Data for Data Science

fredhutch.io
Fred Hutchinson Cancer Research Center

Sign in here: https://hackmd.io/@k8hertweck/introData

Course outline

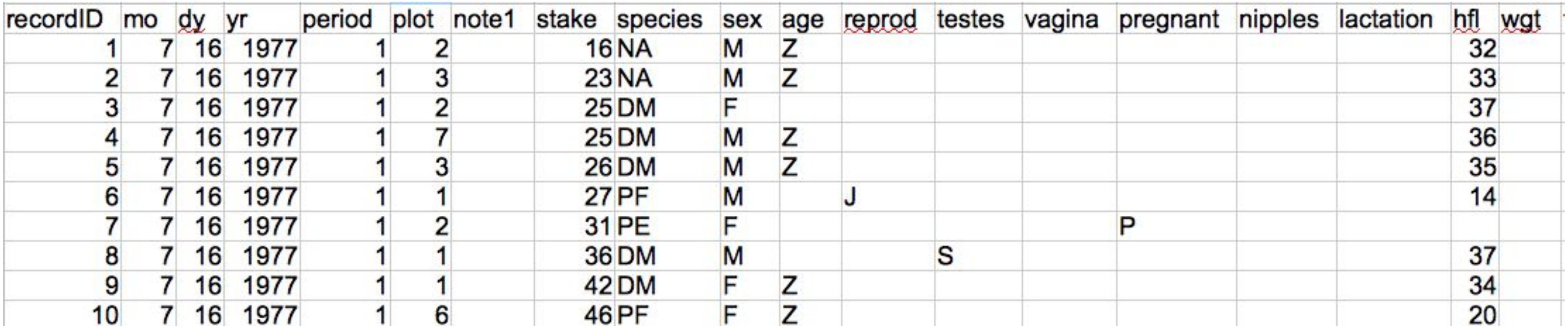
- Class 1: Data entry and creating spreadsheets
- Class 2: Organizing data and project files
- **Class 3: Documenting data with metadata**
- Class 4: Data manipulation and reproducibility

Today’s objectives

After today’s class, you should be able to:

- Understand the utility of metadata, both during active research and after project completion
- Identify different types of metadata and their utility
- Outline metadata for your own research project

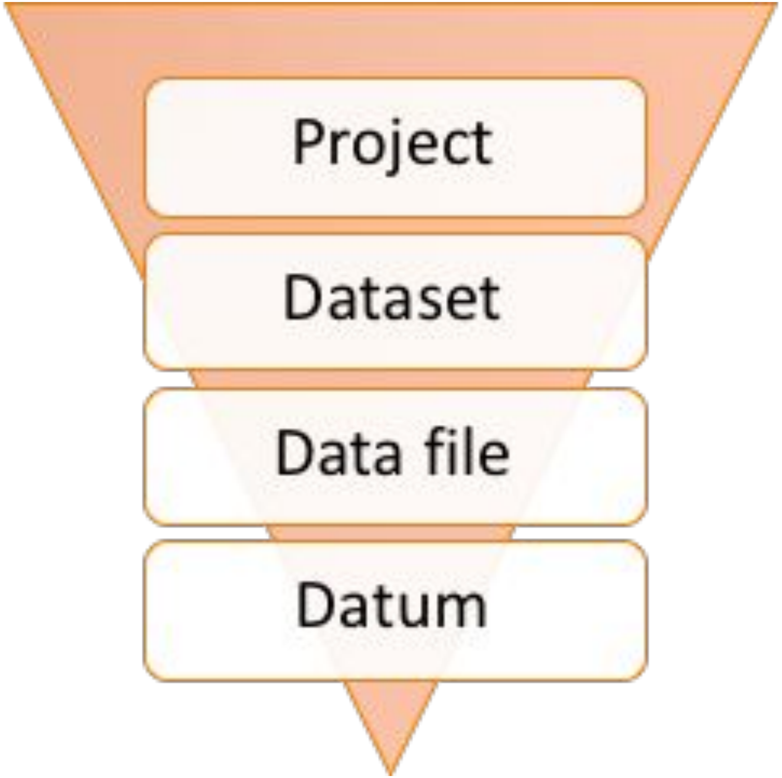
Have you ever tried to use someone else’s data?



recordID	mo	dy	yr	period	plot	note1	stake	species	sex	age	reprod	testes	vagina	pregnant	nipples	lactation	hfl	wgt
1	7	16	1977	1	2		16	NA	M	Z							32	
2	7	16	1977	1	3		23	NA	M	Z							33	
3	7	16	1977	1	2		25	DM	F								37	
4	7	16	1977	1	7		25	DM	M	Z							36	
5	7	16	1977	1	3		26	DM	M	Z							35	
6	7	16	1977	1	1		27	PF	M		J						14	
7	7	16	1977	1	2		31	PE	F					P				
8	7	16	1977	1	1		36	DM	M			S					37	
9	7	16	1977	1	1		42	DM	F	Z							34	
10	7	16	1977	1	6		46	PF	F	Z							20	

Activity: What would you need to know to use these data?

