

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
DATE: Thursday, January 19, 2017
RE: Request for Information on Strategies for NIH Data Management, Sharing, and Citation

Submitted online at <http://osp.od.nih.gov/content/nih-request-information-strategies-nih-data-management-sharing-and-citation>

To Whom It May Concern,

On behalf of the Association of College and Research Libraries (ACRL), I am writing to offer comments on supporting data management, sharing, and citation.

Section I. Data Sharing Strategy Development

NIH recognizes that many factors must be considered when determining what, when, and how data should be managed and shared. These factors include, for example, the purpose for sharing, supporting data re-use and reproducibility, maturity of the science, the infrastructure uniqueness of the data, and ethical considerations.

The NIH seeks comment on any or all of the following topics to help formulate strategic approaches to prioritizing its data management and sharing activities:

1. The highest-priority types of data to be shared and value in sharing such data (Maximum: 250 words)

The Association of College & Research Libraries is the higher education association for librarians. Representing nearly 11,000 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. As reflected in our previous support for governmental policies and legislation that facilitate open access and open education -- including the NIH Open Access Policy, the Office of Science and Technology Policy mandate, and the Fair Access to Science & Technology Research Act and Federal Research Public Access Act bills -- ACRL is fundamentally committed to the open exchange of information to empower individuals and facilitate scientific discovery. Too often, the data and articles resulting from research remains locked behind paywalls or siloed in proprietary computer systems. In order to unleash the power of this information and truly accelerate discovery, we need to ensure that research outputs are made immediately available to the global public, and that people are fully empowered to use it in new and innovative ways.

At present, data underlying publications are of high priority. Sharing the data underlying publications in open and machine-readable formats through trust-worthy repositories (e.g., DSA/WDS certified, <http://www.datasealofapproval.org/en/news-and-events/news/2016/11/25/wds-and-dsa-announce-uni-ed-requirements-core-cert/>) should meet the minimal level of data sharing

requirements. These data should include the associated documentation necessary for access, interoperability and reusability, such as data dictionaries, code and computational details. Ideally, the workflow associated with the project should be made available in a format that encourages reproducibility.

2. The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications (Maximum: 250 words)

To some extent, the question of how long the data should be accessible will depend on the area of research and the characteristics of the resulting data. As such, these time frames should be community-driven determinations.

There are important implications for sustainability in this recommendation. It would be regrettable to see a situation develop where NIH-funded investigators must in future purchase access to data that was produced through NIH-funded research. Ultimately, a viable solution may require a creative consortial approach to maintaining long-term access to these data resources.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers (Maximum: 250 words)

Currently, cultural norms and perceived cost are significant barriers to data stewardship and sharing. While multiple approaches to this issue may increase odds of success, a significant factor is the expectation of the funder. In order to help shift researcher behavior, funder incentives, support, and requirements that encourage good data management and sharing practices is required. For example, grants using established standards for metadata, and best practices for documentation (both which facilitate the reuse of data) should receive continued funding as they offer evidence or responsible management of research data.

4. Any other relevant issues respondents recognize as important for NIH to consider (Maximum words: 250)

ACRL encourages NIH to continue to recognize and promote librarians as partners and sources of expertise with respect to the documentation, organization, preservation, stewardship, and curation of data.

SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

Currently, NIH grantees are required to report “other products of the research,” including data, databases, and software, in section C5a of their annual RPPR submission (http://grants.nih.gov/grants/rppr/rppr_instruction_guide.pdf). However, limited guidance is available on how data, databases, and software should be reported or cited.

NIH recognizes that data and software citation indicates proof of productivity that translates to publications and patents. More thorough reporting of data and software products in the RPPR and in Competitive Grant Renewal applications may strengthen documentation of productivity and may also identify projects and investigators who most effectively share data and software.

The NIH seeks comment on any or all of the following topics:

1. **The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing** (Maximum words: 250)

No comment.

2. **Important features of technical guidance for data and software citation in reports to NIH, which may include:**

a. **Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI) *** (Maximum words: 250)

ACRL encourages NIH to sign on to the FORCE11 Data Citation Principles:

<https://www.force11.org/group/joint-declaration-data-citation-principles-final>. Data citation is an important component of data sharing and incentivizing the practice. Additionally, NIH should explore the work available on issues of data citation implementation as presented in Starr, et al. (2015), <https://doi.org/10.7717/peerj-cs.1>.

* (DOI: <https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en>)

b. **Inclusion of a link to the data/software resource with the citation in the report** (Maximum: 250 words)

Barring instances when the data are protected, this should be a requirement. In keeping with the recommendations in 2a on implementing data citation, a persistent URL should be associated with a persistent identifier. This link should be part of the citation, and should resolve to a landing page specific to the resource (Starr et al., 2015; Fenner et al., 2016, <http://dx.doi.org/10.1101/097196>). This practice enables the findability that should be associated with all data/software resources and permits the gatekeeping that may be necessary in the case of restricted resources.

c. **Identification of the authors of the Data/Software products** (Maximum: 250 words)

In line with recommendations from FORCE11, it is critical that data/software creators be identified and credited for their work. This is one crucial step towards larger cultural changes within disciplines towards recognizing data/software as scholarly products in their own right.

d. **Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately** (Maximum words: 250)

This is an evolving topic of discussion as the systems that host these data evolve and become more sophisticated. Documentation that clearly explains the level of granularity is one way to allow for diverse practices, and in the absence of an agreed upon standard, is necessary. Research fields or communities may also have their own requirements for granularity unique to their research needs.

e. **Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed** (Maximum words: 250)

We encourage NIH to recognize institutional and subject repositories as acceptable sharing platforms for data. Recognize that repositories that align with ISO 16363 provide a level of trustworthiness that is ideal.

3. Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications
(Maximum: 250 words)

While multiple approaches to this issue may increase odds of success, a significant factor is the expectation of the funder. In order to help shift researcher behaviour, funder incentives, support, and requirements that encourage good documentation and sharing practices that make the foundational findings of research findable, accessible, interoperable and reusable is required.

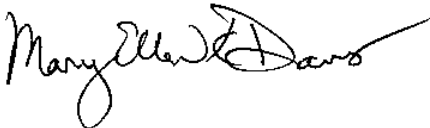
A possible incentive could be funding projects that are focused on data and software reuse, or otherwise rewarding projects that produce data and software that meet the highest levels of effective sharing as demonstrated by their degree of reuse.

4. Any other relevant issues respondents recognize as important for NIH to consider
(Maximum: 250 words)

As noted in Fenner et. al (2016) scholarly data repositories, which are often within the remit of the University Library, deal with issues of data and software citation on a regular basis. We encourage NIH to reach out to members of this community who have developed practical expertise in these areas, and to consider librarians as active partners in their efforts to implement effective data and software citation. ACRL is happy to work with NIH as a bridge to the academic and research library community, helping to build effective collaborations and partnerships between communities.

On behalf of the Association of College and Research Libraries, I urge you to seriously consider these recommendations so that the NIH can increase the re-use of data created through its funding. 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.

Sincerely,



*Mary Ellen K. Davis
ACRL Executive Director*