Anticipating the Future of Biomedical Communications

Meg White
Rittenhouse Book Distributions

Patricia Flatley Brennan
National Library of Medicine

Meg White and Patricia Flatley Brennan, "Anticipating the Future of Biomedical Communications" (2019). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284317191

Anticipating the Future of Biomedical Communications

Meg White, Director, Technology Services, Rittenhouse Book Distributors

Patricia Flatley Brennan, Director, National Library of Medicine

The following is a transcript of a live presentation at the 2019 Charleston Conference. Video of this session is available at https://youtu.be/xxMpezF2sCo.

Meg White: I'm up here on the stage and we said we were ready to go and the mics came live and the music went down, so, good afternoon and welcome. My name is Meg White and I'm here this afternoon to kick off our plenary session.

A few housekeeping items: first, please silence all your cell phones and electronic devices. Just for your information, and Big Brother's still listening, probably, the session is being recorded this afternoon. We're going to endeavor to leave time for questions. Please use the microphones that are in the audience and please be sure to state your name and your affiliation when you ask a question. All right, well, I'm very pleased to be joined on the stage today by Dr. Patty Brennan. She is director of the National Library of Medicine. Dr. Brennan's career and her background are extensive. She has a unique combination of engineering, information technology, and clinical care. Since being named director of the NLM in August 2016, Dr. Brennan has led the development of a strategic plan for the National Library of Medicine and will speak to us today about the future of biomedical communication. Dr. Patty Brennan.

Patricia Flatley Brennan: Good afternoon. You must be wondering what's a medical library doing in this meeting today? Well, how many of you have had a health problem in the last 10 years? Okay. You need us. I will say not much more just yet, but I will tell you that the National Library of Medicine is the world's largest biomedical library. We are a part of the National Institutes of Health. We have 1,700 women and men that work together every day to bring you PubMed, MEDLINEplus, clinicaltrials.gov, and to conduct research and data science and in the application of biomedical informatics technologies to clinical data. I'm extremely proud to be a part of a federal library and I'm very delighted to be addressing this audience because I see intersections between the concerns of publishers, librarians, and federal libraries. We are not a lending library. We are a repository library. We do participate in interlibrary loan. We do have library cards still, by the way. You can get a library card to the National Library of Medicine, but mostly we serve as a resource to the world. We provide the largest bibliographic repository, PubMed; many of you, I suspect, have run into that at one time or another. There are overall 30 million bibliographic citations in PubMed. We're adding 1 million a year. We also, though, as you look on the screen above you, have a number of other products that are important to the world. Our genomic repositories, our clinical variance, our dbGaP, these are data repositories that have billions of genomic sequences in them and I'm going to tell you some exciting news about those as we go forward.

But we also provide information to communities. TOXNET provides important information about the environmental quality in a community. At clinicaltrials.gov a registry of clinical trials where patients who are worried and wondering can they find a treatment for complex illness can go there. DOCLINE is a way that we use to deliver biomedical knowledge to anyone who may need it anywhere in the world. WISER up in the right-hand corner is an in-the-moment smartphone-held application that gives specific toxic management information to first responders.

We touch many lives, many places everywhere in the world, but we know that biomedical communications is changing and it is no longer your mother's library. It can no longer be your mother's library because we must keep pace with communication and serve the public. So, I'd like you to take a minute and watch a brief film that gives you our vision of where we think biomedical communications is going. (Video: *Anticipating the Future of Biomedical Communication* plays.)

I'm very proud to be the director of an engine that powers data-driven discovery and data-powered health. We have just launched our strategic plan. Our 10-year vision of what we will be doing is grounded in three key pillars: first, to accelerate discovery and advance health through data-driven research. Second, reach more people in more ways through enhanced dissemination and engagement; and third,

https://doi.org/10.5703/1288284317191

to prepare a workforce, including a citizen workforce, for data-driven research and health. We envision a future where a library becomes not just a building or a website but an ecosphere, as you see on the right-hand side, of interconnected digital research objects where literature, protocols, study data, funding opportunities, pipeline and pathways for genomic and assessments all are interconnected in the library sweet spot, as many of you know, is in structuring underneath each of those ovals and in building the interconnections between them.

