Association of College & Research Libraries
50 E. Huron St. Chicago, IL 60611
800-545-2433, ext. 2523
acrl@ala.org, http://www.acrl.org

TO: National Institutes of Health (NIH)
RE: NIH Strategic Plan for Data Science
DATE: Monday, April 2, 2018

Submitted online at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-18-134.html

To Whom It May Concern,

On behalf of Association of College and Research Libraries (ACRL), I am writing to offer comments on the National Institutes of Health (NIH) Strategic Plan for Data Science as solicited in NOT-OD-18-134.

ACRL is the higher education association for librarians. Representing more than 10,000 academic and research librarians and interested individuals, ACRL (a division of the American Library Association) develops programs, products and services to help academic and research librarians learn, innovate, and lead within the academic community. We enhance the ability of academic library and information professionals to serve the information needs of students and researchers. For example, through a one-day workshop, ACRL presenters travel to campuses across the U.S. and train liaison librarians to enhance their skills with research data management.

Section 1: The appropriateness of the goals of the plan and of the strategies and implementation tactics proposed to achieve them

GOAL 1 Support a Highly Efficient and Effective Biomedical Research Data Infrastructure

ACRL applauds NIH's commitment to making research data usable to as many people as possible (including researchers, institutions, and the public) and to ensuring that all data-science activities and products supported by the agency adhere to the FAIR principles, which are essential for open science.

Objective 1-1 | Optimize Data Storage and Security implementation tactics.

The Strategic Plan calls for leveraging existing federal, academic, and commercial computer systems for data storage and analysis. This is a good tactic, but it should be executed with an exit strategy for each partnership in case a commercial partner changes its business model and it is no longer advantageous or efficient to use that partner's service. By the same token, federal and academic partners are sometimes threatened with loss of funding. NIH's partnerships should include mirroring to ensure continuity of service and the ability to easily change partnerships without being locked in with any one partner.

Support for technical and infrastructure needs for data security, authorization of use, and unique identifiers to index and locate data needs to address socio-technical factors and improve incentives for

researchers to share data. Deposited data that are not accessible due to their commercial potential should automatically be made openly accessible if the researchers cannot realize commercial viability within a reasonable time frame.

Data should regularly be exported and backed up to offline tape storage in a secure location to prevent loss during natural or man-made disasters, or in the event of large scale cyber attack.

Finally, the NIH Data Commons needs to ensure the security of highly sensitive human subjects data, including those data from small clinical trials, where an underrepresented population is put at risk due to the potential for re-identification.

Objective 1-2 | Connect NIH Data Systems

We strongly support aggregating and connecting NIH data and systems and Data Commons as described in Objective 1-2. These connections will add value to data in NIH systems. However, the implementation tactic "When appropriate, develop connections to non-NIH data resources" does not go far enough. We further recommend compliance with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and Object Reuse and Exchange (OAI-ORE) specification, as well as allowing APIs to access Data Commons. These will facilitate further potential connections that the NIH and its stakeholders can not yet envision. The NIH should not be the only arbiter of which NIH data connections are possible, and allowing these levels of access in the information architecture will free NIH resources from the burden of having to prioritize which connections to develop.

Finally, the Goal 1 Evaluation Metrics that rely on the quantity of data and computing resources by themselves do not fully address the effectiveness of the infrastructure. It is important that the linkages between data resources are accurate and precise. Determining if linkages are utilized and if resources are downloaded would be additional measures of forward progress on this goal and should not only be considered separately in Goal 2. Certifications such as the Data Seal of Approval, ISO 16363, and evaluation against the National Digital Stewardship Alliance Levels of Preservation, etc., and scheduled audits to demonstrate the reliability of the infrastructure and governance should also be evaluation metrics. Governance and continuity of service is especially important for federal sites. There should be contingency plans to ensure site availability even during federal government shutdowns, and system availability should also be another evaluation metric.

GOAL 2 Promote Modernization of the Data-Resources Ecosystem

Objective 2-1 | Modernize the Data Repository Ecosystem

The Strategic Plan states that NIH will focus "funding priorities on the utility, user service, accessibility, and efficiency of operation of repositories." These are all good goals, but sustainability of access to repositories must also be considered along with accessibility and user service. Plans to modernize the data repository ecosystem should also include contingency plans maintaining access to data and tools even during federal government shutdowns. This should be possible through the public-private partnerships and collaborations described in the *Objective 1-1* of the *Strategic Plan*.

Objective 2-2 | Support the Storage and Sharing of Individual Datasets

Objective 2-2 addresses an economy-of-scale problem with a solution that serves smaller data sets with pooled resources. It is important to note that even in individual laboratories, research can generate hundreds of terabytes of data, so we hope that the NIH Data Commons is prepared to absorb data on this scale. Yet, this only solves the storage part of the issue and does not address other challenges associated with data sharing. One major dilemma, not mentioned in the *Strategic Plan*, is that many researchers do not want to share their research data. A recent study by Victoria Stodden et al. (http://www.pnas.org/content/115/11/2584) published in the *Proceedings of the National Academy of Sciences* explored the effectiveness of journal policies that require authors to make their data and code available upon request. The researchers contacted 180 authors of published articles, and only 36% fulfilled their obligation to supply their research data and code. The article acknowledges that some data cannot be made publicly available for a variety of reasons, but also cites recent progress that enables greater sharing of sensitive data. The NIH should explore methods at its disposal to address the variety of factors that inhibit data sharing. In the interim, the *Strategic Plan* should clarify who reviews requests to access sensitive research data.

