DOI 10.3233/ISU-220165
IOS Press

NISO plus 2022 Miles Conrad lecture: The role of a library in a world of unstructured data

Patricia Flatley Brennan[*]
Director National Institutes of Health National Library of Medicine, 8600 Rockville Pike, Bethesda, MD, USA

Abstract. Throughout its nearly two hundred year existence, the National Library of Medicine (NLM) (https://www.nlm.nih.gov/) has advanced biomedicine and public health by acquiring, organizing, preserving, and disseminating knowledge essential to health and medicine. NLM has devised many innovations including standard terminologies and messaging formats such as the Journal Article Tag Suite (https://dtd.nlm.nih.gov/) to organize and manage biomedical literature. While scientific communication largely relied on books and journals over the last two hundred years, digital data are quickly forming the substrate of scientific communications. Data come in forms with much less structure than that afforded by publications, and these can vary from observations made during carefully controlled clinical trials to streams of genomic sequences to the counts of footfalls captured by personal devices. Coincidently, an increasingly diverse set of users – from clinicians to laypeople to public health to big pharma to scientists – bring unique perspectives as they draw meaning from new sets of scientific output. How does a modern library meet its mission to acquire, organize, preserve, and disseminate the many outputs of contemporary science? What role do standards play? How does NLM help this diverse set of stakeholders derive meaning from its resources?

Keywords: AIM-AHEAD, biomedicine, Bridge2AI, ClinicalTrials.gov, Common Data Elements, COVID-19, dbGaP, FAIR Data Principles, Fast Healthcare Interoperability Resources, FHIR, GenBank, Global Health Events archive, International Committee of Medical Journal Editors, Journal Article Tag Suite, JATS, LitCovid, medical subject heading, MeSH, Medline 2022, NCBI, Nigam Shah, NLM, NNLM, PubMed, PubMed Central, Resource Description Framework, RxNorm, SARS-CoV-2 sequences, Sequence Read Archive, SRA, US Core Data for Interoperability, Patricia Flatley Brennan

1. Introduction

Let's start from two points: First – libraries will persist, but the units of scientific communication, such as the digital objects that must be labeled and connected will continue to change, and keeping up with the standards that promote the ability to locate and share them will continue to be a challenge. Second – NLM cannot do this alone. Even within the biomedical sciences, NLM must partner with publishers, authors, distributors, technology companies and, most importantly, with its stakeholders. These stakeholders will increasingly include a larger and much more diverse group than the scientists, clinicians, and policymakers we've had to this point and will also include patients, mothers, children, and high school students.

Standards bring order to complex information by supporting efficient automation of complex information to foster communication and ensure shared meaning. This is critical to the lives of individuals. NLM not only focuses on acquiring, collecting, preserving, and disseminating scientific communication, but also provides the tools – including standards – to make available this scientific communication.

[*] E-mail: patti.brennan@nih.gov.

2. NLM: Serving science and society since 1836

NLM is a research enterprise for biomedical informatics and the world's largest biomedical library. There are three critical epochs in NLM's history.

2.1. 1836 to 1968: NLM was established

In 1836, the Army Surgeon General requested funds from the U.S. government for medical books to refer to in the field, and the growing collection officially became the Library of the Office of the Surgeon General of the United States Army. It was only in 1956 that an act of Congress transferred the library to the Public Health Service and named it the National Library of Medicine. NLM became a component of the National Institutes of Health (NIH) in 1968 under the Department of Health and Human Services.

2.2. 1968 to 2000: NLM developed the foundation of a modern library

In the 1960s, the impact of the Information Age launched an era of information digitization and expansion. Miles Conrad, one of the founders of the National Federation of Abstracting and Indexing Services (NFAIS) and pioneer of accelerating the speed of disseminating, discovering, and acting upon scientific knowledge, predicted this change ten years before, and through congressional action, NLM developed the beginnings of the Lister Hill National Center for Biomedical Communications (https://lhncbc.nlm.nih.gov/) and the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/).

In 1984, Donald A. B. Lindberg, M.D. (https://nlmdirector.nlm.nih.gov/2019/08/20/remembering-donald-a-b-lindberg-a-visionary-giant-with-a-personal-touch/) became Director of NLM. Dr. Lindberg shared his vision for how automation would address both the burgeoning future of genomic knowledge and how to address the ways that patients and laypeople can access this information. By 2000, the NLM had the foundation of a 21st-century digital library that uses its collections to offer literature, data, analytical models, and new approaches to scientific communication that are accessible, sustainable, and available twenty-four hours a day, seven days a week.

