So the question is how do we avoid some of the scenarios we just saw. So there's some basic best practices that you can adhere to which will help not only you manage the data as you are doing the research but other people to use it down the line. Which is really what the issue here is.

So the first thing we're going to talk about is files - file names. So what I would ask you is what is the date of this file? If you look at that date well you got a number of possibilities. It could be the 12th of June 2011. It could be December 6, 2011. It could be January 26, 2011. So without a convention of how to list a date in a computer file you might know what it is and you might remember it for a while, but again down the road when somebody else wants to access this they will not know. So I'm curious, does anybody actually know what the convention is for putting a file date? That's great. So there's actually an ISO standard. And what an engineering I also do standard stuff and I always call it the international standards organization. That is not what it stands for but that's how that's the English-ised version of it. So they've actually got a standard, and what they want is four digits for the year, two digits for the month, two digits for the day. That way you can always be consistent you will know that 2-0-1-2-0-6-1-2 is always June 6, 2012. So this is a really good standard to adhere to if you want to put dashes or colons between that's fine. Underscores whatever. But adhere to that if you need to time date it you can just add that minutes and seconds, hours onto the end. But always adhere. Four-digit year two-digit month, two-digit day. That will help with that. So, which filename follows the ISO standard for the date December 5, 2011? There you go. Easy enough. So try to keep that in mind that will help.

Additionally, who knows what "Sam" means? She talked about this in the in the video. So that could be an acronym for scanning acoustic microscope. It could be the acronym for a drug name. It could be Sam the postdoc. You don't know. Now you might know, but again when someone tries to access and use this data 5-10 years from now they're not going to know. So you want to have some kind of logical convention and it's a little bit harder here because of course the names might be longer, you might not have space. So you have to work within the parameters of the system you're using. But it's just a good idea to try to come up with some logical naming sequence. Here's a good example of why this could be a problem. You're doing research on rat hearts. You've got a rat heart. You cut that rat heart into hundreds of slices. You take those hundreds of slices and you make hundreds of slides out of those slices. You now create thousands of TIFF images out of those slides. Each TIFF image is separate file. It will have a separate file name. So you've got to have some logical sequencing to keep those file names in order. You might make hundreds of huge files out of those smaller TIFF files. So now you've got other images. To complicate things you've got three postdocs working on this, so you've got three different people manipulating hundreds of thousands of images. And each of these postdocs is doing five to seven experiments a week. So you can see the problem where if you don't have some logical conventions for naming and keeping track of things it would be very easy to get confused.

It's interesting I just did some fairly in-depth interviews with some of the civil engineering faculty to find out about their data management practices. And a lot of the research they do especially in transportation engineering, their output isn't really a lot of files. Because usually what they're doing is getting data and then manipulating it to create models for DelDOT or something like that. So their output is usually more & more often, not usually, but more often a model. They don't have these tens of thousands of files that's coming up with but I'm assuming in medical research it's the other way around. It's very easy if you'd end up with a lot of files.

So here are some basic rules. The file name should embody the content including major parameters. So after rat would be the name of the experiment. We can assume here and we'll get to this later that maybe this is experiment 12, it's the 128 TIFF file, we're dealing with lipitor levels. Be non-cryptic. Use intuitive names as much as possible. So put in data validation not DV, or da-val or da-to-val. Again you're gonna be limited by space sometimes do the best you can. Try to make it as intuitive as you can. Always leave extensable endings. Do not use whole numbers at the end because when you go to sort these in the computer file it will not list them correctly. So for instance if you just did one two three up to ten you're going to get 1 and then 10 and then 2. Most systems will sort that way, so do 01, 02 or if you know you're gonna have thousands of files leave as many leading zeros as you need. That way the files will always be sorted correctly in your folders. Be unique where possible. Avoid 20 different files with data.xlsx in different folders. If you start mixing them around in different folders you're gonna lose track of them. Don't use special characters. The reason is you might transfer this to another system that doesn't accept those characters. So as a general rule try to keep your filenames down to numbers, letters and underscores. That way if it gets transferred to another system or an updated system you're not gonna lose anything in the file name. Underscores as much as possible do not use spaces. Again different file systems might not like that. And use consistent documental rules.

So as I mentioned earlier something we looked at that earlier name we didn't quite know what things meant. What you basically got here is an entry you would put into a data dictionary. Anybody actually ever created a data dictionary? Excellent. Wonderful. So the idea is that you can create just a simple basic text file which explains the file structure. You can always attach that text file to the folder that the data is in. Again you're probably going to remember, but somebody 10 years from now who wants to look at one of these file names can simply bring up the data dictionary and it'll tell them. This is the experiment name, this is the experiment number, this is the sample number, this is the stain used, this is the coordinates of an image... It'll explain each element of the filename. And this might be helpful to you. Now a lot of funders (not a lot that's not true) but there's one we will use later on demands that you supply this. Because they just are assuming people are going to be using the files later on and they really need to know this.

