Transcript

MIKE: Thank you all for joining us for another NLM Office Hours. My name is Mike Davidson. I'm from the Training and Workforce Development Team, a part of the User Services and Collections division here at the National Library of Medicine. My pronouns are he/him/his. The goal behind these Office Hours sessions is to give you a chance to learn more about NLM's products and to get your questions answered by our trainers and members of our product teams.

Today's focus is on a newly launched product, the NLM Dataset Catalog. We're going to kick things off with a brief presentation from a few of the folks who are working on the Dataset Catalog, including Peter Seibert of the Controlled Vocabulary Services Program, and Alvin Stockdale and Nancy Fallgren of the Metadata Management Program. Following the presentation, we should have plenty of time for our panelists to answer any questions that you have about the Dataset Catalog.

A few quick logistical notes before we get started. We are recording today's session. That way we can share it with those who are unable to attend, or you can review it later if you like. Everyone who registered for the Office Hours will get a link to that recording. You'll also be able to find it via the NLM Technical Bulletin. We have a pretty substantial crowd here today, so we've muted all attendees to cut down on the background noise of crosstalk. However, we are eager for your questions, and please feel free to submit those questions as you think of them, throughout the session, using the Zoom Chat feature, which you should be able to find near the bottom of your Zoom window. It looks like a little speech bubble. Make sure when you send your questions in the chat, you send them to Everyone. That way all of our panelists can see them and we can make sure that we can get your questions answered by the right person. When we get to the Q&A portion of today's session, I'll direct your questions, the questions that you've submitted, to the right panelist and then they can answer verbally. We may also occasionally be using the chat feature to share some links to helpful resources. But before we go into any of that, I'm going to hand things over to Peter Seibert to help bring us up to speed on the NLM data set catalog. Peter.

PETER: Thanks, Mike. As Mike said, my name is Peter Seibert. I'm from the National Library of Medicine and I'll be presenting the Dataset Catalog beta version, along with my two colleagues Nancy Fallgren and Alvin Stockdale.

So what is the Dataset Catalog? It's a catalog of biomedical data sets selected from publicly available repositories. We're considering it the PubMed of datasets because it allows for federated search across multiple repositories and will connect users directly to datasets. As I said, the catalog was designed for search across multiple repositories. Why did we develop this? First, it supports the NLM Strategic Plan for Data Science and the NLM Strategic Plan that was released in 2017 and that we're still working on for the next couple years. It's designed to allow discoverability of biomedical data sets across many repositories. And we're also looking to drive the adoption of acceptance of the NLM metadata standard that we developed for this product called DATMM (or, DATaset Metadata Model), which my colleague Nancy will discuss in a moment. How are we going about this? Well, we've launched the Dataset Catalog that we're going to learn about a bit more using the schema that we developed here. And we're looking to receive user feedback from this, which I'll go into in my presentation, how you can provide useful feedback to us. We're also looking to drive the adoption of DATMM by showing the usefulness of it in this catalog.

A committed team has been working diligently behind the scenes to develop this Dataset Catalog. So you see my name, Nancy Fallgren and Alvin Stockdale, who will be presenting today, as well as Jeff Beck and David Hale, who are colleagues here at NLM. And this is a hybrid team. So we're also working with a boutique contract IT company to provide some of the data science support as well as programming

support for the product. Now I'm going to hand it over to my colleague Nancy Fallgren to talk about DATMM.

NANCY: Thanks, Pete. So as Pete mentioned, the DATaset Metadata Model (or, DATMM) is the metadata scheme created and underlying the Dataset Catalog. DATMM is a linked data scheme, more formally known as Resource Description Framework (or, an RDF scheme). RDF is a very simple metadata model for expressing pieces of information in triples. Triples are comprised of a subject, a predicate, and an object, just like a simple sentence such as "Fritz is a cat." The value of using RDF lies in pushing the metadata out to the web, where it provides a means for expanding upon information or resources found when you search the web. So, for example, when we provide a unique ORCID identifier for Anthony Fauci in our RDF data and we push that out to the web, we're enabling the web to find and link our data to more information about him. Dr. Fauci's ORCID iD not only links to information about him at the ORCID site, but it also affords the web the ability to link to and display other sites where his ORCID iD is present. As an RDF scheme, DATMM was designed precisely to provide this ability to connect with other like things on the web. Next slide, please.

