Summary of the MLA ’21 NLM Update Question and Answer Session

May 27, 2021 | 11:15 am – 12:15 pm CT

NLM Update at MLA ‘21

Panelists:

Patricia Flatley Brennan, RN, PhD
Director, National Library of Medicine

Dianne Babski
Associate Director, Library Operations

Olivier Bodenreider, MD, PhD
Acting Director, Lister Hill National Center for Biomedical Communications

Amanda J. Wilson, MSLS
Chief, Engagement and Training

Moderator:

Kathel Dunn, PhD, MSLS
Associate Fellowship Coordinator at the National Library of Medicine

---

This is a summary of responses provided live in the NLM Update Q&A and responses to questions we were not able to get to in the session. Please use NLM Customer Support to give us feedback on these and other NLM products, services, programs, and activities.

Q: Does NLM listen to librarian feedback on PubMed? The algorithms for best match are very frustrating. How does NLM obtain input and guidance from the hospital library component of the library community? All members of the advice and guidance group work at university libraries.

A: NLM does pay attention to the feedback that we get about PubMed. Concerns and questions about searching do get listened to and heard. We have sessions at meetings like this so that you can give us direct interaction and direct feedback. We also provide feedback opportunities through the PubMed website. We do know that the process of moving from the 20-year familiar PubMed to the current PubMed caused a lot of changes and raised concerns in a number of areas. The current system reflects feedback received from the full set of PubMed stakeholders.

We are in the process of evaluating algorithms used to look at its impact on our search and search results. This is a critical activity for NLM to continue to do, and we would like to hear your anecdotal stories about where you're running into difficulties, please use our customer NLM Support Center at: https://support.nlm.nih.gov/

With respect to the advice and guidance group, this is referring to the librarians, the members of the medical librarian group that we invite to participate as members on our federal advisory committees (https://www.nlm.nih.gov/about/advisory_page.html). They do work at university libraries, but our advisory committees are composed of members in many different subject areas, and we do always try to include someone to represent medical librarians and medical libraries in our advisory committees.

Q: Are there specific plans with regards to indexing to promote diversity, equity, and inclusion?

A: Our indexing utilizes Medical Subject Headings (MeSH), which is updated annually. We are currently working with experts to review MeSH terms with respect to diversity, equity, inclusion, and accessibility. We are also currently reviewing opportunities to increase the bibliodiversity of the content indexed in PubMed -- taking into consideration geographic diversity, language diversity, and representation of minority communities.

Q: Many publishers forbid interlibrary loan (ILL) of advanced online publications. Is there any discussion of making ILL mandatory for NIH grant funded articles, especially advanced online publications?

A: This might be a time to talk about the NLM and the National Institute of Health (NIH) preprint pilot. Beginning in June of last year, the NIH asked us to include preprints in our PubMed Central (PMC) collection. We started that with preprints related to the coronavirus that were specifically supported by NIH publications. We found that that preprint experiment was successful. Several of you heard the report from Katie Funk at this meeting, and we expect to be expanding it. One way we will be able to make these materials accessible is to get investigators to submit them to a recognized preprint service and made available through PMC.

NIH's public access policy requires investigators funded from the NIH to submit to PMC the electronic version of their final peer reviewed manuscripts upon acceptance for publication. NIH is supportive of efforts to accelerate access to the results of its funded research and has already implemented approaches to facilitate public access publication policies in certain research programs, such as the Cancer Moonshot and the HEAL Initiative. ILL is a current consideration, but it's something that we can think about going forward.

Q: With the end of recent activity in PubMed, discontinuation of Pillbox, and end of the dental journal subset, how does NLM determine what stays, what goes, and how can they make these decisions more transparent?

A: We look at data on usage, how the scope has changed over the years for our products and services, the ways that our users are coming to our products and services, and how they are finding the information that they need. We used a lot of analytics, and we review our product line to ensure that it

aligns with our strategic plan
(https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.html).