I'm going to talk with you for the next 20 minutes or so about what the National Library of Medicine is doing to support biomedical communications in the 21st century and I'm going to explain, address three key aspects. First of all, improving the usability and access to the research literature. Our sweet spot, our strong point in the National Library of Medicine as a part of the NIH is to bring research knowledge together for the world. Second, to promote open science and data sharing; and third, to guide 21st-century scientific communication. Now, the National Library of Medicine is almost 300 years old and we're preparing for this through many conversations around the world. What we began and we will continue to be is fundamentally a collection, but our collection is changing, and I heard some marvelous models this morning about the new collection. We have been custodians. People think of a library as a place that holds things and certainly preservation for the future for future access is a critical role of the library, but increasingly we connect to our connections and in the future we have to focus on discovery on the fly and, as I heard this morning, building new collections based on the patterns of those discoveries. The National Library of Medicine is primarily bringing literature and knowledge and data into the hands of those who need this, but we need to improve the usability and access to the research literature.

This is our interface to PubMed. I suspect many of you have looked at PubMed for your patrons, for yourselves, maybe for someone in your family. PubMed is an amazing resource: 30+ million articles, two and a half million daily users, two and a half million users every day, most of them not coming to our building, by the way, coming to us electronically. About 3 million searches and 9 million page views, but what we know is that 80% of the searches that are done on PubMed have more than one page of search results done and 90% of the people who search PubMed never go to the second page. So, if the article that you need or the article that you've published is on page 2, you're dead. We are now committed to improving search and improving retrieval by artificial intelligence and machine learning approaches. We want to first and foremost improve the search quality, make sure we bring the most highly relevant resources in an efficient manner to our users. But, second, we need to improve usability. So, what we are doing is adopting, out of the research from some of our intramural researchers, what we refer to as a learning to rank algorithm. Let me walk you through the diagram on the screen for just a moment. On the left-hand side you see our 30 million citations, always in partnership with the publishers in an XML format, that are tagged with keywords, maybe, and then have human indexing applying the MeSH terminology to make the metadata useful for search, but we also are now tagging it with experience information, how often was this particular citation pulled up, for example, or what other citations were pulled up with this one. An individual launches a query at step one. That query is exploded through our PubMed interface and a set of series of what we call hits/matches are drawn up, mostly running about in the area of between 500 and 2,000 site hits for each one. We need to sort them better. Currently we present them in reverse chronological order. That's not enough. We also, from that zone, then in step three apply our new AI algorithm to the first 500 certain hit search returns. The first 500 results are then resorted to create a best match of what the user has learned, is looking for, and then from that we monitor experience data, how often are those units searched on, how often are they clicked through, to make sure we're actually improving our algorithm and have ways to lead experience into improvement.

So, here's what is returned to an individual now. If you look at this screen, this is a standard PubMed search screen. On the right-hand side you see a green box that has best match or most recent. We currently return most recent searches, but we are getting the community ready for the fact that we have this best match algorithm available so the red box that you see in the center alerts the user to some other articles that they may not find on the first page. Here's our best match. Here's what we think might be useful to you. Beginning in about 90 days our default search will be the best match search, although you will be able to always get reverse chronological order searches returned.

Here's what our new interface is going to look like. There's three things I want to call your attention to

here. First, again, look in the upper right-hand corner. You may toggle between best match and reverse chronological order but, second, each of our citations now has a small series of phrases underneath it. We call these "snippets" that show the match to your search phrase so that you don't have to click through to read an abstract to determine whether or not you want a particular article. On the left-hand side you see a histogram. There's a slider bar underneath that histogram. It's a reflection of how many articles per year were published in that according to that topic that you requested. You're able to constrain your search to certain periods of time, thus allowing the user to have a better experience with PubMed. We're excited about this. You can see this today, actually, if you Google PubMed Labs but we believe we'll be going live in about 90 days for this. We are still open to your input.