What kind of metadata do we need for each level of data complexity?



Project

Dataset

Data file

Datum

Metadata: data about data

- **Metadata**: the information we create, store, and share to describe things
- NISO (National Information Standards Organization) defines three types of metadata

  a. **Descriptive**: finding and understanding a resource

  b. **Administrative**: technical, preservation, and rights (licensing) information

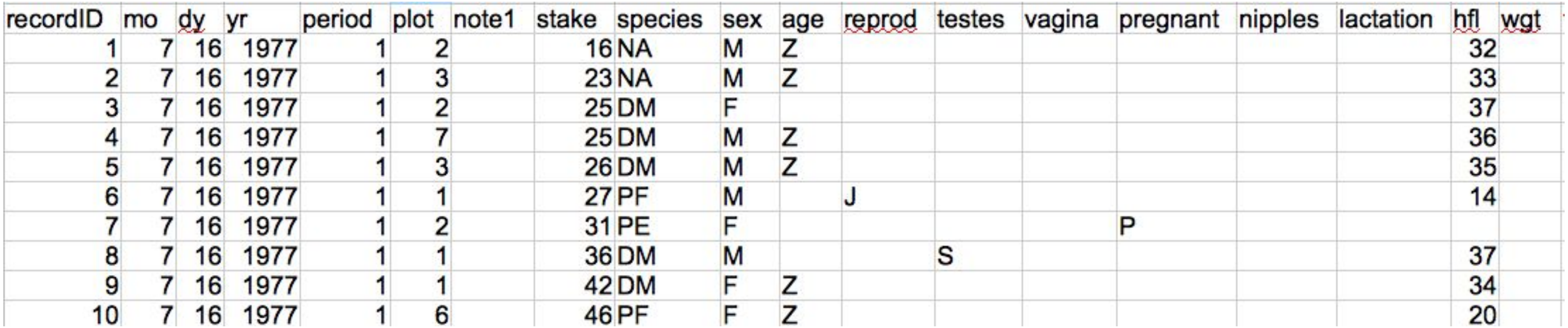
  c. **Structural**: relationships of parts of resources to one another

Understanding Metadata

Metadata for scientific research projects

- **Intent:** what is the project?
- **Collection method:** how was the experiment conducted?
- **File structure:** where are files located?
- **Data type:** what is the file type and format?
- **Data description:** what do columns/rows mean?
- **Provenance:** how have data been manipulated?

What do the metadata for these data include?

