Realizing the Promise of a Digital Ecosystem for Science and Scholarship[i]

Michael F. Huerta, PhD, Associate Director for Program Development and NLM Coordinator of Data Science and Open Science, National Library of Medicine, National Institutes of Health

What is the National Library of Medicine and What Does It Do?

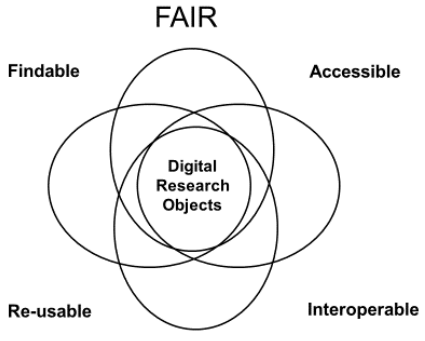
The National Library of Medicine (NLM) joined the National Institutes of Health (NIH) in 1968. As an NIH Institute, NLM conducts and supports research and training in information science, informatics, and data science. The NLM is also the world's largest biomedical and medical library, tracing its origins to the library of the Office of the Surgeon General of the Army in 1836[ii]. Today, in addition to its large collections of physical items, including books, journals, manuscripts, photographs, and other items, NLM is also home to hundreds of digital data and information resources. These include major resources, such as ClinicalTrials.gov[iii], which houses information about and from hundreds of thousands of clinical trials, and MedlinePlus[iv], which provides authoritative consumer health information, as well as smaller resources serving important niche purposes, such as TOXNET, which is a collection of databases and information products related to toxicology, environmental health, and hazardous substances[v].

Every day, NLM receives more than 10 terabytes of digital content from more than 3000 users, and delivers more than 100 terabytes to more than 4 million users, often through application programming interfaces. Users of these resources include researchers, healthcare providers, and the general public. The Library supports activities that engage all categories of users to make its resources known, understood and used. For example, to facilitate and enhance health information access to the general public throughout the country, including many rural areas, the NLM supports the National Network of Libraries of Medicine[vi]. Through its eight regional medical libraries, the Network reaches more than 6500 points of presence across the country in academic health science, community college, tribal college and public libraries, as well as other organizations, such as community health centers.

With medicine and biomedicine as its substantive scope, the NLM has been paying attention to that literature for more than 180 years. In 1879, John Shaw Billings, Director of the Library of the Surgeon General of the Army, and Robert Fletcher compiled and had published *Index Medicus*, an index of medical books, journals, and pamphlets[vii]. Stewarded by NLM, *Index Medicus* continued to be the authoritative index of the medical literature until 2004, but by 1964 the Library had started compiling citations and indexing much of the biomedical and medical literature digitally, in a database system called MEDLARS. In 1971, this database became available online (mostly

through university libraries) as MEDLINE[viii]. And, in 1997, PubMed, which included MEDLINE bibliographic data and more, was launched for free use by anyone on the World Wide Web. Today, PubMed contains more than 27 million bibliographic citations[ix].

In addition to bibliographic data, NLM has established databases and resources for particular types of research data. These include GenBank, a database containing all publicly available DNA sequence data with annotations[x], and others[xi]. As biomedical research becomes increasingly digital, the NLM will likely pay attention to research objects beyond research data and bibliographic data. Such digital research objects[xii] (DROs) might include software used to generate or analyze research data, as well as models, workflows, etc., used in research. (It is important to note that "pay attention to" covers a broad range of possible activities.) To DROs, such as citations or datasets, NLM applies: information science to curate acquired objects, informatics to compute in context on these objects, and data science to extract insight from these objects. After this, the DROs are findable (e.g., by having had metadata assigned), accessible (e.g., through publicly accessible databases), interoperable (e.g., by having adopted common data-related standards), and re-usable (e.g., by linking one set of objects, like publication citations, to another set of objects, like the datasets reported on in those publications). Thus, the processes of NLM applied to DROs make those objects compliant with the FAIR Principles[xiii]. In addition, NLM is interested in making DROs attributable (e.g., through PubMed Identifiers, PMIDs, or GenBank Accession Numbers) and sustainable (NLM considers this carefully before committing to hosting DROs).



