Lisa Wissbaum: Welcome, and thank you to the many MRSEC faculty, students and post-docs for taking the time to come and learn about archiving MRSEC data. My name is Lisa Wissbaum, the MRSEC manager, and I'm joined today by Meghan Lafferty and Wanda Marsolek from the University Libraries. They are liason and subject matter experts with the data repository at the University of Minnesota, known and referred to going forward as "DRUM". I'll first tell you what's new at MRSEC and why we need your help in archiving your MRSEC data. Meghan and Wanda will lead you through how and where to archive that data to meet our NSF commitment. We anticipate this training will take about 30 minutes, with ample time for questions at the end.

Today's Agenda will cover

- data management in the MRSEC
- the process
- why care about data management?
- DRUM best practices

We'll have questions and answers at the end but if you have questions during the presentation feel free to use the Zoom chat and we'll try to address those at the end. Also a fuller slide doc of this presentation will be sent along with a recording of today's session to all MRSEC-supported folks.

So again I'm going to explain what data and why it needs to be archived. In compliance with the current NSF policy on data management plans, our recent MRSEC proposal and subsequent newly awarded six-year grant commits that all data supporting any publications will be published in a stable, citable format using a well-established data repository, which is DRUM for us.

MRSEC's priorities are: to make this data archival process manageable for participaing researchers (all of you); to comply with FAIR data standards; to protect the confidentiality of the peer-review process while giving reviewers access to the data files; and to protect data integrity.

Simply put, this is what we need for every MRSEC supported publication where primary or partial MRSEC support has been received. Each publication needs to include a MRSEC acknowledgement to our new award, DMR-2011401 and a publication author must submit data sets to DRUM.

What's a MRSEC supported publication, you might ask. MRSEC defines 'primarily' or 'partially' supported as follows: 'Primarily supported' means that the lead student or postdoctoral co-author receieved 50% or more of their salary through the MRSEC. 'Partially MRSEC supported' means that the lead student or postdoctoral co-author received less than 50% of their salary through the MRSEC. By definition each of you here today is, or will be, a MRSEC primarily or partially supported researcher. This means on any publication one of the authors must submit the data sets to DRUM. These definitions and how to acknowledge the MRSEC in publications and data sets are also found on the MRSEC web site.

Our objective is clear. We want to archive data associated with MRSEC publications and our motivation is to improve data sharing and re-use. We'll do this with your help and that of our DRUM partner. Now I'd like to introduce you to Meghan from the UMN Library. She'll talk to you more about how to archive your data sets.

Meghan Lafferty: Good morning! I'm going to start by talking about why data management and sharing is important and then talk about some specifics about how you do that and then I will talk to my colleague Wanda who some of you may already have worked with if you have archived any data yet. They are one of our curators. Some of you may have spoken with me before because I am a liason with Chemical Engineering and Materials Science as well as Chemistry.

Have you ever read a paper and wanted to look more closely at the data and so you contacted the author and found out they did not know where the data was? Have you ever tried to make sense of your own data from several years ago and struggled to understand what you did? Good data management practices can help save your future self time and can also help other people make sense of your data. If your data is freely available online it can potentially increase your research impact as people will easily be able to cite it in papers and possibly re-use your data or build on that in their own research. And it can also help preserve your data for the long term.

Now I want to talk about documenting and describing your data. The goal with this is complete and accurate records. One of the ways we do this is through having you complete a [readme.txt] file template to more fully describe your data and I'll be elaborating on that later. It's the kind of things you might be tracking in your lab notebook like describing the conditions of data collection as their relevant or the equipment that you used or any kind of changes you make to your data, those are the kind of things that you might be recording. I'm going to be highlighting some of the important things for you to keep in mind as you're preparing the data.

So without standard ways of collecting and recording data you can lose important details about the data and that can potentially diminish the value of the data. So you need the information that helps make sense of it and if you look at this example [slide] this first table, there are no labels for the columns and maybe you know what it means but if anyone else looked at it would they know what it means? And in the second example there are labels for the columns but maybe someone else looking at this doesn't know what your abbreviations mean so you'll want to do things that explain what your abbreviations mean even if it seems very obvious to you. Your audience will be someone else who is a similar kind of researcher but not necessarily someone who has been working on the same project. It may be a new grad student in your lab in a year or two so you're helping those future potential readers.

File naming practices is something that's really important with data management. You see here [slide] an old PhD comic - maybe you don't name your data like this but a lot of people do... where you just kind of have these... [final.dat] or [this is the right file.dat] - these kinds of names. If you've ever downloaded a pdf of an article and forgotten to change the name then you have this string of letters and numbers that might not mean anything and then do you... you know if you can avoid opening a file to tell what's in it it makes your life a little easier. So that's the kind of thing that file naming best practices does. You want your files to be findable and also make clear how they relate to each other.

