Transcript

Thank you all for joining us for another NLM Office Hours. I'm Mike Davidson, from the Office of Engagement and Training here at the National Library of Medicine. My pronouns are he/him/his. The goal behind these 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.

The focus of today's session is on PubMed and we've got a great roster of folks with us who know a lot about PubMed. We're going to kick things off with a brief presentation from Susan Schmidt (who you can see on video right now) from NLM's Index Section, who will share some information about our recent transition to automated MEDLINE indexing. We're then going to use the rest of the session to have our panelists answer your questions.

So in addition to Susan, our panel of NLM experts includes:

- Jim Mork from our Lister Hill National Center for Biomedical Communications.
- We have Kathi Canese, Jessica Chan, and Amanda Sawyer from the team responsible for developing and maintaining PubMed over at the National Center for Biotechnology Information (NCBI).
- And we also have my colleagues Kate Majewski and Catherine Staley from the Office of Engagement and Training.

Between all of these incredibly knowledgeable folks, we should hopefully be able to answer any PubMed questions that you might have.

But before we get started, a couple quick logistical notes: We will be recording today's session - as I just mentioned, those of you who came in a little bit later might have missed that. We will be recording today's session to share with those who are unable to attend and so that you can review it later at your leisure. The recording will be posted along with the slides, shortly following the Office Hours. We've muted all attendees on entry, but if you have any questions about PubMed, about the presentation, about anything, you can put them in chat at any time. Please make sure you send those questions in chat to "Everyone," so that we can make sure our whole panel can see them and answer them. I and my colleagues will be taking note of the questions that come in during the presentation and then during the Q&A portion, we'll direct them to the correct person.

But before we start addressing your questions, I'm going to hand things off to Susan Schmidt to bring you up to speed on what's new with automated indexing. Susan.

>>Thank you, Mike.

Like he said, my name is Susan Schmidt. I'm a technical information specialist in the Index Section where I've worked for about 16 years. I have been very involved in the transition to automated indexing, including overseeing the logistics of journal transitioning to automation,

contributing to algorithm refinement, and developing the curation process. All of these are really a team effort so it's not just me doing all that.

I look forward to taking your questions in a little bit. I'm going to actually turn my video off while I go through these slides to minimize distractions. I've asked my dog to stay on mute - hopefully that goes well.

My agenda is to give you (in about 10 minutes) the who, what, when, where and why on automated indexing from MEDLINE. But not actually in that order.

I want to begin with the why and move on from there. Specifically I want to briefly address the question of why the change to automated indexing? When was indexing automated? What algorithm is used for automated indexing? Who reviews automated indexing? And where is algorithm development going? And where can you find more information and report problems?

Beginning with the why the change to automated indexing?

The problem was that manual indexing no longer scaled with the volume of literature publication. With more than one million articles published a year in the approximately 5,300 journals indexed for MEDLINE, our backlog of articles in process for indexing had grown to approximately 600,000 articles as of early 2021. This was despite significant resources being devoted to manual indexing. This resulted in an extended and unacceptable time to index of many months and the recurrent question from users of "When will my article be indexed?"

The solution was the focus of the MEDLINE 2022 Initiative. An initiative that was by a steering committee composed of stakeholders from across NLM. And the solution involved leveraging a well performing algorithm for automated indexing: the medical text indexer - the current version of which is called MTIAuto (MTIA) and I'll talk about that further in a few minutes. And importantly the solution involved leveraging our indexers with domain expertise to provide human quality assurance of automation results.

I'll note that with the transition to automation the backlog of citations in process (represented here in blue) was effectively eliminated and the time to index was reduced from many months to one or two days. The orange line represents the average number of days that articles waited after going through the data reviewing process to be assigned for indexing, which in January 2021 was more than 100 days. As I said, articles are now generally indexed within one or two days from the time that the citation for the published article is loaded to PubMed.

When was indexing automated?

Well, the short answer to this is all MEDLINE journals have been transitioned by early April of this year, but this wasn't something that happened overnight. There was a long progression toward automated indexing. Indexers had MTI indexing suggestions available to them starting two decades ago.

