Data and Donuts

Participant Guide

Facilitators:
Nicole Vasilevsky, PhD
Jackie Wirz, PhD

OHSU **| BD2K Skills Course**

Data and Donuts

Table of Contents

About Your Facilitators:

Nicole Vasilevsky, PhD, Senior Biocurator, Ontology Development Group, Library

Nicole has been with the OHSU Library for 6 years, and she does biocuration and ontology development for semantic data integration projects for improving disease diagnosis, meaning she deals with lots of data and tries to make it more structured so it is machine readable and more meaningful and accessible to researchers. Her research and teaching interests are in data management, scientific reproducibility and data sharing. She has a PhD in Cell Biology from OHSU. Her graduate studies involved a lot of mice and trying to optimize her Western blot protocol.

Jackie Wirz, PhD, Assistant Dean, School of Medicine

Jackie recently joined the School of Medicine as the new Assistant Dean for graduate affairs and founded and directs the Career Development Office. Prior to her new position as Assistant Dean, she worked in the OHSU Library for 6 years where she was a liaison to Biomedical Researchers. She received her PhD from OHSU in Biochemistry. Her best memories from graduate school included working with toxic chemicals and breaking expensive instruments.

Instructional Designer:

Bjorn Pederson, Instructional Designer, DMICE

Bjorn joined the DMICE department as the Instructional Designer and Project Manager for the BD2K projects in 2015. He has over 10 years of experience in education and learning technology working with K-12 and adult learners. He assists with development of the BD2K Open Education Resources and implementation of the BD2K Skills Courses. He works closely with the stakeholders and content specialists to design strategies for discovery and direction, as well as assess presented information and skill sets. Bjorn enjoys spending time with his toddler and watching his wife watch tennis.

These materials were derived from prior work by developed by Melissa Haendel, Shannon McWeeney, David Dorr, Nicole Vasilevsky, Ted Laderas, Jackie Wirz and Bjorn Pederson.

Funding:

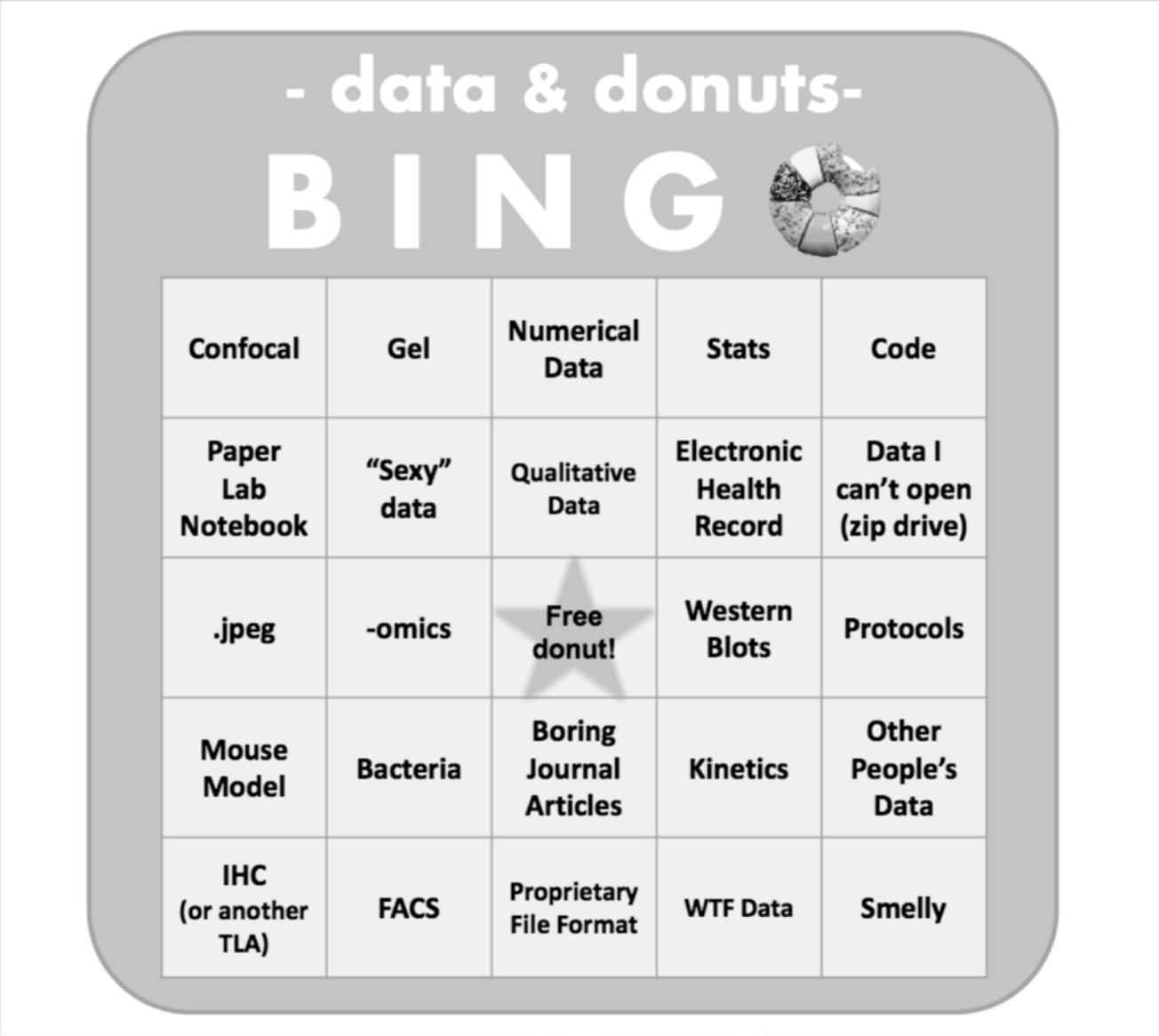
These materials are supported by the National Institute of General Medical Sciences, funded by the NIH Big Data to Knowledge Initiative, under Award Number 1R25GM114820. The principal Investigators are Melissa Haendel (Library and DMICE), Shannon McWeeney (DMICE) and David Dorr (DMICE).