2.3. 2000 to 2036: Accelerating biomedical and computational health data science

NLM is leading innovative data science research to accelerate the mission of the NIH and providing scientists and society with trustable health information. Through the creation of strategies to support efficient and accurate exploration of large genetic and literature databases, NLM is generating new analytical methods and models to gain new insights from clinical data. Additionally, NLM continues to leverage the Network of the National Library of Medicine (NNLM) (https://nnlm.gov/) to provide a human connection in communities across the United States to advance health equity through information and increase community engagement in NIH research programs. Information services created and maintained by NLM will continue to provide essential tools for researchers, clinicians, and the public allowing more people to operate with greater efficiency and speed by connecting them to the resources of a digital research enterprise that is essential for science and discovery.

3. The changing nature of research across NIH

The COVID-19 pandemic has changed the nature of research at NIH, including NIH-supported research conducted nationally and internationally. During this time, NLM became an important engine at NIH to address the changing nature of research across NIH. NLM accelerated the application of informatics to clinical research by encouraging NIH to publicly endorse a set of health data standards in the United States Core Data for Interoperability, as well as promote the use of the Fast Healthcare Interoperability Resources (FHIR) standard in its funded clinical research. NLM helped inspire advances in analytics for generalizable solutions and supported engagements related to research, design, and implementation in communities across the country.

NLM has more than eight thousand points of presence in communities around the country through the NNLM. The mission of the NNLM is to advance the progress of medicine and improve public health by providing the public with equal access to biomedical information to make informed decisions about their health. This was particularly important during the COVID-19 pandemic, when the NNLM partnered with critical NIH-initiatives such as the All of Us program and the NIH Community Engagement Alliance (CEAL) initiative to bring critical information support to communities in a manner that was congruent with community trust.

Research at the speed of the pandemic goes best when it leverages existing community investments and establishes research assets such as standards to promote emerging forms of scientific communication. NLM facilitated early access to research results from NIH-funded research specifically related to COVID-19 with the launch of the NIH Preprint Pilot (https://www.ncbi.nlm.nih.gov/pmc/about/nihpreprints/), which tests the viability of making preprint articles discoverable and available through PubMed Central (PMC) (https://www.ncbi.nlm.nih.gov/pmc/), NLM's full-text archive of biomedical literature. NLM collaborated with more than fifty publishers and scientific scholarly societies to develop the Public Health Emergency COVID-19 Initiative (https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/), which made COVID-19 and coronavirus-related articles freely accessible through PMC in machine-readable formats that permitted research reuse and analysis. NLM also developed the Global Health Events archive (https://archive-it.org/collections/4887), which captures and preserves more than twelve thousand pieces from a broad range of web-based content about the COVID-19 pandemic in an internet archive that will be available for future generations to study.

4. Accelerating discovery and data-powered health

NLM is guided by a ten-year strategic plan (https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.html) focused on accelerating discovery and data-powered health. NLM's strategic plan enumerates three goals:

(1) Accelerate discovery and advance health through data-driven research.
(2) Reach more people in more ways to enhance dissemination and engagement.
(3) Build a workforce for data-driven research and health.

Like other NIH Institutes, NLM is a research engine comprising an Intramural Research Program (IRP) (https://www.nlm.nih.gov/research/index.html and an Extramural Program (EP) (https://www.nlm.nih.gov/ep/index.html) that fund studies and projects all around the country. NLM's IRP comprises a computational biology branch and a computational health research branch to develop and apply computational approaches to a broad range of information problems in biology, biomedicine, and human health. Each of these branches rely heavily on the ability to make known and share common labels for clinical data and biomedical phenomena. NLM plays an enormous role at NIH in supporting artificial intelligence research and its application to biomedical discovery and clinical care.

Each goal relies heavily on NLM's ability to have common ways of labeling challenges using standards.

One example of work conducted by NLM researchers includes LitCovid (https://www.ncbi.nlm.nih.gov/research/coronavirus/), a resource that applies Artificial Intelligence (AI) and machine learning to scientific and medical literature. LitCovid's search engine and results reporting structure is now the core that runs NLM's PubMed searches. After an individual initiates a query, LitCovid returns a series of matching citations. LitCovid uses a learning-to-rank algorithm driven by AI to match these extracted citations to the citations that would be presented so what is presented to the user is a list of citations ordered by best matches. This is important because previously a typical search of PubMed generated tens of pages of results most of which were never viewed. Some NLM intramural research leaves the library and goes into the community; in one instance, NLM's researchers partnered with researchers in the community to improve wastewater-based surveillance. You may ask what role a library might have in regard to wastewater, and the answer is that analytical tools developed by NLM researchers helped estimate the amount of circulating SARS-CoV-2 variants found in the genetic fingerprints in wastewater.