Objective 2-3 | Leverage Ongoing Initiatives to Better Integrate Clinical and Observational Data into Biomedical Data Science

ACRL supports NIH's commitment to data privacy and the ethical application of data science methodologies to health data sets.

Goal 2 Evaluation

In the evaluation considerations for Goal 2, it might be informative to also consider the proportion (rather than just quantity) of eligible publications in PubMed Central that have datasets deposited, measured over specific time periods (i.e. in decade measurements) for publications already deposited, as well as the proportion of datasets that are openly accessible, and the quantity of previously deposited datasets that are switched from restricted access to open access.

GOAL 3 Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools

Objective 3-1 | Support Useful, Generalizable, and Accessible Tools and Workflows Implementation Tactics

We support the range of partnerships and incentives described, but it should be clarified that private sector partners contracted to "refine and optimize prototype tools developed in academia to make them efficient, cost-effective, and widely useful for biomedical research" should also be committed to the FAIR principles described elsewhere in the Strategic Plan.

Objective 3-2 | Broaden Utility, Usability, and Accessibility of Specialized Tools

The *Strategic Plan* discusses the development and adoption of tools from outside of the biomedical sciences. An additional way to implement dissemination of advanced data management, analytics, and visualization tools is through existing infrastructure. Software Carpentry Workshops (https://software-carpentry.org/) and Data Carpentry Workshops (http://www.datacarpentry.org/) are useful for teaching

appropriate computing skills and data skills to researchers, whether this would be part of the curriculum for new professionals, or for continuing education and technical development for more established researchers.

Objective 3-3 | Improve Discovery and Cataloging Resources

The NIH clearly understands the importance of good indexing and a good discovery layer.

It would be useful if this section of the *Strategic Plan* linked out to examples of approaches for making data findable and accessible as it does not quite clarify how the search and analysis workspaces for authenticated users improves findability or accessibility in the Data Commons pilot.

More tools for interaction might be attractive to active users, but discoverability relies on understanding use cases, and how different segments of the targeted populations will search the Data Commons, and matching that with how data producers describe the research data and how data curators normalize and refine the metadata. We agree in the importance of collaboration to this approach towards a community-driven process for identifying and implementing optimal standards to improve indexing, understandability, reuse, and citation of datasets. To this end, we suggest that the NIH fund user studies as well as applied research in information seeking behavior with the explicit goal of improving the Data Commons.

Goal 3 Evaluation

The Evaluation metric described in the Strategic Plan suggests counting new tools developed or tools adopted from other fields, but these implementation tactics should also include plans for disseminating such tools, or for developing a community to update the code for the tools.

GOAL 4 Enhance Workforce Development for Biomedical Data Science

The use of data science approaches to potentially facilitate the NIH's ability to monitor demographic trends and address diversity gaps in its workforce is indeed promising, but lessons that NIH can take from Safiya Noble's recent work, *Algorithms of Oppression* (https://nyupress.org/books/9781479837243/) indicate that quantitative analyses also risk forcing individuals into demographic categories that do not match their identities. For example, when trying to understand gender ratios of professionals in medicine and life science, it would be a mistake to rely on data that only provided research subjects with binary male/female options for gender, or even to provide a third option such as "other" or "prefer not to say," as all of these options limit the range of human experience into two or three pigeonholes for the sake of sufficiently normalized data to satisfy an algorithm and an overly-specific research question. Therefore such research should recognize intersectional identities and non-binary gender identities or risk data inaccuracy and the codification of new inequitable structures.

Objective 4-1 | Enhance the NIH Data-Science Workforce

We support the idea of training programs to improve knowledge and skills of NIH staff in areas related to data science. To that end, as described in our response to *Objective 3-2*, we again recommend an examination of existing programs such as Software Carpentry Workshops and Data Carpentry Workshops. These are useful for teaching appropriate computing skills and data skills to researchers, as curriculum for new professionals, or for continuing education and technical development for more

established researchers and program managers. The NIH Data Fellows Program also sounds like an effective method of promoting interdisciplinarity, in addition to the benefits described in 4-1.

