Data for Data Science:
Data manipulation and reproducibility

fredhutch.io
Fred Hutchinson Cancer Research Center

Course outline

- Class 1: Data entry and creating spreadsheets
- Class 2: Organizing data and project files
- Class 3: Documenting data with metadata
- **Class 4: Data manipulation and reproducibility**

Today’s objectives

At the end of today’s class, you should be able to:

- Maintain records of how data have been manipulated or altered
- Plan ahead for publishing/archiving data
- Identify tools and support at the Hutch for multiple parts of the data life cycle

Completing the data life cycle

Funding agencies and scientific publishers require us to submit our data to repositories.

How do we plan ahead to make sure this doesn't simply meet requirements, but also maximizes the use of data later?

Primer on Data Management

How might data change during the course of a project?

Discovery (discovery.fredhutch.org)

Welcome to Discovery

Discovery allows you to search for your stuff in the Hutch's Fast, Economy local (Swift), and Economy cloud (AWS S3) storage systems. This tool is currently in development. Please try it out and let us know how we can improve it. Please send feedback to bret@fredhutch.org
Thanks!

Maintaining data integrity

See this lesson for a simple model for applying quality control and assurance

- **Quality assurance**: require data to be entered appropriately
- **Quality control**: checking for erroneous data

Specify in your metadata what efforts are made at quality assurance! Look here for some ideas of best practices

Cleaning large-scale datasets

Example tool: OpenRefine (lessons here)

- Clean, transform, and extend large, messy datasets
- Allows export of clean data AND reproducible methods used for cleaning

Version control

Tracking changes and additions/deletions to documents and files

- Can be used for data files (during the course of a project)
- Provides a record of what has changed
- Allows reversal back to earlier versions
- See the wiki entry for more information

See the fredhutch.io course Intro to Git and GitHub!

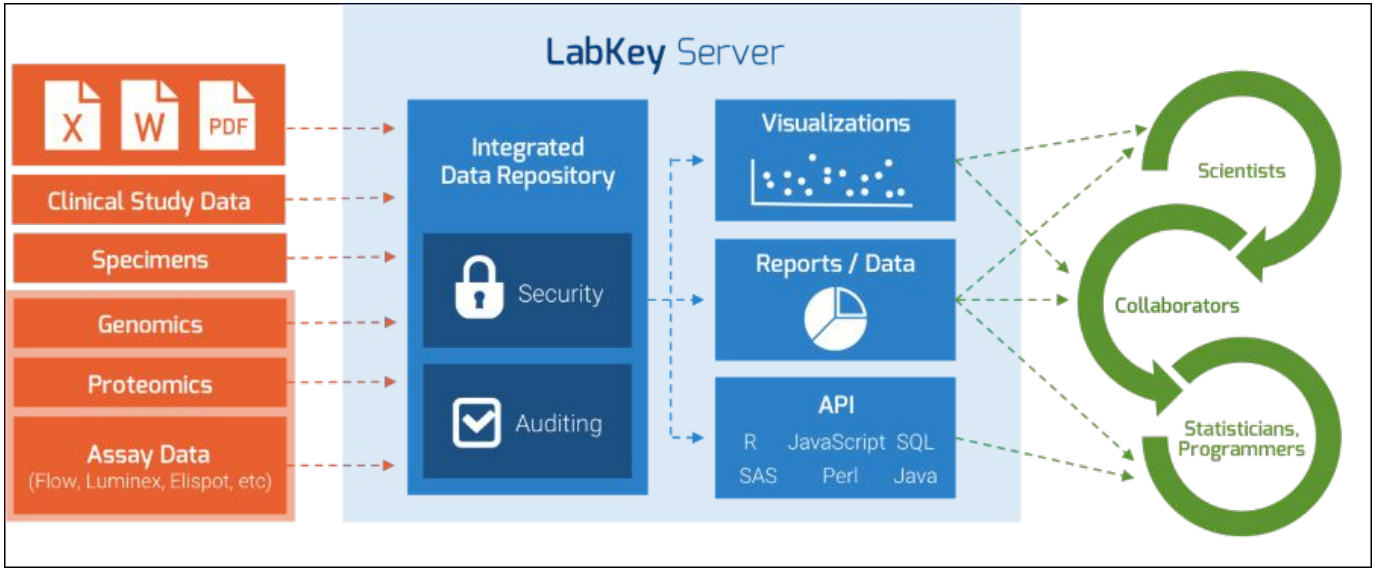
REDCap (redcap.fredhutch.org)

REDCap is a secure web platform for building and managing online databases and surveys.

See the wiki for information on office hours and the user group!

LabKey (redcap.fredhutch.org)

Lab-based data repositories, data showcase, and electronic laboratory



Join the Coop Community Slack to participate in the user group!

Open Science Framework (osf.io)

How OSF supports your research

Search and Discover

Find papers, data, and materials to inspire your next research project. Search public projects to build on the work of others and find new collaborators.