The Importance of Being FAIR

When DROs are findable, accessible, interoperable, re-usable, and attributable, they make possible a more data-centric and open paradigm of science and scholarship, where the products and processes of research can populate an ecosystem that allows for others than those who produced specific DROs to add value to the science and scholarship around them. The starting point for bringing DROs into this more open ecosystem is to share DROs, especially data. Benefits of sharing data and other DROs can include (depending how interoperable they are with other data and tools): providing a deeper understanding of the publications and ideas with which they are associated, gaining additional insight by reanalysis of the data, boosting statistical power to answer particular questions by aggregating multiple datasets, ability to apply big data analytic methods, broadening opportunities for collaboration, enhancing accountability

(e.g., assessing reproducibility) and increasing return on research investments for research participants, science, and society.

Of course, there are also objections to such sharing, including: the costs incurred by making data and other DROs FAIR and sharing them, the possibility of others using the shared data to publish before the lab that produced the data does, concerns about intellectual property, patient privacy, and confidentiality, the fact that credit is accrued to investigators for papers published but not for data-related activities (so efforts not directed at publishing represent a net loss), the concern that data will not be understood and that the data will be misused. Most of these objections can be, and in some cases have already been, overcome (e.g., support of funders for data-sharing activities, use of embargo periods so sharing data happens after publication, and a host of policy and practice solutions to address intellectual property and patient privacy concerns). Yet, some, such as the lack of incentives for data-sharing, will require broad changes in the enterprise, while others still, such as the misuse of data, may never be fully resolved.

Conducting biomedical research in a more data-centric and open paradigm has repeatedly been shown to add significant value to science and scholarship. This was perhaps most famously demonstrated by the Human Genome Project, a 13-year project launched in 1990 that fundamentally changed the direction of biomedical research, and transformed our understanding of health and illness[xiv]. The spectacular success of the Human Genome Project was powered by collaborations and interactions of investigators in the ecosystem wrought of its findable, accessible, interoperable, re-usable data, tools, and infrastructure. Since then, many large-scale data-centric and open research initiatives have proven the value of these paradigms, including the Human Connectome Project and its subsequent related initiatives[xv], the NIH Human Microbiome Project[xvi], and the Genotype-Tissue Expression initiative[xvii]. And, the use of data-centric and open paradigms continues today as projects like All of Us[xviii] and the Adolescent Brain Cognitive Development Study[xix] get underway.

Data-centric & open paradigms have proven successful

From Concept-Centric and Closed to Data-Centric and Open

Despite many examples of data-centric and open approaches being used, most biomedical research is not conducted that way. For most research, the major public products are scientific papers reporting conclusions *about* the data, but the data themselves are almost never seen by others, much less shared with others. Thus, the currency of most biomedical research is not the data, but ideas and concepts about the data; it is concept-centric. And, since data are not made available to others, most research remains closed rather than open. This, however, is about to change. As society increasingly expects data from federally

funded research to be broadly accessible, as computational and communication technologies become ever more powerful, as the scientific opportunities afforded by open and data-centric paradigms become more obvious, and as bipartisan policy directives from executive and legislative branches of the federal government encourage data sharing, it is likely that data-centric and open paradigms will soon be used beyond the confines of one-off initiatives.

An important policy directive was issued on February 22, 2013 by the Director of the White House Office of Science and Technology Policy (OSTP), Dr. John Holdren, wherein federal agencies with annual research and development budgets exceeding $100 million were directed to increase public access to the results of the research they conduct or support, including both the publications and the data underlying those publications[xx].

The National Institutes of Health issued its plan for meeting the OSTP directive in February 2015[xxi]. Regarding access to research publications, the NIH already had a policy in place, and NLM had already developed PubMed Central as the infrastructure to provide public access to them. Starting in 2008, NIH required "scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to PubMed Central immediately upon acceptance for publication"[xxii]. Since the OSTP directive, several other agencies across the federal government have opted to use PubMed Central as the infrastructure for publications supported by those agencies. The NIH plan for making research data more accessible in response to the OSTP directive and in the interest of better science, is described below.

