Organize Your Files

HARVARD LONGWOOD MEDICAL AREA
RESEARCH DATA MANAGEMENT WORKING GROUP

The Great Data Cleanup Campaign Summer 2020

Welcome to today's short online seminar on organizing your files.

This session is being recorded, and you will receive an email when the recording and slides are available.

If you have questions at any time throughout the presentation, please enter them into the chat and I will try to answer them at the end.

Julie Goldman

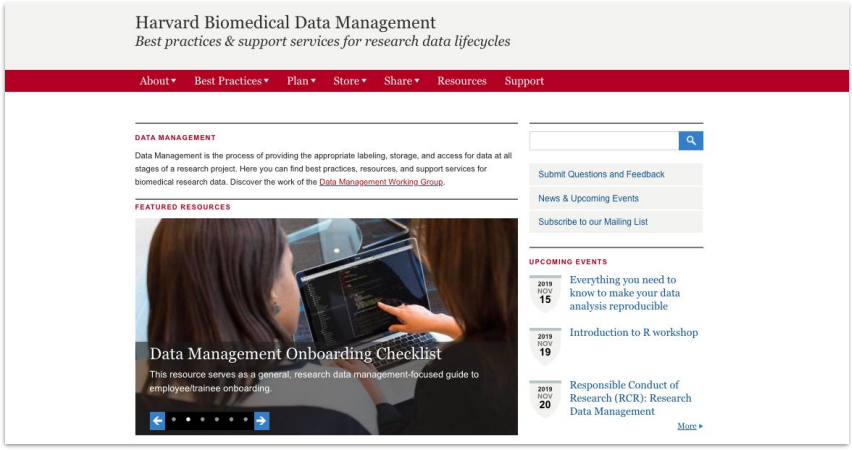
Research Data Services Librarian
Countway Library

julie_goldman@harvard.edu

I'm Julie Goldman, Research Data Services Librarian at Countway Library.

I offer a lot of data management training classes, but I am also available for consultations on identifying best practices, creating a data management plan or finding data.

https://datamanagement.hms.harvard.edu

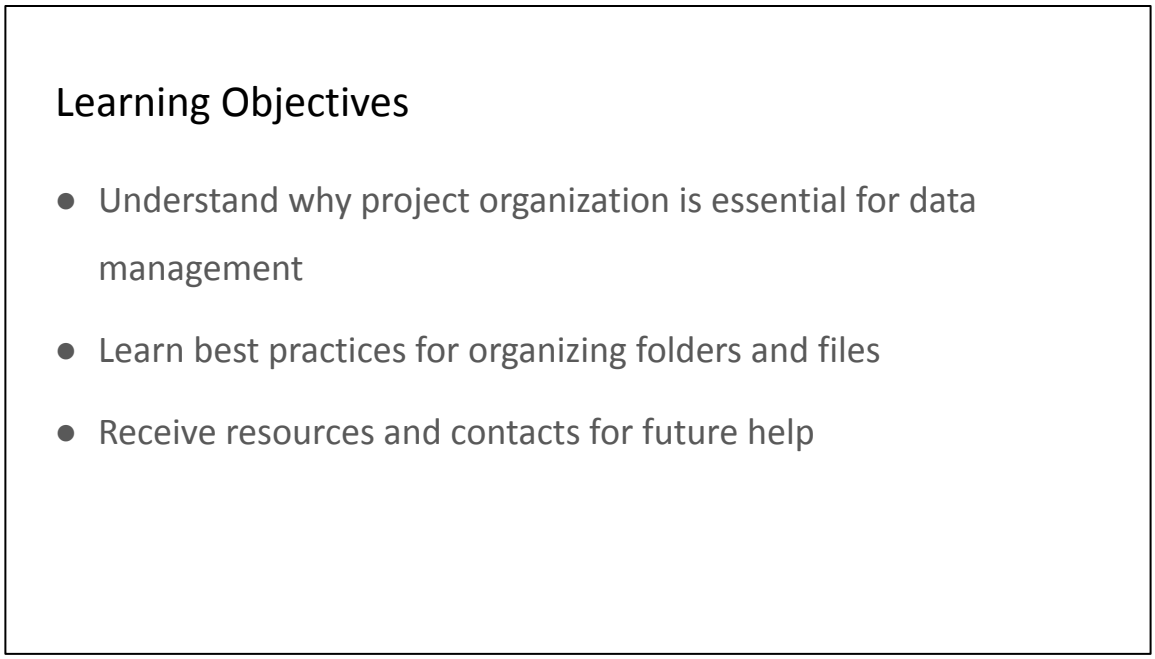


I also co chair The LMA Research Data Management Working Group.

This working group includes members of the LMA community from many areas such as the library, information technology, research computing, research cores, and affiliates hospitals.

And we maintain a website full of resources and information.

We suggest you bookmark the website and reference it as needed.



Let's review today's Learning Objectives:

- Understand why project organization is essential for data management
- Learn best practices for organizing folders and files
- Receive resources and contacts for future help

In this session I will present some best practices for organizing files.

These are simply some principles and practices to get you started, and thinking about what will work for you.

Depending on you research area and type of research, you may find a more optimal way to organize your work.

So I invite you to take the principles shared today and adapt them as you need.

What are some reasons for systematically organizing research and data files?

Let's start off with a chat question.

So in the chat, type in why you think it would be beneficial or useful to organize your research data files.

What are some reasons for systematically organizing research and data files?

Participant Answers

These are all great reasons!

- Easier to find files
- Findability
- Locating them later
- Preservation
- Auditing
- Easier to find and reference related files
- Find the last version after more than a month.
- Reproducibility and analysis tracking
- Avoid double files
- Others can make sense of your system.
- Easier to access
- To avoid missing files by forgetting its location
- Easy access, sharing and building of publication grade figures
- Easy to deal with them
- Adds to consistency and stability of project
- Easy to retrieve!
- To expedite project workflows
- Save time searching
- Just because you remember which files are which now, doesn't mean you'll remember later
- Easy to keep track
- To facilitate finding
- Easy for collaboration
- Version control
- Other team members could use in future
- Easy to share with lab members
- Would help to get a sense of the data and information long after having done the project.
- Makes life easier when you come back to it
- Reproducibility and reuse
- Cut clutter

Let's start off with a chat question.

