Versioning of data and code using Git

Introduction to Data Management Practices course

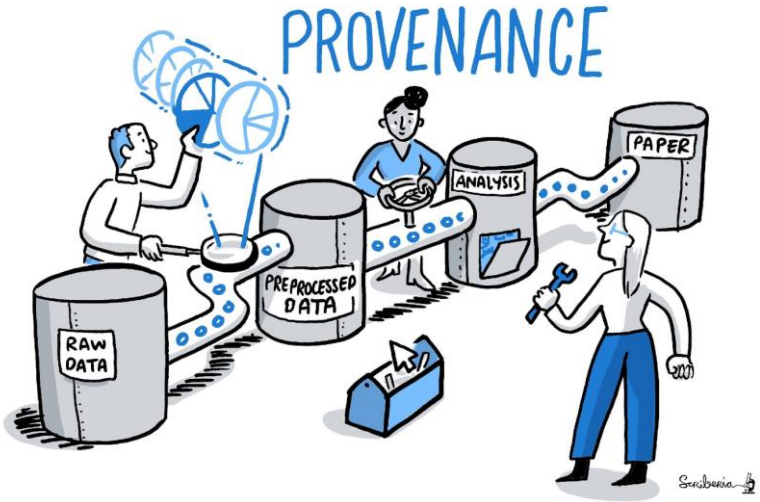
NBIS DM Team

data@nbis.se

https://nbisweden.github.io/module-versioning-dm-practices/

Outline

1. The fundamentals of version control
2. Local versioning on your computer using Git in RStudio
3. Remote versioning on the web using GitHub.com
4. Workflow between local and remote versioning using Git in RStudio and GitHub.com



Exercise 1

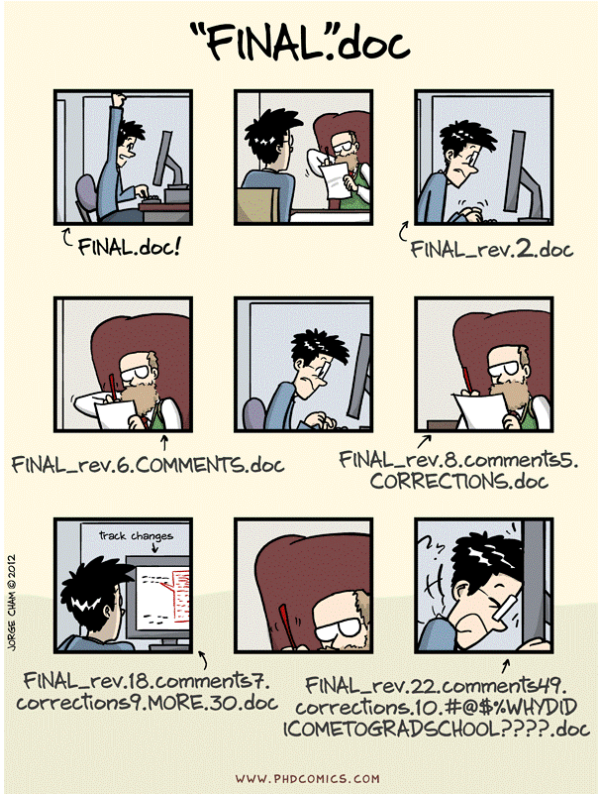
- Problems with change. Which of the following issues can you relate to?

Version Control

- Making changes to files
  - We risk losing content
  - Unintended side effects
  - Breaking analysis pipelines

- Collaborating with others
  - Coordination between multiple devices
  - Resolve conflicts

- Addressing these issues by tracking changes is called **version control**



"Piled Higher and Deeper" by Jorge Cham, https://phdcomics.com

Version Control Systems

Version Control Systems (VCS)…

1. Maintain a repository with all versions of files, along with metadata (author, timestamp, unique identifier)

2. When making changes, VCS create a new file version, rather than overwriting the existing one

3. Provide features to enable multiple lines of development

Version Control Systems

Version Control Systems (VCS)…

4. Compare different versions of a file, track changes, revert to previous versions if needed

5. Facilitate collaboration, mechanism for resolving conflicts

VCS provide a systematic and organized approach to managing changes to files, preserving the history of changes made to a project.

Manual versioning?