NLM supports extramural research projects through its Extramural Program division. Take the NLM-funded work of Stanford University's Dr. Nigam Shah (https://reporter.nih.gov/search/OcWe4F9wSEeqy5_xsq81Jw/project-details/9759984). He wanted to automate responses to the very commonly asked question in clinical care, "What happened to your patient who looked like mine?" Dr. Shah brought together information not only from the clinical experts, but also from libraries, randomized controlled trials, guidelines, and algorithms to expedite that answer to clinical practice, an effort that is now operational at Stanford University Hospital.

An NLM-funded project run by Dr. Quynh Nguyen (https://reporter.nih.gov/search/e3Zubm02PEaaHvBWFQWYXg/project-details/10217256) at the University of Maryland leveraged Google Street Maps to better understand the community, characterize the built environment, and look at the relationship between the built environment, the natural environment, and health outcomes.

5. Supporting trans-NIH projects

NLM is vital in supporting projects across the entire NIH, including the NIH Common Fund's Bridge to Artificial Intelligence (Bridge2AI) (https://commonfund.nih.gov/bridge2ai) program. The goal of Bridge2AI is to generate new flagship biomedical and behavioral data sets that are ethically sourced, trustworthy, well defined, and accessible. The program has contributed about $125 million into national research to develop software and standards that unify data attributes across multiple data sources and across data types. Teams are creating automated tools to accelerate the creation data in accordance with the FAIR Data Principles, a set of guidelines that ensure that data is both Findable, Accessible, Interoperable, and Reusable and ethically sourced.

When running AI on data generated in studies involving volunteer participants, it is critical not to exploit those individuals' rights to their privacy. A project just released in summer 2021, referred to as the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) (https://datascience.nih.gov/artificial-intelligence/aim-ahead) project, helps investigators build partnerships around the country to leverage data science, clinical research, and community engagement to specifically address the challenges of health disparities. Researchers are looking to detect and improve health equities by providing and recommending more appropriate systems for scientific discovery. The AIM-AHEAD program will create the infrastructure to support interoperability at scale, which in turn will facilitate that standards are applied in a uniform fashion to research and clinical data across the country.

The goal of the NIH Common Fund's Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa) (https://commonfund.nih.gov/africadata) program is to harness data science for health discovery and innovation in Africa by leveraging technologies and existing investments in the sub-Saharan African region. These technologies and investments would develop solutions to the country's most pressing medical and public health problems by including the expertise of academic, government, and private sector partners. This program is valued at $75 million and has made a total of nineteen awards. A data coordinating center and open data science platform at the University of Cape Town fosters in-country data science to rapidly return discoveries to the community. Additionally, this program is learning more about languages used across this diverse continent and how to build structures to make these data sets fair and accessible while also preserving their original meanings.

6. NLM is a trusted provider of literature and data products and services

ClinicalTrials.gov (https://clinicaltrials.gov), NLM's repository of registered clinical trials and results, hosts more than four hundred thousand research studies and the results of over fifty thousand studies. Many of these never make it to publication, and including these results in its repository allows NLM to foster the public accountability and engender public trust in science by directly providing research results. PubMed and PMC are NLM's literature repositories that respectively provide access to bibliographic citations and full-text articles. Currently, there are 30 million citations in PubMed and more than 7 million full-text articles in PMC. NLM's genomic resources include dbGaP (https://www.ncbi.nlm.nih.gov/gap/), the database of Genotypes and Phenotypes, which brings together specialized studies for researchers to examine the relationship between genes (genotype information) and physical characteristics (phenotypes). NLM resources such as GenBank (https://www.ncbi.nlm.nih.gov/genbank/) support fully-computed annotated gene sequences, and the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra/) holds over thirty petabytes of original raw genomic sequences and make them available for studies. NLM had the first SARS-CoV-2 sequence available in its GenBank repository within a month of the first documented case of COVID-19. NLM is working to be as responsive to its users as possible and to make information more discoverable and accessible.

Currently, NLM is creating the Comparative Genomics Resource (https://www.ncbi.nlm.nih.gov/comparative-genomics-resource/) to support comparative analyses of sequenced eukaryotic research organisms. In the past, a specific genomic resource was created for every single different model organism from rats to zebrafish. Contemporary science is best supported when it is possible to traverse the genetic structures of these different organisms. NLM is looking to build pathways both between and within organisms and developing ways to make this genomic data accessible. Not only is NLM creating a robust interconnected system by using modern commercial cloud technologies and open data, but it is also breaking down silos to accelerate the generation, analysis, and sharing of data by fundamentally applying standards that make this genomic data fair and accessible.