So in the chat, type in why you think it would be beneficial or useful to organize your research data files.

What are some reasons for systematically organizing research and data files?

- Easier to locate a file
- Find similar files together
- Moving files becomes much easier
- Easy to identify which files you want to back up
- Keep organized in the long-run
- Increases productivity
- Helps you to keep and maintain a record of the project
- Projects can easily be understood by others

What are some reasons for systematically organizing research and data files?

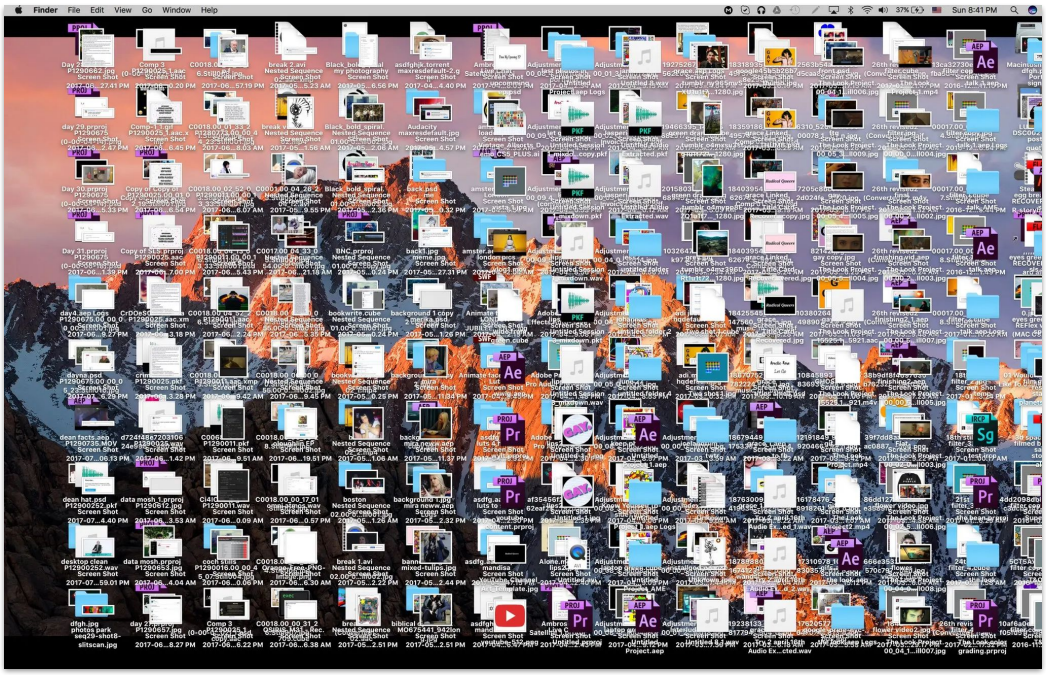
- Easier to locate a file
- Find similar files together
- Moving files becomes much easier
- Easy to identify which files you want to back up
- Keep organized in the long-run
- Increases productivity
- Helps you to keep and maintain a record of the project
- Projects can easily be understood by others

Let's start off with a chat question.

So in the chat, type in why you think it would be beneficial or useful to organize your research data files.

What are some reasons for systematically organizing research and data files?

- Easier to locate a file
- Find similar files together
- Moving files becomes much easier
- Easy to identify which files you want to back up
- Keep organized in the long-run
- Increases productivity
- Helps you to keep and maintain a record of the project
- Projects can easily be understood by others



I don’t mean to call anyone out or shame anyone, but does your desktop look like this?

In this situation, it is very likely there are multiple copies of files, multiple versions files.

Making it very challenging to find anything!

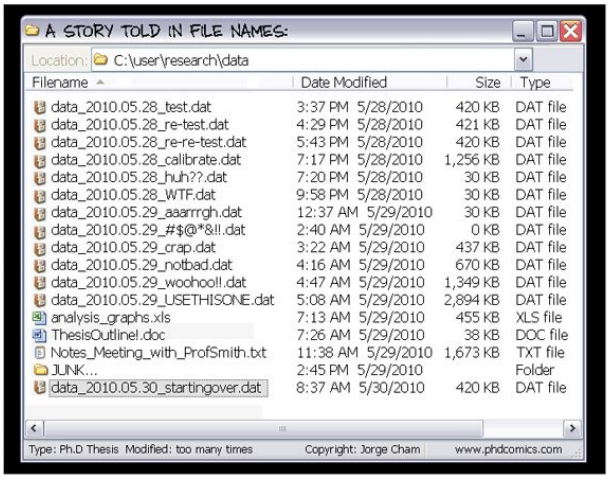
So we need to either take a bit of time to clean this up.

Or, even better, plan for organization ahead of time!

In this session, of course we are going to focus on digital research project files and data, but you could potentially apply the practices talked about today to your real life and personal desktop!

Case for File Organization

- Classifying and organizing files to make them more useful
- Enable “someone” to look at your files and understand in detail what you did and why
  - ***Hint: that someone is you!***
- Establish systems, document them, and use consistently
- Any system is better than none



And don’t just move all those files from your desktop into a folder! Like we see here, it is still impossible to find or know what anything is!

Bringing this back to data management, managing your data is all about organization across a project that allows for accurate communication about that project.

Data organization refers to the method of classifying and organizing data sets to make them more useful.

Once you create, gather, or start manipulating data and files, they can quickly become disorganized.

And poor organizational choices can lead to significantly slower research progress.

But, the core guiding principle for file organization is simple:

Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why.

This “someone” could be any of a variety of people:

Someone who read your published article and wants to try to reproduce your work, a collaborator who wants to understand the details of your experiments, a future student working in your lab who wants to extend your work after you have moved on to a new job.

Most commonly, however, that “someone” is you.