NLM had over 300 different products at one point. In feedback received, especially from RFIs, many people have stated that the amount and variety of NLM's products made it confusing when they were trying to figure out how to use or where to go for their information. We're trying to streamline our platforms to make it easier for our users to find the health information they are seeking.

With the products that we are sunsetting, we integrate the information from those resources into other highly used products. The data does not go away. In many cases, we also make the data available through APIs or through FTP downloadable files for our users to access.

With regard to the dental journal subset, that and many other specialized search queries are no longer updated. They were originally created to assist searchers in retrieving comprehensive results. PubMed offers an amazing, robust feature for creating your own searches and saving those through My NCBI. We've made the queries available for you copy, add to your My NCBI account, and build on those from there for continued use.

Q: How does NLM know that these regions will better serve people? Less population and more land mass does not equal better serving people.

A: We definitely want to give each of our regional medical library (RML) subsets an opportunity to meet and engage with communities where they are. We did that in the new regional configuration by balancing workloads in two ways: focusing on the amount of population and the number of Network of the National Library of Medicine (NNLM) member libraries and organizations supported. To do this most equitably, the number of regions was reduced from eight to seven.

The value of this model is that each RML can best determine how and where to staff and offer programs to meet communities within their region. This could be over a large geographic area or a densely packed smaller area, and every configuration in between. The good news here is that NNLM regions have always covered large areas. NNLM staff are experts and have long histories at making sure that they're reaching populations in each component of their region. There is also a greater distribution on a per capita basis which means distribution of those funds will be more equitable. The NNLM has always covered large geographic areas and NNLM staff are experts at making sure they reach all these areas.

Q: In the last year, we've recognized rich sources of information outside of traditional journal articles. NLM is clearly engaging with preprints, which is great, but other forms of grey literature, which used to be curated by Disaster Lit are no longer easily discoverable. How could NLM help librarians and library users discover high quality grey literature about health, health care, and social determinants of health? I'm thinking of documents like the Johns Hopkins University May 2021 School Ventilation Report, or the October 2020 CAP Housing Instability White Paper.

A: NLM has acquisition and selection experts that are scanning the internet for grey literature that falls within the scope of NLM Collection Development Guidelines. Select information is included in other

NLM resources such as Digital Collections and Bookshelf. Other resources include HHS ASPR TRACIE for grey literature on disasters and public health emergencies.

Q: What are the NLM plans for preprints after the April report on the NLM Beta on Preprints?

A: We plan to continue to run in Phase 1 for the foreseeable future. We will be completing an assessment of the impact of the pilot in the first 12 months before determining next steps. We'll continue to post updates in the *NLM Technical Bulletin* to keep folks abreast of the assessment and next steps.

Q: When will NLM restart document delivery interlibrary loan (ILL) of its physical materials? When do we have an estimate of when NLM will be able to fill ILLs?

A: We do not have a timeline at this point. The NIH campus continues to be on maximum telework, and we have not reopened our collections area to staff. For now, we continue to fulfill ILL requests via our digital collections access.

Q: Can you give more details on the expansion of automated indexing and PubMed? What will be indexed automatically and starting when?

A: We have been using the Medical Text Indexer (MTI) for many years. At first when we started using machine assisted indexing, it provided suggestions for MeSH terms that the human indexer would select. For the past two years or so, we have been applying fully automated indexing with some level of human review for about 10% of our MEDLINE journal titles. MTI has been around since 2002; and we continue to refine it so we can apply it to more journals in MEDLINE.

Q: Indexing is already not as good for population subgroups and social science concepts as it is for clinical medicine concepts. Will machine indexing improve MeSH performance for the needs of users searching for public health and social determinants of health (SDOH) information needs? Or will it just speed up low-quality indexing?

A: NLM's Medical Text Indexer (MTI) system was started in 2002, and we continue to improve and refine the algorithms.