Good file naming practices can make your life easier. So you want to use file names that make obvious what the content is. In this first example, if you name the file [2019FederalTaxes.pdf] that doesn't depend on what folder it 's in to convey meaning. if you name it [Taxes/2019/federal.pdf] that depends on it being in a certain folder to make sense. It's not necessarily going to mean anything to you unless you open it and that takes time. And you want to use consistent terminology and, particularly within your research group and with your collaborators, think about shared terminology, how you're describing things. And there's some things where you might already have that but you might think about how to describe this thing or that thing. Also, rather than using 'final' or 'last version' or 'useThisOne' or those kind of things use version numbering like 'version 1' 'version 2' and also we recommend making use of numerical dates in your file names or folder names becuase then it sorts by year then month then date. Because if you have something with the month name in it, it's not going to sort nicely by the calendar. Also there are tools for file naming so it doesn't have to be like, you going in one by one to rename things. There are ways to make this easier. I mentioned that you'll want to list your versions alphanumerically and I would recommend using something like 'v01' or 'v002' etc. depending on how many total versions you think you will potentially have. If you're doing an article I hope you don't have 100 versions but maybe you will. That is also helpful just for sorting. And you want to decide how many versions of a file that you're going to keep and how often you're going to delete versions and then it's a good idea to assign someone that responsibility. Identify milestone versions to keep and keep those in a single place.

Directory structure for naming your folders - the goal is something that's helpful and predictable for identifying what information is in a folder. And when we've worked with people who are putting their data together to upload into DRUM people use a lot of different kinds of ways to name things. Some people are naming things by the material they're working with... it might be by the type of data like the instrument or what kind of measurement, it could be like this is all the data associated with this figure in your paper... Whatever system makes sense to you, you just want to pick a system and stick with that. But you do want something that doesn't depend on a lot of layers like I mentioned previously it doesn't depend on something being in a certain combination of folders and subfolders in order to understand what you're looking at. So there's not right or wrong way but just pick a sytem and stick with it.

[slide] This is an example of our template for the [readme.txt] file. We worked with the Center for Sustainable Polymers previously and worked on a template that we had that was more general and we took out some of the things that very rarely apply for people in chemistry, chemical engineering, materials science. You don't have to look closely at this, as you can tell it's very small. This is something we'll have you complete and this is what a lot of the back and forth is about when you're working on curating your data.

Some of the kind of information that you'll include is: who contributed to the project; what kinds of data and analysis, a full list of file names, when the data was collected, any other kinf of pertinent information like that; where it was collected; what questions you're trying to answer; what you would usually include in the abstract

of the paper. Will reference the published article if there's one associated with it.
The goal of doing all this is to make sure to meet what people in the library world refer to as FAIR data standards so they want the data to be findable, accessible, interoperable and reuseable. And now Wanda is going to talk more specifically about some of the things that apply more to DRUM.

Wanda: Thanks Meghan. Good morning everyone, or good day, I don't know where everyone is, potentially you could be spread across the world. I'm going to talk to you a bit more about when you're submitting your data to DRUM and the process that we go through and how we make this work. It's very collaborative. Here is a screenshot [slide] of DRUM, DRUM accepts data sets that meet standards and those standards are primarily useable and available to the public, so open. And what's great about this too is that all of these data sets and this information are exposed to Google and it's fully indexed so it's more easily findable by people across the world. DRUM is free for all UMN faculty, students and staff. You're not charged which is really great. We accept all data types and topics and we go through a process of pulling out key words and making your data accessible and relevant for folks. It's curated by experts and there is the public data sharing. We have a data seal of approval which just kind of makes it, it's that stamp that we worked really hard with other organizations with other organizations across the US on. And like I said it's accessible by Google so if you're an author or helped put the data together, it makes more impact for you, which is always good when you're trying to tell your story, right?

This is an image of the MRSEC DRUM collection page [slide]. Data sets make up the collection and so all of your data sets will be linked here as long as we know that they are to be associated with it so sometimes you'll have to let us know. We'll usually be able to tell becuase of that NSF funding number that you provide along with it. And this is to help your administrators and your teams tell a story at a glance. There is a manual process we have to go through to get the data linked to this page. It's not difficult but we are human and sometimes that gets left out so if you are looking at this page and don't see your data set please let me know and we'll get it linked. It doesn't take much but like I said we're humans and please share grace with us.