A few months from now, you may not remember what you were up to when you created a particular set of files, and may end up having to spend additional time reconstructing your previous experiments.

Or even more simply, you may actually need to do something again!

You may discover a flaw in your initial analysis, or you get access to new data, or you discover a better technique.

However, if you have organized and documented your work clearly, then repeating the work will be much, much easier.

And remember, one size does not fit all.

So find and establish a system that works, document the system so everyone can follow and use it!

What are some some strategies to organize research project folders and files?

So now we understand why we are talking about this today!

Next, I invite you to think about how you could organize your files. So type in the chat...

What are some some strategies to organize research project folders and files?

In other words, what elements would be helpful for organizing files?

Participant Answers

These are all great ways to organize folders and files!

Organize Folders

- Use date
- Date, project, experiment
- Create folders with an experiment code as the title and put all relevant files in that folder
- Date, pi, subject
- Title or topic, date
- By project, version number, name of last writer.

Organize Folders

- By components to the project
- Date, noun that describes file, date and file number that may be another version
- Folders divided in type of files
- Subdivide in projects; type of experiment, date and add information in the name
- Based on categories - raw data, code, etc

Organize Files

- year_month_day in the filename
- Date at beginning of the file
- ISO standard for date
- Descriptive titles
- Chronologically
- No special characters or spaces
- For revisions R1, R2, etc

So now we understand why we are talking about this today!

Next, I invite you to think about how you could organize your files. So type in the chat...

What are some some strategies to organize research project folders and files?

In other words, what elements would be helpful for organizing files?

What are some some strategies to organize research project folders and files?

- Year
- Type of document
- Project
- Project stages
- Team member
- Experiments
- Instruments
- Time period
- Geographic location
- Institution or project site

So now we understand why we are talking about this today!

Next, I invite you to think about how you could organize your files. So type in the chat...

What are some some strategies to organize research project folders and files?

In other words, what elements would be helpful for organizing files?

- Year
- Type of document
- Project
- Project stages
- Team member
- Experiments
- Instruments
- Time period
- Geographic location
- Institution or project site

Top 5 Organization Practices

Establish systems and use them consistently

One project, one folder with nested folders

Create separate folders for data or project stages

Determine organization of files within folders

Create README.txt files in higher level directories

Keeping track of research data and documentation is critical.

Here are five top practices we are going to walk through, and then I will show some examples of these in action.

Establish a System

- Spend time planning out folder hierarchy and file naming conventions in the beginning of a project
- Consider how you or others will look for and access files later
- Establish a system as a group and prioritized for implementation
- Provide a method for easy adoption
  - A shared dropbox with the folder hierarchy in place
  - README file in onboarding documentation for new contributors

First, establish a system.

Think hard at the beginning of your project about how you are going to organize your data as it grows.

Consider how you or others will look for and access files at a later date

Remember when we talked about data management plans, it is important to include plans for data organization, such as your directory organization and file naming conventions.

So establish a system as a group and prioritized for implementation when the project starts.

Also provide a method for easy adoption, so that people actually follow it!

Providing a template with the folder structure already planned out, and a README file for documentation when someone joins the lab.

And then of course refine any workflows or processes as the project evolves.

One tip for designing your system is to start small.

Consider just a project or just an experiment and then expand from there.

When developing a file structure, the first step should be to evaluate all of the files and create a structure that the files logically fit in rather than making the files fit the structure. This will allow for faster recall of the files. Create the entire folder hierarchy with empty folders before adding any files to the structure then you can start moving the files from their old location to the new file structure. It will quickly become clear if the file structure is meeting the needs of the files, if not, new folders will need to be created.

Organize Folders Hierarchically

- Arrange folders and files hierarchically
- One project, one folder
- Create directories that follow a consistent pattern
- Good at representing the structure of information
- Avoid overlapping categories
- Don’t let your folders get too large
- Don’t let your structure get too deep

For your folder structure, arrange your files hierarchically, meaning folders within folders, sometimes referred to as directories.

It is a series of nested folders, each containing files, folders, and executables.

Directories that are inside other directories are often referred to as subdirectories.

Put each project in its own directory, which is named after that project.

Ideally, keep the project name under 32 characters and include a combination of: the project title, a unique identifier, and the date.

So establish a folder hierarchy that aligns with the project.

Within the project folder, you should separate additional folders and files based on attributes, which we will talk about next.

This allows you to keep similar items stored together.

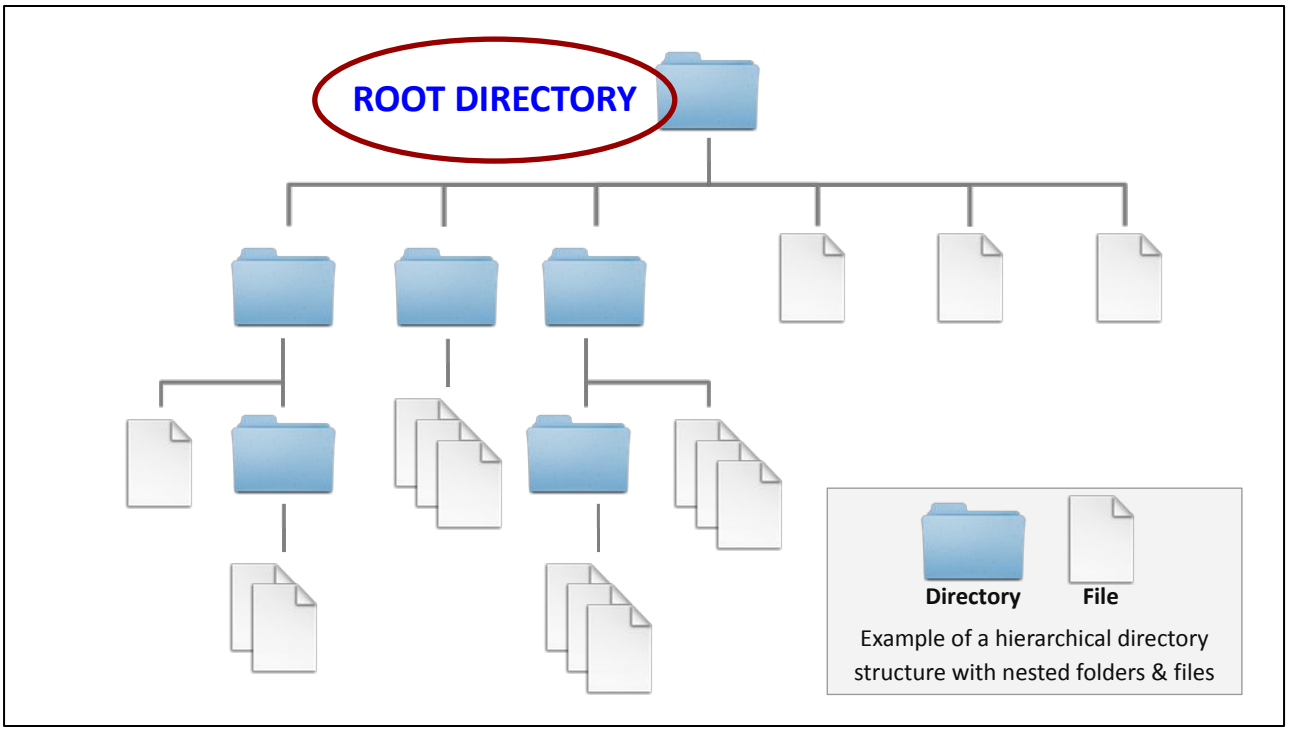
This organizational structure is good because it is familiar and widely used.

