KATE: Hello and welcome to NLM Office Hours. If this is your first time, NLM Office Hours is a chance to learn more about NLM products and services directly from those who provide them, and to get your questions answered. NLM Office Hours is co-hosted by the Network of the National Library of Medicine, the education and outreach arm of the NLM. To provide broader and continuing access to the information, these sessions are recorded and posted on NLM's website.

This month's Office Hours we have David Anderson and Miranda Jarnot from the Unified Medical Language System (UMLS) team, joining myself (Kate Majewski), Mike Davidson, Michael Tahmasian, and Brittney Davis, all from the User Engagement Program here at the NLM. David and Miranda will be providing a brief presentation today before they address your questions. Please input your questions at any time in the chat panel and please be sure to use the blue drop-down to be sure that your message is directed to everyone. All right, so with that, I think we can get going. David to you.

DAVID: Thank you, Kate. Let me just make sure that I've got my audio going. OK, good.

Good afternoon. My name is David Anderson and I am the Product Owner for the Unified Medical Language System (UMLS), a product of the National Library of Medicine. And today we're going to be speaking to you specifically about the UMLS Metathesaurus. We'd like you to leave here today with some idea how to get started with the UMLS Metathesaurus. At first glance, I think the UMLS seems like a very complicated thing, and I really want to dispel that notion because at the end of the day, it's really actually pretty simple. We are grouping names from different terminologies into concepts or units of meaning and providing some relationships between those concepts. The UMLS Metathesaurus does other things, but at its core, that's where most of the value of the Metathesaurus lies.

So we just want to provide you with some information to get you started. I'm going to talk about the Metathesaurus data and my colleague Miranda Jarnot will talk about how we edit UMLS Metathesaurus, with hopefully plenty of time to answer your questions. So please do provide your questions in the chat and we'll be happy to answer them.

So what is the UMLS Metathesaurus? It is a set of files that brings together many biomedical terminologies and standards to enable interoperability between computer systems. The UMLS brings some amount of order to the very complex body of medical language. And by medical language, I mean names of drugs, disorders, procedures, lab tests, and other biomedical concepts used in healthcare systems and research systems. UMLS takes biomedical terminologies and integrates them into one system. National Library of Medicine has been producing this system since 1990, when we first released the UMLS to the public on CD-ROM.

We bring a variety of biomedical terminologies into the UMLS. Some of the logos are shown here just to give you a sense, and these terminologies are used in both production systems, like the ones used by your doctor's office or hospital, but they're also heavily used in research. And one of our goals here is to promote interoperability between the clinical healthcare world and the research world, and terminologies help to do that. UMLS is a useful piece of that puzzle.

So who uses the UMLS? Researchers are primary users of the UMLS. Researchers are often looking for patterns in healthcare data, or they work on producing better methods for finding patterns in healthcare data, and they use the UMLS as a knowledge base to accomplish this. Health application developers also use the UMLS, and these include developers of large electronic health record systems used in hospitals, as well as developers of research databases, terminology services, and other specialized applications in the healthcare space. These applications might support clinical care, provide access to electronic health record data for research purposes, enable searching of the biomedical literature or other records, or just target specific problems in the healthcare space. Health service providers include hospitals and other healthcare providers, and they are typical users of the applications I just mentioned. So they may rely on UMLS for displaying terms in their system or performing complex queries, improving clinical documentation, automated coding suggestions, clinical decision support, and reporting information to registries and regulatory agencies. And also educators use the UMLS as a teaching tool for natural language processing courses and medical terminology courses. Not everyone falls into these categories, but they represent most of the users of the UMLS for using the Metathesaurus as a knowledge base for application development or conducting research.

So let's take a broad look at the UMLS data. The UMLS Metathesaurus is fairly large, 28 gigabytes. And so I want you to kind of picture in your mind a list of 16 million names. And these names come from 187 different source terminologies in 27 different languages. And the names are grouped into 3.3 million concepts or units of meaning. And in addition to the names, there are also 60 million relationships, both hierarchical and non-hierarchical. There are also definitions and other attributes as well as other information.