Now, the world is moving toward data and the world is moving toward openness. That is not a surprise to the audience here. I want to talk to you a little bit about what the National Library of Medicine is doing to improve and promote open science and data sharing. First and foremost, in our PubMed Central literature repository we hold the full text of over 5 million articles, most of them from federally funded research or historically valuable articles. These are freely available. About half of them are available for machine processing and the other half are available downloadable for human reading. Within this PubMed Central repository now, though, we are able to link data sets directly to articles so the data in support of an article can be made available to an individual. We have two key repositories: our citation repository, PubMed, and our full-text literature repository, PubMed Central. Each of these provides a pathway to data. Within PubMed, our citation repository, we make link outs to Figshare, we make link outs to various public data repositories as you see on the left side. Within PubMed Central we have data citations and other supplementary material connected directly to the article. We know this is not the final solution, but it's now a pathway to get direct access to data. It's been a bit of an experiment to get our researchers to curate their own data. I will tell you they need a little help with this. Libraries have lots of work to do in the future. Almost all of what I've described to you so far is housed at the National Library of Medicine and can be downloaded by an individual through our FTP sites or through industries that we also have partnerships with, but our genomic databases are growing so rapidly that the download is becoming impossible to support, so we have recently started to move some of our high-valued genomic data sets into cloud instances, into commercial cloud storage within Amazon, AWS, and Google Cloud. We have—the first repository we launched was the Sequence Read Archive. The Sequence Read Archive is about a 12 PB repository of annotated sequences, genomic sequences that are now available that can be interrogated and operated on in cloud instances. We are also in the process of modifying our search and analysis algorithms including the blast algorithm, which is a sequence alignment algorithm to make these possible to run in cloud instances. Those of you who have started to migrate into cloud instances know it's not merely a matter of lift and shift. You have to restructure things, you have to find new pathways in. Fundamentally, we are committed to making data accessible, and clinical data are accessible through the fast interoperability—fast health care interoperability resource. This is one step we are taking toward the goal of making all data findable, accessible, interoperable, and reusable.

Now, I've talked to you about literature; I've talked to you about data. But what brings knowledge out of data are the analytical tools, the models we apply to data. The National Library of Medicine is questioning how do we build a library of models? What characteristics? What is the grammar for models? What are the metadata for artificial intelligence models or machine-learning algorithms? We are beginning to explore how to code and document and make accessible the analytical tools that one can use to operate on our resources. We believe that models must be described by key metadata or data elements including the type, the purpose, the assumptions built, when it was made, and also most importantly, what was the intended use and can this model scale? Building analytical models takes millions of dollars of research investments and if they're only useful for solving one problem, we've wasted federal dollars, so we need to know how do we make models scalable and when do we do that? That is a process of model verification and validation, and the library is becoming a partner with scientists and methodologists around the world to be able to make this happen.

But we are deeply in a period of change and I'm bringing the library along to improve the discoverability of the biomedical literature and through the literature data and through the "data to models." I'm happy to be saying today we're announcing the launch of MEDLINE 2022. Those of you who have

been involved in the library know we have been quietly working on this for a while. MEDLINE 2022 is an approach to make our literature more available more quickly in a way that will accelerate scientific discovery. We are focusing on three key areas: first, curation at scale; second, expanding metadata; and third, creating strategies that are efficient and connected. Our activities for creation at scale are all targeted toward that upper right-hand corner. Having a citation indexed within 24 hours of being deposited. We cannot do this without partnership with the publishers. We recognize that your ability to send us XML and properly tagged articles allows us to accelerate our indexing strategy. We must also recognize that human curation, a very expensive effort, a very important effort, has to be preserved for those articles most in need of that. Second, to develop expanded metadata. We are focusing on optimizing metadata to support interconnectedness across our various resources and registry. We are focusing also on expanding the use of funders' metadata in our literature services. We heard this morning how important the funders are in open access and open data sharing. We need to make sure we have a way to support the funders in looking at and evaluating the impact of their resources. And finally, our emphasis on creating things in an efficient and connected way. We are enhancing the MeSH vocabulary. We're working on adding authoritative vocabularies to MeSH to increase it, and MeSH, by the way, I'm sorry, is a medical subject heading. I apologize for talking federal jargon in front of you. This is the most essential terminology that we use to do our indexing and metadata.