Learning Objectives

After this course, you should be able to:

- Understand the diversity of data and begin to understand the issues with data reuse
- Apply best practices for creating a file directory and naming your files
- Understand what version control is and when and how you should use it
- Apply best practices for data backup and storage
- When your manuscript is accepted to Cell (or some other awesome journal), you will know how to make your methods section is structured, transparent, and accessible
- When you graduate/leave the lab, ensure your data is findable and reusable by your lab mates and others

Data Bingo



Confocal	Gel	Numerical Data	Stats	Code
Paper Lab Notebook	"Sexy" data	Qualitative Data	Electronic Health Record	Data I can't open (zip drive)
.jpeg	-omics	Free donut!	Western Blots	Protocols
Mouse Model	Bacteria	Boring Journal Articles	Kinetics	Other People's Data
IHC (or another TLA)	FACS	Proprietary File Format	WTF Data	Smelly

The Gummi Bear Exercise

Go to link:

Add link

Image credit: http://speedynebula13.blogspot.com/2010/10/gummy-bear-anatomy.html

Current Data Management Challenges

1. Do you have previous lab experience?

2. Some troubles we have encountered with (lack of) data management (check if these apply to you too):

- ☐ Lab notebook “handoff”
- ☐ I won’t need to access this data in [one year, two years, etc] from now
- ☐ What is metadata?

Growth assay.xlsx

	Day 1	Day 2	Day 3	Day 4	Day 5
GFP + dox	104.3	168	260	428	772
	98.3	156	212	360	800
	116.7	172	280	420	364
	91	192	192	140	292
	113	224	260	356	1004
GFP - dox	172	140	212	184	456
	116	228	308	412	576
	125	180	228	372	512
	135	136	220	268	612
	78	192	204	204	596
Mad4 + dox	131	156	152	168	340
	130	156	124	172	360
	132	140	152	264	360
	100	160	152	152	468
	124	184	208	188	540
Mad4 - dox	151	148	216	228	588
	116	228	212	312	608
	137	208	216	298	612
	142	268	216	320	808
	129	220	348	404	740

Your files on your computer

How many icons do you have on your desktop?

- ☐ My desktop is completely covered with icons
- ☐ My desktop is half covered with icons
- ☐ Less than half of my desktop is covered with icons
- ☐ I have only one icon saved on my desktop and it is solitaire (or your favorite game)

Within "My Documents" are there individual files or is everything filed in folder?

- ☐ Yes, I have tons of files saved under "My Documents" (or whatever folder I used to mainly save my documents and files)
- ☐ No, all of my files are neatly organized in file folders
- ☐ None of these apply because all my files are saved on my desktop

Do you use the built-in search function whenever you need to find a file?