We began experimenting with MTI as the first line indexer (or, MTIFL) with humans curating the indexing in 2011 and moved some journals to that approach over time. We processed comments with a version of MTI beginning in 2017. We used a modification of MTI First Line, called MTI Review, beginning in 2018, that included automated indexing of publication types and that served as a staging ground for transitioning journals to automation. We began transitioning journals to full automation with MTIA in early 2020.

To speak a bit more about the current algorithm used for automated indexing, MTIA, this schematic shows the data input and processing steps involved in the MTIA algorithm. There is a lot of detail here and I share this to give you a sense of the complexity involved. But let me walk you through this. In terms of data input the algorithm processes the citation, title, and abstract. That is important to note. We are not currently processing the full text but may in the future. It maps entities in the title and abstract to MeSH but also uses a machine learning component for some metadata, such as subheading. It may include high frequency terms occurring in the PubMed similar articles. It then applies ranking, filtering, and boosting rules to produce automated MeSH indexing which includes MeSH terms, supplementary concept records, check tags, subheadings, IM designation, and publication types. Humans are then involved in the curation and quality assurance of these results.

To give you a sense of performance, these data come from a random 10% sample that we draw every day and a set of journals where we curate all citations and assure in our review that the indexing represent accurately what the article is about based on the title and abstract. These data reflect approximately 20,000 citations that have been curated with this approach, between March and August of this year. Precision (the blue bars) is the measure of accuracy and recall (the yellow bars) is the measure of completeness. Our performance numbers are very good with overall precision at about 94% and recall at 86.

There are specific areas where we are working to improve. Automated indexing of supplementary concept records (SCRs) that represent proteins or gene products is one such area. This is why we emphasize the curation of articles that involve complex chemicals, like proteins, and are working to improve the algorithmic recognition of chemical entities.

Who is reviewing automated indexing?

We have a team of curators with domain expertise - all are former indexers. They review results of newly indexed citations on a daily basis, focusing their curation efforts on certain types of citations. For example, those involving metaphorical usage or ambiguity, those involving genes and proteins (as I just noted), and those involving specific publication types (like systematic reviews or trials).

We also review that random set of 10% of our daily load (as I mentioned on the previous slide). Curators are looking at results in a display like this, which illustrates term derivation, and may highlight certain types of index terms with special messages for the curator.

Regarding the first part: where is algorithm development going next?

I'd mentioned that current MTI contained some elements that are provided by machine learning but we are expanding machine learning to all MeSH terms and we are referring to this machine learning based algorithm as MTIX. I want to emphasize that continued refinement of the algorithm is inherent in this process. We are working to develop retraining sets, which machine learning depends on, comprised of human index articles to maintain the algorithm's performance over the long term. In conjunction with colleagues across NLM, we're working on a chemical entity algorithm that we expect will improve our indexing of chemicals.

And in terms of finding more information or reporting problems, these are links to the NLM Technical Bulletin articles about MEDLINE 2022 and automated indexing and to our current FAQs. I think we'll drop these links in the chat.

To report a problem, use the NLM Support Center link - which is available by the link at the bottom of every web page. Then, click on "Write to the help desk" to submit feedback or questions. I want to stress that if you see a potential indexing error, please report it because we'll definitely act on those reports.

Finally, I'm adding an additional what question here in terms of the main points I'd like you to take away from this presentation.

With automated indexing, we've eliminated our large indexing backlog and are providing indexing with immediacy. We have a well performing algorithm that we will continue to refine and we are leveraging our human curators' experience and domain expertise to provide quality assurance.

I think it's now time to take questions and I'll hand things back over to Mike.

>>Thank you so much, Susan.

We already have some questions coming in which is fantastic. Thank you all for popping your questions in the chat during the presentation. We also have some questions that some folks submitted ahead of time so we'll work those in as well. But, if you have questions about automated indexing, about Susan's presentation, or really about anything to do with PubMed or training we have PubMed experts, we have training experts, we have algorithm experts, we have indexing experts. We've got all sorts of great folks on this call so we should hopefully be able to take any questions that you have. So just go ahead and drop those in chat.