So if you want to learn the UMLS, here are some steps that you can take. This is the progression that I would recommend. So first sign up for a free UMLS license and account if you have not already done so. This will get you access to the UMLS Metasource browser downloads and our API. But the first thing to do, I think is really to explore the UMLS Meta Thesaurus Browser. This will give you a nice introduction to the data that you will find in the UMLS Metathesaurus files if you choose to download them, or the API if you choose to use that. So I encourage you to explore this, try some searches, click around and see what you can find. We'll take a little look at this in a live demo here.

And so this is just a search interface. And what this is going to search is it's going to search the UMLS concepts. And so we can try a search, I'm going to try Addison disease and we get a list of results, which is a set of concepts that have the words Addison disease in them. And so I can click one and it's going to give me some basic information about this UMLS concept for Addison disease, including a unique identifier. The concept unique identifier, also known as a CUI. A semantic type, which is a broad class that we assign, we assign at least one semantic type to every concept in the UMLS. You also find definitions from the different source vocabularies in the UMLS, as well as a list of names that we also call Atoms. And it's a long list of names, different ways to refer to this particular concept. You can also browse broader concepts or narrower concepts. And so this is a good way to sort of click around the UMLS and different concepts and see what's out there. You can also take a look at specific codes from specific vocabularies or terminologies in the UMLS. If I want to filter this list of names by SNOMED names, I can click on the SNOMED code and find more information about this particular code from SNOMED, including its position in the hierarchy, some attributes, and some relationships that SNOMED asserts about this particular concept. For example, Addison's disease is a cause of Addison melanoderma, etcetera.So that's just a brief look at UMLS Metathesaurus Browser. I encourage you to explore this further if you like.

But if you after exploring it are interested in thinking about, well, I'd like to actually see and work with the data that is behind this. I recommend that you go ahead and download it and take a look at it. And so you can click Download UMLS, and there's several options on this page, but the ones you want to pay attention to are UMLS Metathesaurus Full Subset. This will give you all of the files that are part of the Metathesaurus downloaded and ready to go in a .zip file. Or also you could just look at the MRCONSO file, which is the most widely used UMLS file. It's the one that includes the names, codes, and concepts from over 180 source vocabularies. So either of those are good options to sort of get started with the data.

So a little bit more about those files. All these files that you get when you download the Metathesaurus are pipe-delimited tables. So meaning each field in the table is delimited by a pipe character ( | ) That's the one that's probably next to the brackets on your keyboard. And while the data is in tabular format, these are tables, the files are usually too large to load into Excel, so as convenient as that would be, Excel only accept something like a million rows, notice these files we're talking about today are much larger than that. But these files can be loaded into relational databases like MYSQL, Oracle, sqlite3, etc. And we do offer some load scripts for assisting with that. And if you want to explore the files without loading into a database, it can be a good idea to get comfortable working in the command line and using tools like Python to examine the files.

So when you download the UMLS Meta Thesaurus, you're going to see this this list of files. And it's hard to know kind of where to start, but there are a couple of files that I think are worth focusing on. First is the MRCONSO file. And this is the most important file in the Metathesaurus because it contains all of those 16,000,000 names that I mentioned earlier and groups them into UMLS concepts using Concept Unique Identifier, also known as CUIs. It also has codes or identifiers for those names and the source terminology for the names. Most of the work that we actually do on the UMLS Metathesaurus involves editing this file and all the other files kind of revolve around this file.

So I just want to give you an idea what the MRCONSO data looks like in a very simplified form. So let's look at a specific concept for Addison disease. And as I mentioned, the UMLS groups biomedical names into units of meaning or concepts. And here's some of the names used to refer to Addison disease. And you can see they're the kind of variation that exists. Each of these names comes from a UMLS source terminology, and each name has a code assigned by its source terminology. These are identifiers that are used in real world systems. For example, you'll find SNOMED CT codes in clinical systems or MeSH descriptor code in PubMed. Each unique string from a unique source with a unique code is called an atom, and each gets a unique identifier. And finally, all of these names are atoms are considered synonyms in the UMLS and they're given the same concept unique identifier (CUI). And this is how we link names and codes from different terminologies together in the UMLS.

