On behalf of AAU, APLU, and COGR, final response to
https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html

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

As recommended by the National Academies in their report, Sharing Clinical Trial Data, NIH should convene multi-stakeholder groups to determine the infrastructure, policies, and workforce needed for responsible data sharing. The convenings should include committees of faculty within different subfields to help determine what data is truly valuable and for what time limit within their subfield/discipline. In instances where data may have value to multiple disciplines, it will be important to convene multidisciplinary stakeholder groups. High priority should be given to those data sets that are deemed to be most useful to the research community and public.

NIH should focus on supporting those communities that have self-organized to share data; they have done so because that data is valuable enough for their efforts. Examples might include epidemiological data critical to public health emergencies such as Zika, genomic data (GeneBank), chemical structures (PubChem), and others.

As an initial priority and as a means of quality assurance, high priority data should be defined as data that has undergone peer review, one form of which is data that underlies published articles. For the long-term, the community will need to identify, assess, and deploy best practices and policies to ensure data sets are of high quality, discoverable, accessible, and usable.

High priority data should also include data obtained from rare, unique and shared equipment and large research projects which involve multiple investigators (i.e. Hadron collider, telescopes, satellites) and data collected by federal agencies as on Data.gov.

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)

There is such variability in data types and potential usefulness of data that NIH should allow researchers in their Data Management Plan to set reasonable lengths of time data will be available.

As for long-term maintenance of data, there should be consortia of the federal government and universities to manage a few repositories. The costs for managing these repositories should be supported by the research sponsors.

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

The monetary burden of first deposit into the repository should be part of the original award and should be direct costs on the grant. These costs are not insignificant; the Royal Society estimated that data sharing could require resources on the order of 1-10% of a funded project.

The burden of cost for long term storage past 5-10 years should not be on the primary researcher/institution, but should have a sustainable funding model, either funded direct (charge to the grant or federal repository), recoverable indirects, or to the secondary user.

One concern for long-term storage of data is the sustainability of repositories. Having a few repositories that are federally funded and managed, perhaps in partnership with universities and the private sector, may help ensure that deposited data continues to be accessible to the public.

It will also be important to ensure there is harmonization in data sharing standards, policies, and practices across federal agencies to reduce administrative burden on research projects. This recommendation is in line with Executive Order 13563 of January 18, 2011 which called for greater coordination across agencies to reduce costs and simplify and harmonize rules; the National Academies report “Optimizing the Nation's Investment in Academic Research: A New Regulatory Framework for the 21st Century”; and the GAO report entitled “Opportunities Remain for Agencies to Streamline Administrative Requirements” (GAO-16-573: Published: Jun 22, 2016).

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

Since most of the benefit accrues to secondary researchers and because these data are seen as a common good, mechanisms should be in place to move some of the cost burden to the community of users. As a common good, it should be the primary responsibility of the federal government to support, fund and manage data repositories.

There should be some formal set of rules defining triage and data discard. Not everything is worth saving and not everything worth saving is worth saving indefinitely. Those rules are probably related to frequency of use and time since last use. That is the essence of curation.

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)

Researchers may not want to share data sets until after accepted publication of their research, which might occur after the official end of the grant. Interim reports would not capture many of these post-publication data sets.

Additionally, sharing data as an "interim research product", perhaps released before peer-reviewed publications, may mislead research. Interim data releases could have data whose interpretation and analysis change once the full data set becomes available. We recommend allowing the researcher flexibility in managing interim data releases, including an option to hold the data until publication or some reasonable time period after the end of a project.

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

Requirement of including a DOI is reasonable.

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

Requirement of a link to the data set is reasonable.

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

Acknowledgement of author contributions is important to the scholarly enterprise. If reporting this information to NIH will be made visible to a wider community bestowing benefits and credit to each author, then it is a worthwhile and reasonable request and may outweigh the associated burden of reporting this information to NIH.

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)

It seems reasonable that data that underlies a publication should be individually cited and reported.

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)

Providing the link to the data set as suggested in #b would seem to be sufficient, but providing the name of the repository and link to the home page would not be overly burdensome.

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)

In new applications, there could be a checkbox asking if the researcher has made all applicable data from past awards available.

NIH could convene journal editors, who have an important role to play in incentivizing data sharing, and help facilitate conversations around harmonizing their data sharing guidelines.

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

Faculty will need training and support to share data responsibly, to ensure that data that has national security implications, privacy implications, or which are proprietary are handled appropriately. There is also the concern that while on its own, an individual data

set may not have national security or privacy issues, when combined with another data set it becomes a risk (i.e. mosaic effect). There should be some oversight to help ensure that this does not occur.

There needs to be investment in the development of standards, governance, metrics, data planning practices, tools, and repositories needed for creating high-quality, reusable data and documentation for the public and scientific community.

The Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) have created a working group comprising campus representatives, including provosts, senior research officers, chief information officers, librarians and compliance officers to review issues relating to opportunities and challenges for universities as they move to implement new public access requirements. We look forward to working with NIH, the OSTP, and other Federal research agencies, over the next few months to further clarify data sharing standards and policies.

The opinions outlined above are on behalf of the Association of Public and Land-grant Universities (APLU), the Association of American Universities (AAU), and the Council on Governmental Relations (COGR). Any follow-up questions regarding these comments may be directed to: Tobin Smith, AAU (toby_smith@aau.edu); Kacy Redd, APLU (kredd@aplu.org); or Jacquelyn Bendall, COGR (jbendall@COGR.edu).