So this drawing represents the core metadata classes in DATMM. One feature of RDF is the ability to reuse classes and properties from other RDF schema, and this is actually what enables linking data together on the web. With that in mind, DATMM was designed to reuse classes and properties from other existing RDF schema such as Dublin Core, BIBFRAME, SKOS, and schema.org. Where necessary, we created our own DATMM classes, specifically the dataset class, the repository class, and the Documentation class. However, all the properties in DATMM come from other RDF schema. Next slide please.

So this is a more granular diagram of the classes and properties in DATMM. The model focuses on describing datasets which are the central feature of the model, along with their related collections, repositories, subjects and agents. In addition, the metadata scheme includes documentation, which are generally articles, associated with datasets. So while researchers may find referenced articles about a dataset useful and interesting, this also provides the option to find a dataset based on an article written about it. So in short, these DATMM classes and properties are expected to offer sufficient information for someone to find a data set, determine whether or not it's of interest, and then go to its home site for further assessment and potentially to access it. Finally, you can see the DATMM is a fairly brief and simple model. We've learned that if you require too much metadata, or if you offer the ability to provide too much metadata, the scheme can feel overwhelming and then inhibit adoption. So with that in mind, DATMM is kept deliberately lightweight in an effort to encourage and facilitate usability. Now I'll turn it over to Alvin to talk about using the data model.

ALVIN: Thank you, Nancy. Before we evaluate and ingest repositories, first we have to find them. We started looking at the list of repositories created by the NIH Biomedical Informatics Coordinating Committee (or, BMIC). Next, we looked at repositories participating in NIH's Generalist Repository Ecosystem Initiative (or, GREI). GREI is an NIH initiative whose primary mission is to establish a common set of cohesive and consistent capabilities, services, metrics, and social infrastructure across various generalist repositories. We also looked at a list of recommended repositories compiled by the Public Library of Science (or, PLOS). Currently we are looking at repositories found using Re3data.org's Repository Finder which includes almost 3200 repositories and has an API that allows us to programmatically query the repository list to find exactly what we're interested in. Next slide.

After we find repositories that might be suitable for the Dataset Catalog, they go through a multi phase evaluation. In the first phase we're focused mainly on the content of the repository. Is the repository

domain specific? If it is a generalist repository, does it contain biomedical datasets? Does the repository contain original research or is it a knowledge base which accumulates original research to add to a growing body of information? Phase two looks at the currency of the data. How often are new data sets added? Have any been added in the last six months? At this phase we also determine how we can extract the data set metadata. Do they have an API or FTP server? Does it seem easy to use? In phase three, we determine if the data set metadata can be mapped to the datum schema. Are data sets uniquely identified? Do they have contributors and concepts? Does the metadata contain all DATMM required properties? Next slide.

We currently have 4 repositories ingested in the Dataset Catalog. The Database of Genotypes and Phenotypes (also known as, dbGaP), is an NLM product. The Immunology Database and Analysis Portal (or, ImmPort), is sponsored by the National Institute of Allergy and Infectious Disease at NIH. DRYAD is an NIH GREI repository. Harvard Dataverse also a GREI repository with an academic focus. We have 27 other repositories that have been mapped to the DATMM schema, but the metadata has not been transformed to the DATMM schema yet. That transformation is a resource intensive process and is a current bottleneck we're trying to solve, and I'll talk about that more in a few slides. Next slide.

Once we have selected a repository for inclusion to the Dataset Catalog, we have to crosswalk that repository's metadata to the DATMM schema. This process can take one to two weeks per repository and is performed by two to three staff members. Each staff member downloads and reviews the metadata for multiple datasets to ensure the fullest mapping. The mapping team provides file paths for source metadata properties that have an equivalent in the DATMM schema. They also provide the metadata value for that property from their dataset. Finally, we also provide instructions the programmer will need when transforming the data. One example is if names for people are given in direct orders such as Alvin Stockdale. We want names of people inverted like Stockdale, Alvin, but we don't want to invert names for organizations. Once the team has completed their mapping, they present the mapping to the DATMM team for approval and talk through any points that prove difficult. Next slide.