So what can this be used for? Well, as I just said, linking names, codes and concepts together. And some examples of that are if you want to crosswalk from one terminology to another, or maybe you want to search for a string and return a set of concepts or codes, or identify synonyms for particular string. Another very common use is identifying meaning in text. So an example of this is annotating clinical notes or other documents with standard identifiers to find co-occurrences of concepts in those documents. And that can tell you that there's a relationship between those concepts or those concepts are similar.

So the other file I want to look at is the MRREL file. This is our relationships file. MRREL aggregates all of the relationships asserted in the UMLS source terminologies into one file. We're essentially taking those assertions and passing them through to our users by representing them in a common format. So we create a few of the relationships in this file ourselves. The vast majority are simply present in the source terminologies that make up the Metathesaurus. And some of these relationships are hierarchical, so broader or narrower relationships or parent-child relationships, like in a taxonomy, some relationships are non-hierarchical. And I'll show you some specific examples in a moment.

So this is a piece of a hierarchy from SNOMED CT, which is one of the UMLS source terminologies. Each of the relationships between broader and narrower concepts here are represented as rows in the MRL file. Each node in this hierarchy is related to the one above and below it, and MRREL provides a standard representation regardless of the source of the hierarchy. So one advantage of that is if we have a concept that like hyperglycemia, we can find a set of broader concepts as asserted by the various source terminologies in the UMLS. And there are many different ways to place a concept within a hierarchy. And the Metathesaurus aggregates all of the hierarchies in one file, the MRREL file, so that you can see all the hierarchical information in one place. Likewise, you can find the narrower concepts for a given concept, and these are narrower concepts according to SNOMED CT, MeSH, Human Phenotype Ontology, NCI Thesaurus, depending on the structure of the hierarchies in those specific terminologies.

In addition to hierarchical relationships, we also have non-hierarchical relationships and these are specific assertions like has\_finding\_site or is\_finding\_of\_disease, may\_treat, has\_manifestation, or related\_to and these can be very useful. It can be worth noting that the Metathesaurus generally does not even come close to capturing all of the possible relationships that are relevant in medicine. So you should not expect to come in and get a comprehensive list, for instance, of, say, relationships between drugs and disease, because we only represent the relationships that are present in the source terminologies. And so it's good to think of the UMLS Metathesaurus as a starting point for making these connections, but it may not provide every answer, every kind of relationship that you want.

So again, these are the two files to look at, especially MRCONSO. So if you want to learn more about the files available in the Metathesaurus, I would recommend checking out the UMLS Reference Manual, which is a comprehensive guide to the UMLS and has some detailed descriptions to those files. And here's some additional links that you can also check out.

So that ends my portion. And now I'm going to pass it over to Miranda, who is going to talk about editing the UMLS.

MIRANDA: Thanks, David. Let me start my screen share. OK, that should be it.

Good afternoon. I'm Miranda Jarnot and I'm the editing lead for the UMLS. So I'm going to tell you a little bit about how the UMLS is put together. My role is to acquire our source vocabularies and then shepherd them through the process of insertion and integration into the UMLS. In addition to performing source editing, I work with our contract editors to manage the editing workload, resolve editing and source data questions, answer user questions, and train new editors. So let's take a look at what goes on inside the Metathesaurus.

What is the UMLS? Well, as David told you briefly, it's a database of information containing names and codes from many different controlled biomedical vocabularies. It has many uses including linking health information and codes across computer systems, aiding and developing and enhancing electronic health records and applications, data mining and search engine retrieval, just to name a few. So this is just a brief overview of the UMLS and what I'll talk about is how the UMLS is constructed, how source vocabularies are added and integrated into the UMLS.