Q: What we see in MeSH, especially in the entry terms, is that there are many inversions, entry terms with a comma in them. This is a remnant of the old alphabetical lists. It would really help if you would remove the inversions from the list of entry terms, as it really makes finding relevant terms very complicated and time consuming.

A: The inversions, or permutations of index terms, are present to increase the searchability in our system. It's actually better to have more of those, rather than limit them, because it increases the possibility of finding the item you're looking for regardless of how a user enters a term. We also use about 27 different terminologies from the Unified Medical Language System (UMLS) to create the PubMed mapping file, so having these additional access points helps with retrieval.

Q: Are there any plans to do work on decolonizing the archival material and/or applying that to revising and revamping the MeSH vocabulary?

A: This is an interesting concept that we will reflect on going forward. We revise and revamp the MeSH vocabulary on an annual basis. There are three working groups that meet every year to consider new MeSH terms that relate to the scientific literature. Applying a decolonizing lens to revising and revamping the MeSH vocabulary is an interesting concept because we're in a modern system for finding current literature.

Q: Automated indexing for curation at scale does enable efficiencies but it will be imperfect. What measures will NLM take to ensure trust in automated indexing and corrected mis-indexed publications?

A: NLM has a team that reviews the application of automated MeSH terms. At the same time, we will be reviewing our algorithms. If there's a consistency issue, we want to make sure that we fix that to apply that better logic going forward.

Q: What data about the automated indexing will be made available to the public to encourage trust?

A: We are already doing automated indexing, and we have established what we consider an acceptable precision level. Our subject matter experts will continue to review and work on improving the algorithms and application of MeSH from the information available.

Q: How will librarians be engaged to identify areas for application of automation for greatest impact?
A: As it relates to MeSH indexing, we have a wide range of staff that work on indexing from scientists to librarians. They all contribute to thinking about how to automate indexing and other things that we can automate going into the future to improve efficiencies.

The MTI algorithm and its implementation are actually available to the public. You can test it if you want and apply it to your collections (just search “NLM MTI” to find it). There's the batch mechanism that allows you to send your data to our algorithms, and we’ll process them for you.

Q: Will the machine indexed citations be reviewed by an experienced human indexer before they're added to MEDLINE?

A: We already have about 10% of our journals going through the automated MeSH indexing process. 100% of these citations are being reviewed by a human indexer, but we expect to reduce that based on meeting our acceptable precision level. We continue to add more journals to automated process and continue to review and enhance the MTI algorithms into the future.

Q: Why do I need a license to search the Unified Medical Language System (UMLS)?

A: The UMLS contains a collection of more than 200 health data standards, each of which have their own copyright and intellectual property terms. The UMLS license agreement outlines the terms of use of all of these terminologies, which includes free access and internal research use for each.

Q: The consumer health vocabulary which is part of the Unified Medical Language System (UMLS) is outdated. Are there any plans to update it and add more items?

A: The short answer is probably no.

The longer answer is that the UMLS metathesaurus is a terminology integration system. The terminology integration system is a system that integrates existing terminologies as they become available, and as they're made available to NLM. With the exception of MeSH and RxNorm, NLM is not developing these terminologies directly and so essentially, we rely on the developers of these terminologies, which we integrate.

There's always been a tension between integrating interesting vocabularies that don't necessarily have long-term support because they provide coverage for some aspect of the UMLS that wasn't covered before, and the issue of what happens when the developers of these terminologies (who are often academic partners, for example) don't necessarily receive the long-term funding needed to maintain these resources.

We do the best we can to maintain these terminologies and to encourage developers to keep them up to date, but there are limitations to this process. Hopefully, there's redundancy among the various terminologies in the UMLS, and some of the terms that were provided by the consumer health vocabulary, for example, have appeared in other terminologies.

Q: Last year, Dr. Brennan stressed that data and publications are now an interoperating system and that traditional opportunities for librarians are changing. But the change requires reinvention. Today, you answered many of my concerns. You talked about data publications, new roles for librarians, any specific observations going forward?

