

# CyVerse\_Workflow\_Notes

stribling

2023-04-04

**General set-up & workflow for shared data, analyses, and coding for the forest fragmentation working group** Overall goal: Store all shared data online, develop code in a shared version-control workspace, and run code & analyses on shared data using cloud computing

**Data storage: We can store data in a common folder in CyVerse**

- Grace's shared folder is our current data home: `"/iplant/home/mgracemcleod/Forest_Fragments/Data/"`
- Our Team for this forest fragmentation working group is **"Forest Fragments"**

**Data notes & questions**

- I think (?) spatial shapefiles need to be stored unzipped within cyverse in order to access the files within the Cyverse's RStudio docker ("Rocker")
- It looks like we need **all** the multi-part files in order for R/Rocker to successfully read in a .shp file.
  - What needs/issues does this generate for data storage?
    - \* How do we best manage multi-part shapefiles?
    - \* Does storing a bunch of files in Grace's folder use up all her storage space? If so, how can we share the storage burden? Sharing with a "Team" sounds like the right solution, but I'm not currently able to share with our Team.

**Code development: We can store and edit code on Github, using a shared group repo, "forest\_frag".**

- We can integrate the github repo with RStudio on personal computers or within CyVerse's RStudio ("Rocker").
  - This allows us to pull the most up-to-date code, edit it, and push code back to main repository.
  - Version control within github allows us to "back up" and find old versions of code if we need to.

**Coding notes & questions**

- Jenny Bryan's Happy Git and GitHub for the useR is a great resource for getting your RStudio fully integrated with GitHub.
- GitHub also allows for data storage, but I've read recommendations against using that feature. When is this feature useful?
- See issues raised below under "Analysis notes & questions"

**Analysis Workflows:** We can run analyses within the “VICE” environment of CyVerse.

1. A Rocker session can be linked to the ‘forest\_frag’ repo on GitHub to pull the code into Rocker.
  2. Data for analyses is stored within CyVerse in a shared folder.
  3. Code is run through the “Rocker” app within CyVerse; should function just like RStudio. + Just need to ensure filepaths point to the cloud storage location.
- Each free Cyverse account permits 200 Core Hours of cloud computing for free.
  - – This allows us to pull the most up-to-date code, edit it, and push code back to main repository.
  - – Version control within github allows us to “back up” and find old versions of code if we need to.

### Analysis notes & questions

- Each Cyverse account permits 200 Core Hours of cloud computing for free.
  - Simply running “Rocker” within CyVerse for ~ an hour with some super simple code used up a LOT of my free processing hours. I don’t understand why; is there a way to avoid this?
  - What is the best use of this cloud computing resource, given the restrictions?
- Given limited processing hours within CyVerse, we may need to do most coding and work on personal computers, and then open CyVerse & Rocker for the big analyses, once code has been fully written.
  - This necessitates all of us downloading all files to our PCs anyway, so we can work off the cloud, tinker around, and then push it back up to the cloud and run the big analysis.

**Visualization of how all of this comes together?**