Design Your Study

Start a project and add collaborators, giving them access to protocols and other research materials. Built-in version control tracks the evolution of your study.

Collect and Analyze Data

Store data, code, and other materials in OSF Storage, or connect your Dropbox or other third-party account. Every file gets a unique, persistent URL for citing and sharing.

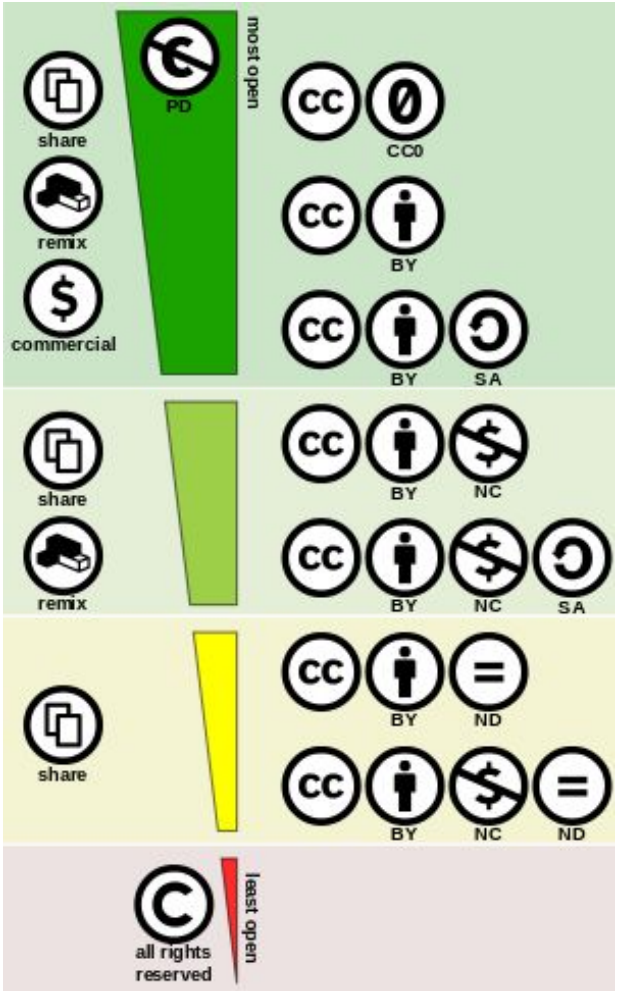
Publish Your Reports

Share papers in OSF Preprints or a community-based preprint provider, so others can find and cite your work. Track impact with metrics like downloads and view counts.

Using and choosing licenses

Licenses dictate how *you can use other people's data*, and how *other people can use your data*.

What license does Data Dryad use?



Shaddim; original CC license symbols by Creative Commons CC BY 4.0

The big picture

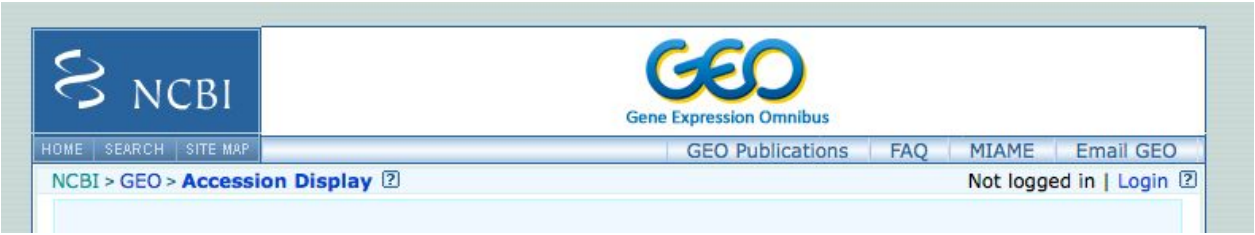
Ten Simple Rules for the Care and Feeding of Scientific Data

“This article offers a short guide to the steps scientists can take to ensure that their data and associated analyses continue to be of value and to be recognized.”

Data Dryad Best Practices for creating reusable data publications

Also see this quickstart guide

Look at the GEO metadata template document. When should researchers start filling in the information? Who needs to be involved in documenting the data?



GEO example

FAIR data principles

Findable, **A**ccessible, **I**nteroperable, and **R**e-usable

Most data management focuses on humans being able to reuse data. FAIR principles assist automated searching and reuse of data.

“Data” includes algorithms, tools, and workflows! (FAIR article)

FAIR data: Findable

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

FAIR data: Accessible

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

FAIR data: Interoperable

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

FAIR data: Re-usable

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

Summary

- Plan ahead for data management to save time and improve the quality of your research
- Document your metadata early and often
- Find tools to help automate and check your work
- Planning for reproducibility will help you and the broader scientific community

Best practices will vary depending on your particular field and project. Feel free to ask Hutch Data Commonwealth (Scientific Computing, The Coop) for assistance!

What are the biggest questions you still have about data management?

What resources would help you further?

Fred Hutch Data Science Wiki:

Data generation

Scientific Computing