Module 1 - Lesson 04

Reproducible Components

Melinda K. Higgins, PhD.

2017-09-29

Outline

1. "Data"
2. Documentation
3. Organization
4. Automation
5. Dissemination

Data - "for example…"

- often thought of as numbers in a spreadsheet
- can be unstructured text
- images
- video and other media
- interview transcripts
- artwork
- and any other "RAW" materials needed to complete your project

Data - "should be…"

- high quality
- reviewed for completeness
- reviewed for mistakes and errors
- checked for changes or updates
    - ideally, your final reproducible workflow will allow up these changes and updates to be automatically updates into your final products

Documentation

- main component is text
- well written
- good organization and flow
- easily accessible
- understood by team members at all levels
- code + text + figures combined [e.g. literate programming]
- at end, formatting styles applied via "markup/markdown"

Organization

- projects grow
- supporting documentation and files - numerous
- relationships change and can grow more complex
- need file organization and naming schemes
- file names should be:
    - readable by the computer, easy to search, easy to sort (especially by date and author if needed)
    - human readable with logical naming schemes and contain enough info so human knows what is in the file/what the file is for
    - and short enough to be reasonably manageable
- consider user-based access and security (partitioned by "need to know"

Research Compendium Example https://github.com/ropensci/rrrpkg

Automation

- at a minimum, a diagram or instructions for workflow should be documented on how the components are to be assembled for your final product
- write code/scripts to automate
  - data raw to processed output
  - creating and removing temporary files
  - creating tables, figures, other components
  - assembling the components into final documents, products
  - rendering documents into multiple/desired formats

Dissemination - Why?

- store and share your data and code – even if it is only for your future reference
- sometimes expectation/requirement of funding agency, publisher
- increased visibility, you as source - default subject matter expert
- speed of collaboration - faster advancement of science, knowledge
- good will with community/public

Dissemination - How?

- Cloud-based "File Storage"
  - Dropbox https://www.dropbox.com/
  - Google drive https://www.google.com/drive/
  - Github (better with version control and tracking) https://github.com/
- Data repositories
  - GenBank https://www.ncbi.nlm.nih.gov/genbank/
  - PDB https://www.rcsb.org/pdb/home/home.do
- In addition to Github
  - Bitbucket https://bitbucket.org/
  - Dryad http://datadryad.org/
  - Figshare https://figshare.com/
  - Zenodo https://zenodo.org/

Dissemination - Who?

- Yourself
- Your organization - internal reports
- Journals - articles, manuscripts
- Books
- Blogs/Websites
- RSS feeds
- Rpubs https://rpubs.com/
- Gitbook https://www.gitbook.com/
- Bookdown https://bookdown.org/yihui/bookdown/

Next in Lesson 05 …

Getting Started with

&