- ☐ Yes, almost always
- ☐ Sometimes
- ☐ No, I fumble around until I can find the file I'm looking for
- ☐ Never

Are your files on your computer organized by:

- ☐ Class
- ☐ Project
- ☐ Chronologically
- ☐ Other

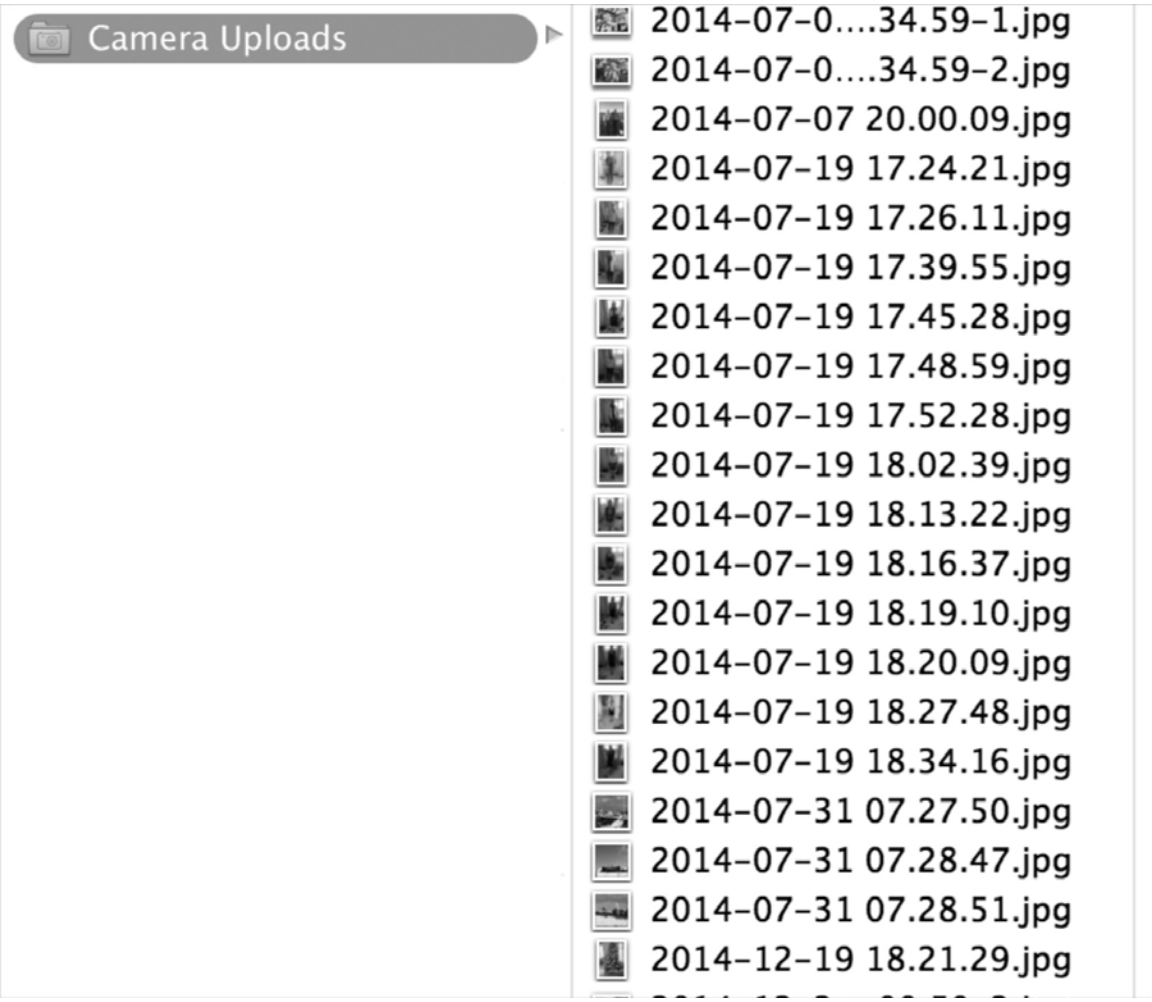
Do you save multiple copies of your files in different folders?

- ☐ Yes, often or always
- ☐ Sometimes
- ☐ Never

Do you ever use ReadMe files?

- ☐ Yes
- ☐ Never
- ☐ I don't know what that is

Searching for files



Do you use the search function on your computer to find your files?

