Hi everyone we're gonna go ahead and get started. Welcome to - can everybody hear me? Okay to get started? Okay great! Welcome to Data Management for Medical Researchers. My name is Sarah Katz, I am the Health Science Librarian here at the University of Delaware, and I just want to give you a little bit of background about the workshop. Since August the University of Delaware Library has been participating in a research data management pilot project that's been funded by the National Network of Libraries of Medicine the Mid-Atlantic Region and this pilot's been hosted by the NYU Health Science Library. This project provides a holistic approach to developing data services that focuses on building a required knowledge base, understanding and connecting with researchers, and promoting effective outreach strategies, and integrating a broader institutional data community. Throughout this project we've conducted interviews with many faculty here at the University of Delaware College of Health Sciences and basically to kind of understand what we're doing here already about data services. In addition, this workshop is another aspect of this pilot. Ultimately we hope to improve data services on campus throughout the participation in this project. For some other information about what we've already been doing with regards to research data management I'm going to turn it over briefly to my colleague Tom Melvin.

My name is Tom Melvin. Why am I here? Well, I'm the engineering liaison to most of the engineering departments. I do not do chemistry or biochemical engineering, but since 1989 I have been here and I represent (I've been in bands I know how to do this). I liaison to civil and environmental, material science, mechanical, electrical, computer, engineering. In the last four years I have also started liaisoning to geology geography and marine studies. Kind of somebody retired they asked "could you do this temporarily?" And I said, "yeah". Temporary became permanent, so. I work with faculty a lot on campus. In this aspect I have, for the last about five or six years, I've kind of been the contact person in our department (Reference and Instructional Services) for any data management issue.

I'm wondering if anybody actually uses research guides off the library webpage? So there is a research guide which is a webpage on data management. If you just click on the research icon, the research guide icon on the library homepage, and type the word data you will get the link to it it. It has a lot of general information which is the nature of these guides. It will have links to some of the data repositories we're going to be talking about, some of the metadata directories that we're going to be talking about. We've filmed one of the writing and data management plan workshops. You can watch that video from there. We are filming the video today, and this will be uploaded there, as well as the slideshow. So I wanted to show you the page - we're having technical difficulties so that won't be possible. Maybe I'll be able to get to it later. As always if you have any questions please contact us. We're easy to find up at the library, and I'm gonna turn it back over to Sarah. We're gonna send out all the links to anybody who's registered to the workshop so we'll have a link to the data management research guide as well as everything we've talked about today. If we can't get to these pages later we will be sure to link to them so that you can see them.

Alright so a little bit of audience-participation. We won't have too much of that, but a little bit. When you think of data management what exactly do you think of? Is there anything in particular that first comes to mind? Spreadsheets. What else? Yes - Security. File cabinets. Red cap. How to organize. Great. Version control. All great things exactly. These are all different types of things that we should be keeping in mind or that come in mind. Data management - all different things that we need to keep in mind, so data management leads to greater organization of data and workflows. Organized data is more comprehensible either to others or to researchers when they look back at it later. Organize workflows lead to more efficient research process. Data management ensures access to data in the future either to others or for the researcher. So all of these things are very very important when we think of data management. Both for you and for others looking at your data.

So just a quick little thing a little overview of what we plan to cover today. Our learning objectives. So you could read them here but I'm just gonna quickly go over this just to give you an idea of what we plan to go over today. Recognize the current and forthcoming requirements that mandate the management and sharing of research data. Identify current requirements and issues around rigor and reproducibility. Apply best practices for creating and documenting file names, variables, and workflows. Identify appropriate options for storing and preserving research data. Locate appropriate standards, if any, and recognize their value for research. Evaluate repositories and determine the best sharing options for data.

When we talk about the research data management climate what we're discussing here is why is this a big issue now? I know in the library world the last 10 years this keeps popping up a lot. That's why we did the courses we did, the workshops, the training. And one of the reasons we did the previous workshops we did was kind of just to get a feel of what's going on on campus. How many people are thinking about this how many people are aware of it. And as you can imagine what we discovered is it's everything's but he's kind of in their own place. There isn't much of a coordinated effort now, which is one of the things we're trying to become part of is moving in a more coordinate and coordinated campus-wide effort because it's becoming more of an issue specifically from grant funders and publishers both. Which we will both talk about.