I'm going to start with this question from, I think, Concepcion. A clarifying question asking: "Automated indexing is only using title or abstract? Not full text?"

Jim, our algorithm expert, in terms of inputs and outputs and what's going on in the middle. Do you want to handle that one?

>> Yeah. I can handle that Mike.

That is correct, Concepcion. The algorithm is only using the title and the abstract. We also use a little bit of the journal information as well but we don't technically have access to the full text articles for computational work.

NLM and the Library Operations is working on getting that added to some of the contracts so that we can expand our access to full text but right now we do not have access to the full text. So, that is the main reason that the algorithm is only using the title and abstract.

>> Alright, excellent. Another common question - and it looks like Steve has asked a couple variations on this question, so I'm going to condense them.

>> Sorry Mike, I was going to chime in on Jim's response.

Sorry, just to provide a little bit more information. In terms of that, Jim was quite correct, that we currently do not have computational access to every journal that is indexed for MEDLINE. As Jim alluded to, our colleagues in the Technical Services Division have done a review of our existing licensing agreements and I know that they are working on negotiating those to provide computational access where they can. I don't know whether we will ever have computational access to all journals that are indexed for MEDLINE.

And then the other issue with the full text - Jim, you can chime in here if you choose again - there are issues in terms of indexing from the full text because, to begin with, there is a lot of noise in the full text so we would have to have an approach that really segments text section that the human index or would have focused on: the statement of purpose, the methods, the results. So that research to do that would have to be put in place.

The other little caveat I want to say about the full text is one of the most important things that we would want to extract from the full text is the description of the study population which we would use as check tags. Unfortunately those are often depicted in a chart or table in the full text which computationally is very difficult to parse.

So there are challenges with the full text. Jim, I don't know if you wanted to add anything to that.

>> The only thing I was going to add Susan is you're right.

The full text is very noisy. One of the things that we're seeing with our next generation of MTIX is that it was trained on humanly indexed articles that were indexed from the full text and we're seeing really good results where it's identifying terms that would have been in the full text but may not necessarily be in the title and abstract.

>> Excellent.

Thank you both for that and Susan, while we have you here and thinking about this is, this is another question (or pair of questions) from Steve about what automated indexing is actually indexing. So we know it's indexing title and abstract, but in terms of what records and what articles it's indexing, clarifying that it's only being applied to MEDLINE items (items from MEDLINE journals) and asking whether this is something that could be expanded to all PubMed items sometime in the future?

>> NLM's focus is the automated indexing of MeSH terms for MEDLINE journals.

At this time, there are no plans to expand it to the other content that might be in PubMed , like preprints and the like.

>> Alright, I'm going to actually throw another one to Jim.

This one is from Lee, asking, "Can other organizations use the algorithm for their own databases that are indexed with MeSH?"

>> That is a good question and yeah.

Right now, we have access on our website and I think they are going to put some links in the chat for you on how to access that. But we have a batch mode where you can send up a large number of items to be indexed with MTI to our systems and we process it and then we send you an email to download the results. We have an interactive facility where you can cut and paste text in and get results back and we have an API that you can call through your JAVA program and get similar results.

>> Those MTI tools online are fun to play around with.

I was digging around in there a couple weeks ago for some stuff and it's really slick the way that stuff works. If you want to get a sense of how indexing happens on your own text that is a great resource and Catherine has dropped in a couple of links with how to access the batch tools, the interactive, and the API and all of that stuff. So that's a great way to approach that.

Let's see, I think we're going to go to Susan now. This is a question from Debra (and this might be a little bit outside your wheelhouse, so if you want to kick it back to us, that's fine.)

"How will the need for new MeSH headings be determined going forward?"

>> Thank you for the question.

Indexers have always had the ability to make requests for new MeSH terms. We have a system we do to communicate with the MeSH team on that. So any time an indexer saw a concept in an article that was not represented in the MeSH vocabulary we could make a request to MeSH for that.