Possible, but time-consuming and error-prone…

To highlight some benefits of VCS over manual versioning:

- Instead of mandating users to create backups of the entire project, version control systems securely stores only the necessary information to recreate previous versions of files on demand.

- Instead of depending on users to come up with meaningful names for backup copies, the version control system automatically timestamps all saved changes, ensuring efficient tracking and organization of revisions.

Manual versioning?

To highlight some benefits of VCS over manual versioning:

- VCS keep a 100% accurate record of what was ***actually*** changed, as opposed to what the user ***thought*** they changed.

- Instead of blindly copying files to remote storage, version control systems perform conflict checks to ensure that no one's work will be overwritten. If conflicts are detected, they are identified, making it easier to resolve and merge changes.

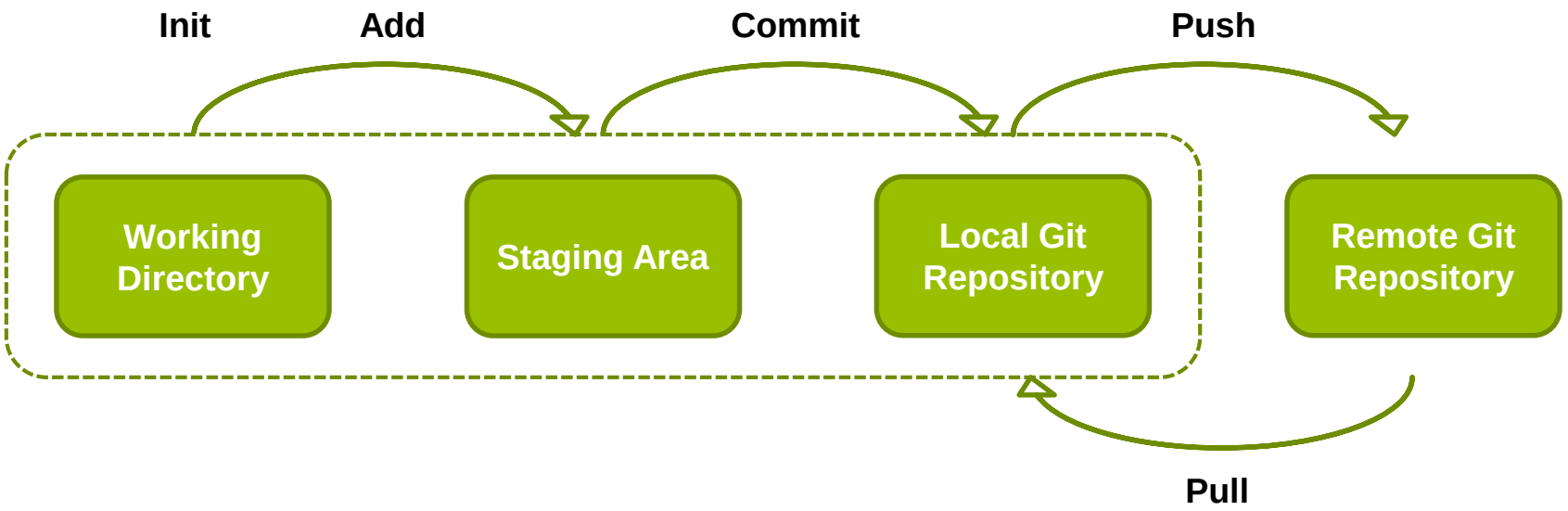
What is Git and GitHub?

Git: version control system, install on your computer. Helps you keep track of changes in files. Acts like a virtual repository where all changes are stored as snapshots, also known as commits.

GitHub: web-based hosting service for Git repositories. Provides a platform for collaborative software development.

- Just like OneDrive can synchronize Word documents across multiple devices, GitHub can propagate changes made in Git repositories across different computers and provide a centralized location for collaboration and version control in a group setting.

Simple Git Terminology



Importance for research

- Vital to keep track of changes made to data
- Promotes reproducibility
- Helpful during various stages of the research process
  - Providing information to reviewers, editors, and readers

- Enhances the reliability and transparency of the research work

Exercise 2

As you now have been introduced to version control systems, we invite you to consider these five reasons for incorporating a version control system into your research workflow.

Which two are the most important for you?