So, basically what does data management mean for your future? Why do you have to be aware of these things? A brief history, and what we're going to cover basically are the NIH data management requirements. We're going to be concentrating on the NIH because we're assuming that's where most of you are getting your funding from. It is the federal agency you deal with the most. When I did these before the engineering people lot of it was NSF. The NSF is actually a little more advanced on this and we'll talk about that a little bit. So we're gonna talk about the data management and sharing requirements that they have proposed. Nothing is codified with them yet, but they have very strong suggestions, as well as publisher data sharing requirements. You have to be aware of both. Because you might have a different requirements coming from who you get the funding from, as opposed to who you published with. And Sarah will talk a little bit about the rigor of reproducibility of the data which is really the main issue that we're talking about. Making sure that other people who want to access your data later can use it to get to the conclusions you came to.

So, where did we start? Well back in 2003 is when this first became an issue. And the NIH came up on the policy that any grant application for over $500,000 had to submit a data management or data sharing plan with the grant application. It is my understanding that there was kind of an out on that that if you had a good reason why you didn't feel you could submit a data management plan you could just write that in "here's why I can't". I suspect a lot of people did that. That's just kind of a guess of mine, but that's a guess based on a lot of experience of dealing with people over the last five or six years. So I don't think they were very stringent about it at the time. And one of the things that's important to remember which I'm gonna mention now is that when they talk... How many people have actually done a data management plan for a grant application? Well you know a couple. So as you know they don't have to be long. And in fact they want them short. So this doesn't have to be when you right the DMP it doesn't have to be a long arduous process. Some of the agencies actually stipulate no more than two pages. They want certain elements but they don't want it to actually be long. So you don't have to think of this as an arduous thing. So it first came up in 2003. For years and years it kind of just stayed the same.

In February of 2015 NIH published a white paper called "The Plan for Increasing Access to Scientific Publications and Digital Scientific Data". From NIH funded scientific research in the white paper they came up with some suggestions. They wanted full descriptions of the data and how the data was collected. All the different types of data you expected to collect and how you collected that data. They want to know what software tools you use to analyze or create or produce final results of the data. That's important because there's been a lot of experience of people going back and trying to use data from 10, 15, 20 years ago and they cannot locate the software that was used at the time. So the data gets a little useless at that point and I've actually read accounts of people who had to spend a good deal of their funding research funds recreating the software that was used to use the old data again. So it's saying saving the software along with the data is an access issue. You can store it somewhere, you also want to make sure people can use it. So they want to know that. What protocols or steps you have used to create the data. Again this is so somebody later on down the line can recreate the research. How you ensure the long-term preservation of your data. Where you're going to store it, what their policies are in terms of backup how long it will be there. And how you provide access to the data. Specifically are you going to limit access for some reason? Now you're allowed to do that, but you have to say why you're going to do it. So for instance if you're going to be using some of this research to apply for a patent you might not want everyone getting into that research. So that's a legitimate reason they just want to know what the access policy is going to be. So this is basically what they've proposed in the white paper.

Then in 2016 they put out an RFI for strategies for how they could implement these ideas into the community. Basically they were looking to solicit the following information: How the data should be managed, how the data should be made publicly available, and how to set standards for citing the shared data and software. We're not going to talk a lot about citing now but that is an issue. Like any other publication you want your data to be locatable. anybody you used the data Citation Index yet? Curious. Anybody know... what about... yeah it's not a lot to use. So there are two databases now basically. The data Citation Index is a product of the people who put out Thomson Reuters the people who put out the web of science. And Base is just a database of datasets and you can go in there and locate. They do a very good job of trawling over 200 different repositories and uploading and indexing the information. So basically citing the data using metadata, which we'll talk a lot a little bit to cite the data is important as well.

So these are the issues they wanted one of the areas they've actually been pretty serious about and enforcing their policies on, is genomic data sharing. I assume and this is just an assumption of mine I don't I'm not involved in medical research, but from what I've learned about data in general, I'm assuming that they went here first because there's a lot of it going on. So that it was kind of an easy place to go first and set some standards. So right now the NIH wants any large-scale data dealing with any of these GWAS, SNPs, genomic sequencing, metagenomic, epigenomic gene expressions... They do want the requirements are that you must submit a genomic data sharing plan with your request and that you must agree to share the data no later than the date of publication. Again I don't know how much they're enforcing this. Has anybody encountered this I'm kind of curious? Not yet, okay. So but that's one of the policies they're actually trying to enforce fairly strictly right now. And again I assume they've just gone there because there's so much of this research going on it's a good place to start.