I think many of the new terms that they have added over the years have been from the Index Section looking at all of the literature. We still have that ability as curators, so we're still making MeSH requests for concepts that we see in the literature, again, not well represented, but I do know that the MeSH section has its own special projects and the like, which I cannot speak to but maybe Mike could address.

>> Sure.

I will mention in addition to the curators having the ability to request MeSH terms, so do you all. The MeSH team accepts suggestions from anyone. So if you go to the MeSH website (Catherine, this might take you a little longer. I blindsided you on this one, but if you could put a link to the MeSH webpage in the chat), there's an option there where you can request terms that you think should be in MeSH or changes to MeSH and the MeSh team reviews that.

If you're interested in more about MeSH our previous NLM Office Hours, which was held in June, was actually about MeSH and this MeSH creation process. So if you have not checked out the recording for that I would recommend you take a look at that as well.

Alright, let's see what else we've got in here that I might have missed and, again, please feel free to put your questions in chat and it's great to have all these indexing questions but any PubMed questions that you have we're happy to take as well.

I think we've got-- maybe Jim will look at this one.

Again we've talked about what indexing is operating on. It's operating on title and abstract. And Steve is asking about, "What about author supplied keywords?

>> So we've looked at author supplied keywords a couple of times right when they first started being allowed in PubMed records and then we came back around to it again this year looking at it - because it would be a great source of indexing if we can use them.

The author is hopefully the best person to know what their article is about. One of the problems we found is that the keywords tend to fall into either "useful," which is the minority of the keywords, or they end up being keywords that the author thinks might draw somebody to come and look at the article, or it's talking about things that they want to do in the future but are not covered in that actual paper, so we found that they were less helpful than we were hoping for in the indexing.

>> Thanks, very much.

And this is another question and this might be right for Susan, if you want to take the first stab at this, or Jim you can handle this as well.

Concepcion is asking about, again, in sort of the same vein of what are we actually indexing, what material are we actually indexing on, "If an article does not have an abstract what happens then?"

>> Then it's basically on the title and fortunately for us most of the title-only citations in MEDLINE are informative titles so we can come up with indexing for it.

I will say that there is a small percentage of citations (well, relatively small) for articles in MEDLINE that are not in English so the only thing that we have, for example, is the foreign language title. We have an internal translation program for that, where we translate it to English and then base the indexing on that.

>> Alright.

We have a couple of questions here that are sort of in a similar vein and they are going to go to Jim.

They are about-- I might give you a couple rapid fire here that are all about comparing manual indexing to automated indexing.

Molly is asking about differences between how precision and recall compares to manual indexing. We have another question here about the number of MeSH terms being attached: "Is it less concise or more concise?" Maybe, more or less MeSH terms being added to a particular record?

So we'll start with that and then I'll come back to you with another one

>> Alright, well thanks.

I think I'll actually hand the "precision and recall" back over to Susan, if you don't mind. She covered some of this in one of her slides really well, where it talked about what we're seeing with the current precision recall based on the human review of the articles.

Susan, do you want to take that?

>> Yeah but I think the question thought was--

So the data that I shared there which were very good precision and recall numbers were based on the curation that humans are doing, and this is part of my job, I am looking at the MTI

indexing and going from there. I am curating it from that point based on just the title and abstract.

I think maybe the questioner was asking about how those numbers compare when you process a citation by MTIA that was indexed by humans and that comparison, which would relate to the 500K study that we did last year

>> Yeah, I think they were looking at sort of the before and after of MTIA now.

>> So Jim, I'm bouncing it back to you really.

>> Sorry, I kind of mushed a couple of questions there. I apologize.

>> My fault for not understanding it.

We actually did a study, as Susan was mentioning, last year. We looked at about 500,000 records that were humanly indexed and were not automatically indexed and we wanted to see what terms the algorithm was missing, what was it adding that maybe were incorrect or inconsistent with the human indexing, and what we found was I think the numbers are around 79 or 80% precision.

The recall - I don't remember what that was. I think it was around 52-60% recall. And what we noticed, because that is a fairly low number-- the 52-60% recall is fairly lower than what we were expecting. But, what we noticed is that it comes back to, what Susan was talking about earlier in that, the check tags (which are the most numerous MeSH headings used) are typically found in the full text. They talk about what are the ages? What are the species? What are the sexes of the participants in the paper? And those tend to be in the full text and not in the title and abstract.

