- Lots of accessory features that make it an attractive platform for development of both public and private projects
  - e.g. Markdown based wiki (Bio 724 website)
  - e.g. Host webpages

Recommendations for effective version control

- Modularize your code / documents – facilitates simultaneous editing, merging
    - Favor organizing your code into functions and submodules
    - Breaking ordinary text up into logical subunits
        - e.g. `abstract.md`, `introduction.md`, `methods.md`, …
        - If using Quarto markdown use `include` mechanism
- Frequent small changes are easier to work with than infrequent, large changes
- Favor plain text workflows (e.g. markdown) and convert to binary/proprietary formats (e.g. PDF, MS Word) late in process