What is NIH Doing to Make Digital Research Objects FAIR?

As data and other DROs become more broadly accessible, it is important that NIH encourage and facilitate these objects being findable, accessible, interoperable, and re-usable. There are existing and ongoing efforts at NIH that do or could support the FAIR principles; some of these are described, below.

Findable – PubMed is a powerful platform for discovery of the biomedical literature, with coverage from 1946 to the present, and more selectively before 1946. The full contents of some 5600 journals are indexed with a curated, hierarchically organized terminology (MeSH[xxiii]) that allows for sophisticated search and retrieval of citations. This infrastructure could be leveraged to make other DROs, such as datasets, findable, perhaps by building on MeSH in ways that would be well suited to categorize and find datasets, with a pointer to the locations of the datasets.

Some publication citations in PubMed already link to datasets[xxiv]. And, it is expected that within the next year, investigators submitting papers to PubMed Central will be able to also deposit in PubMed Central the datasets associated with those papers. Both mechanisms allow data to be findable via the literature.

Data repositories make their constituent data findable and NLM supports a portal with information about, and links to, some 70 data repositories that are supported by NIH[xxv], and that allow data egress and ingress. This portal can be used to identify data repositories contain-

ing data of interest, and the search mechanisms available for the respective repositories can be used to find specific datasets.

Looking forward, NLM is now exploring specifications that could be used to describe datasets with appropriate metadata, ideally in ways that would be widely applicable across much of the diverse data landscape of biomedical research.

Accessible – As was mentioned, the 2013 OSTP directive to increase public access to research results supported with federal funding included increasing access to both research publications and research data. Highlights of the NIH plan to make research data more accessible include that policies would apply to all NIH mechanisms of research support, including grants, contracts, and intramural projects, and would apply at all levels of support, regardless of the amount of budget.

Data management plans would be expected from all applications and proposals for research support, and would provide information such as the type and amounts of data expected to be generated or collected, the data-related standards to be used, how data might be made available to others, provisions for re-use, etc. The data management plans would be part of the review process, with the review of the plan being able to affect the merit score.

Peer review of the data management plans would allow plans to be reviewed on a project-by-project basis, with the expertise and norms of that particular research community brought to bear on the assessment. Peer review of data management plans, with that review affecting the overall score of the review, will also raise the salience of the plans with applicants and reviewers, encouraging an appropriate level of consideration being paid to them both parties.

Interoperable – Data, and other DROs, are made interoperable with other DROs, tools and data resources through the use of standards. The NLM develops, supports the development of, and stewards widely-used standards, particularly for biomedical literature, healthcare information technology, and certain types of research data. These standards include[xxvi] terminologies, such as the Unified Medical Language System[xxvii] and SNOMED CT, coding systems like LOINC[xxviii], and metadata tagging specifications like JATS[xxix].

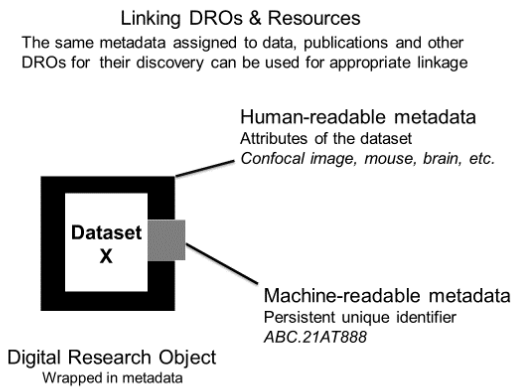
Across NIH, data repositories and major research initiatives supported by NIH, its institutes and centers, also specify data-related standards. Some of these, such as the Human Connectome Project, have incentivized investigators beyond that initiative[xxx] to adopt their data-related standards as their adoption allows investigators not supported by the initiative to rigorously compare their data with the initiative's data.