NLM has made a sizable investment in terminology standards and health data standards to promote common approaches that support the NIH, science overall, and health care delivery. NLM curates, disseminates, and promotes critical health terminology standards, including those that support clinical care and research. One important effort that provides standardized nomenclature for clinical drugs intended for humans is available through RxNorm (https://www.nlm.nih.gov/research/umls/rxnorm/index.html).

The medical subject heading (MeSH) thesaurus (https://www.nlm.nih.gov/mesh/meshhome.html) is a repository of NLM's key terminology for understanding, classifying, and organizing medical and scientific literature. With MeSH, NLM leverages a large ontology and inheritance properties, the relationships of which improve the ability to locate and retrieve relevant literature.

One way NLM ensures that a hospital's clinical information system is achieving good health care is by providing access to internally-built normalized naming system for clinical phenomena and therapeutics. NLM is the designated U.S. repository for three separate terminologies that support the U.S. government's efforts to advance meaningful use of health information technologies. These resources include the standard nomenclature of medicine (SNOMED CT) (https://www.nlm.nih.gov/healthit/snomedct/index.html), the Logical Observation Identifiers Names and Codes (LOINC) (https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/LNC/index.html), and RxNorm (described earlier). In addition to these basic terminology standards, NLM also develops composite standards such as the Value Set Authority (https://vsac.nlm.nih.gov/), that are clusters or lists of codes and terms. These Value Sets help hospitals and health care providers to either make appropriate billing determinations or measure the quality of service by bringing together certain diagnostic features.

NLM contributes to the development of and helps disseminates the Fast Healthcare Interoperability Resource known as the FHIR standard. NLM supports the NIH to extend the research and clinical use of FHIR through a broad national partnership. When considering the complex interaction between a care provider and a patient, or what a radiology image looks like and how it can be used in comparison to prior images, or the challenge to organize the millions of books or thousands of clinical records, it becomes clear that investing in standards provides purposeful expressions of a common worldview, brings order to complex phenomena, and supports the exchange of meaning.

By combining automation and recognized standards, NLM is accelerating the application of standards to its core operations. For example NLM's MEDLINE 2022 (https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html) initiative is designed to make data-driven discovery more efficient and more responsive by implementing an "automated first" approach to assigning index terms from MeSH to citations for inclusion in its PubMed citation repository. This new process speeds up the indexing of citations in the biomedical literature from weeks and months to hours and days. At the same time, this effort allows NLM to re-deploy subject matter experts to the challenges of gene and chemical concept curation, thus preserving scarce human resources for specialized areas requiring human attention. NLM is continuously improving the automated indexing algorithm with human review and oversite.

NLM supports the use of Common Data Elements (CDEs) (https://cde.nlm.nih.gov/home), a strategy to promote research rigor by implementing consistent naming conventions for essential concepts. Using CDEs is a way to purposefully label data in a research process, and could encompass biological measures, relevant items, or scales, or describing to whom the research applies. It also supports better connection of findings across research activities. Like any other change in the way things are typically labeled, this change requires that researchers buy-in, which is advanced by demonstrating the value of addressing research topics in a more comprehensive manner. This extends the value of a research study beyond the resolution of a single hypothesis to build a research data infrastructure that can be reused to answer new questions in the future.

7. Addressing the challenge of unstructured data

What is NLM's future? The most important upcoming challenge to NLM is addressing unstructured data. NLM's history of building terminologies around the structure of an article, a digital record, or a sequence is now being taxed by the fact that they are increasingly facing unstructured and novel data types. This brings the challenge to consider the development and application of standards in a new way for the future. If standards are what helps humans make sense of phenomena and share that meaning with others, then it is critical to develop standards on increasingly atomic data. How can atomic data be formalized so NLM can bring together structures for information, ontologies, and sense-making tools? NLM believes its primary purpose is to support discovery and knowledge building from a foundation of core data, whether that is literature, clinical observations, or sequenced biological data.

To do that, NLM must create the building blocks to interconnect information into a resource that can be used and leveraged by society. The responsibilities of NLM are and will continue to be the acquisition, organization, preservation, and dissemination of information important to health and biomedicine. But biomedical and health information is emerging in more and more granular ways. More importantly, more people – be they scientists or research participants – want to be able to use this information to write their own story. Traditionally, the world expressed its data and other information through professionally-determined specific terminologies and standards. The ontology of organizing specific data units and connecting them created a story. Now, however, there are many stories, and those stories change over time. In addition, in this era of patient-centered care, the stories must be expanded to connect an individual's understanding of their own health needs and professionals' understanding of the illnesses those people have. While these understandings can overlap, they are not identical. NLM's opportunity as a library now is to provide the tools that help people write their stories. Partnerships and engagements are necessary to accelerate discovery to improve ways to make resources available – this cannot be done alone.