Two important things to note about the UMLS. First of all, its scope is defined by the combined scope of its source vocabularies. There is no overall scope of the UMLS itself. Second is it’s concept centric. This means that synonymous terms are grouped together into concepts, each concept with its own distinct meaning, and the goal of the UMLS is to establish synonymy between source vocabularies. In creating this concept centered approach, we preserve specific information from the source vocabulary, such as definitions, hierarchies and relationships asserted by the source. We only add new information that will allow the source vocabulary to fit into the UMLS data model. This new information includes semantic types, relationships between concepts from different source vocabularies, and specific attributes that enrich the source vocabulary's representation.

How does source data get into the UMLS? There are five steps in the process of adding a source to the UMLS. First is source acquisition, followed by pre-inversion, and inversion. Then there's test insertion and real insertion, and we'll look at each of these steps in the following slides.

The first step is source file acquisition. Source updates vary, so not all sources are updated for each release. Some sources update daily, some yearly, and some very seldom. The method by which we acquire source files varies as well. In some cases, the files are emailed to us from source contacts. For internal sources, such as RX Norm, we get a notice when the files are ready and download them from internal NLM sites. Other sources require us to check for file updates on the source’s website. Once source files have been acquired, they undergo a process called pre-inversion, a series of steps which prepare the source files for the inversion process.

The third step of adding a source is inversion. During inversion, the source data is processed into a common format that can then be loaded into our editing system. The inversion process is specific for each source because no two source files have exactly the same format. Some examples of common formats include XML, Excel, and flat files such as OBO. We also can use PDF and Excel files.

After inversion, the source is inserted into a test database that's loaded with a snapshot of the current production environment. We review this test insertion to make sure the new source items with new source terms integrate correctly into the UMLS. The main thing we pay attention to is synonymy. As we review, we ask, are newly merged terms in the correct concepts with other terms that mean the same thing? Are there newly created concepts that should have merged with existing concepts but didn't? We then check other data. Are semantic types present and correct? Are the context hierarchies, attributes and definitions present and correct? We also consider how much editing and QA editing will be required. For example, are there concepts with ambiguous terms that will have to be disambiguated? Are there concepts where the system can't decide where to put a term and a conflict results? And if the source has many terms that will need review, are there patterns that we can use to approve these terms without having to look at each one individually? For example, if NCBI loads and there are 23,000 new concepts for bacteria that don't merge with any other sources or other concepts, we may be able to approve them without editing. Once the test insertion looks good, the source is inserted into the production database. This is called real insertion. The source is reviewed again to make sure all steps are complete, and the source is integrated properly into the UMLS. At this point, QA editing and regular editing can begin.

Editing sources in the UMLS involves one main principle: respect the source. This means that we rely on the source to tell us what terms mean so that each source's data can be represented as accurately as possible. Although we may add new data, such as semantic types and relationships, we don't create meanings in the UMLS. We rely on what the source tells us. The UMLS only takes a position about meanings when sources disagree about synonymy. Sometimes the meanings are not clear, or it becomes a judgement call. And in these cases, we can go back to the source to ask for clarification or discuss among ourselves to devise the best solution. In addition to basic editing principles, each source has its own rules for editing. Editors have comprehensive resources and guidelines for source editing that we refer to often during editing.

As I mentioned before, UMLS is concept centric, and editing is all about maintaining synonymy within concepts. So let's take a look at the structure of a UMLS concept. First, again to reiterate concepts are the central component of the UMLS. Each concept has its own distinct meaning and includes terms that represent that meaning. Concepts are created during the insertion of the source, when there is a term that does not currently exist in the UMLS. They're also created during editing, when terms are split from an existing concept to make a new one. Each concept has a unique identifier, the CUI (concept unique identifier). It's assigned during production and remains with the concept through release.

Another continuation of concepts in the Metathesaurus. This is an overview of the main parts of the concept. Atoms are the unique terms, as David mentioned, that are found in the source vocabulary. All the atoms in a concept should have the same meaning. Contexts are the hierarchical representations that show relationships within the data. They're provided by the source, but not all sources have them. They're very helpful for editors in determining meaning. And, an important point, there is no overall Metathesaurus context. Semantic types are another part of the concept. At least one is required for every concept, except chemical concepts generally have more than one to indicate both structure and function of the chemical attributes are extra information that's sometimes included. These can be very helpful also for editors in making editing decisions where the meaning of a term may not be clear. These can include definitions, mappings, lab and trade names for chemicals, just to name a few.