Tips on organizing your files

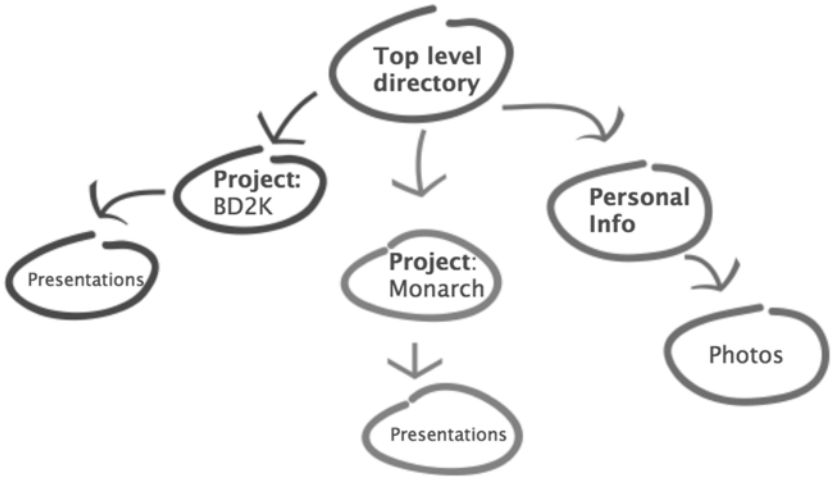
Using the search function will not necessary work if you are trying to perform the following tasks:

- **Find files manually**
  - When I use the search function on my Mac, I get a lot of irrelevant results
  - A lot of my files have similar names, so I get a lot of results that match the file name I'm looking for
- **Find groups of similar files**
  - If you don't have your files organized into a file directory, it can be hard to search for groups of files, such as 'work' files, 'music' files, etc.

Tips on organizing your files:

1. **Decide on a file directory structure**
   a. Think about what kinds of files you have on your computer
      i. Raw data
      ii. Analyzed data
      iii. Presentations
      iv. Manuscripts
      v. Coursework
      vi. OHSU related files
      vii. Personal files
      viii. Photos
      ix. Music
   b. Write down some of the types of files you have below and/or circle the files above

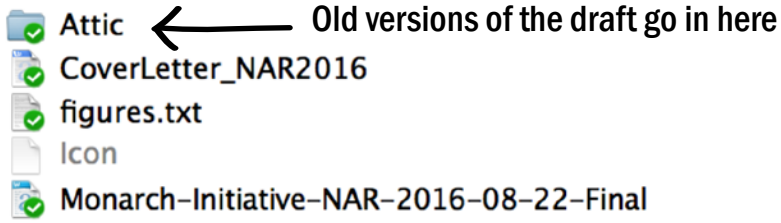
2. **Create a mindmap**



   a. What are some strategies you could use to organize your files? For example, organize by project, or organize by date. How do you organize your directory? (Or how do you think you could improve your directory organization?)

3. **Use subfolders.** Create some top level folders and then create subfolders. However, think about how often you need to access particular files, if they are buried deep in subfolders, it could be annoying if you need to access them frequently. You could create a 'working' folder for something you are working on often, then move it to an 'archive' folder when you are done. (Or you can create a shortcut or alias, see #4 below).

   Example: My colleagues and I were working on a manuscript and we kept the current version of the manuscript in the folder NAR2016. As we created new versions of the draft, the old files were moved to a folder called 'Attic'.



4. **Create shortcuts/Aliases.** A shortcut allows the file to be in two places at once. You can create a copy of your file and save it in your nicely organized file directory, and then keep a shortcut (Windows)/alias (Mac) somewhere more accessible like your desktop or your top level folder. You can delete the shortcut without losing any data. You can create shortcuts for working copies

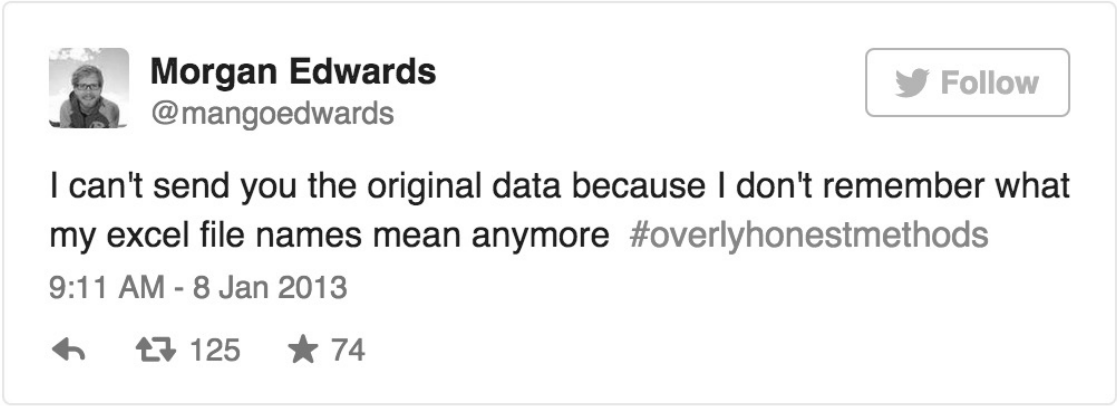
of your files or just if you would like your file to be organized into more than one folder at once.