It also is good at representing the structure of information and the files generated throughout a project.

However, you need to be careful of a few things:

Since items can only go on one place, avoid overlapping categories.

And don’t let your folders get too large or your structure too deep.



Here is an example structure of a directory hierarchy.

We have a root or parent directory, this may be the top level directory such as your home directory or a collaborative lab folder many people are working together in.

This directory has subfolders or subdirectories and files.

Folders may continue to nest, based on categories your choose to organize by.

This could be the project folder scaffolding you create before a research project begins.

Create Folder Attributes

- Identify ways to divide files into categories or attributes
  - Project
  - Time
  - Location
  - File type
  - Raw data
  - Analyzed data
  - Code
  - Methods
- Dependent on number of files and what aspect is most important
- Top level organization is the most important attribute

So in order to create this folder structure, or directories, we identify ways we can divide our files into categories, or attributes.

Top level organization is the most important attribute since it will dictate how the rest of the folders and files are divided.

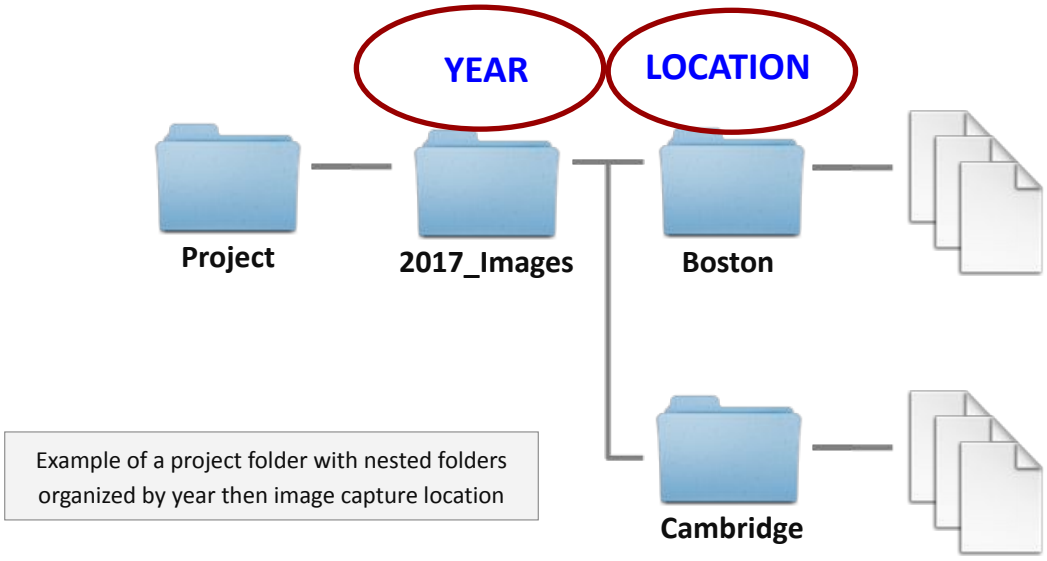
Attributes may be high level projects, your data or project stages, data collection location, based on time, or file type.

Ideally broader topics will be at the top level of the hierarchy and more specific topics lower in the structure.

How files are nested is dependent on the number of files you are working with, and what aspect of those files is most important for analyzing or re-using the information in them.

As long as everyone on the project is using the same principles to organize and name files, then the system is up to you and will be dictated by the type of project and data you are working with.

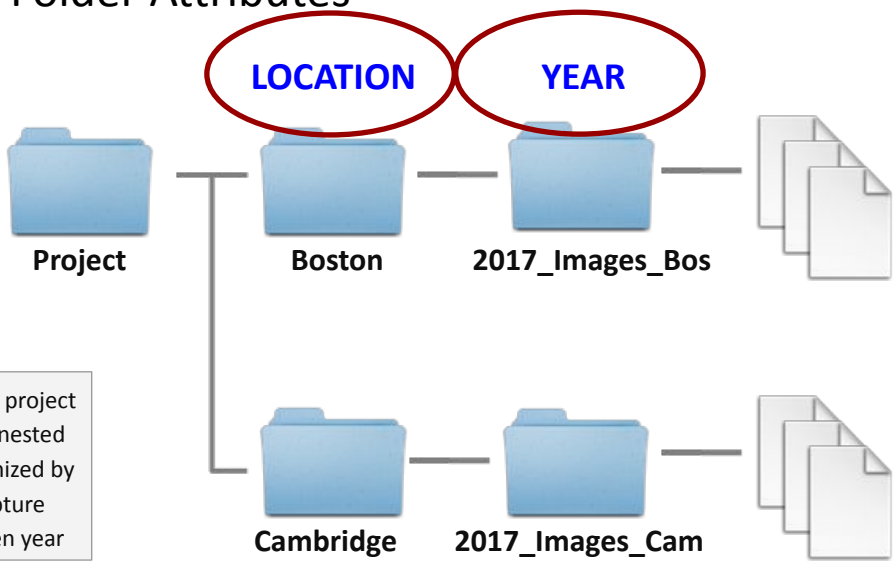
Create Folder Attributes



Example of a project folder with nested folders organized by year then image capture location

So for instance, if you have hundreds of thousands of image files collected over many years from many different locations, you may want to organize first by year then location.