This complicated mapping work is stored in the DATMM metadata application profile. The metadata application profile contains each property from every class in our model and you can see a very small portion of it represented in this slide. The mapping team fills out the file path, metadata value, and question/comments columns for the repository they are mapping so the programmer can write a Python script to transform the source metadata to the DATMM schema, allowing it to be ingested in the Dataset Catalog. Next slide.

The National Library of Medicine is over halfway through a six-month pilot investigating generative AI, using a safe and secure environment. For example, we can upload and query our own data, but the large language model will not use our data in its training corpus. I'm a participant in the pilot and I'm trying to solve the bottleneck I spoke of a few slides earlier where it takes a few weeks for our part-time programmer to write Python scripts to download and transform source repository metadata to the DATMM schema. I'm trying to generate these scripts myself using generative AI. The first phase is retrieving the data set metadata from the repository. To do this, I upload previously created Python scripts. Then using chat prompts, I instruct the chat bot to generate Python script for a new repository. I validate the generated script by running it and resolving any errors. Once the download has been verified, the data is stored. The second phase is more difficult. This is where we transform the source metadata to the DATMM schema, using the mapping shown on the previous slide. The process is really the same as the first phase though. Upload existing Python transform scripts for repositories we've already ingested. Also, upload the mapping sheet for the repository I want to transform. Using chat prompts will generate the Python transform script. Data will be validated by staff who perform the mapping. Once validated, the

dataset metadata will be ingested into the Dataset Catalog. The ultimate goal for phase one is to create a download tool that will generate Python script to download datasets from a repository. The tool will ask for the URL that contains the repository's API documentation and will ask for the location you want to save the files. After the information is provided, the chat bot will provide a download script for the user to run to download dataset metadata for that repository. The goal for phase two is to create a transform tool that would generate Python script to transform source repository metadata to the DATMM schema. The tool will ask you to upload the repository's mapping spreadsheet and it will generate the transform Python script needed to convert source metadata to the DATMM schema. And now we'll go back to Pete.

PETE: Thanks Nancy and Alvin, now that we've discussed the schema used to organize and form and store the metadata that we use for search, and we've talked about the process and challenges of obtaining and making available for discovery this metadata I'm going to do a demonstration of the Dataset Catalog. Before we get started on that, I just want to go over a few kind of key points about this. It is a federated search tool, discovery tool. It's a catalog that searches across multiple repositories. It is a catalog though, so it does not house the actual data sets. It makes them available for searching, such as PubMed, which is why we call it the PubMed of datasets. The underlying schema is DATTM, which we hope to promote as a a standard for the transport of this metadata. And during this beta period we will continue to add repositories and the datasets associated with those repositories. So you'll watch it grow, the collection grow through the beta period. And we're really looking for user feedback to drive further development of this tool and inform us of the corpus, the collection that we're building. And this beta period goes through June of 2024.

Speaking of that timeline, I said that we had launched the Dataset Catalog in beta format earlier this year. It was around January 23rd internally that we launched it and then externally in early February. We plan on having this in beta through June and then move into a formalized evaluation period with our leadership who will make an assessment of the sustainability of this product. Depending on how that goes, I see us taking the tool and doing some retooling for about a four-month period and then hopefully putting this back into production early in January of 2025.

So with that, I'm going to move over to a live demonstration of the catalog, the Dataset Catalog, so I'm hoping everyone can see my screen. I'm going to take you on a little tour quickly through the Dataset Catalog. As you can see the look and feel should be very familiar to those who use NLM resources. It's intentional. We're attempting to provide a consistent digital presence for our users. So this should be very similar to PubMed or PMC or the NLM catalog LocatorPlus. We have a simple search bar here in the middle. It allows for term searching, word searching. You can also do some logical searching by amending 2 concepts together using Boolean operators AND, OR, NOT. And then we do have some single phrase searching. If you put that in quotes, you can retrieve just that individual citation. Right now with, as you can see, a little over 880,000 data sets. That's not super useful, but as we grow this corpus, like I previously said we're going to throughout this beta, I think that'll become much more usable.