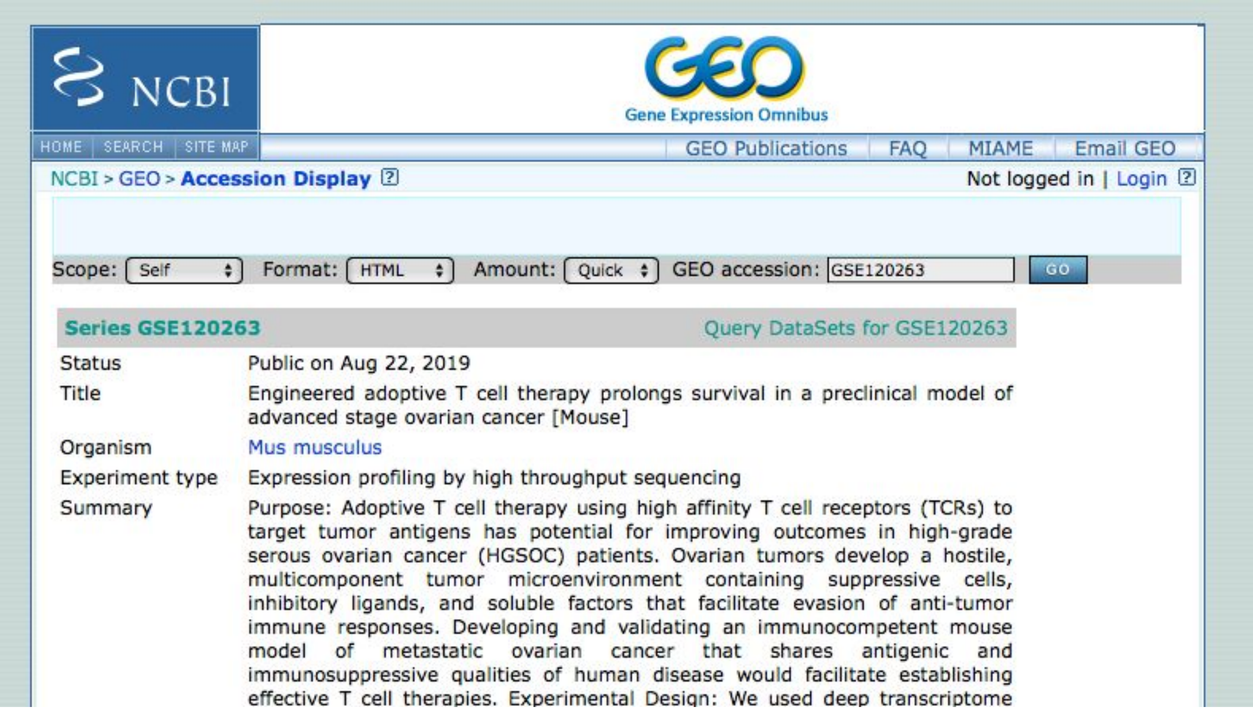


recordID	mo	dy	yr	period	plot	note1	stake	species	sex	age	reprod	testes	vagina	pregnant	nipples	lactation	hfl	wgt
1	7	16	1977	1	2		16	NA	M	Z							32	
2	7	16	1977	1	3		23	NA	M	Z							33	
3	7	16	1977	1	2		25	DM	F								37	
4	7	16	1977	1	7		25	DM	M	Z							36	
5	7	16	1977	1	3		26	DM	M	Z							35	
6	7	16	1977	1	1		27	PF	M		J						14	
7	7	16	1977	1	2		31	PE	F					P				
8	7	16	1977	1	1		36	DM	M			S					37	
9	7	16	1977	1	1		42	DM	F	Z							34	
10	7	16	1977	1	6		46	PF	F	Z							20	

Rodent metadata

Metadata for the entire project (rodent, plant, ant, precipitation)

Activity: What are the metadata, and where are they found?



NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE120263

Series GSE120263 Query DataSets for GSE120263

Status	Public on Aug 22, 2019
Title	Engineered adoptive T cell therapy prolongs survival in a preclinical model of advanced stage ovarian cancer [Mouse]
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	Purpose: Adoptive T cell therapy using high affinity T cell receptors (TCRs) to target tumor antigens has potential for improving outcomes in high-grade serous ovarian cancer (HGSOC) patients. Ovarian tumors develop a hostile, multicomponent tumor microenvironment containing suppressive cells, inhibitory ligands, and soluble factors that facilitate evasion of anti-tumor immune responses. Developing and validating an immunocompetent mouse model of metastatic ovarian cancer that shares antigenic and immunosuppressive qualities of human disease would facilitate establishing effective T cell therapies. Experimental Design: We used deep transcriptome

GEO example

How do we document metadata?

Include READMEs!

- One README per data file/set, as .txt file, named to reflect data file
- Format README files consistently
- Use standardized dates, and vocabulary that is convention for your field

Guide to Writing READMEs

README: General information

1. **Provide a title for the dataset**
2. **Name/institution/address/email information for**
    - **Principal investigator (or person responsible for collecting the data)**
    - Associate or co-investigators
    - Contact person for questions
3. **Date of data collection (can be a single date, or a range)**
4. **Information about geographic location of data collection**
5. Keywords used to describe the data topic
6. Language information
7. Information about funding sources that supported the collection of the data

Guide to Writing READMEs

Items in **bold** are most important

README: Data and file overview

1. **For each filename, a short description of what data it contains**
2. Format of the file if not obvious from the file name
3. If the data set includes multiple files that relate to one another, the relationship between the files or a description of the file structure that holds them (possible terminology might include "dataset" or "study" or "data package")
4. **Date that the file was created**
5. Date(s) that the file(s) was updated (versioned) and the nature of the update(s), if applicable
6. Information about related data collected but that is not in the described dataset

Guide to Writing READMEs

Items in **bold** are most important

README: Sharing and access information

1. **Licenses or restrictions placed on the data**
2. Links to publications that cite or use the data
3. Links to other publicly accessible locations of the data (see best practices for sharing data for more information about identifying repositories)
4. Recommended citation for the data (see best practices for data citation)

We'll talk more about licenses next week!