5. **Create an 'inbox' folder** for files that have yet to be filed. Be diligent about clearing out your inbox folder.
    - I use my desktop as my inbox.

6. **When you decide on your system, stick to it!** Being consistent and disciplined will help you stay organized going forward. It is worth the time to file your files and data into the appropriate folders, it will save you time in the long run. You can always start being more organized going forward and slowly organize your old files over time. You don't have to clean up your entire computer all at once.

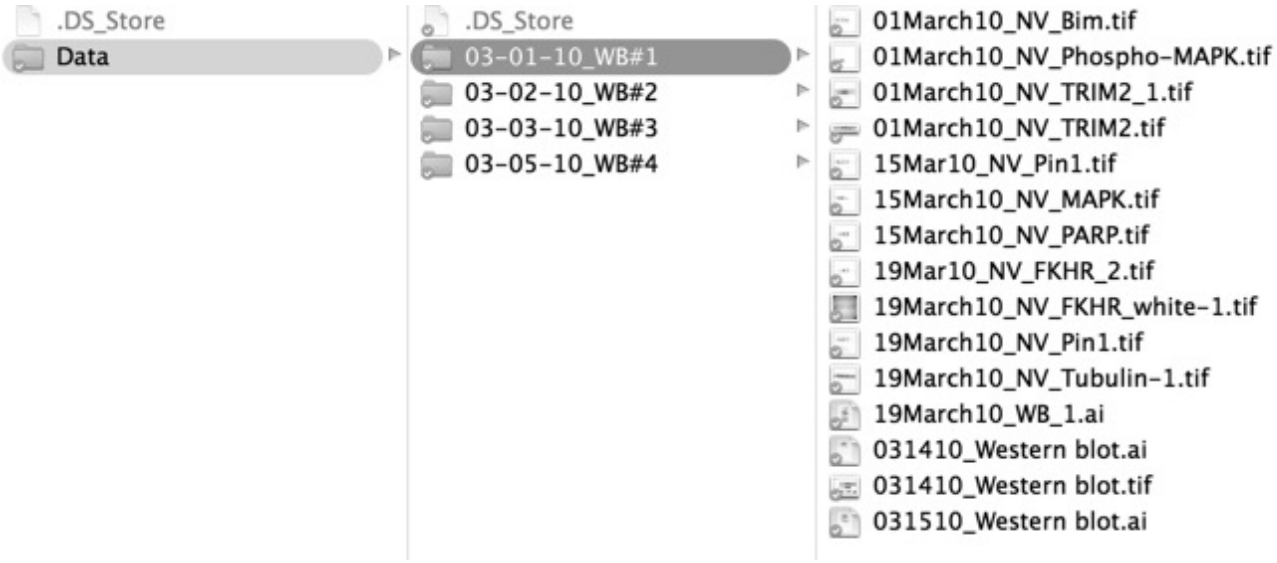
Reference: http://www.howtogeek.com/howto/15677/zen-and-the-art-of-file-and-folder-organization/

File naming



Case Study:

You arrive in your new lab for rotation and you are told you need to find the old files from the previous graduate student and review their previous data. Review the screenshot below and answer the following questions:



1. What do you think the subfolders of Data are referring to?

2. What other type of information can you gather from these files? Do you have any idea what these file names mean?

3. What if the researcher had included a ReadMe file that included the following information:

4. What does the first file name mean? (01March10_NV_Bim.tif)

5. What are some other issues with the file names? How could NV improve the file names so they are easier for her and others to interpret at a later date?

Recommendations for file names:

1. Avoid special characters and spaces

*/ \ : * ? “ < > [ ] $ &*

Why? Special characters can have specific meanings in your computer’s operating system and using them in file names can cause issues

2. Add version numbers at the end, ie, v1, v2, every time you make significant changes to the file. Add ‘final’ to the end when the file is finalized.