So in addition to any requirements that might be issued by the granting agency and one of the things we'll do later on to show you how to get that information it's pretty easy to get. But we're again we're concentrating on the NIH here just because we assume that's the big fish in the pond here. Many publishers are now also coming up with policies. They usually will not conflict with any policies put out by the granting agency. You just simply want to be aware where you're going to be publishing and if they have any kind of data sharing policy associated with their publication practices. One of the first ones actually come up with a strict one was PLOS - the Public Library of Online Science - which is an open access platform where a lot of science stuff gets published. And they basically have said refusal to share the data in accordance with their policy will be grounds for rejection. Their policy is that the authors must show proof that they have shared their data somewhere by providing the journal with a unique identifier. Basically a DOI. If the proof is not provided PLOS can reject the paper outright or retract it if it's already been published and an author removes the data from public view. So you must specify that the data are deposited publicly and list the name or names of the repositories along with the DOI or an accession number. And they're pretty strict about that. Which makes sense they're open access, they're going to be one of the first people to go here. But other publishers are coming on board with this. Nature Science, Cell have language written into their policies that state that data necessary to understand, assess, and extend the conclusions of a manuscript must be shared and proof of this must be required specifically for genomic data. All other kinds of data they don't require proof yet like PLOS does. PLOS wants it for everything. Nature, Cellpress Science are just asking for but now but they're probably going to be moving more in the area of... there's basically broad discussions going on in the publishing community that are going to point to greater enforcement of this for all data sets.

A good example of publisher requirements that are coming on board now that you need to be aware of, there's one that came from the International Committee of Medical Journal Editors that released a proposal in January 2016 to require that de-identify patient data from clinical trials that underly a journal article must be made public within six months of publication. So this group consists of JAMA, The Lancet, and the New England Journal of Medicine, these are all prominent journals. This is just a good example of the policies that are coming on board that you're going to need to be aware of if you want to publish in these journals.

So a quick quiz. Just shout out the right answer. According to multiple publisher requirements, all data necessary to which of these must be available? A) Understand, B) Assess, C) Extend, or D) All the above. D - see you paid attention that's great! So the point here is that your funding agencies are going to have requirements increasingly so in the future so will the publishers. My experience is that they don't normally conflict. They're pretty much going to be the same. You just have to be aware that you have to deal with this on each end. It's fairly easy to find out that information. We learn one of the things we can do if you are having trouble finding any information about that contact us we can help you get to the right place.

In January of 2016 the NIH added the requirement that grant applications must address a number of issues related to the rigor and reproducibility of research. So we've got these new guidelines that were started in January of 2016. And the most important thing is to start paying attention here to some of the percentages that we're going to talk about here, and the numbers that we're addressing. These requirements grew out of a growing concern about a lack of reproducibility in many areas of science particularly biomedical research. So the first example here is when Bayer sought to validate published results on potential drug targets they determined that the data collected was consistent with the published literature only 21% of the time. So our first example here is 21%. Our second example here is Amgen. So Amgen, another drug company, tried to reproduce findings in 53 landmark preclinical cancer studies and they were only able to confirm the findings of 11% of those studies. So first we had 21% now we have 11%. And these are not just in drug trials. The open science collaboration group together to duplicate a hundred experiments published in 2008 in three high-ranking psychology journals. They found that they were able to reproduce the exact same results in only 39 cases. Of those cases that didn't match, some were significantly off. And 15 showed that the results were not similar at all. These studies and many others raised concerns at the NIH and plans emerged to adjust the NIH grant submission guidelines to enhance reproducibility. Whether you believe there's a reproducibility crisis or not you will have to adhere to NIH's new guidelines regarding rigor and reproducibility. From a data management perspective it's a guideline for scientific rigor that's relevant.

In NIH's description of what's needed to ensure scientific rigor they state that it includes full transparency and report in reporting experimental details. What this means is that saving the data produced by experiment is not sufficient. The data will not allow others to reproduce your work. In thinking about data management it's now necessary to take a broader perspective. So we have to preserve the lab notebook. And the experimental workflow must also be documented and preserved. And now also the computer programs must be saved including the information about parameters, versions, operating systems, and anything else that's necessary to reproduce the results. So for a clinical trial the protocol must be saved including the study design, the intervention, the inclusion, and exclusion criteria. So, in short, the data management encompasses ensuring that everything needed to reproduce a study has been preserved.

So there's now some data danger zones we have to talk about. So let's see if I can get this to work our next portion. So I have a little video that hopefully will play...

