Colorado State University Libraries
CSU Libraries
Training and Instruction
Transcription of Data organization, 10/12/2016

Collection: Training and Instruction (10217/195518)
Title: Data organization
Date: 10/12/2016
File Name: FACFLIBR_DaD-DataOrg_TM_20161012.mp4
Date Transcribed: November 2024
Transcription Platform: Konch AI

BEGIN TRANSCRIPTION

[00:01 - 01:51] Tobin Magle: Hi, and welcome to Data and Donuts. I'm Tobin Magle, the data management specialist at the Morgan Library at Colorado State University. Have you ever tried looking at someone else's files to find a piece of information? Have you spent too much time trying to reformat a data to make a table or graph? Proper data organization can help with these problems. Today, we're going to talk about how to organize your data using folders, file names, and spreadsheets. To review, the research cycle generally goes as follows. You come up with a hypothesis, design experiments, and hopefully write a data management plan, collect and analyze data, and finally publish these data, which can be used to generate new hypotheses. Data organization is initially important during the data collection phase and through analysis, but will also make analysis, archiving and sharing easier. As I mentioned before, we're going to be talking about folders, file names, and spreadsheets. First, let's talk about folders. Your computer's file structure is inherently hierarchical, or arranged in order of rank, or put another way, the file system is organized as folders inside folders. The very top folder of your computer is called the root directory, and this symbolized by a backslash. Inside root, you have folders like bin, data, and temp that store files that are important for the operating system. Finally, the users folder contains a separate folder for each user. Your user folder contains the folders that you're used to working in on your computer, like documents, desktop, etc. Dividing your data into folders helps you stay organized. A good way to approach this is to separate your files by the most important attribute, such as operating system files versus user files, and further separate things inside those folders by less important attributes like desktop and documents. When thinking about how to organize your data, it's a good idea to go back to the data inventory you created in writing your data management plan.

[01:51 - 03:46] Tobin Magle: This document describes the type of data you're going to collect and listen to other research outputs. So, as you're looking through your data inventory, ask yourself what file types do I have? How can I group these files by project, by time, location, file type? Finally, of these attributes, what are the most important ones? You'll base your folder structures on the answers to these questions. To illustrate these points, let's look at a hypothetical grad student named Lou. He's trying to use data from someone who has left the lab, who collected weight and cytokine data from 16 mice. Half of these mice were infected with the parasite. Don't worry too much about the details of the experiment. If you're not familiar with this type of research, it's meant to be a high level example. Now, answer what are the attributes of this project. Which seem like the most important ones? One attribute of this experiment is time. Is the data are being collected longitudinally over the course of a year? Another attribute is infection status, because some of the data are from infected mice and some from uninfected. Additionally, we have two data types: weight and cytokine data. Are there any other attributes that I haven't identified? How would you rank the attributes that we've identified? Which is the most important? Are you being overwhelmed by the possibilities? Don't panic. There are many acceptable answers to these questions. The important thing is to try a strategy and be consistent. If it really doesn't work out, you can always switch strategies later. It's also important to record the strategy and readme file so other people know how your files are organized. Let's go back to Lou's project. Download the files from the link on the slide. Using the attribute to identify in the last exercise, create a file structure for Lou. If you have time, you can describe your strategy and Lou's readme.txt file for bonus points.

[03:46 - 05:30] Tobin Magle: Think about your own project and decide how you should organize your files. Even if you have good folder organization, bad file naming can make your data hard to use and interpret. The goal of file naming best practices is to make the file names human and machine interpretable. The first rule of good file naming is to use descriptive names. These details help human readability because the name says something about the content. Bad file names like file.txt are not recommended, even if they're in a well described folder, because they lose this information if they're moved. Better names include collection dates and some text to indicate the contents of the file. The Ok file name tells me the dates collected and that the data come from mice. The best file names, however, get really specific about the contents and the format of the file. The good file name tells me that the date collected, and that the file contains mouse weights and tab separated value format. The second rule is to name files from general to specific. Let's use the example of gene expression experiments that have several replicates. First, separate out dates and content within the date. Put the year first, then month and day from general to specific. This allows the computer to sort the files in a meaningful way. Then, let's look at the content. The replicate number is more specific than the type of data, so it should go last. That way the sorting would be meaningful even if there is

non gene expression data in the folder. The third rule for file naming is to avoid abbreviations, because your abbreviations might not be intuitive to other people. Unless they are widely known, avoid them completely. Another option is to include a detailed readme file for what your abbreviations mean. The fourth rule for file naming is to avoid spaces.

[05:31 - 07:18] Tobin Magle: Spaces are a bad idea because computers use spaces to separate out file names, so typing a file name with a space in it in the Unix terminal will likely make the computer think that there's two different files. Alternatives to spaces are dashes, underscores, or CamelCase, which is the capitalization of each word in the file name to delineate words and make it more human readable. The fifth rule for file naming is to avoid special characters like the ones listed on this slide. They are best avoided because certain programming languages have special meanings associated with these symbols. For example, the tilde tells the Unix shell to return to the home directory, alternatives to these symbols, are underscores and dashes, as with the space. The final and most important rule is to be consistent with your file naming conventions. This will allow your data to be more findable. Even better, try to establish common conventions for your research group so you can work better together. You probably aren't starting from scratch on your research project and already have files that are inconsistently named. Luckily, there are programs out there that will automate file renaming for you. Let's get back to helping, Lou. Now that you separated Lou's files into folders, rename his files with descriptive names that adhere to best practices described here. Now that our files are in folders and named properly, let's work on the interior of the files or how to organize your data efficiently in spreadsheets. Many people use their spreadsheet programs like a lab notebook. In addition to data, these spreadsheets contain color coding and formatting that often contain information about the data. They also have notes, calculations, graphs, and tables. This is a very human readable and intuitive way to use a spreadsheet. However, this type of organization has its downsides. For one, computers are dumb, so they won't understand the meaning of your notes or color coding, or formatting.

[07:18 - 09:05] Tobin Magle: Also, making graphs, and tables, and spreadsheets programs is inefficient, and it can be done quicker using scripts. Thus, taking the time to use spreadsheets wisely will save you time in the long run. To use spreadsheets wisely, here are some tips. Only include one table per sheet, because computers can't tell where one table ends and the other begins. Don't use multiple sheets because computers can only read in one at a time. Use descriptive variable names so that it's easy for humans to understand what the variables are. And finally, don't mix notes and data in the same cell because this isn't machine readable. Even if you follow all of these recommendations, there's still one more step to make your data truly machine readable. Tidy it up. There really only two rules to making your data tidy. Number one, each column should be a variable.

Make sure not to combine multiple variables in one column. And number two, each row is an observation or one measured value. Let's look back at Lou's spreadsheets to see how tidy his data are. Open MouseInventory.xls. Is he using spreadsheets wisely? Is each column a variable? Now, look at the January data for both weight and cytokines. What variables are being measured, and what are the observations? Finally, do you think there's any way to combine these files into a smaller number of files? Thanks for listening. I hope you found this data organization session to be helpful. Please email me at the address on the slide if you need help. Also, check out our newly organized data management pages for more information. If you want to learn more about data organization and automation, check out the lessons on the Data Carpentry and Software Carpentry websites. I've used both of these lesson sets for inspiration for Data and Donuts.

END TRANSCRIPTION