NIH Data Management and Sharing

0:00
We're going to get started.

0:02
I am Michelle Kennett, Associate Vice Chancellor for Research, and I'm here with Stephen Pryor and I will let him introduce himself.

0:14
Hi.

0:14
I'm Stephen Pryor, Director of Digital Initiatives for University of Missouri Libraries.

0:19
And that includes working on data management plans and repositories and other digital research projects.

0:29
All right.

0:29
So I'm going to share my screen here and we're going to get started, I hope.

0:38
OK, Can, can you see this?

0:45
I guess the question is, can you right the right screen?

0:48
Yep, we're good.

0:50
All right, very good.

0:51
So we're going to talk about the new NIH data sharing requirements here this afternoon.

0:59
And one of the caveats I will say about this is that certainly as we've been having more and more discussions about this, it is certainly come to my attention that especially for certain disciplines and different folks, there are a number of open-ended questions out there, especially about certain kinds of data and concerns about that.

1:26
I think you you are not alone in that.

1:30
There are a lot of questions out there that, that are bubbling around.

1:34
We probably may not have the answers to those today.

1:38
But two things I can tell you is we are watching and and listening to where those questions are going with NIH and others.

1:46
And also working to create some conversations internally to discuss what those barriers are for investigators and what potential solutions we might be able to come come up with or at least explore where the national conversation is going around really significant questions.

2:06
So just we have an awareness of that today.

2:10
We're going to go through what we know about the requirements and expectations and as I said, we are, we are aware that there are some other kinds of issues out there, but hopefully maybe we can at least provide you with some resources and things here to to start along this process.

2:31
So first we're going to talk about what what is this?

2:36
So where what is it?

2:38
Where do you come from?

2:39
What are the requirements?

2:40
And then Steven will get into some details around what things with repositories and data plans and things that assistance that you can get from some of their resources.

2:55
And then he is going to touch on the issue around orchid IDs also.

3:03

It is tangential to this conversation, but it is an important issue as it is something that is being discussed and I think will become more and more important in terms of requirements down the road.

3:17
So I think again those of you publish are familiar with orchid IDs but we're going to talk a little more about what that, what that is and making sure that you're properly in tuned in terms of the the future.

3:33
So again we have the the data sharing policy that came out went to into effect January 25th.

3:42
It really is an expansion of previous data sharing policies that NIH had to promote the sharing of scientific data for the purposes of accelerating biomedical research discovery and in part by enabling validation of research results, providing access to data sets and and promoting data for future studies.

4:07
And so those are the high level reasons why this conversation, why NIH has gone down this path.

4:15
And I will say that while this does call out biomedical research, the conversations that I've seen and heard would make me believe that in the future you know there are certainly plans to make this more expansive.

4:33
So I think you know that is something that's really across the board and as I said this is a really an expansion of an existing policy that you know there has been some things in place around genomic research around you know the the cut out $500,000 cut off and things.

4:57
Now again the application is across the board.

5:04
So as I said the policy is going to apply to all research funded or conducted in whole or in part by NIH and it results in the generation of scientific data and I'll talk in a minute about what they are defining as scientific data.

5:21
And again it funded or conducted extramural grants contracts, intramural research projects, other funding agreements regardless of level of funding.

5:32
What it does not apply to is research or other activities that do not generate scientific data of training, infrastructure development, other non research activities.

5:46
So again there there is a that is defined scientific data.

5:54
Again what they're talking about data commonly accepted in the scientific community to validate or replicate research findings.

6:03
Again, this is sort of the impetus behind some of this is issues around replication and reproducibility.

6:14
Again, they kind of reiterate that what they do say is it does not include laboratory notebooks, preliminary analysis, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues or physical objects such as laboratory specimens.

6:35
So they have defined that.

6:38
So one thing I will say about this is again those of you who have questions about again scope and issues at this point in time, I would encourage you to get with your program officer to discuss those concerns.

6:56
The program officers are going to be the vetting process for these plans.

7:01
And so I think that if for until other information comes out or any guidance or otherwise addresses those concerns, those are things I think should be taken up with a a program officer.

7:20
So in terms of the policy there needs to be a investigators should plan and budget for managing and sharing of data And I realized that that it sounds very easy and I realized what some of the barriers to that are.

7:37
But again, you ultimately you can budget for that, submit a plan again that there is a template on the NIH website for that and you also can find information some of these resource informations on our sponsored programs website.

8:00
Also it's expected that there is compliance with the approved plan.

8:07
There may be again depending on what centers or offices you work with through NIH other caveats that they expect and hopefully those would be communicated.