So for all the reasons that came out in Sara's presentation earlier try to be trying to follow conventions, document those conventions, be as intuitive as possible, use the dating convention. Try to make it as easy as possible for somebody many years from now which is again is what we're talking about to use this data. This is what you're trying not to end up With you know. "Huh", "WTF", I love that… "Crapdat". They might be in dated order but that's about it.

So this is what you want. You want a list of very logical file names that you know what they are. You know what the dates are, you know what the image numbers are, whatever. Very logically arranged. Again not just so you and the people you're working with in your research team can figure it out people down the road can figure it out. Document it. Create a data dictionary. Keep track of all that as you go through. It's a real good idea in a project with multiple people to have one person in charge of this so you don't end up in a situation where people are wondering "Did you do that? Did I do it? Did I backup the files? Did you write the data dictionary?" So it's always a good idea, in a multiple member research team, to assign someone to be the data person. That really helps avoid confusion. So Sarah's going to talk about data collection now.

So another aspect of good data management is the clarity around variables. As you can see in this spreadsheet a variable names are a little bit cryptic. So one can guess here that SID might be "subject ID". WGT is probably "weight" though "sam" is probably anybody's guess here. And smoking is pretty clear but the meaning of smoking is not really clear. Is it "Have you ever smoked?" Is it "How much you smoke?" If the meaning of the variables is not clear everybody may be entering different information in. So basically here we have to make sure that everybody has the same understanding so that we can end up with consistent units of measurement. So if you see here we have SID, everybody seems to kinda have the same understanding of SID. We've got a one, two, three, four, even though we're kind of guessing that SID is "subject ID". Under "weight" "wgt" it seems to be weight we're guessing here. But our variables here not quite consistent. Smoking - we've got a Y, we've got two packs, we have N, we have never. No consistency there. Name okay where everybody's entering it differently. Somebody has a last name somebody has Sam Jones. we have Read, Kevin and then we have Emma Banks. No consistency there as well. Under "sam" we have 13, we have 37, we have A21, and we have January. There's no clear understanding here as well. So we have a lot of inconsistency with the units of measurement here. So this is what we're trying to avoid on variables. So we need to document our variables.

All of these problems can be addressed with proper proper documentation. Documentation should include the name of the variable a description what type of data it is such as a date or an integer. What unit of measurement are for that variable, and if there's a restricted set of possible values, what are those values? And it's helpful if the name chosen for the variable are is intuitive and meaningful as possible to avoid confusion. So now we have some study_ID the field type is text. Description a unique ID of the study and possible value is an 8 digit number. We're much clearer there. Now we're not guessing of what that means. You know date_enrolled field type is a date. Initial subject enrollment date and now we have the format so that we can't enter something incorrect there. And now "Weight" type integer - weight of subject unit is pounds. We've answered all and cleared up our confusion here. So we're kind of answering here and fixing all problems here to avoid confusion. And unlike our previous variables here documenting our variable names is going to fix that issue.

So now our data dictionaries. This type of documentation is known as a data dictionary. Data dictionaries are more typical in clinical than basic science research, but documentation of variables is really important for any type of research. And will help clear up all types of confusion so that others can really understand your research in the future.

So a little bit now about workflows. Documentation of workflows is just as important as documentation of variables and file naming conventions. The meaning of data depends on the context of how it was collected. For clinical trial protocol this means documenting the study design the primary and secondary outcomes the inclusion and exclusion criteria and so forth. Clinicaltrials.gov is a website for registering clinical trial protocols and posting results so that there's a well established way to document protocols. Documenting of those types of experimental protocol ensures that others can fully understand and reproduce the research done. However there is another side to the workflows involved in the study or experiment. One could describe this as the internal workflow of the study or lab team. Documenting these would most easily be described as making sure things don't fall through the cracks. So we always want to make sure that everything is accounted for and everybody knows which part of the team they're responsible for. All right so within the lab who is responsible for the data management in the lab? Is it the PI? Is the lab manager? Is that the graduate student or is it everyone? Think about your current workflow. PI, everyone, lab manager... I would kind of say everyone. Some people say that means no one. You can kind of there's some arguments here. But also there's some you know, some people say it's everyone somebody says it's certain person like the PI, the lab manager, the postdoc. But the problem is if it is really everyone that kind of means it's no one. So it is a really good idea to make sure somebody really is in charge. Because when it is everyone things do fall through the cracks. So do make sure that somebody really is in charge, and make sure that they're in charge for quality control purposes. So whenever possible it's important to assign one person to truly be responsible for ensuring that any naming conventions that have been established for files or variables are being adhered to. And that some minimum level of documentation is being recorded, that existing version controls are being followed, and of course that data is always being backed up. So if you leave it to everyone nobody is really kind of watching out for everything. So I think it is kind of a good idea to make sure that somebody really is overseeing everything.