Create Folder Attributes



Example of a project folder with nested folders organized by image capture location then year

Or organize by location, then by year.

Again, how you organize folders depends on what attributes are important to your project, and how many files you have.

Organize Files Systematically

- Determine organization of files within folders
- Consider sorting to decide what element of the file name will go first

Chronologically (ISO 8601 date standard)	**Classification or code** (standardized)	**Alphabetically** (depending on type of files)
20171028_001.tiff	USNM_379221_01.tiff	bos_20171028_001.tiff
20171028_002.tiff	USNM_379221_02.tiff	bos_20171028_002.tiff
20171028_003.tiff	USNM_379221_03.tiff	bos_20171029_001.tiff
20171029_001.tiff	USNM_379222_01.tiff	cam_20170922_001.tiff
20171029_002.tiff	USNM_379222_02.tiff	cam_20170922_002.tiff

Next, by organizing files systematically, it will minimize the time you spend looking for things.

So within your folder structure, determine the organization of files.

Consider sorting when deciding what element of the file name will go first.

Will this be chronologically, by classification or code, or alphabetically?

File names starting with YYYY-MM will sort differently than files starting with the MM-DD-YYYY form.

Best practice, if you are working with dates, is to use the ISO 8601 date standard, which is YYYY-MM-DD.

This allows your files to successfully sort chronologically.

However, you may have a classification or code system, or alphabetically works if you have a group of them same kind of files.

So yes, this sorting will really be dependent on your file naming convention, which we are going to save for next week!

But I include it here because it is an important aspect to consider when developing your directory structures as this organization may be dependent on those folder attributes we previously identified.

Create Documentation Files

- Need to make sure we are capturing metadata/information about the contents of folders and files
  - Project & contact information
  - File naming conventions
  - Who made it, when, and where
- Create this documentation or README.txt files across level directories briefly describing their contents

The final practice for creating your directory structure is to create documentation files.

What if we give data access to a collaborator, a new student joins the project, or it's three years later and we have forgotten when the various stages of the project
happened?

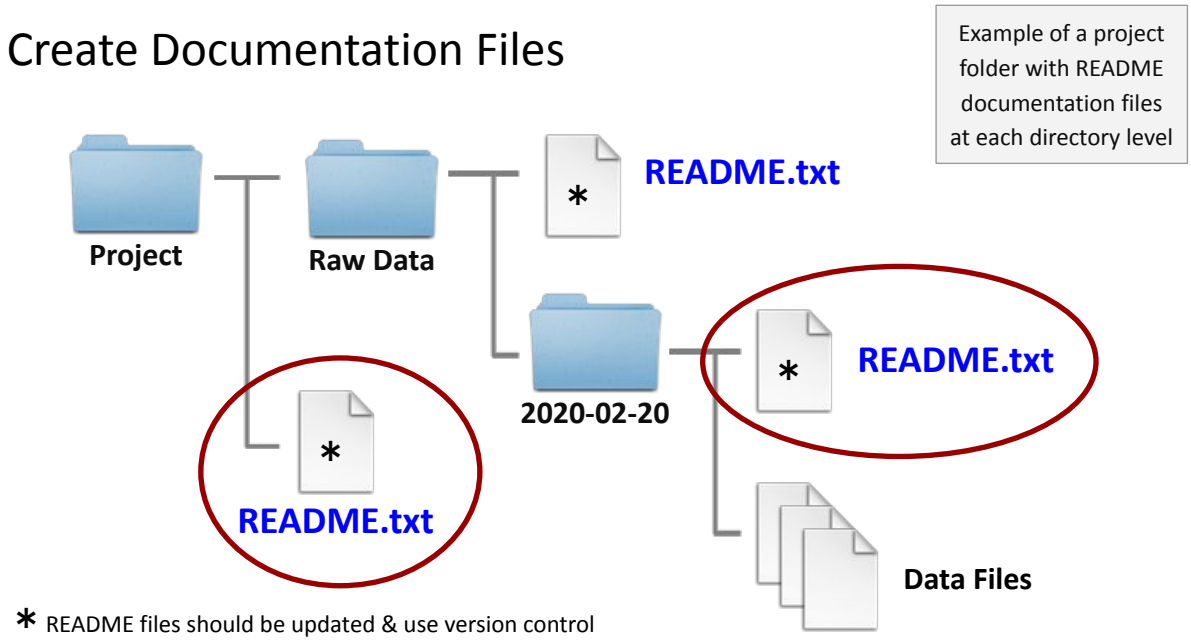
We need context, and we usually add it with some metadata in the form of descriptive text files, or README files.

A good way to think of README files, or metadata, is as information kiosks.

You place a small number in strategic places so that people who are lost can find what they want.

So create this documentation or README.txt files across level directories briefly describing their contents.

Create Documentation Files



For example, a README at the project level will contain some information about the project, including the names and contacts for each of these researchers, and the project folder file structure.

However, if you have more than a few data files, it quickly becomes critical to add metadata files to describe all the files in each directory.

So for example, at the experiment level, a README should include naming conventions, and briefly describe all the contents such as how data was collected, data versions, and any other important information.