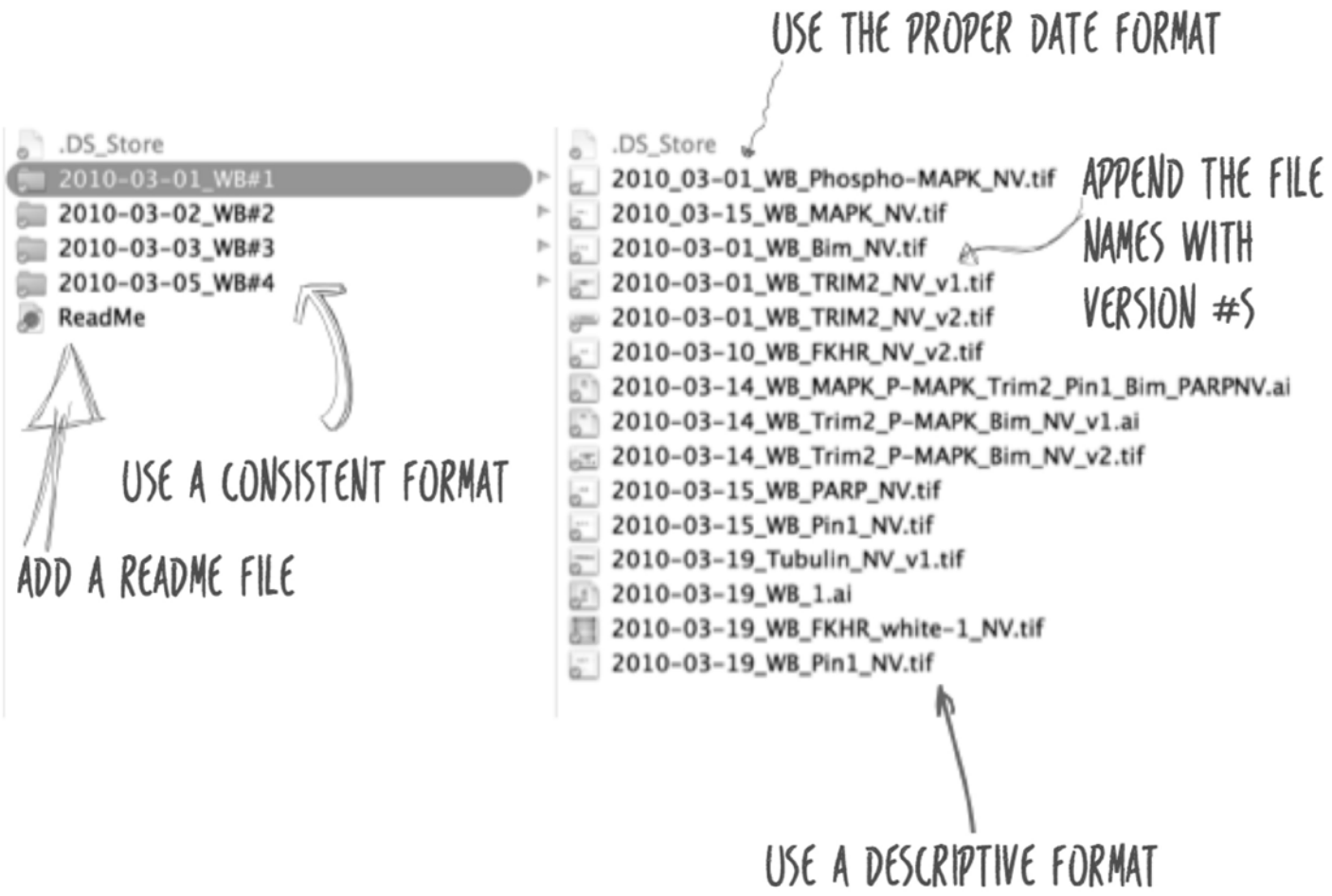
Example:

DataManagement101_Handout_v1.docx
DataManagement101_Handout_v2.docx
DataManagement101_Handout_Final.docx

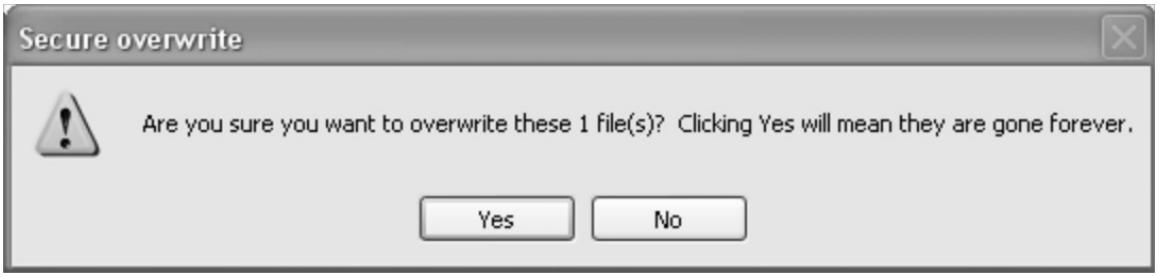
3. **Use the ISO data format**

YEAR-MONTH-DAY
DataManagement101_Handout_**2016-09-15**.docx

Example of improved file names



Version Control



Do you use any of the following version control software?