The last advance I want to talk to you about today I think might be the most commercial and I'm quite interested in your reactions to this. The National Library of Medicine, in cooperation with the National Institutes of Health, wants to increase the discoverability and use of preprints as part of biomedical scientific communication. We're announcing the beginning of a pilot for examining preprints and making them accessible through our existing citation repositories and literature databases. The National Library of Medicine preprint pilot has as its goal to improve the discoverability of preprints, which downstream should accelerate discovery in science and, frankly, make science more efficient so we're not replicating studies that are already done, which is a waste of federal dollars. In order to make this happen, certain conditions have to be made. The preprint gets deposited into a preprint server. The preprint is then shipped to the PubMed Central, our repository literature database, and an associated bibliographic citation is made in PubMed. All of the records of the preprints will be clearly marked as the material has not been peer-reviewed. It's critical that our readers understand this. Our experiment is limited to NIH-funded research or preprints coming out of NIH-funded research. The preprints must be fully available in an XML markup. That's the responsibility of the preprint service and the repository must have a license that allows for the inclusion of that particular document into a noncommercial repository. So, our plan is in partnership with existing preprint services that we will make their materials more discoverable through our interface.

The experiment is designed to accelerate scientific output of NIH funding. Clearly, that's what we're interested in. The National Library of Medicine is part of an enterprise that spends almost $40 billion a year of tax money to bring health to society. We must be good stewards of that. NIH recently has been encouraging, but not requiring, its investigators to use preprints in the service of conveying their information and the results of their research early. We have a public access policy that requires that all archival articles supported by NIH be made freely available to the public in short periods of time, but we want to encourage investigators to start using preprint repositories that can accelerate access to the information that they have available and fundamentally that help us to support the guidance that NIH is giving to our researchers of how to communicate. Now, in this era of publication we know scientists have many choices of where to put their articles. The National Institutes of Health remains committed to expanding access to the scientific literature, but we want to do that in a way that it maintains the integrity of the scientific literature and yet, as a federal body, we do not give direct advice of which journals are acceptable to publish and which are not acceptable to publish and we do not maintain a whitelist or a blacklist, but rather the NIH has released guidance to its community to say that our publication policy is to look for journals, look for outlets that have rigorous editorial policies, have clearly defined ethics, have an emphasis on communicating in a scholarly way to the public. The National Library of Medicine is committed to working with partnerships. As we develop the preprint experiment it is absolutely essential that we maintain a relationship with the preprint service. We will not become a publisher. We will not become a preprint service on our own. As you notice, the preprint services that we are willing to partner with have the same criteria of good publication criteria

that NIH has listed that we need to use—have appropriate licensing, there must be rigorous transparent policies and practices to address plagiarism, competing interests, misconduct; and fundamentally the preprint service itself must maintain accurate records for the preprints and allow us to have a way to coordinate, that is, to align archival publications with various forms of preprint activity, including the development of and increasing expanding use of open peer review.

The National Library of Medicine is and will continue to be a trusted source of health information. We do not produce the information. We do not publish the information but we facilitate the access to that information. Participation in the National Library of Medicine's experiments with PubMed, our new interface, our experiments with the preprint service, and our use of our new and improved MEDLINE 2022 indexing strategies we hope will increase the accessibility of trustable health information to the public at large. We recognize that in an era of machine learning, machine engagement, artificial intelligence, the concept of trust must move beyond a human level of developing an interpersonal agreement of what trusted information is to computable machine/machine interfaces that remain trustable and private. As a library, we are committed to the values of library science, which indicate full, unfettered access to health information in a way that is unsupervised without unnecessary oversight so that the examination of ideas, the examination of science to create new ideas becomes a tool that is useful for accelerating science everywhere in the world. I thank you for the opportunity to talk with you about how our focus on improving biomedical science communications will accelerate health for all and I'm ready to hear your comments and questions. Thank you very much.