A: I come from a field that found its base being changed very quickly in the 1990s. Nursing was really going under a lot of changes; there was movements in of physician's assistants. There was a redistributing of tasks to less skilled workers and team models. There is this conversation about how we keep a profession alive when the occupational space where the professional work is changing. I think the NLM wants to partner with the MLA and with all of you in ways that we can advance this forward. What I am recognizing is a couple of things.

First, close partnerships with the training program is absolutely essential. In our extramural programs, NLM supported a partnership between our training programs that are housed at institutions that also have schools of library and information science to see if there could be better crosswalk between them. We do have a need to work with the curricula in medical libraries training programs. Terminologies, organization of information, distribution, unrestricted search for information are critically important as we move towards data.

Second, finding ways to repackage these areas and reimagine what one's career could look like. We also have found that it is important as we look at career change and mid-career development to consider what kind of education is needed for people. Many of the questions that have been brought to us have to do with what can we do to support the mid-career librarians who are unable to interrupt their jobs, to leave their work for 3 months or for a years' worth of training, how do we best support them? I would see this as a conversation we want to continue to have through Amanda Wilson and the Office of Engagement and Training (OET) to understand if there are ways NLM can support what I consider in-stream career reorientation.

The last part that is important to me is that we think about where the essential need in society is for medical librarian knowledge and find ways to make sure that it gets integrated more broadly. The way we're doing this at NIH is to work with the research grant process to make sure that some of our concepts like using common terminologies or common data elements actually become a part of the research application. Now, the downside of this and one of my biggest observations is the realization that reaching out to one community doesn't mean we're reaching out to every community.

When we encourage our training groups and our training programs to reach out to the schools of library science and their institutions, that doesn't always carry through. We're learning to have multiple means of communication. As these opportunities become open, they can be initiated by either side, as opposed to waiting for the institution. Please keep comments coming to me about how we can best support the life cycle of the career of a library scientist, particularly one with this specific focus in medical librarianship, because we know that the work of the NLM does not occur only in Bethesda. It occurs around the country, and these partnerships are really important.

Q: Could you speak more to the TRACE variant reporting program? Specifically, whether or how NLM is engaging in data sharing with international partners.

A: The TRACE program is specifically an activity that is housed within the United States. It is a broader government activity including NIH and other agencies. NIH received funding from HHS for this program. NLM worked very closely with Dr. Collins, the National Center for Advancing Translational Sciences (NCATS), and others on the NIH response. NLM houses the sequence strings. NCATS identifies the targets on those sequence strings—for interventions for vaccine activity, for example. NCBI houses several genomic repositories. We are part of the International Nucleotide Sequence Database Consortium and that allows us to partner with a site in Europe and a site in Japan, where we update our holdings periodically to make sure that we are having broad access to genomic information.

You may have seen several articles in the last couple of months, most recently in *Nature*, questioning the manner in which we should work with GISAID, which is a private organization that was established in 2008 to house nucleotide sequences in a manner that protects the intellectual contributions across the globe, particularly in our southern hemisphere countries. GISAID's information is not publicly available and open access. They do require permission and have very strict conditions of use.

The NIH is working very closely with our partners internationally, specifically, with something referred to as the HIROs, the Heads of International Research Organizations, to make sure that these kinds of arrangements—which are important to protecting the efforts of our scientists around the world—do not also interfere with the public health needs of the world. The conversations are continuing. The issues are complex, and we're very grateful to have both Steve Sherry from our operation here at NLM, and Dr. Collins personally involved in these conversations.

Q: How will you ensure search strategy reproducibility? If someone reruns or tries to replicate a search a year later or more, how do you ensure that the search will retrieve the same results? This seems problematic presently and as metascience continues to grow, do you have any plans to make this work better?

A: Live PubMed is updated regularly. Without a specific context for this question, no one should expect to use the live PubMed for their research in terms of reproducibility of search, because we know that if we search tomorrow morning, we're going to get different results as the system is updated with new records.