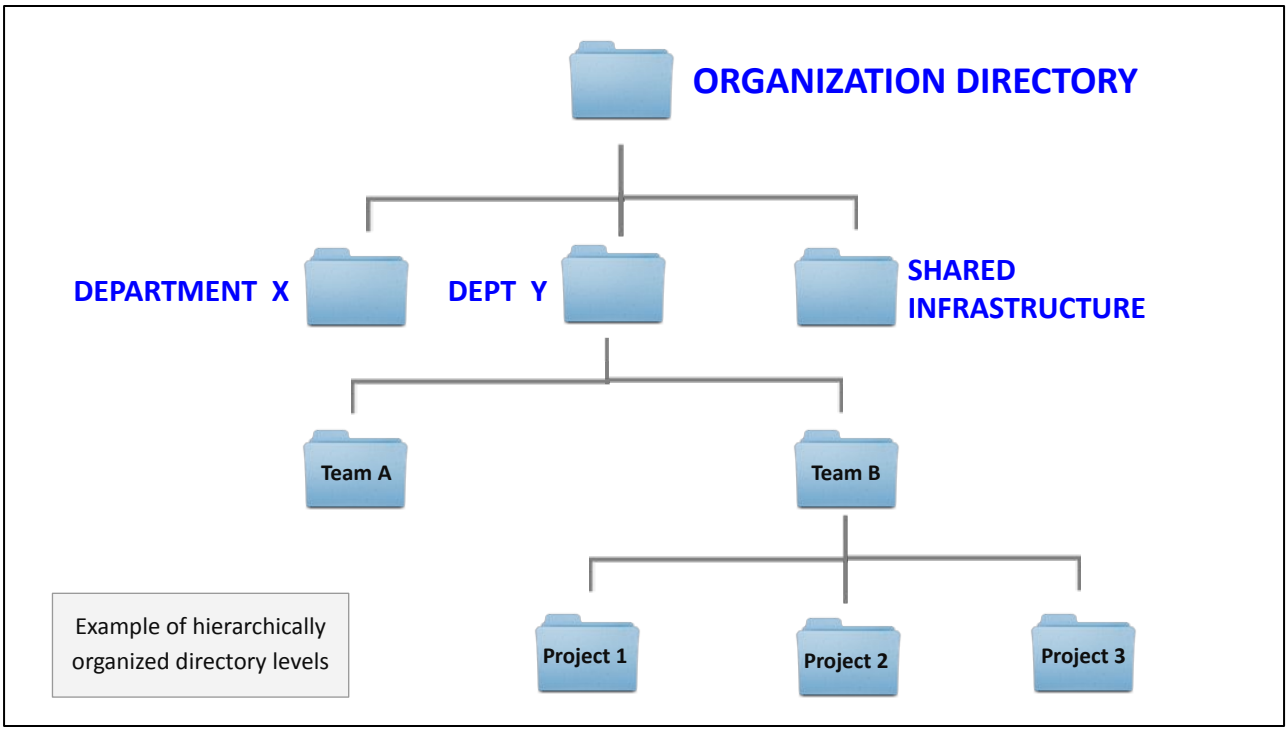
We will be talking more about READMEs in a future session, so tune in for that.

Examples

Let's look at some examples at different directory levels and with varying level of detail.

These will all use the principles we just sent through.

Hopefully these will help you determine the best structure and organization for your type of group or project.

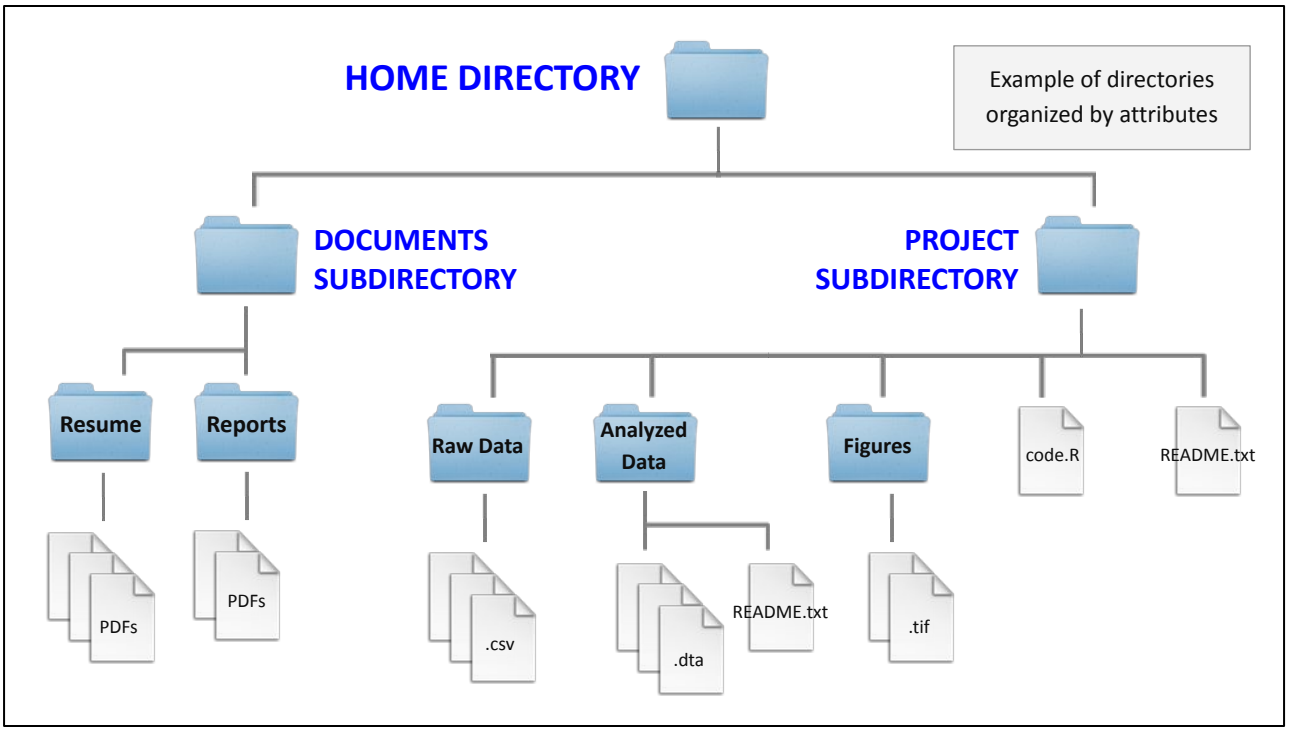


Here is a high level look at what you may see for available directories.