As the value of the use of common standards becomes more evident, institutes and centers of the NIH are increasingly communicating about and coordinating such efforts. For example, the NIH Clinical Common Data Elements Task Force maintains a conversation among all institutes and centers on this topic and is currently considering how to best harmonize related infrastructure, as well as developing and documenting best

practices for standing up common data element initiatives. The NLM has also developed web resources on behalf of the Task Force, including a portal to NIH collections of common data elements and related resources[xxxi], and a repository for common data elements used in research conducted or supported by NIH[xxxii]. The repository allows users to search for specific common data elements, by topic, funding opportunity announcement, etc., as well as serving as a platform to compare and harmonize similar but distinct common data elements. Such harmonization mitigates the unnecessary proliferation of these types of standards.

NIH is now launching pilot projects to create cloud instances as virtual spaces for data, analytic tools, repositories, and other DROs. Digital research objects that populate this NIH Data Commons will need to be compliant FAIR principles and certain standards[xxxiii], enhancing their interoperability.

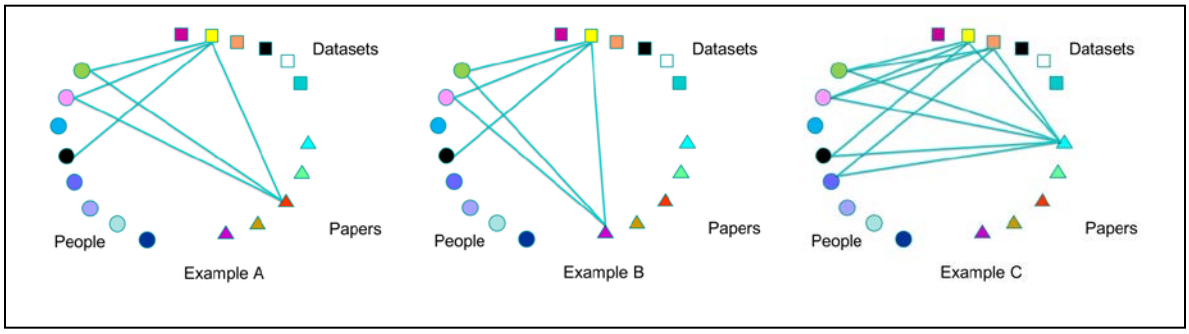
Re-usable – Digital research objects are re-usable and most useful when they are linked to each other. Such linkage depends upon metadata of the DRO (metadata are data or information about DRO). For example, if the DRO is a research dataset, it might have both human-readable and machine-readable types of metadata. The human-readable metadata might be a set of descriptors such as "confocal image, mouse, brain" to reflect the instrument source of the data, and the organism and tissue type from which the data were collected. The machine-readable metadata might be a string of alpha numeric characters. Ideally, the identifier is unique so it resolves to the intended object, and persistent so



that version of that object will not be lost over time, and its provenance tracked (i.e., changes to the object can be monitored by relating the identifier to identifiers of subsequent, modified, versions of it).

Linkage of DROs forms the basis of a digital ecosystem because such linkage allows DROs to interact in an automated and dynamic way. A simple example of such an ecosystem is shown below, where each disc represents a particular person, each square a particular dataset, and each triangle a particular scientific publication. And, each of these objects is associated with other specific objects through linkage of their respective persistent unique identifiers, shown as lines connecting them.

In Example A, below, three individuals participated in activities resulting in a dataset (i.e., they are the data authors), and two of them were authors on the paper. In Example B, the same dataset produced by the same three data authors resulted in a second publication by the same two paper authors. In Example C, the dataset and data authors from examples A and B, as well as an additional dataset authored by a subset of data authors in previous example and a



new author, served as the basis of a scientific publication for which all authors of both datasets served as paper authors.

At this time, through some of NIH's research data repositories, bibliographic data platform (i.e., PubMed), administrative systems for grants, investigator's identifiers, and other DRO identifiers, simple linkage like that illustrated here is possible for certain sets of DROs (e.g., as does the National Database for Autism Research[xxxiv]).