And again, in looking at the statistics for the new algorithm - the MTIX - we think a lot of that is going to be taken care of by the new algorithm since it's trained more on the human indexing from the full text.

>> Yeah, I'll add there Jim, because we have been doing some analysis lately, specifically about that.

What we're really seeing is that the MTIX machine learning algorithm is better at predicting materials from the full text because the body of millions of citations that it trained on were human indexed, representative of the full text. So we do expect to see improvements in that as we transition to MTIX.

I think there was-- No, I've just had a senior moment. There was another question.

>> There was a second part to the question.

>> I've forgotten what it was.

>> We'll actually get back to that in just a minute.

We have a PubMed-specific question so I want to make sure we get our PubMed team panelists in here as well.

This one is from Sara: "Is it ever worthwhile to include non-English search words in a search strategy in PubMed or is PubMed entirely English?"

And I think Kathi, you wanted to take that one.

>> Sure. Thank you for the question.

We actually accept non-English language abstracts from publishers so they may send them to us. Currently there is not a huge corpus of that (non-English abstracts in PubMed), but we do have them and they are indexed in the abstract field. So if there is a non-English language abstract, those terms are included in the abstract field. So you can search with those terms if you would like. Your retrieval will probably be fairly small, but it could be throughout the years that corpus will grow. Thanks.

>> Thank you so much for that, Kathi.

I've got another. The other question that we were asking, that I was mushing together before, there were a couple of them.

One is, Sara said, "I've noticed some recent records containing much fewer subheadings than I'd expect. Will the use of subheadings change in the future?"

And I think maybe, Jim, you had your hand up for this one.

>> I think so.

Yes. The current algorithm is using some older machine learning to identify subheadings and as we move into the new MTIX algorithm it should be expanding the amount of subheadings that we're adding. And also do a little bit better at putting them together with the drugs and the diseases as we move forward. Because it is trained on a larger corpus than the system that the MTIA is using right now.

>> Yeah.

That is one of the things that I think you all will keep hearing in Jim and Susan's answer and is something that you've heard from us a lot. There is always more work to be done. There's

always things that we're working on improving. That is true of automated indexing. It's true of everything we do at NLM.

If you've been to one of these Office Hours where we've talked about PubMed (or really anything where we've talked about PubMed), one thing that we keep saying is PubMed is never done. There's always more work to be done. There's always ways to make it better and to keep improving on things.

I'm going to actually throw this next one to Susan and this was from Concepcion: "How can I know whether a record was indexed by a human vs automatically?"

>> That's a great question.

And there's information about that in one of the Technical Bulletin articles that was in my slide (and that I think has been added to the chat). Since 2018, we have actually been indicating whether an article was automatically indexed in the XML that's provided. So, that is where that information lies.

And, of course, anything that is indexed since the beginning of April has been automatically indexed - so anything with the DCOM date after that point.

>> Right, so there's your definite point.

That XML can obviously be retrieved from our bulk data downloads or via the API. You can access that to review that XML and see that flag. But again, as Susan said, anything after April is definitely automated.

Let's see. Oh, this is the third part of my three part question about comparing human indexing to automated indexing. I'm finally getting to the last bit of it. Does automated indexing -- what is your experience:

"Does automatic indexing double the number of MeSH terms compared to human indexing? Is it less concise than humans?"

>> Well if we let the automatic algorithm just go and do what it feels like doing, we have seen that it could double, if not triple.

At one point, there was a time where any country - if it was mentioned in the title or abstract - was added. And there are times where they are talking about 20-30 countries and it would just add every one of them. So we've had to tune the algorithm to fix that. We right now have a fairly concise number of MeSH headings. It did not double the MeSH headings from human indexing. The current algorithm and the new algorithm is even based on history, it's based on what the human indexers have done over time. The current algorithm tries to mimic what a human does when it's looking at the article. It tries to identify what is most important. It tries to identify things

that maybe are there but less important and it takes all of this into consideration when it puts the list together of what terms it's going to present as indexing.