Here, each one of these tiles will take you to another resource page. You can read more about the Dataset Catalog to include the governance of it, how repositories were added, why they were added and where we're going, how we handle delete bins and any other governance of the of the tool. Here's a simple, lightweight user guide. I said this is in beta, so there's limited functionality, but we do provide a user guide to help get people started. Again, we framed this around PMC's User Guide so it'll look very familiar. We have a list of the repositories that are for search right now. Alvin said that we have kind of a bottleneck of repositories, so we're going to start putting-- We have mapped over to a little over 25 other repositories and I'm going to put those up here in the next few days and you can view upcoming repositories that we're going to be adding to the catalog. And then here's an area that you can learn more

about DATMM, includes downloading the schema in RDF. So one of the things that we really want to do with this beta launch is receive user feedback. So we have this very simple little tab here on the side. It's on almost all the pages. You can come in and it provides just a little text box. You can write any comments that you want. If you want to, you can include, but it's not required to include contact information. If you do provide contact information, I will tell you I read every single one of these comments that are made and if you provide your contact information I will get back to you pretty quickly.

So I'm going to do a very simple search, but I am going to use some Boolean operators. You can use layman's terms or you can actually use MeSH terms. We've indexed all of the datasets using MeSH RDF, so we also map over to those terms if you use a layman's term or an entry term for that MeSH term. So I'm doing a search of hypertension and I'm interested in datasets that also are related to the treatment. Again, you'll see a very familiar looking results page. On the left hand side you'll see some facets. This is a beta product, so we've only provided a few options for filtering, but you can filter by repository, time frame that the result set is from (and you can divide that up into decades), and then you can also limit that down to some of the MeSH terms that were used that returned these results. We were dealing with hypertension, but there's also other data sets in here that have other MeSH terms associated with them and that's what you're seeing on this left-hand side. You can click here and see more. You can see we have this set for 10 results on each page. You can also change the sort order by date.

I keep mine on relevance and I'm going to then interrogate, I think this one, this very first one looks interesting so I'll click on here and then I get the full citation for this dataset. So eventually what we would like is for this to be able to export all this information into a tool that would allow for the user to export it into citation management system. Something like Zotero to allow for the citation of datasets within a bibliographic record. We provide the title of the dataset, the entire description, the URL to locate it. Here you'll see those MeSH terms that it was indexed under, the keywords that the user associated with the dataset themselves in the host repository, the host repository that this dataset is stored, as well as the contributors and any associated publications. So if I click on this one, it'll take me right over to PubMed and you'll be able to see the journal article that was derived from this research object. You can also search directly in the Dataset Catalog by these MeSH terms and you can look up the term in the MeSH Browser where you can search by contributors and find other datasets that they have contributed to.

Lastly, I'll take you over to this right-hand column. This is the access information. So here this will take you outside of the dataset catalog so that you can further interrogate the dataset, see if you're really interested and download it. I'll demonstrate that in a minute. We also try to provide any licensing or rights information. If we don't have the licensing or rights information directly from the dataset, we'll provide the rights information that the repository is providing. And here, if I click on this, you'll see that it does indeed take you to the import documentation and there are all the rights statements on what you can do with this dataset. And I'll click on here. This will take me outside of the catalog so that I can actually go here, interrogate the dataset a bit more, and I can download the data set if I want.

Here's also the feedback. So I'm going to push that a couple times. If you click on here, you can provide feedback. These are on almost every page and that really takes me to the end of my tour of the Dataset Catalog. I'm going to take it back over to Mike and see if we have any questions.