Hello my name is Dr. Judy Benign. I'm an oncologist at NYU School of Medicine. Hello Dr. Judy Benign. I read your article on b-cell function. I think that I could use the data for my work on pancreatic cancer. I am NOT an oncologist. I know but I think I could use the data for my work on pancreatic cancer. Do you have the data? Everything you need to know is in the article. No. What I need is the data. Will you share your data? I am not sure that will be possible. But your work is in PubMed Central and was funded by NIH. That is true. And it was published in Science which requires that you share your data. I did publish in Science. Then I am requesting your data. Can I have a copy of your data? I am not sure where my data is. But surely you saved your data! I did. I saved it on a USB Drive. Where is the USB Drive? It is in a box. It is in a box at home. I just moved. But can I use your data? There are many boxes...so many boxes! I forgot to label the boxes!

Hello again. Thank you for sending me a copy of your data on a USB Drive. I received the envelope yesterday. You were welcome. But I will need that back when you are finished. That is my only copy. I did have a question. What is your question? You might find the answer in my article. No. I received the data but when I opened it up it was in hexadecimal. Yes that is right. I cannot read hexadecimal. You asked for my data and I gave it to you. I have done what you asked. But is there a way to read the hexadecimal? You will need the program that created the hexadecimal file. Yes. I will. What is the name of the program? Saito synth. I do not know this program. It was a very good program. The company that made the program went bankrupt in 2007. Do you have a copy of the program? I do not use this program anymore because the company that made it went bankrupt. Maybe you can buy a copy on eBay?

I have good news. You again? I talked to my colleague. She knew a person with a copy of the software. Then why do you need me!? Everything you need to know about the data is in the article. I opened the data and I could not understand it. If you have the program you will find it as clear. Well, I noticed that you called your data fields Sam. Is that an abbreviation? Yes! It is an abbreviation of my co-authors name. His name is Samuel Li. We call him Sam. I see. And what is the content of the field called? Sam1? Ah, yes! Sam1 is the level of CXCR4 expression. And what is the content of the field called Sam2? That is logical if you think about it. What is the content of the field called Sam2? I don't remember. What about Sam3? Is there a guide to the data anywhere? Yes of course. It is the article that is published in Science. The article does not tell me what the field names mean. Is there any record of what these field names mean? Yes. My co-author knows what the content of Sam2 is, and Sam3. And Sam4. Can I talk to your co-author? I'm not sure. I would very much like to talk to your co-author. Well he was a graduate student. He went back to China two years ago. Can I have his contact information? He is in China. His name is Sam Li. I think I cannot use your data. You could check the article to see if what you need is there? Please stop talking now.

All right. But it almost showed a little bit without having even less information the confusion that sometimes by not recording, you can end up with. But I will be sure to - when we send out our email with our links and everything - to include the full video so that you can not only see some of those links but you can see the full video as well. But that does illustrate some of the common challenges that you can face by not properly managing all of your data. And some of the snafus that can occur after the fact when data is not properly managed.

Alright so when you think about all of the steps you take when creating, processing, and analyzing your data you want to avoid a scenario where all your work is tied up in a USB Drive that only you and sometimes not even you can understand. You don't want your data to end up with Sam Li in China. In particular. And so now what I'd like you all to do is think about your current workflow and ask yourself the following questions. Can you easily locate your raw data? Can you understand it? Can you understand the processes that you took to get the raw data to the processed data? And what tools or versions of software did you use to get it to those stages? Can you connect the different types of related data you collected? If those collected forms of data and imaging data that you that are related is it all easy to track down? So can you connect your steps? Could you figure out how you got things from one step to the next? And are your naming conventions consistent with others on your team? And finally do you name your subjects, animals, and specimens, in a consistent way just as everyone else on your team does? Is it a documentation that would allow you to share your data with someone else, and would they know exactly how you named something? So could you share it and have somebody actually understand it not just know what Sam1 is. You want to ask yourself these questions because you don't want a simple data management error to be a cause of something really bad to happen.

So take this article from the New England Journal of Medicine - the editors found multiple errors within a table in the paper. Even though the errors didn't alter the conclusion of the article, the authors couldn't find the primary data they collected. Because they couldn't locate the primary data, their paper was ultimately retracted. And this is a perfect example of where poor data management had really significant consequences.

[Title slide indicating that this is the end of Part 1 of this series. Please continue with Part 2 of 3.]

660
00:28:45,333 --> 00:28:47,399
[Title slide indicating that this is the end of Part 1 of this series. Please continue with Part 2 of 3.]