So in general I would say the number of descriptors that we're presenting in all the automatic indexing is going to be similar to what you saw before. As one of the listeners mentioned, the subheadings are going to be fewer than what you had seen in the past.

>> Excellent.

Thank you for that. I have another question for our PubMed team. Kathi, I think this one should go your direction.

"Will proximity operators ever be added to PubMed?"

>> Right. Thank you for that question.

As you know, we don't currently have proximity operators in PubMed - we do have phrase searching - but we are continually monitoring the feedback from users and we understand that that is a feature that is very important to our advanced users so we are looking into perhaps offering something in the future but you'll have to stay tuned - always to the NLM Technical Bulletin and New and Noteworthy. Thank you.

>> Exactly.

PubMed is never done.

Alright, there is a question that I want to ask a little bit of clarity on. If you want to go ahead and clarify a little bit in the chat or we can actually unmute you if you want. Somebody was asking about "I appreciate very much how PubMed indexes retracted articles. I don't know how this works." When you say "indexes," in this case, are you talking about MEDLINE indexing or are you talking about adding to the database? If you could explain a little bit more about what you're looking for we can probably get a better answer for you.

And let's see, I had another question in here. Here we go. I think Susan, this might be for you:

"Can automated indexing be applied for retrospective indexing of already-indexed records? For example, when new MeSH terms are added to the vocabulary would we use MTIX to retroactively index older records?"

>> Just as we have never done that with manual indexing, we are not planning to back and re-index everything with the new terminology.

I mean obviously we're still going to be doing what we classically call "the year end processing," where we normalize the citation data for changes to MeSH. So that we will, if the preferred term

changes, we update that on the citation record - or if a supplementary concept record now becomes a MeSH term. But in terms of retrospectively going back and adding the new MeSH terms, not intending to do that right now.

>> From the perspective of somebody who trains people on how to search PubMed, that would be a seismic change in the way that PubMed operates in terms of MeSH terms.

So if that was something we were even going to consider, we have to think about it carefully to make sure that we don't upset the apple cart.

Alright, let me see what else we've got that has come in here. This is a question that I'm not sure anybody else knows on this panel better than I do. I'll throw it open if somebody wants to answer it and I can do my best.

"Earlier this year, you were recruiting teams to check how well PubMed indexing did with DEI. What have been some of those outcomes or is it too early to tell?"

I know that algorithm bias is a concern that we have that NLM has done a lot of thinking about both in terms of search retrieval and in a variety of different products but it's always something that we're concerned about.

And I don't know whether, Susan or Jim, if you have anything you want to share on that. I can see if I can answer this broader question too.

>> I can add a little bit to that but I think you covered the vast majority of it, Mike.

The bias is a huge problem in the machine learning and deep learning community.

It's the fundamental questions: Does our data have bias? What is bias? If we have bias, how much bias do we have? Is it only one type of bias? Multiple types of bias?

And then the bigger question ends up coming: machine learning and deep learning require lots and lots of training data. If the training data is biased the algorithm results are going to be biased. So then it becomes a question of how do we clean that historical ground truth data to get to a point that is reasonable in the bias areas and not reduce or affect how the algorithm performs overall.

And what we're doing with the algorithms is trying to follow where the community is going with this, looking to see where some of these answers are coming from.

>> And I think with the specific efforts that you're talking about, I believe those activities are still underway.

Obviously, we will be looking into this one way or another for quite a long time. But I believe that those are still underway and have not had published results yet.

We're coming down towards the end of our time and I have a few more questions here in the hopper to address. This will be your last chance to get any questions in. Lightning round questions would be great or if there's a question you asked at the beginning and we missed it - please put it in again, so we can make sure to get to that.

I'm going to throw this one to Amanda. This one is also from Steve.

"Could the phrase index ever be removed so that all phrases could be searched for as in other databases?"

>> So that is not something we're currently looking at.

As you may know already, the phrase index is the most efficient way that we're able to provide phrase searching while maintaining our system's speed and performance for all 3.4 million of our daily users.