MIKE: Thanks, Peter. We do have actually quite a few questions and we're now going to spend the rest of this session answering as many of these questions as possible. Hopefully, we'll get to all of them, but if you have questions, please feel free to keep them coming in the chat. We also have a couple of questions that were submitted ahead of time, so we'll try to sort of filter those in as well. Again, just put your questions in the chat, make sure you send them to Everyone so that we can all see them and we can make

sure that we get answers to your questions verbally. I'm going to start off with a question, it's probably for Alvin from Kimberly. **Does the generative AI and the Dataset Catalog support used to develop and test large language models?**

ALVIN: No, what we're doing right now is making sure that we're working in a safe, secure environment. So actually the large language model is not informed on the information that we upload to it. We're also only using publicly available information from these repositories themselves and really the goal of my GenAI project is for me to just get kind of programming skills that I don't currently have to be able to download datasets from these APIs and then transform them to the DATMM schema.

MIKE: Great, thank you. And I guess this one from Pam, Peter, this one might be for you. **Are there any plans to include public health datasets?**

PETER: Yes. As we grow the collection itself, our collection will align with the NLM's overall collection management policy. As we're in a beta, that's why you're only seeing the four repositories and they were mostly chosen for location and the structure of their metadata, if that makes sense because we're trying to test DATMM itself within the Dataset Catalog. As we grow this corpus, yes, we will make sure that we are going into all areas of the NLM's collection areas. I hope that was clear enough.

MIKE: Absolutely. And I think we might have some other collection related questions as we go forward. But if folks need more clarification, please again just ask for a follow up in the chat. Another one from Kimberly. Again Peter, this is probably for you. **How are you addressing copyright licensing and contract issues? I know you touched on that a little bit, but if there's more that you can say about that.**

PETER: Yes. So when we go through, Alvin spoke at a high level about the inclusion criteria that we go through for each one of the repositories. That is part of the inclusion criteria to make sure that the data is presented in a FAIR manner (findable, accessible, interoperable and reusable), but that it's also open that there are no embargoes. We do not include embargoed datasets and then the copyright in the licensing. As you heard at the end we attempt to pull that out from the data repository. The data needs to be open, but there are sometimes licensing issues or copyright, any of that type of information. We intend to keep that over in that access area to inform the user before they go to the repository as much as we can about what to expect about being able to reuse that data.

MIKE: Excellent. All right, here's a question, actually a couple of people were asking about this Marcus and I think also somebody else who's named I've lost, but Marcus asked it first so I'll give Marcus the credit, probably for you Alvin, **will you be able to search programmatically with E-utilities or I guess any other API-based programmatic searching ability?**

ALVIN: Yeah, that's a great question. Pete, I don't know if you want to talk about that because this is kind of a future thing that's not currently part of our beta launch.

PETER: Yeah. So as part of the tool in the future, it's on the on the development path, to provide API access - an API, this is a triple store so we would have like a SPARQL endpoint that we're developing but that would go both ways. So we want to see programmatic access to be able to download our metadata collection as well as allow for users to be able to point to where their datasets are being housed. And this aligns with the NIH Data Management Policy to make publicly available where grantees have placed their data. So that would be future development that we would look for in 2025.

MIKE: Yeah, I know that that's often a popular request, but it's not usually the first thing that we deal with. We got to get things working properly first on the on the ground floor. All right. It looks like we

have a couple of questions here about selection or inclusion of certain things. So I'll go through them. I think Peter most of these are going to be for you, but if others want to jump in please feel free. From Angela, **will this include other government sources such as HCUP, which I believe is the Healthcare Cost and Utilization Project I may be wrong about that, or the Value Set Authority Center?**

PETER: That's an interesting question and I specifically like the last part about the Value Set Authority Center. The scope of the Dataset Catalog is original research objects. So we do not we do not have within our collection criteria to collect knowledge bases or databases such as MeSH itself or even the Value Set Authority Center. We're really looking for the original research data from data repositories such as the ones that you see or some of the generalist ones that we work with figshare, Mendeley. There's a whole bunch of them. But that's really where we're focused in our collection right now.

MIKE: On a similar collection related question and, again this might be for you Alvin or possibly Peter, **do you do any selection of specific datasets within an ingested repository? For instance, do you ingest all dataset records from Dryad or only a subset of biomedical ones?**

ALVIN: That's a great question. So some repositories like dbGaP are fully medical in scope and so there's no inclusion criteria that we run, we just take everything. But for instance, Dryad and Harvard Dataverse are both generalist repositories. So we have a multi-phase inclusion criteria to only get the biomedical datasets. So the first inclusion criteria is does the dataset have a related publication in Pub Med. The second is to see if it has funding from NIH. The third is to see if any author supplied keyword matches to a UMLS term. If it meets any of those criteria then it's included.