Objective 4-2 | Expand the National Research Workforce

We strongly support the diversity enhancing efforts in data science, like the example (https://www.ncbi.nlm.nih.gov/pubmed/28439180) cited in the *Strategic Plan*. We also support workforce development through data management education for students as well as training for established professionals. Data management services in academic libraries are growing, as are efforts to increase data management and data science skills of new professionals entering the workforce. ACRL currently offers relevant professional development through its Research Data Management Roadshow (http://www.ala.org/acrl/conferences/roadshows/rdmroadshow). Furthermore, librarians at Purdue, Virginia Tech, James Madison University, and University of Illinois at Chicago, to name but a few teach for-credit graduate level courses in data management and visualization. Courses such as these should be explored for inclusion in curriculum for certain professions working with life science data. Finally, there are many established graduate level courses to teach the medical community to use GenBank and Protein Data Bank. Partnering with experienced educators, trainers, and facilitators already teaching these tools, methods, and skills would be an efficient and effective means of disseminating advanced data management, analytics, and visualization tools.

Objective 4-3 | Engage a Broader Community

ACRL supports NIH's commitment to community engagement through expanded access to non-research academic organizations, community colleges, and citizen scientists.

GOAL 5 Enact Appropriate Policies to Promote Stewardship and Sustainability

Objective 5-1 | Develop Policies for a FAIR Data Ecosystem

A fairly comprehensive study by Cox et al. (https://doi.org/10.1002/asi.23781) finds that “libraries have provided leadership in [research data management], particularly in advocacy and policy development.” These areas of governance and policy development are a particularly strong area for libraries, and this would be a good area to develop partnerships. This is especially true given the alignment of ACRL’s core values and the NIH’s dedication to FAIR principles.

In order to promote long term data accessibility we encourage NIH to mandate that data from funded research be deposited in the Data Commons, or in the case of larger or highly specialized data sets, deposited into an NIH-approved dedicated data resource (as described in the *Strategic Plan* 2-2) as a condition of funding so that institutions do not try to assert sole rights to research data.

Objective 5-2 | Enhance Stewardship

The *Strategic Plan* cites the essential nature of data-science approaches for NIH to achieve its stewardship goals. Data stewardship is a key area for the kind of library partnerships described under Objective 4-2. Librarians have an active role in data stewardship in a number of data repositories and digital preservation consortia operated by across the United States as well as globally (e.g. the Edinburgh-based Digital Curation Centre http://www.dcc.ac.uk/).

This is a dynamic area, but the Cox study cited in our response to *Objective 5-1* above gives a sense of current and projected data management activities in academic libraries.

Section 2: Opportunities for NIH to partner in achieving these goals

There is significant overlap in the Strategic Plan with the goals of the Institute of Museum and Library Services National Digital Platform (https://www.imls.gov/issues/national-issues/national-digital-platform). This program addresses digital capability and capacity of libraries and museums across the US. It is the combination of software applications, social and technical infrastructure, and staff expertise that provide digital content, collections, and related services to users in the US. Given the differences in user communities, the overlap should not be construed as redundancy, but just as the research data themselves should not be siloed, nor should the strategic plans of various federal granting agencies when they are addressing the same challenge. One of IMLS's recently funded projects issued a report on strategic planning in data science for research libraries (http://d-scholarship.pitt.edu/33891/1/Shifting%20to%20Data%20Savvy.pdf).

Among other issues, the Coalition for Networked Information (www.cni.org) explores institutional, national, and disciplinary-level strategies and systems related to effective storage, preservation, and access to research data.

Section 3: Additional concepts that should be included in the plan

In order to ensure long-term usability of research data, it is not enough just to leverage cloud providers. While these providers usually provide effective redundancy and economy-of-scale, it is necessary to couple these with stable and transparent governance. When government services are transferred to the private sector there needs to be tangible benefit to citizens.

The NIH Data Commons should be developed with adherence to ISO 16363, also known as the Trusted Digital Repository (TDR) Checklist. This will make the NIH Data Common more resilient in the face of changes to business models by commercial partners, and in the face of changes to laws and regulations by the Legislature and the Executive Branch.

The ethical concerns that underpin most data science work should be given more attention in the *Strategic Plan*. Publications such as *Algorithms of Oppression* and *Automating Inequality* provide some examples.

Section 4: Performance measures and milestones that could be used to gauge the success of elements of the plan and inform course corrections

The NIH Data Commons should be immediately undergo a self-audit with the Trustworthy Repositories Audit & Certification (TRAC) tool, to be completed in no more than two years, followed by an

independent ISO 16363 audit within four years. These performance measures will inform course corrections to ensure the reliability, commitment and readiness of NIH for long-term data preservation.

Section 5: Any other topic the respondent feels is relevant for NIH to consider in developing this strategic plan

No comment

Section 6: Respondent information

Name

Mary Ellen K. Davis

Email Address

mdavis@ala.org

Type of Organization

Academic institution
Scientific research organization
Private sector
Health professional
Professional society
Advocacy group
Patient community
Government agency
Member of the public
Other

On behalf of the Association of College and Research Libraries, I urge you to seriously consider these recommendations so that the NIH can best support next-generation data science challenges in health and biomedicine.

If you have any questions about these recommendations, please do not hesitate to reach out to me at mdavis@ala.org or 312-280-3248.

Mary Ellen K. Davis
ACRL Executive Director