Standards could support ways to share worldviews, not necessarily having the same worldview, but aligning worldviews will be critical as NLM develops new ways of doing business. NLM is taking on the challenge to help people find meaning from data driven by their needs.

No longer is there a clear distinction between what constitutes health data and is considered non-health data; NLM should build tools that bring this all together. It requires thinking in new ways and scaling these ideas from cells to society while recognizing that there are overlapping ontologies. NLM should partner with users and patients, as well as scientists, policymakers, and clinicians, to support their mental models for reconstruction. This represents new challenges in new ways: How to determine what is truth? And what constitutes accuracy? Who gets to be privileged to say how one specific definition or problem should be labeled? And how can standards emerge from ephemera? The thousands and thousands of things that happen to a patient on any given day could, in and of themselves, be important. Sometimes they are very transient – consider a mother gazing at her child. Sometimes they are very enduring, such as childhood trauma exposure. It is important to understand how standards can help capture and label those things to make them meaningful.

And yet because of the size of data sets and the speed with which data is being generated, past approaches to generating and using standards may not scale. Not every data point will be equally relevant. This is a time for new models, including those based on probabilities of present value and future use, to guide investments in data standards.

Finally, engaging with scientific and lay communities to promote access to individual articles, scientific data resources, and standards is essential to help people make data meaningful in their lives. One of NLM's strategic goals is to make resources accessible to more people in more ways. This requires rethinking how to increase relevant access to the literature by applying appropriate data element descriptions and increasing the alignment between how a questioner might describe what they are looking for and how that item is defined or described as a resource in NLM's holdings.

NLM recognizes that two hallmarks of a trusted resource are how it answers a particular question and how often that same answer is repeated. The chain of trust must rely on both the correctness according to a given perspective of an answer and the strategies used to acquire the answer. NLM must evaluate the impact of algorithm tools such as machine learning algorithms on searches to be sure that the presentation of information remains true and trustable. Flexible strategies of exposing the literature will avoid privileging one view over another. And yet this requires user engagement to determine how the community creates and evaluates believable knowledge.

NLM's responsibility to create a library that inspires a future of public engagement, rapid response, highly-personalized treatment, and the development of novel therapies requires that it bring the concept and the values that standards enable to structure and to retrieve whole data types with many new meanings. Many partners are needed on this important journey.

About the Author

Patricia Flatley Brennan, RN, PhD, is the director of the National Library of Medicine (NLM) at the National Institutes of Health. NLM is a leader in biomedical informatics and computational health data science research and the world's largest biomedical library. Since joining NLM in 2016, she has positioned it as a global scientific research library with visible and accessible pathways to research and information that is universally actionable, meaningful, understandable, and useful. This ensures that scientists, policymakers, clinicians, patients, and the public can access biomedical information when and where they need it.

Dr. Brennan is the first nurse, industrial engineer, and woman to be NLM director. Her unique career path guides her approach to integrate health information management with artificial intelligence, machine learning, and deep learning to advance the future of health care.

NLM meets current and future challenges by combining information discovery support with cutting-edge research, training programs, and biomedical data and literature resources. NLM's resulting successes include data management tools to optimize patient care and health services and data-driven discovery to characterize the human genome for comparative and evolutionary analysis.

Under Dr. Brennan's leadership, NLM has grown its intramural and extramural research enterprise, extended stakeholders' access to credible and reliable health information, and further acquired and preserved globally available biomedical literature using modernized approaches to digital research and outreach. Its future prioritizes discovering new analytic and data science advances and recruiting and retaining talent that reflects societal diversity.

Dr. Brennan has received numerous accolades recognizing her contributions to her field. In 2020, she was inducted into the American Institute for Medical and Biological Engineering (AIMBE) College of Fellows. She is also a member of the National Academy of Medicine and a fellow of the American Academy of Nursing, the American College of Medical Informatics, and the New York Academy of Medicine.

Her blog is NLM Musings from the Mezzanine (https://nlmdirector.nlm.nih.gov). She can be reached by email at patti.brennan@nih.gov or by phone at (888) 346-3656. Twitter: @NLMdirector

This work was carried out by staff of the National Library of Medicine (NLM), National Institutes of Health, with support from NLM.