Now, imagine an ecosystem where all DROs have persistent unique identifiers and are associated with (linked to) all DROs appropriately. So, in addition to identifiers of particular datasets, publications, investigators and their affiliations, also present and appropriately linked in this ecosystem are identifiers such as those for the specific instrument used to collect the data (e.g., the particular magnetic resonance imaging scanner), the specific data-related standards (e.g., NIfTI-1[xxxv] data format), the software used to statistically analyze the data (e.g., AFNI_17.2.05[xxxvi]), pre-registered experimental protocols[xxxvii], etc. For any given DRO (whether a dataset, paper, data author, paper author, software tool, etc.), such an ecosystem would allow a person - or a computer - to be aware of all of the other DROs directly linked to it. Of, course, this awareness of associations need not be limited to the first degree of association, but higher levels; analysis of such higher dimensional relationships across networks of DROs could provide interesting insights about the nature of the science, itself. Comprehensive awareness of associations in and by the ecosystem could be maintained by something like a blockchain approach[xxxviii],[xxxix], with the DRO links representing the transactions tracked. Such an approach could provide a way to characterize the DROs and maintain provenance at-scale in an open, distributed, and reliable manner, adding significant value to the ecosystem of science and scholarship.

Data Science and Open Science at NIH: Looking Forward

With the retirement of Dr. Donald A. B. Lindberg as the Director of NLM, the Director of NIH asked a working group of the Advisory Committee to the Director to examine the nexus of NLM's purview and expertise and the future of data science and open science in biomedicine[xl]. Later that year, the working group issued a report[xli] with six recommendations, all of which were adopted by the Director of NIH. One of these was that "NLM should be the intellectual and programmatic epicenter for data science at NIH" and another that "NLM should lead efforts to support and catalyze open

science, data sharing, and research reproducibility, striving to promote the concept that biomedical information and its transparent analysis are public goods." Soon thereafter, Dr. Patricia Flatley Brennan accepted the position as Director of NLM.

Since Dr. Brennan's arrival at NLM, input has been solicited and received from many, including NLM leadership and staff, leadership of NIH institutes and centers, external experts, and the general public through meetings, workshops, committee deliberations, town halls, and published requests for information (some of these activities have been undertaken as part of the decadal NLM strategic planning process). Informed by these ideas from diverse perspectives, a view has emerged of the key issues that need to be addressed as NLM assumes the leadership role for data science and open science at NIH.

A clear and urgent priority for biomedical data science and open science is to engage with others across NIH to develop solutions for sustainability. As more large scale, large cohort studies, initiatives, and research programs are launched, the valuable data they produce must be housed, curated, and disseminated. Economies of scale and experience can be realized with a strategic enterprise approach, solving the same problem once rather than multiple times, converging on common standards, common architectures, coordination of acquisition activities around compute, and developing best practices for implementing and maintaining data assets and related infrastructure. Of course, it is important that trans-NIH approaches are flexible enough to meet the needs of particular studies, initiatives, and programs.

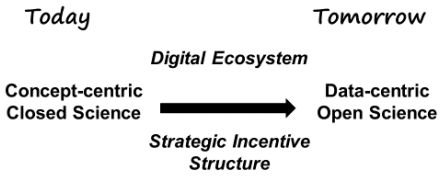
Another important contributor to sustainability is the use of evidence-based value assessment (e.g., cost/benefit analyses). Decisions such as those about: which data to keep, at what level particular datasets should be curated, how long specific datasets should be kept, which infrastructure should be invested in, which policies should be implemented and at what level of compliance, would all benefit by having an empirically-derived base of evidence for support. Such evidence could be used to develop criteria and heuristics for guidance about future investments in data, infrastructure, and policy.

Other priorities include the: 1) Strategic engagement beyond NIH, as data, science, and scholarship do not respect borders of nations, economic sectors or disciplines. 2) Development of a data-savvy workforce, including not only data scientists, *per se*, but data scientists cross-trained in biomedicine, biomedical scientist cross-trained in data science, both intramurally and extramurally. And, as data science and open science figure more prominently in NIH portfolios of extramurally-supported research, program officers, scientific review officers, and scientific policy staff of NIH must become more familiar with these areas. 3) Promotion of open science through changes in policies, engagement with the public around data and open science issues, and the development of tools designed specifically for use by public to facilitate their participation in research activities. 4) Research and innovation in data science and open science, developing new analytic approaches and tools,

solutions to challenges of curation at-scale, and exploration of various flavors of artificial intelligence to harness the dynamic and expanding ecosystem of science and scholarship.