All right we're gonna talk a little bit in this section about storage and preservation and the difference between the two. So first up is storage, and a little bit about some of the storage options available at UD. So we have a few available things available here at UD and first I just want to briefly mention REDCap which is a secure and encrypted web application for building and managing online surveys and databases. This is a tool that can be used for data collection and can be configured in a way to de-identify patient data, and can be used in HIPAA compliant environments, and though it's typically used to collect biological or medical information it can be used for virtually any type of study. I also want to kind of mention from the library perspective there is UDSpace which is our institutional repository. UDSpace is not currently used for collections of large data sites, but it is possible that this may change in the future. And for more information you can contact William Simpson who is our librarian responsible for our institutional repository in the library. And I'll bring up Alicia Morgan here for just a moment. Okay, you go I know Tom just handed out the sheet we had put together. And I think the key point is the reason for this is to make sure that you are aware of some of the resources that we have available for the College of Health Sciences - specifically for, and this is really more for the working data storage as opposed to the long-term preservation. But we do have several options, and so therefore and make sure if I've noted this in beginning, it's important to evaluate the particular data you're using as far entered before you select a resource. Because some data certainly in the medical world will have PII or PHI content or health information that we need to be treated with a special you know with more sensitivity and protection. So the first of which is the one resource we have here for general data storage is the WynDFS server which is the server space that available. Actually it's created and managed by IT, but the College of Health Sciences makes it available to or, not without a fee, but to faculty and staff and researchers. And they don't even make a profit on it. They charge researchers what I teach at the college. And again it's available for important data that you want to make sure that it's backed up reliably and sort of a more enterprise business level of service. It includes a service called shadow copy so if you make changes you can recover your data from the last time. It takes copies throughout the day in the week so you can recover data more easily. But it is not not free.

Then we get to the two cloud services that that UD has arrangements with. First the Google Drive that I'm sure most of you use? Okay. And that's and that's a great resource it's been used on campus for a long time, and it doesn't it's much different than using the personal Google Drive because it does ensure more the requirements for hosting that make sure it's more appropriate for business data internally. It's still though is not really appropriate for what we call level 3 data which is secure report with data requiring additional security provisions. Again confidential privacy related data. And that's where we get to OneDrive. I don't know if you mean are any of you familiar with the Microsoft OneDrive service? It has only recently been rolled out on campus in January. It is part of the Microsoft Office 365 bundle that the university has gone into an agreement with. And it is essentially similar to the Google Drive service except that it carries protections. While it is a more well protected system. It can support file encryption, volume encryption, and it is the really the solution we have on campus that is the global solution that is available on campus, that can handle the storage and preservation storage of PHI health information and privacy information. So that's, and also in conjunction with the Office365 email system, can handle the transport of secure files. So that's what really makes it, distinguishes it from everything else on this on this sheet.

REDCap also has some has some security capabilities in a similar way. But that's where if you have issues if you need protection for your data if it's not de-identified, this is the one place that you might investigate as a solution. And certainly I wouldn't start putting data up there without talking to - Richard Wisk or Bill Flynn or some of the other folks who are working with this kind of content. But it is really a new thing and it's a probably the best resource we have for this purpose. I think the big piece is to be aware that these are available which and offer a viable alternative to using your personal Dropbox your personal Google Drive or your personal OneDrive. And then make sure that it's and requires because it requires authentication via UDEL IDs and so it's much more clear who has access to the to your data, and who can who can use it. So last, I did note that the REDCap storage and access to this where to go for help. And then also for those faculty or researchers who are using the high performance computing services that we have at UD and there are some. A reminder that that comes along with some data storage space. At the same time it doesn't have the same kind of protections as what you would what you can configure on OneDrive. So that's this is just a quick cheat sheet and thing to remember is interesting. This is as of April 2018. There are things coming down the pike that haven't rolled out yet. But the OneDrive is a is a new solution that I think would be worth investigating if it's appropriate. The one thing we don't have yet with all with these solutions is, other than REDCap, is an easy way to share with colleagues at other institutions. And that's what's being discussed right now. So look keep checking back on the space… Thank you.

[Title: This is the end of Part 2: Data Management for Medical Researchers. Please continue the workshop video in Part 3.]

516
00:22:16,900 --> 00:22:18,966
[Title: This is the end of Part 2: Data Management for Medical Researchers. Please continue the workshop video in Part 3.]