8:18

But and again NIH website is actually a very valuable resource for a lot of detail around the expectations and examples of that.

8:30
They have FAQs.

8:31
There actually is a lot there.

8:36
And as I said, awardees are expected to file the plan is approved as a term and condition of the award provide updates in annual progress reports.

8:48
If there are plans to change the the plan, then again it's working with the NIH program officer to obtain approval for those modifications.

8:59
So there is an expectation that what you say you're going to do is what's going to happen.

9:09
These are the elements of the the data management and sharing plan.

9:15
I'm not going to go into great detail.

9:16
I'm going to turn it over just to Steven here in a second who will talk a little bit more about repositories and plans.

9:24
But again these are the the topics that are expected to be addressed on the NIH website.

9:32
They go into more detail in terms of each of those requirements.

9:37
One thing I do want to point out is on the bullet oversight of manage data management and sharing that piece.

9:45
So the the what should go in there for the University of Missouri studies is that the Pi will be responsible for the oversight and this is tailored specifically to your project and I did put a link there.

10:01
The sponsored programs has some language, an example on their website of how that could be written and and what that would entail.

10:13
So again be careful not to take if you're using the template from NIH or other places sometimes again they have different mechanisms for oversight.

10:26
NIH has said we don't, we're not going to prescribe how that is, but we're not doing it it.

10:32
It needs to be done.

10:34
And again for our purposes, it's API responsibility of to make sure that you are following your data management and sharing plan.

10:49
And again we've given you some tools there and language to to be used in those plans.

10:58
All right.

10:58
I will let Steven take it.

11:03
All right.

11:04
Can we switch the screen sharing?

11:07
I'll share mine.

11:10
Yes, can get there.

11:26
Oh, here we go.

11:27
All right.

11:31
Thanks.

11:31
So excuse me starting off with the elements of a data plan, so.

11:38
So as you're building your data management plan and you're following the guidance from the NIH, they list pretty clearly that there are six things that they expect to be addressed.

11:50
We have several online guides through the library.

11:55
I just found a couple more pages that needed that Michelle's and so I finished, finished that and so they've been updated for the new guidance they outlined.

12:04
These six areas have little snippets of additional information, plus the NIH page that's linked from those and and I'll be copying in some more links here.

12:16
In a moment.

12:19
Outlines all of this information and so in our online guides through the libraries we have links to to several available tools, example language and so on.

12:38
This particular one, our data management guide, we have links to a tool called DMP tool.

12:43
You should be able to use your University of Missouri single sign on log into that.

12:49
Let me know if you have any problems with that.

12:52
It seems like it's working fine for people.

12:56
Sometimes there are issues because we're a multi campus system and so I've seen people be associated with the wrong campus.

13:09
Let me know if you have any issues, I can fix all of those.

13:12

The DMP tool lets you select what program you're applying for and then gives you sort of boxes to fill out for all the pieces of the plan that need to be completed.

13:23
We have links to some examples.

13:25
We have links to Orchid information that Michelle mentioned and then information about various things that you need to think about as you're working through your proposal and thinking about what types of data you're going to be producing.

13:42
Who needs to be involved with collecting and storing it?

13:45
What are the file formats?

13:47
Is there special software involved?

13:50
All of these things are going to link up to those six issues that we looked at, related tools, software and code, data type standards, How are they going to be preserved?

14:02
OK, so we have tools for more details.

14:05
With that, contact your IT Professional Research Support Services group if necessary, and your grant program officer.

14:16
As you're thinking about your data that's being generated, think about confidentiality, personally identifiable information, whether IRB needs input into human subjects, data involved, and and documenting your data.

14:38
So these are things that you'll have to think about addressing with the goal towards sharing this data.

14:45
OK, so if there's confidentiality issues is how do you plan to de identify the data for example?

14:55
Or is it even possible to de identify the data?

14:59
I there there are examples of social sciences studies that that work with a small town population in a

certain area or something like that, and and they're interviewing people and in order to provide the relevant details for the research that it can't, it couldn't be sufficiently de identified right?

15:24
And so there are sometimes tricky issues that you need to think about with regards to your data.

15:29
Be prepared to explain any issues around sharing the data.

15:35
Another issue around collecting your data set is documentation.

15:40
So going along with the sort of whole picture of data management is, it is not particularly useful and I think the the individuals reviewing your data management plan are going to want to see that the data is is useful to people who are going to be accessing it, right.

16:10
And so if you were providing a data set, it's helpful to provide documentation about the data.