Here the top level directory is for an organization, then divided by departments, teams and then projects.

We will talk more about this during the Dropbox class in a few weeks, but you may have some shared subdirectories that are available for collaboration across departments and groups.

It is important to understand how your department or teams operate, because that should be reflected in your file structure.



This home directory has two subdirectories, documents and project.

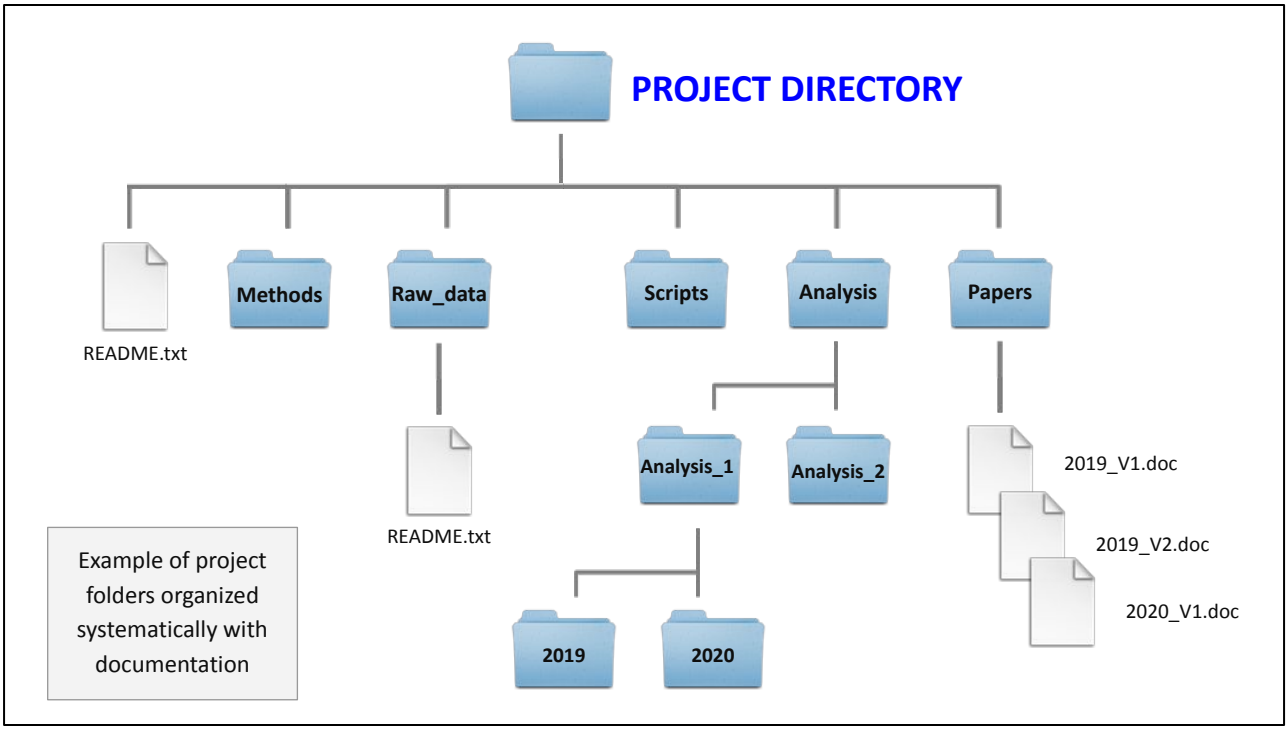
Within the project folder there are further subdirectories for raw data, analyzed data, and figures.

Remember to always keep raw data separate.

Another good practice is to make the files or folder read only so that someone doesn't inadvertently alter it.

Here we see the project's code.R file, and examples of both a project level README and dataset level README text file.

In addition, this group is also maintaining documents for resumes and reports, keeping administration files separate from the working research files.



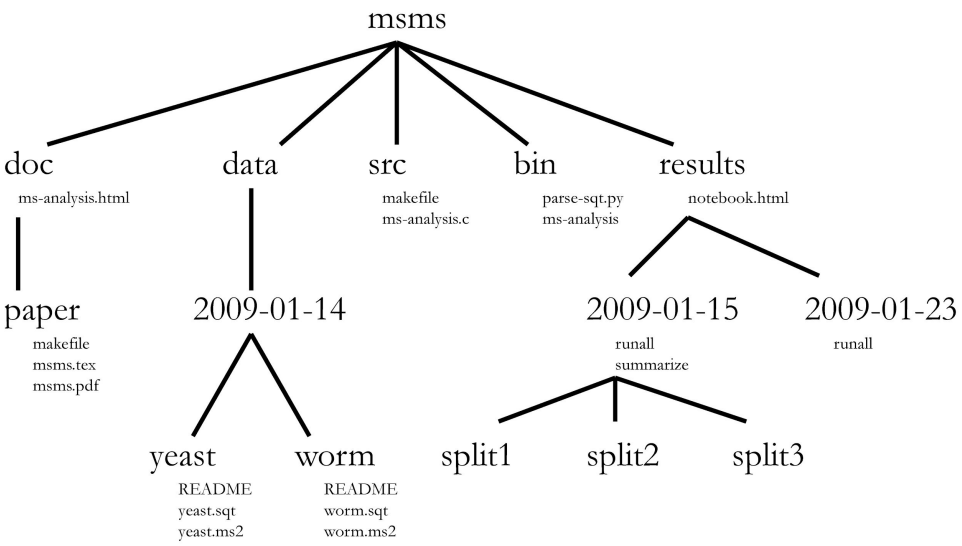
Looking at a project with a little more going on, we can see that the number of subdirectories and nested folders will grow based on many contributing factors like the complexity of a project, length of a project, or number of team members workings on a project.