Guide to Writing READMEs

Items in **bold** are most important

README: Methodology

1. **Description of methods for data collection or generation** (include links or references to publications or other documentation containing experimental design or protocols used)
2. **Description of methods used for data processing (describe how the data were generated from the raw or collected data)**
3. Any software or instrument-specific information needed to understand or interpret the data, including software and hardware version numbers
4. Standards and calibration information, if appropriate
5. Describe any quality-assurance procedures performed on the data
6. People involved with sample collection, processing, analysis and/or submission

Guide to Writing READMEs

Items in **bold** are most important

README: Data-specific information

1. Count of number of variables, and number of cases or rows
2. **Data dictionary: Variable list, including full names and definitions (spell out abbreviated words) of column headings for tabular data**
3. **Units of measurement**
4. **Definitions for codes or symbols used to record missing data**
5. Specialized formats or other abbreviations used

Include a README like this for each file/dataset

Guide to Writing READMEs

Items in **bold** are most important

README: Data-specific information

We often use data published by other researchers (such as the human genome sequence). Metadata for your analysis in these cases should include:

- Where the data can be found, including a URL and data citation (if applicable; this information may be referenced in the license)
- How the data (from the original URL/source) were filtered/transformed/manipulated as a part of your analysis

Guide to Writing READMEs

How do we document biomedical metadata?

- Types of metadata:
    - **Reagent**: clinical samples, biological reagents, chemical reagents
    - **Technical**: auto-generated information from research instruments/software
    - **Experimental**: protocols, conditions, equipment
    - **Analytical**: data analysis methods, including software name/version, quality control parameters, output file types
    - **Dataset**: project objectives, personnel, publications, funding

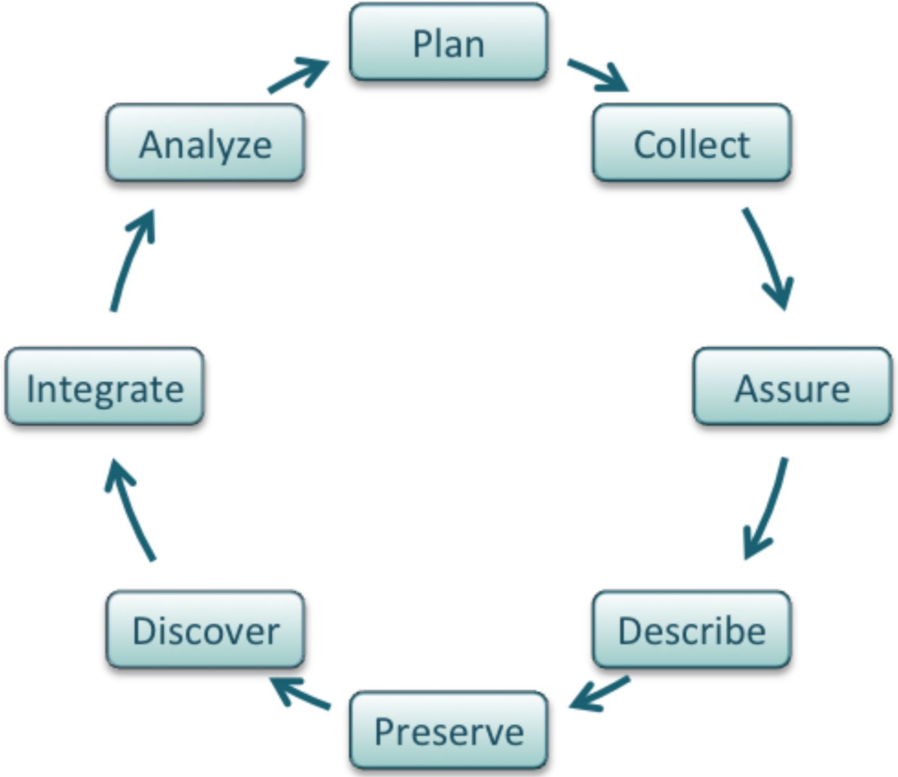
Harvard Biomedical Data Management

How do we document biomedical metadata?

- Include information about IRB approval
- Data security: consider where data are stored and who has access
- Document data that exists, but may not be stored with a given project
- File naming conventions can assist in documentation
- Recommended directory structures for large-scale projects can reflect components of metadata

Harvard Biomedical Data Management

Metadata and the data life cycle



Documenting information about a dataset and/or project is a part of the life cycle

Metadata facilitates all parts of the life cycle

Primer on Data Management

Activity: What data are available from the GDC Data Portal?

Where can you find the metadata?

Who can use these data?

[GDC Data Portal]

Why does metadata matter?

What questions do you have about documentation of metadata for your projects?

Summary

- Metadata includes multiple types of information, some of which may not seem important while conducting the research, but are essential for others being able to understand it
- Every data file should have an associated metadata file that describes it
- Some metadata is better than no metadata!
- Most metadata represents things you’ll need to include in a publication (or other deliverable), so document early and often

Next time: data manipulation and reproducibility

Resources

- Guide to writing “readme” style metadata from Cornell University
- Harvard Biomedical Data Management: Metadata overview, READMEs