Finally, it is important to note that the behaviors of individuals and the practices of research-related organizations in a closed, concept-centric paradigm of science and scholarship are very different than those required for an open, data-centric paradigm. For example, in the former data are not shared, while the latter depends upon sharing data (and other DROs). The flip from the former paradigm to the latter will require changes in incentives to both people and organizations comprising the biomedical research enterprise (e.g., universities, funders, publishers, professional societies, regulatory agencies, etc.). Ideally, these incentives would be distributed across the entire enterprise and would be strategically aligned with each other to be mutually and maximally reinforcing, and avoiding unintended consequences. Due to wide variations in how poised various biomedical research domains are for adopting a data-centric and open paradigm (e.g., genomics already is largely thus; not so for epidemiology), such strategic incentive structures would likely be best designed and developed domain-by-domain, rather than across all areas of biomedicine at once.

In closing, the cumulative biomedical knowledgebase and breathtakingly powerful scientific technologies available today present significant opportunities to understand health and mitigate illness. A digital ecosystem wrought of data science and open science promises to multiply these opportunities many-fold. With the right incentives in place, this promise could be realized in the foreseeable future.



References

[i] Supported by the National Institutes of Health, National Library of Medicine

[ii] https://www.nlm.nih.gov/about/briefhistory.html

[iii] https://clinicaltrials.gov/

[iv] https://medlineplus.gov/

[v] https://toxnet.nlm.nih.gov/

[vi] https://nnlm.gov/

[vii] https://www.nlm.nih.gov/services/indexmedicus.html

[viii] https://www.nlm.nih.gov/pubs/factsheets/medline.html

[ix] https://www.ncbi.nlm.nih.gov/pubmed/

[x] https://www.ncbi.nlm.nih.gov/genbank/

[xi] https://wwwcf.nlm.nih.gov/nlm_eresources/eresources/search_database.cfm

[xii] Defined here as digital instantiations or representations of products and processes of research

[xiii] https://www.nature.com/articles/sdata201618

[xiv] https://www.nature.com/news/human-genome-project-twenty-five-years-of-big-biology-1.18436

[xv] http://www.humanconnectome.org/about-ccf

[xvi] https://commonfund.nih.gov/hmp

[xvii] https://commonfund.nih.gov/gtex

[xviii] https://allofus.nih.gov/

[xix] https://abcdstudy.org/about.html

[xx] https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

[xxi] https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

[xxii] https://publicaccess.nih.gov/

[xxiii] https://www.nlm.nih.gov/mesh/

[xxiv] https://www.ncbi.nlm.nih.gov/projects/linkout/

[xxv] https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

[xxvi] https://www.nlm.nih.gov/healthit/snomedct/

[xxvii] https://www.nlm.nih.gov/research/umls/

[xviii] https://www.nlm.nih.gov/research/umls/loinc_main.html

[xix] https://jats.nlm.nih.gov/index.html

[xxx] http://www.nature.com/neuro/journal/v19/n9/box/nn.4361_BX2.html?fox-trotcallback=true

[xxxi] https://www.nlm.nih.gov/cde/

[xxxii] https://cde.nlm.nih.gov/home

[xxxiii] https://datascience.nih.gov/BlogFAIR

[xxxiv] https://ndar.nih.gov/

[xxv] https://afni.nimh.nih.gov/node/11

[xxxvi] *Ibid.*

[xxxvii] http://www.sciencemag.org/careers/2015/12/register-your-study-new-publication-option

[xxxviii] http://www.tandfonline.com/doi/full/10.1080/02763869.2017.1332261
[xxxix] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163477
[xl] https://acd.od.nih.gov/working-groups/nlm.html
[xli] https://acd.od.nih.gov/documents/reports/Report-NLM-06112015-ACD.pdf