Relationships (RELs) are another fundamental part of the concepts of in the UMLS. They can either come from source vocabularies or be created during editing to link related concepts or disambiguate potentially confusing concepts. There are two types. Source level RELs, these are assigned by the source and cannot be deleted or changed by editors. There are also Concept level RELs which are created by editors during editing. Once these are created, they can't be deleted by editors, but they can be changed from one type to another. So relationship types relationships are either: related to, broader than, narrower than, or not related. This slide shows you a few examples of each type. One thing to note is that Not related RELs are not released. They primarily serve to prevent editors from having to repeatedly review concepts that have similar strings but don't mean the same thing. The presence of an XR relationship doesn't necessarily mean that two concepts are not related, but they should not be merged.

As I said before, the most important job of an editor is to make sure that all terms in a concept are synonymous. When a source is inserted into the UMLS, synonymy is first determined by the source data, such as source codes. Algorithms and merge functions are then run to find synonymy between terms and concepts so that new terms end up in the right place. Editors then review to make sure asserted synonymy is correct. For editors to determine synonymy, they must follow the prime directive of respecting synonymy asserted by a source. In terms of synonymy, in the UMLS, lexical variants are generally synonymous, as shown in this slide. These include singular and plural forms such as feet and foot or apple and apples, direct and indirect forms such as neoplasms, breast and breast neoplasms, and punctuation variants such as insulin dash like receptor and insulin like receptor.

But sometimes lexical variants are not synonymous. Home nursing and nursing home, mushroom poison and poison mushrooms. Editors use clues from the source to determine synonymy, including contexts, definitions, and other attributes. The editor's own knowledge is also valuable, and when needed, editors do research to find meanings and determine synonymy. But it's important to remember that there are no foolproof rules to apply in determining synonymy, and This is why having human editors is critical.

At the end of March for the AA release and September for the AB release of the UMLS, the production database is locked for editing or frozen. About a week prior to freeze, regular editing stops and the editing leads perform QA to get numbers of concepts remaining in bins to 0 or as near 0 as possible so the data will be as clean as possible. Once QA is finished, the production database is closed for production of release files. Editing remains down for two to three days while the files are prepared. Once this is done, the database comes back up and production production begins on the next release. The whole cycle starts over. We start editing sources for the new release cycle while the current files are being prepared for release to the public. Once the release files are completed and QA, they're released to the public. Currently, we're in the 2024 AB release cycle and are scheduled to freeze on Wednesday, September 25th, 2024. So that is how the UMLS is put together.

Thanks for listening and I hope this has given you a deeper understanding of UMLS source acquisition, insertion and editing.

KATE: And thank you so much, both David and Miranda for those informative presentations. And it's time now to turn to questions. And again, please enter your questions into chat and please direct them to everyone. If you send questions directly to our presenters, they may not see them. So please send your questions to everyone so that we can verbally repeat them for our presenters. And just to note that I may not be getting to the questions in the order they were asked in order to answer those questions we think that are of use to the most people. But I hope we can get to all of your questions today. So I'm going to start with a couple of questions for David to give Miranda a chance to catch her breath.

So we're going to start with this question from Jeffrey. **David, could you speak to the UMLS license, what it is and what it allows?**

DAVID: Sure. Yeah. So the reason why we have a license is because we have many source terminologies in the UMLS. Some of them are freely available, you could just download them from their website. Others are actually proprietary and so require a license of their own. So what we've basically done is negotiated the rights to distribute those terminologies along with all of our other terminologies, but under the condition that we have users agree to this license. So there are different license restrictions that you can take a look at if you actually look at our license. And some of the more restricted terminologies, basically you're restricted to internal use at your site for research, product development and statistical analysis only. So what we're doing with the license is we really want to provide access to this data so that researchers can use it and so the license is a way to just have kind of a one step, you know, a single step that everyone takes to get access to the UMLS and also provide access to these proprietary terminologies.