But what I will say on this topic is that the way we're maintaining the phrase index, we're looking through citations daily. We are adding new phrases to the phrase index twice a month, so that's happening. The phrase index is currently millions of phrases strong and we also take recommendations from users for phrases that are not included in it. So if you have something that you search regularly or something that you think should be included please write to the help desk, let us know, and we'll do what we can about adding that to the phrase index.

>> Thank you for that, Amanda.

Actually, I'm going to go to one of the questions that had been submitted ahead of time because it's kind of related to that sort of reporting customer issues. Susan, I think this one might be for you. You asked, during your presentation, to encourage people to report issues indexing if they see them.

"With the transition to automated indexing has there been a significant increase in the number of those reported errors?"

>> No is the short answer to that.

We have a Customer Service Specialist in our section who provides me with data on this regularly. And I would say that since the beginning of 2021 we have indexed more than 2 million citations and, in that time, we've gotten 147 customer queries reporting potential indexing errors. So, obviously, that is an extremely small number of customer queries given the volume of work that we do. There has not been a trend in an increase.

It's a pretty standard average of about 8 queries per month. In the month of August, I know that we got a total of 14 - 11 of them being about automated indexing. But again, in the month of August, we indexed 85,000 citations. Anyway, so that was my long answer. My short answer was no. Thank you.

>> Actually, I have a follow-up for you that's sort of on a related topic, on a “what happens when there's a problem" topic.

"If automated indexing doesn't work for a particular record how do you recognize that? How does the human curation aspect of this work? What do the human curators do to correct that or make adjustments?"

>> So as I said, we are focusing our curation efforts on particular categories of citations and when we run across an error - a completely wrong term, for example, has been indexed (this article's not about that) - we obviously make that correction and depending on the circumstances of what triggered that error we'll give feedback to Jim, the MTI developer so that we can prevent that error going forward.

If it seems sort of suspicious, like some acronym triggered a drug that it was not talking about whatsoever, we would then also go back and do an analysis (Jim would) to see where else we have indexed that incorrectly and clean up all those errors too. And again, I want to reiterate that it's really important to us that if you run across an error please write to the help desk and let us know and we will fix it. Thanks.

>> Excellent. We have just two minutes left.

So I want to get to a couple of quick ones. This one might not be a quick one, but I'm going to give it to you anyway Jim. If you can't answer it quickly, just give us the short version.

"How are citations selected for the training sets for machine learning in MTIA?"

>> The short answer is randomly.

The long answer is we tried to look at the last 5-10 years of articles that were indexed and in particular we're looking at human index and not automatically indexed articles. And the reason we don't want to go back any more than 5 or 10 years is MeSH changes every year and indexing rules change every year so the last 5-10 years are a better representation of what the current indexing policy and the MeSH vocabulary looks like at this point. What we'll do is identify the set, we use roughly 80% for training and the other 20% for testing and evaluation.

>> Alright and we have time for one last question - and this is one that was submitted ahead of time.

I'm going to throw this to Susan, which is something that is on a lot of people's minds, I think: "How can we trust automated indexing?"

>> Obviously, I realize that is a concern - “can you trust MEDLINE indexing?”

I would say that based on the data we see where human curators are reviewing MTIA indexing and assuring that the indexing is accurate and reflecting what an article is about based on the title and abstract, I would say yes that you can trust automated indexing to represent what an article is about.

And as has been stated repeatedly during this presentation, we're going to continue to refine and improve the automated indexing as we move forward. Thank you for that question.

>> Thank you for that answer, Susan.

Thank you to you, Susan, to Jim, to Amanda, to Kathi, to Kate, to Catherine, to Jessica, to all of our tech support teams, and thank you to all of you who showed up on your Wednesday afternoon to watch us and listen to us and ask us some fantastic questions.

As always, if you have other questions that didn't get answered you can always write to the help desk. We are always eager, all of our product teams, for your feedback and for your questions - because we don't know what you don't know, unless you ask us. That way we can help you. Thanks again and enjoy the rest of your afternoon.