So in this single project directory, this is an example of a multi-year project that may result in lots of folders and lots of files.

This is to encourage you to establish a structure ahead of the project, adapt the workflow as needed, and try to remain organized throughout the duration of the project.

Also, remember we don't want to let this get too deep, so think of ways to streamline folders where you can.

How you organize folders depends on what attributes are important to your project, and how many files you have.



Example from: Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. *PLoS Comput Biol* 5(7): e1000424. https://doi.org/10.1371/journal.pcbi.1000424

Okay, this is a specific example for organizing a computational biology project.

A nice practical example maybe for many of you.

We have a project named msms with a few subdirectories.

The doc directory has additional subdirectories for manuscripts.

The data directory is for storing fixed datasets.
Datasets are organized by date, using the ISO standard, year, month, day.
You can see README files in each of the yeast and worm data folders.

The src folder is for source code.
The source code ms-analysis.c is compiled to create a script file and is documented in that documents folder.

The bin folder is for compiled binaries or scripts.
The python script is called by both of the runall driver scripts for the results files.

And finally the results directory is for tracking computational experiments performed on the data.

The driver script runall automatically generates the three subdirectories split1, split2, and split3, corresponding to three cross-validation splits for this project.

This is a good structure to follow if you are running computational experiments, and also highlights how automation can help keep you organized as well!

Example: Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424. https://doi.org/10.1371/journal.pcbi.1000424

Local Project Directory	Github Repository

▪ Project plans/objectives
▪ Project datasets
▪ Project codes
  ○ Jupyter notebook
  ○ R scripts
  ○ Python scripts
▪ Output files
  ○ Visualizations
  ○ Tables
  ○ Other useful outputs
▪ Project report
	▪ README file
	▪ Project datasets
	▪ Project codes
	  ○ Jupyter notebook
	  ○ R scripts
	  ○ Python scripts
	▪ Output files
	  ○ Visualizations
	  ○ Tables
	  ○ Other useful outputs
	▪ Project report

Example from: Obi Tayo B (2019) How to Organize Your Data Science Project. *Towards Data Science.* https://towardsdatascience.com/how-to-organize-your-data-science-project-dd6599cf000a

Lastly, and related to the previous example, is a strategy for keeping data science projects consistent across all the systems and platforms you may be using.

So here we see a practical example for how you would structure your local project directory on your computer, and how that structure should be mirrored on your GitHub repository.

This makes going between the two platforms seamless, and also allows you to easily sync, push, and pull files when working with them.

Also, keeping good records of all your current and completed projects enables you to create a repository where you can save all your projects for future use.

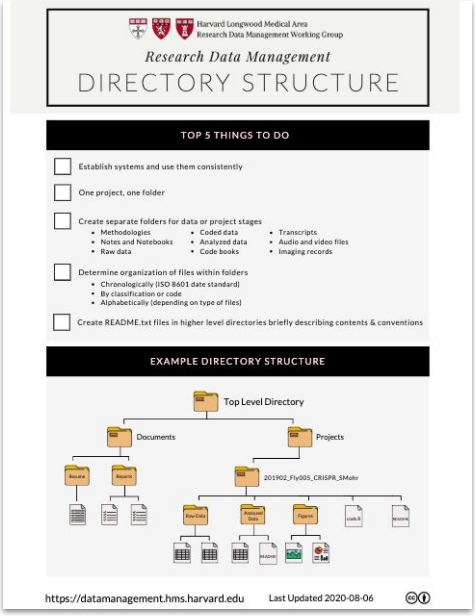
This is also emphasizing good practice to maintain two versions of your project, one locally, and a backed up copy.

Directory Structure

To save time and prevent errors later on, you and your team should decide how you will structure folders and organize your files. Do this at the beginning of your project and modify as your data grows!

Download Interactive Form:

PDF **osf.io/fp9j5**



Here is a checklist you can use to follow the top 5 practices we outlined today.

This worksheet also has one of the generic examples we looked at.

Hopefully this is a useful resources to include in your group or lab protocols to ensure everyone is following these best practices.

Takeaways

- Organization is a key aspect of data management and will help keep the project on track by saving time, storage, and data loss.
- Think hard at the beginning of your project about how you are going to organize your data as it grows.
- Try to structure project folders hierarchically and divide data into categories.
- Most importantly, consider what makes sense for your project and research team, and how people new to the project might look for files.

To summarize,

- Organization is a key aspect of data management and will help keep the project on track by saving time, storage, and data loss.
- Think hard at the beginning of your project about how you are going to organize your data as it grows. Something that works well for one file, or for two files, won't necessarily work well for a hundred files.
- Try to structure project folders hierarchically and divide data into categories. Directories can be organized in many different ways.
- Most importantly, consider what makes sense for your project and research team, and how people new to the project might look for files. And establishing a system from the start, allows for standardized data collecting and analysis by many team members.

Upcoming Summer Seminars

~~7/06~~	~~What is Data Management?~~
~~7/14~~	~~Data Management Spelling Bee~~
~~7/22~~	~~Let's Talk Data~~
~~7/30~~	~~What is a Data Management Plan?~~
~~8/07~~	~~Organize Your Files~~
8/13	How to Name a File
8/19	What's in a README?
8/25	Project Management in Dropbox
8/31	How to Cite Data

Register: bit.ly/RDM-Seminars

Remember we have a full line up of webinars every week this summer.

Next week we will build on the structures we talked about today, and focus on naming files and folders within a system.

Thank you!
Questions?

Contact me:
julie_goldman@harvard.edu

Resources:
datamanagement.hms.harvard.edu/resources

Please fill out this survey

bit.ly/rdm-sum20

Thanks for joining!

The recording and the slides will be posted on our website.

I would appreciate your feedback on today's session, so please fill out this survey at the link provided: http://bit.ly/rdm-online

If you have any questions, enter them into the chat!