So here [slide] is an example of two data sets under curatorial review. You'll notice the thumbnails for each of these says 'submission under curatorial review'. When you import your data or upload it rather, this is what the world sees. So this gray thumbnail 'submission under curatorial review' [the world] knows that it is not finalized and so there could be some changes. There are persistent links or URLs that you will get right away with the data set once it's uploaded so you can share that. Once we have finalized the data set we'll then mint a DOI for you and you can share that. So there's actually two persistent links that you'll get. Sometimes in academia the DOI has more leverage or is thought more highly of than a URL; really technically it's the same thing but there are little nuanced differences. You'll see in the submission on the left - or maybe you won't because it is small - that there is no description or abstract in the first data set and in the second one there is. And what's important about including a description and abstract or one or the other is that it gives context to what this data set is all about. That's really important. When I'm curating your data that you upload I use the lense of a future user or future researcher and how they will use this data. So I focus on the [readme.txt] that Meghan talked about so that's really important and if I can find a publication I look through that to make sense of the data. What I'm looking for is how I can, as a future researcher or future user, how I can reuse or reproduce this data.

If it includes code, how I can run the code. How I can use it and reuse it. Once you submit your data I or another curator will start a conversation with you back and forth via email and we're just trying to, we'll make recommendations about how or if you should change your data and it's not necessarily about changing your data it's about the documentation, again, how can others use this? An example of a suggestion that might be made is... I'm gonna say like, "a clear detailed description of the work flow would help me and others make sense of this large set of data files." "What order should the scripts be run in?" "What libraries must be installed?" "What is expected output?" Things like that. Again, I'm saying it over and over again because it's important. I'm trying to think about how others will use this and what information they need. Sometimes when we are collecting data and analyzing it we just know it and so there's steps we leave out. So I'm taking that outsider's point of view. And then sometimes there will be extra files in the data set that maybe you were using for testing but don't necessarily need to be in the shared data.

Here is an example [slide] of a finalized data set. You'll notice that that gray 'under curatorial review' is gone and now there is a thumbnail, an image from the manuscript, most likely. And you will see there is also under the thumbnail is both the URL and the DOI that we minted for you. Here you will see that the files that are stored is the [readme.txt] as well as a ZIP folder for all of the data files. Other things that we have included are instruments and data dictionaries. A lot of times though even if there is a dictionary we'll still request a [readme.txt]. It's really important that we look at data that is saved in file formats that can be used by others for presentation and accessibility.

The three step process relates to MRSEC. These are the steps that you'll want to think about and take. When preparing data, use the best practices of data management that Meghan was talking about. Also, when you receive this slide deck, there are slides that didn't make the cut because we're trying to be observant of time and things like that but they're still important. Make sure that the data that you're submitting is limited to what you are using for your manuscript. You don't need to just include everything, but focusing on your manuscript and the data that you use for that. Make sure that you're labeling your data clearly and, as well as, units, sometimes, Meghan mentioned that some of those files didn't have headers and so, we need to know what are the variables and if there's a measurement, a unit of measurement we also need to know that and organize your files into a logical directory structure.

So it's all prepared now you're going to upload it to the data repository. We have a pretty smooth workflow I think. It walks you through all of the things you need to do. If you are submitting data that is going through peer review process currently you need to let us know up front because we won't want to post it on the open web, right? We want to make sure only your reviewers can access it. And we'll do that we'll put it in a google drive and share that drive link with you so you can share it with your peer reviewers. You'll also want to provide that [readme.txt] and think about 'Creative Commons' ("no rights reserved"). And so again... if your manuscripts are in peer review please let us know that. And if you need to share a DOI right away with the publisher we can get that DOI for you but it will not be live it's kind of like a placeholder. You can share it but it will not go anywhere but we will then mint it and make it live once your peer review is done and we've curated the data fully.

In that final step, which I kind of talked about just a minute ago, is including the DOI in the manuscript. So we'll help you with that either before or during your peer review or after if your data is not going through peer review. And, this is back to that 'creative commons'. Your data is not copyrightable. The reason for this 'creative commons' is so that when others want to use it they know what's appropriate as well as to cite your work. In our DRUM upload there is a tool that you can use to make sure you get the license you are wanting. Whether that be, people can just use as-is, there's no editing of it, or you can remix it, whatever you're wanting to use with your data.

And with that, that is what we have from the libraries and we are just super excited to be involved in this process with you, in this journey and if there's any questions we'd love to hear them.

Lisa Wissbaum: We could certainly open up to questions now. What's the best method for that, Wanda or Meghan.

Meghan: We can have people ask...

Lisa Wissbaum: Just have them unmute themselves?

Chris Leighton: Could I ask a quick question? I'm wondering for people who aren't here, will this be posted somewhere on the MRSEC web site or emailed out so they can look at it at another time.

Lisa Wissbaum: It will be in one or both of those places for sure, Chris. Everyone involved in the MSEC will receive a copy of the full slide deck which had some information that's valuable but we just didn't include it and then also a recording of this will be shared as well.

Chris Leighton: Great, thank you.

Lisa Wissabum: Any other questions? Ok, well fee free at any point.. the folks at the Library have been very helpful through all of this, very responsive with emails and of course I'll field or try to help in any way as well. So just reach out if you need anything. And thank you very much for your attendance.