KATE: Excellent. Thanks, David. Another question for you, a question from Henry. **Are there any plans to provide LLM training data or pre-trained models to support the UMLS?**

DAVID: That's a great question. I would say, LLM specifically is a fairly new technology and so we have been doing some investigation into different things that we can do with LLMs. As far as actually providing a pre-trained model, I believe that at present we don't have any plans to do that, but I do know that there are a number of examples in the biomedical literature that you can check out that integrate the UMLS in various ways to either produce a pre-trained model or I think, you know, some folks have actually been using LLMs to do various things in conjunction with the UMLS. And internally we've also looked, you know, taking initial steps to look at how can LLMs help to actually identify synonymy and help us to build the UMLS. So I think that's sort of early stages at this point, but there's some interesting stuff going on and I would urge you to check the biomedical literature.

KATE: Thanks. We have a couple of questions about the relationship between UMLS and a couple of other tools and maybe you could address those questions. One, I think actually Miranda answered this question, but just to maybe reiterate, **what is the relationship between the Rx Norm and UMLS?**

DAVID: Sure. So UMLS came first and then RxNorm came along later. And RxNorm is a terminology for drug names. So normalized drug names. And RxNorm actually adopted the same data model as the UMLS. So if you're looking at the files for UMLS and RxNorm, they basically use the same data model in terms of how we represent names and also relationships and all the various components of the data. So RxNorm is also a source terminology for the UMLS. So it's one of those many source terminologies that we bring in to the UMLS. And so we update RxNorm, you know, on a weekly basis, but we bring RxNorminto the UMLS twice a year when we release the UMLS. So that's kind of the relationship between the two.

KATE: Great, thank you. And then this one I'm actually not familiar with at all. So just sort of throwing this out there. **What is the relationship between UMLS and Mondo?**

DAVID: So Mondo is somewhat similar database to the UMLS in that it aggregates ontology content from a variety of sources. It actually does link to the UMLS. We actually don't link back to Mondo, but there are ways to make connections between the two and they can sort of complement each other. So I would say they are very similar resources, but Mondo has a specific focus. UMLS, you know, they both have different focuses, I would say.

KATE: OK, great. Thanks, David. Let me give you a break for a moment. We'll go to a question for Miranda. Miranda, **what tools do you use to ensure that UMLS edits make for better and not worse matches to the automated technologies?**

MIRANDA: That's always the goal. We have, in addition to algorithmic steps that we take as a routine based on, you know, many, many insertions for a given source, we also use bins, we call them bins in our system that we can assign whatever particular attributes you want to them to search for things. For example, split codes. Split codes from a particular source. We can look at pairs where, say, one is in one concept code. The code from a particular source is in more than one concept. So that sort of QA.

We can look at concepts that need relationship relationships based on string. Again, we look at demotions, which are places where the system doesn't know where to put something because merge constraints say it can go here or it can go here and the system doesn't know where to put it. There's a lot of QA that goes on behind the scenes that's algorithmic.

And again, careful looking at things. I was editing this morning and there's a couple places, a couple of different concepts, where strings matched, but the meaning is very different. And that's a place where you need a human editor to go in and really look at the meanings of all the atoms in a particular concept to make sure they mean the same thing. Using the hierarchies, using relationships, using other attributes that might be helpful to tease those meanings out. Hope that answered your question.

KATE: Thanks Miranda. Not sure if this one's for you, but I'm going to give it to you and you can pass it over if I'm wrong. So Balaji's asking, **do all diseases in UMLS have ICD-10 codes?**

MIRANDA: No, there are plenty of disease concepts that don't have ICD-10 codes because ICD-10 doesn't necessarily cover all diseases. And there are also cases where the names that ICD-10 may cover or include synonyms for a particular disease. There may be other synonyms that may be older that aren't wouldn't be in the same concept. Sometimes we need to find those and merge them. But short answer is no, they're not. All of them contain ICD codes.