PETER: And I would just like to add on to that. That is one of kind of the key differences between-- you know, the Dataset Catalog is a catalog first of all. So it's a finding tool for these datasets, it does not house the actual datasets. But also when you're thinking of like a generalist repository that allows search across many different such as Datasite or Dryad or figshare, the ones that Alvin had mentioned, we're actually parsing down to just provide access to biomedical research, not the entire scale of scientific research or even just data. So that's a key difference there too.

MIKE: And I think you also sort of answered Matt's question about inclusion/exclusion criteria for moving forward. But Matt if you have further follow up on that, please feel free to ask it. Question for Anil, and we might need to get a little bit more specific on this we'll see what we can do about it, **asking whether participation in the dataset project is open to the non-NIH community. So I'm not sure exactly what the question is there, but is this either from a from a dataset perspective or from a user perspective, open beyond NIH?**

ALVIN: If the question is about at the repository level, whether it's funded by NIH or not, we are focusing on NIH ones first, but we are open to repositories that don't have affiliation with NIH that have biomedical datasets.

MIKE: And from the user perspective, it's obviously open to anyone.

PETER: That's correct.

MIKE: All right. Well, speaking of funding, Joe was asking **can you filter or search for datasets that are funded or administered by a specific IC, specific NIH and Institute and Center?**

PETER: Not currently. That is not part of the beta launch though grant administrators were one of our use cases, a persona that we developed the tool for. So that functionality we will be leaning into very hard when we put this into production because that is one of the key uses that we see for the for the Dataset

Catalog. We just couldn't get there to get this tool actually out and start getting it into the public hands this year. But that is one of our highest level development projects for putting this into production.

MIKE: Gotcha. And there's a related question which you may have sort of already answered in in terms of just like where we are in the process about the fact that **including funding information in dbGaP studies is notoriously spotty, are you linking those sets to funding in another way?**

PETER: So funding I think could probably be an entire webinar in and of itself and I think we'd have to bring the PubMed folks in here too to talk about it. But it is a completely unstructured field in almost every repository that we go to. So we spend quite a bit of time trying to parse out, but there is not a definitive or standardized list. That's why having the organization that funds is very important for it to be able to specify itself and uniquely identify itself or even the users when they can provide unique identifiers such as ORCIDs. So we struggle with grant and grant numbers just like any other repository. We do some massaging. You know I said we index with MeSH terms. Alvin spoke about how we invert names. So we do do some massaging of the data, but we are trying to do this at scale at the same time, so there's only there's a limited amount of time that we can go clean up bad metadata in repositories.

MIKE: That's a perfect segue actually to Isaac's question in terms of labor and how much we can do. **Is the indexing of datasets with MeSH terms labor intensive or are we able to automate that process?** Alvin, that might be something for you to answer.

ALVIN: Yeah. So that process is already automated where we use author supplied keywords and try to get a one-to-one match with a MeSH term. It's automated, but it took time for the programmer to make that automation happen. It also takes time. So every time you ingest a new repository or you go get an update from a repository, what takes most of the time is doing that mapping from author supply keyword to MeSH. So it was labor intensive to create the process and it's time intensive to continue it.

MIKE: Oh wait, we have a comment from Lisa going back to our funding discussion. Funding PubMed overlap is something that we're talking about at our institution to be able to track datasets over time from the beginning of a data management plan creation to ingestion of research into PubMed. This is a larger constant conversation on an institutional level, just FYI. So that's good to know and it puts a good context on sort of how thorny this issue can be.

Let's see what else we got here. I know I missed some, so I'm going to try to scroll up here from Christy and there it is for Peter. **Are there ways to import these datasets into a computing cluster, say at a university?** And I suppose your mileage may vary depending on the circumstances.