16:17
And so this screenshot here is a a screenshot of a data dictionary, a simple data dictionary that just lists all of the variables.

16:29
So if you are producing a spreadsheet or a CSV file of data, you know what all the rows and columns are, but nobody else is necessarily going to know that unless you specify it.

16:45
And so this example data dictionary has all of the the variables participant ID number, the variable name or the column name is ID, what's the measurement unit, what are the allowed values, and a text description of what it is.

17:02
OK.

17:03
And that helps people understand what your data is and how to use it once you share it.

17:10
And the other thing to consider in terms of sharing your data and putting it out there, what kind of licensing and rights do you want to apply to it?

17:19
So when you are asked to upload it to a repository, the repository may give you options.

17:27

Your grant program might specify that it needs to be Creative Commons 4 point O public domain or things like that.

17:41
I know that some of these like how open it is, whether you can restrict commercial uses, whether you can restrict adaptive works, things like that.

17:53
Some of those, some of the availability.

17:57
Where you're able to choose those restrictions depends on the grant program.

18:02
So a lot of times the grant program will have will say this has to be completely open or you can maintain a certain amount of control over the the work after it is released.

18:22
And then where are you going to share the data or how are you going to share the data?

18:29
NLMNIH often has a list of specified repositories where certain programs specify that you have to deposit a copy, say in Pub Med Central for example.

18:45
If your data set is smaller than two gigabytes in size, it has to be deposited in into Pub Med Central specifically.

18:55
There are other discipline specific repositories that are appropriate.

18:59
I know that there are some for genomics and some for different specifications and programs.

19:09
There are general repositories ICPSR for sort of social science interview.

19:15
There are lots of sort of recordings and video data in that database.

19:21
Databrary is another general Dryad.

19:25
And then we the university has an institutional repository MO space which is appropriate for any sort of scholarly output, including research data pub published or generated by faculty at the University of Missouri to be made public.

19:44
And Open Science Framework is another general repository.

19:52
The screenshot from the data dictionary I showed earlier was from an OSF project that lets you sort of organize your data files and then eventually also publish them.

20:07
So those are some options for data repositories.

20:12
So check first whether your program specifies which specific repository it has to go into.

20:18
And then you have lots of options for domain repositories or the University of Missouri, a little bit more about the University of Missouri Repository.

20:28
So anything deposited in the University of Missouri Institutional Repository is eligible for a permanent identifier or ADOI.

20:39
These deposits are Open Access, indexed, widely accessible.

20:44
The metadata is harvestable in a standard metadata schema.

20:50
The repository software that we use is an industry standard for article and data repositories, and it's backed.

21:01
So the MU Libraries has a commitment, decades long now for preserving and perpetuating this repository into the future and and maintaining and curating that data in that repository.

21:19
So it's intended to be a permanent record of the scholarship of the University of Missouri.

21:25
And it just so happens that all of those things in that list that I just read off are things that the NIH considers desirable characteristics for repositories when you go through their documentation on how to choose a repository to document your data.

21:41
So most space would be eligible for what they consider to be a desirable repository.

21:49
Assuming again that your program or call for proposals doesn't specify that it has to go to a specific repository, this is what a most space data deposit page might look like.

22:05
There are actually 3 files attached to this object and one is called data.

22:10
One is additional information about the FILA data and one is operational taxonomic units data.

22:17
And so you can attach multiple files that include your data file and then descriptions about the data like metadata, the data dictionary, those kinds of things and have them all together in one object.

22:31
This particular the object has a handle which is a unique Ida unique permanent URL, and we can also do Dois for these objects.

22:44
And then with each deposit or item in most space, if you want, you can view statistics showing how many times it's been viewed and downloaded and those kinds of things.

22:58
And then finally a word about orchids.

23:01
So Michelle mentioned orchids.

23:03
There are some of the NIH and NSF programs that are requiring bio sketches or online sort of CV objects that that is a a record of your scholarship and work that you've produced.

23:34
One way to keep track of this because there are there are NSF and NIH grant required ones.

23:44
There's from My Vita here on campus, there's lots of different systems where you might have to upload and share this kind of information about all the things you published before, grants you've applied for or been awarded, books, publications, chapters, things like that.

24:02
Orchid is a central place where you get a unique identificate.

24:05
You get a unique ID number as a researcher and it stores that record and it identifies 1 Stephen Pryor from all of the other ones that are out there.

24:18
And there are at least five or six out there publishing things and so.

24:27
So when you apply for a grant or submit your article to an Elsevier journal and you include that orchid ID, that data will automatically feed into your orchid.