KATE: OK, thank you. All right, so this one's from Mickey. **Often we find synonyms, sometimes one that has just literature references and no concepts relationships. Should we be reporting those and, if so, how?**

MIRANDA: So I'm looking at that one. And the second part of that question was, **"To clarify my synonyms question. It's about different CUIs that seem like synonyms."** That's something we'd like to look at. And if you find a situation like that or a case like that, please let us know and report it on our user page.

David, do you have the link for that? For user questions? Or I see Jenny is in the in the audience. She can probably put it up faster than either of us. There we go!

DAVID: Write to the NLM Support Center, NLM help desk. And that's a good way to submit those things.

MIRANDA: Yeah. We like to hear things like that from users, so please do let us know.

KATE: Great. Thank you. OK. So a question for David, this was actually asked earlier, but I want to get back to it, from Tareq. **Does UMLS provide solutions for customized requests. And Tareq provides an example, they're focused on technology. So looking at specifically MRI technologies. Can you retrieve a group of technologies? Is that something we provide?**

DAVID: So it really depends on whether that area is represented in one of the UMLS source terminologies. And if it is, for example, you know, if there is a branch of Medical Subject Headings is one of our source terminologies that we create here at NLM, if there's a branch of that that addresses those specific kinds of technologies, UMLS may be a source of information about that, but the only way to really know is to, you know, search for what you're looking for and see if you can find it.

So there are lots of ontologies that have been created for research and we don't have every ontology. Another resource you might look at is BioPortal where there are a number of ontologies that have been created just for research purposes that might cover a particular area like that. So for the UMLS, it really depends on, you know, we are not adding, we're not creating our own content for the UMLS. We are incorporating content from other terminologies and passing it through. So it really depends on whether that content is in one of those terminologies.

KATE: Thank you very much. OK, question from Leila for David. **Could you explain a little bit more about semantic groups?**

DAVID: Sure. So we didn't really get to this, but there are three components of the UMLS. One is the Metathesaurus, another one is the Semantic Network, and another is the SPECIALIST Lexicon and Lexical Tools.

So the Semantic Network is basically a hierarchy of semantic types that we use to classify the UMLS concepts and one of the files along with the semantic network, ff you search for Semantic Network, you will find the semantic group file and the semantic group file does is it kind of groups all of the semantic types into even broader categories. And this can be helpful if you are, you know, say trying to really broadly classify the concepts in the UMLS or really broadly classify in general. We use some of the semantic groups actually as filters in the Metathesaurus Browser. So it can be a useful way to sort of further filter the semantic types and the concepts in the Metathesaurus.

KATE: Excellent, thank you. Question just came in about retrieving using hierarchies. **Ammar would like to know if there are standard queries, like SQL queries, to retrieve the children or parents of a specific concept.** Is that for you David?

DAVID: Sure. So we do have some example SQL queries in our documentation. So if you search for UMLS SQL queries, you will find some examples of those. I'm not sure if there are any that specifically address parent-child or broader-narrower. But yes, I would look at specifically at the queries that address the MRREL file. That's the relationship file that we talked about, and that should at least get you started in sort of figuring out how to do a SQL query for broader-narrower.

KATE: OK, I think we have time for one question if it's a quick answer. So Leila's asking, **can we group STIs into our defined semantic groups?**

DAVID: Yeah. So there's a semantic group file that basically links the semantic group to all of the semantic types that are in the Semantic Network, which is part of the UMLS. And so I think that's the file that you would use. And also if you wanted to extend that file, I think that would be the way to go. You know, there might be additional information that you want to capture, but I think starting with the semantic group file is how you would do that. It's essentially just a table with, in one column the semantic group and the other column the semantic types.

KATE: Fantastic. Thank you so much. It's actually time to wrap up, but I would point everyone to the chat where we've had lots of helpers today providing some great links for you, including the link to the support center. If a question occurred to you and we didn't get to it today, please go ahead and send your questions to our support center and this wonderful team will get back to you with some helpful guidance.

So thank you so much again to David and Miranda from the UMLS team and their helpers in chat for a wonderful presentation today and answering your questions.