PETER: Yeah. With the beta product right now, we don't have that programmatic hook, that export. So it would just be very labor intensive. And remember this is a catalog. So these are metadata records that they will only ever take you to the location of the dataset itself. So if you're actually looking to programmatically download these datasets, that's not the intention of the of the catalog.

MIKE: All right, yeah, let's see what else we got here. I know again, I said I missed some. Oh, here again, going back to the funding question from Sun Young, **what PID do you use for funding agency, ROR or Funder Registry ID?** Alvin, I think you might have an answer to this with somebody who's been definitely more in the weeds than some of the rest of us.

ALVIN: No, it's a great question. So identifier for funders is not in our model. I would say that probably at this point, I've reviewed metadata for maybe 40 or 50 repositories, and I've only seen one that actually even includes the ROR identifier for a funder. What we typically see is just a raw string of a funder's name, and if you're lucky, also a funding ID for the actual grant. And we do record both of those.

MIKE: All right. OK. So we have **a couple of questions that I'm not sure whether we'll be able to answer right now, but I'll throw it open to our panel. They relate to GREI to how datum DATMM aligns with the GREI metadata recommendations. And then a further follow up question, GREI metadata recommendations are based on Datasite metadata schema 4.4. Does the NLM Dataset Catalog team work with major players in metadata schema to incorporate DATMM as the minimal core data to be included in the more widely used metadata schema such as datasite, Dublin Core, or CEDAR Workbench?**

I know that was a mouthful but for those of you who are involved in this on the sort of more model policy level, any thoughts on that?

NANCY: So we looked at some other models as we were creating this model, as we were creating the DATMM model but we do not incorporate any other models in ours. We kind of look at this as this is a really, really lightweight model. And so if you are creating a metadata model, and again keeping in mind that this is a linked data model, if you are creating a linked data model, you can incorporate this one into your data model. But we are not looking, at least at this time to incorporate our model with other models.

ALVIN: So the great metadata recommendations that one user asked about and another user kind of mentioned, those are really just properties from Datasite's model. So Datasite has asked us to map our schema to their schema, which Nancy and I did. We provided that to them about a year and a half ago. When I look at the metadata recommendations for GREI, it's really just those important properties from Datasite's model that we also have an equivalent for. So when I say important ones, I mean not just recording a contributor's name, but having the ability to record their ORCID if it exists in the metadata. Not only having a subject, but having the possibility to put in a URI for the subject from an ontology. So even though some of our property names are different, we're achieving the same things.

PETER: Yeah, and I just wanted to add on to Nancy saying, you know, we did investigate many schema, and if you look at the DATMM schema itself, the model, it incorporates classes and elements across other ontologies. So we're pulling in elements from DCAT to inform us of certain concepts and the majority of our classes are pulled from other ontologies as well. So we're using a standardized kind of way of describing these individual specific classes and elements of our model.

MIKE: All right, well we're just about out of time. I'm just going to try to hit two sort of future looking questions real quick. So panelists, keep that in mind in your responses. **Is there a target for frequency of updates to the catalogue? If so, how often will the datasets in the catalogue be re-evaluated to ensure they still meet the criteria for inclusion?**

PETER: So we're looking at, during this beta, we're updating the data weekly or every two weeks and then the frequency for into production obviously is still going to be derived from this beta but it's often driven by the host repository itself and how often it refreshes its data.

MIKE: Fair enough. And last one, Leslie asks, **can you give some examples of data repositories that might be included in the future?** Alvin, you want to take a quick crack at this?

ALVIN: Yeah, we are about to have a kickoff meeting to start working on figshare. As we're working on figshare, which is a massive repository with over 1.2 million datasets, There's also other dataverses from institutions other than Harvard that we're going to be able to add as well.

PETER: TCIA too.

ALVIN: Yeah. The Cancer Imaging Archive is one we’re really excited about. And we've been working with some of the people who work with that repository.

MIKE: Excellent. All right. Well, we are just about out of time. I think we got to most of the questions. But if we didn't remember, there's that feedback tab on the side of the page, so on the side of the data set catalog pages. So go ahead and click on that. You can put your question in there. Just make sure you include your e-mail address so that we can get back to you. And I think that that should just about do it.