What happens in general in the research community is that they come up with test collections, and the idea is not new. Test collections have been used for a very long time. It means that it's a frozen set of documents with relevance judgments that have been made, that are used as the gold standard, and the search strategies or search engines are actually tested against the test collection. That also ensures for reproducibility, not only longitudinally for somebody to test their algorithm over time, but it also allows many different groups to test their algorithms against the same collection.

Q: Does NLM have a list of products that are using any artificial intelligence (AI) technologies from machine learning, or natural language processing (NLP), or chatbots language features? If not, can you provide one for the public? There are concerns about bias and algorithms in the general marketplace, and this may help with understanding and trust.

A: I think at this point, it would be easier to make a list of the NLM products that do not use AI as opposed to the ones that use it. There's been a lot of use of supervised machine learning techniques, including in the Medical Text Indexer (MTI). Part of the MTI actually uses it; it's a hybrid of the processing that I described and a native base, which is a form of supervised machine learning in there, and of course, everybody at NLM is working on some form of AI.

Dina Demner-Fushman, in our intramural research program, is researching the use of AI techniques for question answering. There's a lot of this going on for search. There's a lot of this going on for image processing, and it is the case because we have access to collections of annotated datasets that are amenable to supervise machine learning and AI techniques. Of course, we need to be vigilant about the relevance and inclusivity of these data sets and the applicability to a variety of use cases.

NIH has a new initiative that you may have heard about, that was launched in the last couple of weeks, called Bridge to AI. This has a tremendous opportunity for information, communication, and scientific communication because the goal of Bridge to AI is to create large scale AI-ready datasets against which algorithms can be tested, and for our development of our resources to be sure we have a test of what constitutes a trustable algorithm.

Q: Machine learning and natural language processing (NLP) algorithms are already shown to embed and invisibilize existing problems in language and classification systems. If there's not a system beyond efficiency to help debias terms, it will inevitably make existing problems worse because they'll be harder to see. How would corrections be applied to past algorithmically assigned terms?

A: NLM has a team that continuously reviews the application of MeSH and makes updates and improvements.

Q: What about Google's new BioMed Explorer?

A: Google's BioMed Explorer is based on data provided from NLM's PubMed & PubMed Central, and CORD-19, of which NLM is a collaborator. We know that many people come to PubMed after a Google search. We have been working very hard with our colleagues at Google to understand their search algorithms to be able to make use of them.

There are innovative technologies coming out of various search companies, and the critical piece where the NLM takes its special role is the openness and the preservation of our data. The literature resources that we have are available without constraint for use by serve the public. That is where our emphasis is.

Q: We know that NIH and NLM are putting increasing emphasis on rigor and reproducibility. Do you think that reproducibility librarians should focus on computational reproducibility, data sharing, reporting guideline awareness/use/compliance, or something else?

A: Yes. The NIH Policy on Data Management and Sharing will go into effect in January 2023. Information professionals will be key to helping the research enterprise in their organizations comply with these new requirements. This is going to be another success for LIS professionals to support research, as was the case with the roll out of the NIH Public Access Policy and identifying the best data repositories.

Addendum: The following question and its answer were added to this summary after it was posted online to the MLA Connect email that was sent Thursday, June 24, 2021.

Q: What are the PubMed user demographics--how many scientists, physicians, students, librarians, etc.?

A: PubMed is accessed by more than 3.3 million users a day. Visits to NLM websites (such as PubMed) are private and secure; we do not collect any personally identifiable information (PII) about you, unless you explicitly choose to provide it to us. The following user demographics were collected from users who indicated their role when taking our PubMed survey:

- Researcher: 30%
- Healthcare professional: 29%
- Student: 26%
- Educator/trainer: 4%
- Patient, family, caregiver, or friend of patient: 4%
- Librarian or information professional: 3%
- Other: 3%