24:40
And then anything that's connected to your orchid, like an NSF bio sketch should automatically pull that data out of there.

24:47
So it's a fantastic sort of intermediary service data pushes and pulls from it.

24:54
It keeps you disambiguated as a as an author, as a researcher and maintains, helps maintain sort of a singular record for you.

25:09
And when you go to orchid.org sign in.

25:14
You may have created one with a personal e-mail address.

25:16
That's fine.

25:17
You can also look through for when you click Sign in.

25:20
Underneath the box there's a Access through your institution button, which is a pretty common thing.

25:26
Now you can go through single sign on and access your account that way.

25:32
And then there are actually ways that you can associate multiple e-mail addresses, recover your record if you've signed up before, if you've done Orchid through another university's e-mail address.

25:46
So yeah, we can help out with that sort of thing if you have used Orchid before and and lapsed on it.

25:58
So for all of these services I would recommend you can either contact me For more information.

26:04
You can go certainly go through your subject librarian and then I have a list of links here and some other information to provide here At the end there's a paste of all of these links so we have a general data Management library guide.

26:30
It links out to a specific guide on the NH, the NIH Public Access plan that Michelle helps and her office helped update for the 2023 plans.

26:45
And we have some information on NSF and general data management basics.

26:52
There's some sample language in there about what most space can do.

26:56
And then as I sort of see more, I've I've seen a few things come through under this new plan.

27:03
And so I will.

27:06
I plan to be updating that with more sample language regarding the oversight section and other other details that are relevant.

27:24
And the other thing that I pasted into the chat is the link to a story that I just received this morning.

27:31
And unfortunately, it was our the hours were have already passed for this morning.

27:38
But the Health Sciences Library, the Health Sciences librarians who also work and consult on data management plans, actually have set up office hours on Friday mornings for helping answer some questions about what what the librarians can do in your data management plan.

28:00
And so if you have questions about about metadata, about what would be an appropriate repository, where can the data go, how can we describe the data?

28:14
Those sorts of things.

28:17
Definitely check that out.

28:23
Yes.

28:23
So there's a question in the chat.

28:25
Is most space able to handle all MU investigator data presentation and management needs?

28:30
Would there ever be a time we can't handle all the data your research is generating?

28:33
So to clarify most space would be.

28:39
So thinking back to the definition that Michelle gave of the scientific data generated during the project, that definition if you're very specifically kind of refers to the final output, at least in in my reading of it so far.

28:58
I think as has also been indicated we're we're sort of figuring this out and learning more about it all the time.

29:03
But it reads to me that that this is the final out.

29:07
You know it doesn't include lab notebooks and working copies of things like that.

29:14
So there are a lot of things that can happen.

29:19
Truly large big data data sets would probably not be appropriate for most space for a number of reasons.

29:28
Most space hosts a file and somebody's going to have to download it.

29:31
So if you're talking about something massive, that just might not be a reasonable way to transfer that data.

29:40
And so we do.

29:42
I do often refer issues.

29:45
They're they're evolving technologies through the Research Support Services group.

29:50
It used to be research computing and high performance computing.

29:55
And so we've had conversations with them before about storage infrastructure or other possibilities to transfer really large data sets.

30:05
And then again, if your project is just generating a lot of incidental data and leading towards a final product, most spaces not the place to archive that sort of intermediary data, it's to publish the final product.

30:27
Does that help?

30:32
Oh, yeah.

30:32
Matthew Keeler's here.

30:33
Thanks.

30:41
Thank you.

30:41
Do we have any more questions?

30:42
Feel free to type in the chat or you're welcome to unmute as well.

30:46
And just ask are there costs if we use most space, is there a cost for hosting that data, for hosting data in most space?

31:07
Again for sort of in the in the general case there there is not a cost.

31:17
Again, we might have to refer or come to another agreement with something really large.

31:26
So Pub Med says, so the the Pub Med guideline says if it's under 2 gigabytes, it can go in Pub Med on the order of a multi GB data file.

31:41
Or I mean we can handle more than two gigabytes, but if it gets larger than that, or if it's a different kind of data, video data or audio, I think we should have a separate conversation.

32:02
But for for most, for the most common cases, there is no charge, no no cost associated with most base.

32:14
I did learn that the Health Sciences library has some medical data consultation services.

32:23
There are issues there again with going through it, making sure that there's no DE identification issues and and they do offer some services that have sort of an hourly cost associated with with those.

32:41
Can I ask a follow up?

32:43
If you go to the pub Med route and the limit is 2 gigabytes, is that 2 gigabytes for data or is it 2 gig for the entire package?