- ☐ Dropbox
- ☐ Box
- ☐ Google docs
- ☐ Github

Advantages of using version control software:

Version control software means it manages changes to a project without overwriting any part of that project.

- Can go back when you make mistakes
- when changes are made
- Share work with other people
    - Both work on things at the same time and merge back together
- Akin to game of telephone- version control can let you see exactly when a change was made

Long-term data storage

1. Do you know what metadata is?

- ☐ Philosphy
- ☐ Describes data
- ☐ Dating site
- ☐ A metal band

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.

Metadata Exercise

1. Get our your phone and take a selfie (or chose a picture already saved on your phone or computer).
2. Now pretend this is an important piece of data for your research project. What kind of metadata could you apply to this picture? Write down the metadata about this file in the table below. See the instructions below as to the types of metadata you may want to include.

Type of metadata	**Your metadata**
Example: Title of picture	Selfie at Data and Donuts course.
Example: Type of file	Image (or jpeg, etc.)

Metadata guidelines:

Introductory information

1. For each file, 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. Name/institution/address/email information for
    - Principal investigator (or person responsible for collecting the data)
    - Associate or co-investigators
    - Contact person for questions
5. Date of data collection (can be a single date, or a range)
6. Information about geographic location of data collection
7. Date that the file was created
8. Date(s) that the file(s) was updated and the nature of the update(s), if applicable
9. Keywords used to describe the data topic
10. Language information

Sharing/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
4. Recommended citation for the data
5. Information about funding sources that supported the collection of the data

A more comprehensive list is available at: http://data.research.cornell.edu/content/readme

Where do you put your metadata?

✓ You can save it in a spreadsheet or create ReadMe files.

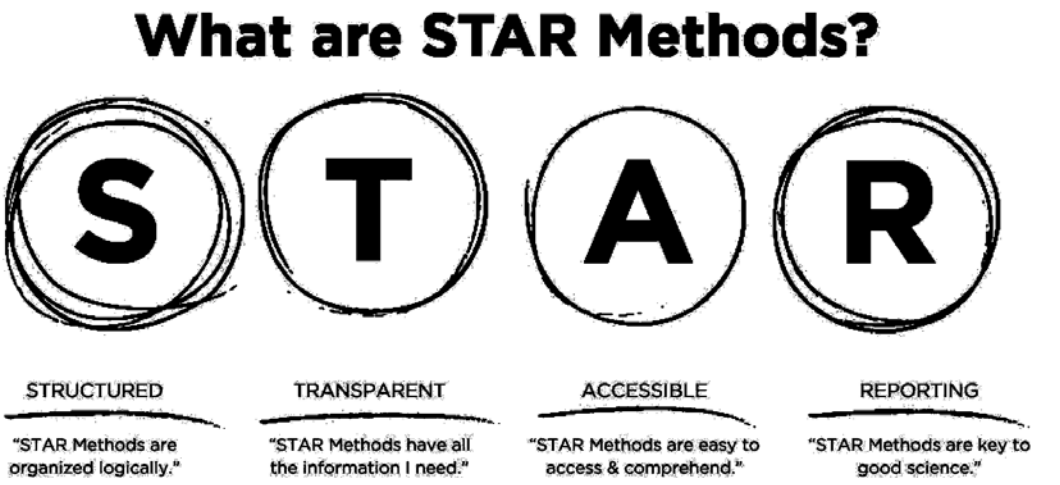
Create 'ReadMe' style metadata:

Recommended practices:

1. Create a ReadMe file in a text editor like Notepad (Windows) or TextEdit (Mac). Save the file in the folder with your dataset. (Or if you are using a spreadsheet, you can create the ReadMe file in the first tab of the file.)
2. Save the ReadMe file using a file name that describes the associated dataset. Example, 'ReadMe_WesternBlot_LymphNodes_2016-09-15.'
3. Use consistent formating in your ReadMe files, ie outline the information in the same order and use consistent terminology.
4. Use the ISO date format: YYYYMMDD.
5. Follow the scientific conventions for your discipline, whenever possible, use terms from standardized taxonomies and vocabularies. Standardized taxonomies and vocabularies can be found at:
    - Biosharing.org
    - Bioportal.org
6. Use unique identifiers when possible for resources like model organism strains or antibodies.
    - Model organism databases:
        i. Mouse: MGI (informatics.jax.org)
        ii. Zebrafish: ZFIN (zfin.org)
        iii. Rat: Rat Genome Database (rgd.mcw.edu/)
        iv. Fruit fly: Flybase (flybase.org)
        v. Yeast: Saccharomyces Genome Database (SGD) (yeastgenome.org)
        vi. Worm: Wormbase (wormbase.org)
        vii. Frog: Xenbase (xenbase.org)
    - Antibody registry (antibodyregistry.org)

Unique Identification of Resources

Journals are increasing their requirements for reporting methods sections and data sharing at the time of publication. For example, the journal Cell recently adopted the STAR methods.



One aspect of the STAR methods is including a 'key resources' table in the methods section. A problem with scientific reproducibility is the ability to recreate another scientist's experiment based solely on the information in the methods section.

Have you ever tried to reproduce a protocol based on the methods reported in a paper?

- ☐ Yes
- ☐ No

If yes, was there sufficient information reported that you could reproduce that experiment?

- ☐ Yes
- ☐ No

If you have published scientific work, did you include the manufacturer, catalog number or other uniquely identifying information for your research resources, for resources like antibodies, or plasmids, or animal strains?

☐ Yes
☐ No

Issue of reproducibility:

Instructions to authors claim that a methods section should include enough information for another researcher to reproduce their experiment, but this is not often the case. An analysis showed that most studies do not provide unique identifiers for research resources such as antibodies, plasmids, model organisms, knockdown reagents or cell lines (*PeerJ* 1:e148).

The Resource Identification Initiative provides a portal where you can obtain unique identifiers for research resources and report them in your manuscript. Go to: https://scicrunch.org/resources

For additional details on recommended reporting practices see: Recommended reporting guidelines for life science resources (https://biosharing.org/bsg-s000532)

Where to store your data

If I gave you data stored on the following device, would you be able to open it?

- ☐ Floppy disk
- ☐ CD
- ☐ Zip disk
- ☐ USB key
- ☐ Cloud (ie dropbox, google drive, etc.)

Data backup

What kind of data backup do you use?

- ☐ Hard drive on my computer
- ☐ External hard drive
- ☐ Cloud drive
- ☐ External storage like USB key, CDs
- ☐ Network drive
- ☐ None
- ☐ Other:

“A hard drive can last anywhere from 6 seconds to 6 years”
- OHSU IT department

What are other ways we can lose our data?

ORCID

DISTINGUISH YOURSELF IN THREE EASY STEPS

ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized.

www.orcid.org

Do you have good data management practices?

- ☐ Are your folders organized on your computer?
- ☐ Do you implement a file naming convention?
- ☐ Do you use version control? (Either version control software or just simple version control in your file names.)
- ☐ Do you know where to find unique identifiers for your research resources?
- ☐ When you are ready to publish your data in Cell, will your data conform to the STAR methods?
- ☐ Do you have a backup of your data?
- ☐ It is recommended that you have three copies of your data, do you have three copies of your data backed up?
- ☐ Do you have an ORCID ID?
- ☐ Will you apply any of the best practices we talked about today to your own data?

Resources

BD2K Open Educational Resources:
https://dmice.ohsu.edu/bd2k/
Also available at: https://github.com/OHSUBD2K/BDK01-Biomedical-Big-Data-Science

Metadata:
Understanding Metadata – NISO:
http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Guide to writing "ReadMe" style metadata:
http://data.research.cornell.edu/content/readme

Relevant Publications:
Haendel MA, Vasilevsky NA, Wirz JA (2012) Dealing with Data: A Case Study on Information and Data Management Literacy. PLoS Biol 10(5): e1001339. doi:10.1371/journal.pbio.1001339

Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, LaRocca GM, Haendel MA. (2013) On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ* 1:e148 https://doi.org/10.7717/peerj.148

Notes

OHSU | **BD2K**

www.ohsu.edu/bd2k

Contact us:
Jackie Wirz: wirzj@ohsu.edu
Nicole Vasilevsky: vasilevs@ohsu.edu