32:52
In other words, the data dictionary, the code, and the data, I believe that that's the entire package.

33:09
OK, thank you.

33:10
I'll make a note to follow up on that.

33:19
So there there was a question about sharing qualitative data and I think that is a good question.

33:27
I know that just from some of my travels through trying to to see where folks are AT and what what is going on in this space.

33:37

There is certainly a lot of conversation around that and I think that's something that we will continue to explore and put up information as we find it if if it leads to helping to better define that.

33:54
One thing I think is a big issue with that too is in terms of what the expectation is in terms of consent from participants and such.

34:06
So again, if if your participants have not agreed then then there's some additional issues or or potentially the the the need for an exemption in your plan to actually put up information that may be identifiable or otherwise go against what you have indicated to your to your participants.

34:36
So, but as we as we find more information or or or get thoughts on that, we'll post those.

34:47
We also had a question about what you might know about other federal funding agencies besides NIH and if they will, you think this is a trend that there are other agencies are going to start requiring this in the future as best I can tell.

35:05
And from the information that I hear from organizations and places that that have been discussing this, I I think it is I think there can be an expectation that funders will will kind of across the board move in these directions.

35:24
So again that is just sort of a, you know, I don't have anything yeah in writing from anywhere that says yes, this is what's going to happen.

35:35
But certainly the the conversations that that I have heard indicate that there is likely to be an expansion of that the IT was 2013, right, that there was the memo from the Office of Science and Technology Policy sort of directing all of the federal agencies to come up with a plan to ensure public access to to publicly funded research.

36:11
And that was sort of when this first wave of the old NIH plan, the OR the previous NIH plan and then NSF which I see in the in the comments and they kind of sort of went along together.

36:28
And then there are a lot of other granting agencies and private granting agencies, Gates Foundation and all of these other ones that that have also gone along and written very similar sorts of requirements.

36:44
And so I I think it's definitely I might be behind I don't know if NSF has has updated their rules like NIH has but I I would expect that they would continue to get more stringent on the the public access

and and I think that is the difference because and I think and you are correct there has been guidance and suggestion and encouragement for some time around this issue.

37:18
But I think that the difference is the movement towards the requirements of what you you need to have and and making it part of terms and conditions and things like that.

37:29
So yeah there's a question that was sent to me as a direct message which is a good question.

37:37
Should researchers have a plan for when they should sunset their data, pull it from the system because it is no longer relevant?

37:49
That would be a good question and I would think that maybe under these new rules.

37:53
That's one of the situations where you would want to work closely with your program officer to check to check with the Funding Agency or or check the guidelines to to see if that is appropriate.

38:11
There are lots of situations where I still get the sense from reading it myself and and maybe you one of the the grant administrators might have a different idea.

38:25
But again you know when confidentiality is a concern or data quality in some way might be a concern.

38:34
You know you your your plan is for how you are proposing to manage the data and if there is a good case or a good reason why data would be no good to anybody after a certain period of time.

38:53
I think if you could make that argument, you could you could try to make that argument.

39:05
And I'm not sure that I can say definitively right now how the the funder or the the reviewers would would feel about that.

39:16
As a as a librarian, of course we go through you know, we well, I'm familiar with both public libraries and academic libraries.

39:26
And so in the Public Library, it doesn't do anybody any good to have a 1988 physician's handbook on the shelf because people will read it and get bad information.

39:40
However, we do keep those things in academic libraries because it might be important to see what people were, you know.

39:47
So that concern might also be addressed.

39:50
If the data might have some issue with relevance in the future, that might be more of an issue of making sure that that's documented.

40:00
So when somebody finds it 10 or 15 years from now, they know what that concern is and to be aware of it.

40:07
And so that kind of.

40:09
Can go along too with that data, data, documentation and making sure that when you share your data it goes out there with the package describing what it is and how to use it.

40:21
Well, thank you everyone for attending this session today.

40:24
We really appreciate you coming out and spending some time with us to learn about this important topic.

40:30
If you have any follow up questions, you're welcome to contact us here at the research office and we can pass it along or if it's fine with you, Michelle and Steven, they can, yeah, contact you directly as well with questions.

40:46
We appreciate it.

40:47
And we are recording this session and we'll send out the recording and slides.

40:54
It will take a couple days for us to get everything edited and ready to go out, but we will send it out next week.

40:59
All right.

41:00
Thank you so much.

41:01
Have a great day.

41:04
Thank you.

41:04
Thank you, Michelle.

41:05
Thank you, Steven.

41:06
Thanks.