



FASEB

Federation of American Societies
for Experimental Biology

Representing Over 130,000 Researchers

301-634-7000
www.faseb.org

9650 Rockville Pike
Bethesda, MD 20814

March 13, 2020

Lisa Nichols, PhD
Assistant Director for Academic Engagement
National Science and Technology Council
Subcommittee on Open Science

RE: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research [FR Doc. 2020-00689]

Transmitted electronically via e-mail: OpenScience@ostp.eop.gov

Dear Dr. Nichols,

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide feedback on [Request for Comments](#) (RFC) seeking input on a draft set of desirable characteristics of data repositories used to locate, manage, share, and use data resulting from federally funded research released on January 17, 2020. As a coalition of 28 biological and biomedical scientific societies collectively representing over 130,000 individual scientists and engineers, FASEB recognizes the critical role of data preservation and accessibility in facilitating scientific rigor and reproducibility.

Our comments in response to the draft characteristics posted by the Office of Science and Technology Policy (OSTP) in this RFC reiterate FASEB positions on core issues of interest to our members: coordination and harmonization of data policies across federal agencies, long-term sustainability of data management, and data accessibility while upholding the standards of scientific peer review.

I. Desirable Characteristics for All Data Repositories

A. Persistent Unique Identifiers:

To ensure that large volumes of data are of the greatest potential utility to researchers, clinicians, and the public, FASEB supports the use of unique identifiers. Consistent with the [FAIR principles](#) (Wilkinson et al., 2016), identifiers such as digital object identifier (DOI), accession numbers, or ORCID ID will aid in researchers' ability to identify and access data even if the metadata URL has changed since its publication. Potential efforts OSTP may want to consider

include: (1) developing tools to improve search functions and the aggregation of data, and (2) creating formatted citations associated with each dataset, preferably including a DOI. These improvements can also incentivize researchers to share quality data. Greater reuse and citation of datasets will encourage investigators to optimize the formatting and organization of their data and metadata for reuse by others, rather than merely fulfilling minimal reporting requirements.

Successful implementation of interoperable data management practices will require training for all research team members. Institutions should also foster an atmosphere where quality data management and appropriate data sharing are standard practice. To establish and maintain such an environment, institutions should encourage investigators to collaborate on improving data practices within their discipline and ensure data management resources can be easily identified and utilized.

B. Long-term sustainability

Responsible data stewardship requires a long-term plan. Data management plans (DMPs) are an important tool for promoting quality data management and appropriate data access.

Consideration of potential opportunities for data reuse at project initiation also ensures retention of all appropriate data. Inclusion of DMPs as a component of grant applications clarifies expectations between investigators and research sponsors. Flexibility and adaptability can be achieved by having individual investigators develop a DMP specific to their research area, data types used, and resources available. Research sponsors may also enlist DMPs for secondary uses of benefit to the research community, such as identifying common resource needs and other barriers.

To attain the benefits of DMPs without creating unnecessary burden, DMPs should be short summary documents that address the most essential aspects of data management and access. In most cases, a brief (one-to-two pages) summary should be sufficient, although additional information could be requested just-in-time for select circumstances. FASEB recommends the following DMP content requirements across federal agencies:

- Description of the data and metadata to be collected
- Overview of data management practices
- Summary of any data sharing restrictions (confidentiality, intellectual property, etc.)
- For shared data, information about when it will be made available, where it will be stored, how it will be maintained, and how others will be able to find, access, and reuse it
- For data that will not be shared, justification for not making it accessible (which many include considerations of feasibility, data utility, etc. as well as sharing restrictions)

C. Metadata

Research reproducibility depends upon rigorous experimental design and appropriate analysis of resulting data. Metadata provide essential information for determining appropriate use.

Unfortunately, robust, consensus-based metadata standards do not exist for many fields or many data types. Furthermore, minimal metadata standards have not been established or deployed

across all scientific agency databases. Therefore, FASEB encourages OSTP to support the development of community-based metadata standards. Scientific societies can support these efforts by identifying and convening subject matter experts and disseminating consensus standards. We also urge OSTP to foster trans-agency development of automated tools for assigning metadata to files and datasets. Development of these tools can begin before or in parallel with the establishment of consensus standards. Automation would streamline efforts associated with tracking and updating metadata to meet current standards, accelerating adoption of new standards and changes to existing standards reducing investigator burden.

Repository tools are also indispensable for promoting data citation and attribution to investigators responsible for generating datasets. Data citation enhances the findability and accessibility of datasets and incentivizes data sharing. Currently, tools supporting citation of journal articles are more robust and readily available than tools for data citation. If researchers must look up a new citation format and manually assemble citation information, they will cite the associated journal article because it is simpler and more expedient. Tools that export dataset information, similar to what is provided for articles indexed in PubMed, lower the “activation energy” for data citation and provide a visible reminder to do so. To further promote such recognition, OSTP may want to consider collaborating with scientific journals to develop manuscript submission tools that prompt, facilitate, and standardize reporting of repository use.

F. Free & Easy Access and Reuse

FASEB understands and supports the development of an IT ecosystem that facilitates access to large, high-value datasets, as this will ensure these datasets are consistent with FAIR principles.

To effect positive change, research sponsors must carefully balance the costs and benefits of data access when developing and amending policies. Making datasets accessible – including the skilled human labor necessary to prepare and maintain data and metadata, technological infrastructure, and continued development of effective search platforms – is costly. Some datasets have little value for reuse or a short “shelf-life”; requirements to share and preserve such data could create inefficiencies in research funding and resource distribution. Therefore, FASEB recommends that sponsors ensure data access policies prioritize data with the highest potential for reuse

G. Reuse

The diversity of data types, research areas, and resources available make it challenging to identify data accessibility strategies that are practical and relevant for all fields of research, challenges that are further amplified within the biological sciences. Regular assessment of data utilization will allow investigators and federal agencies to evaluate usage and outcomes in the context of past performance and project future needs. Such utilization assessments would be further enhanced by the creation of time series data, when feasible. Analysis of user communities may also reveal patterns in how usage expands to new disciplines, thus informing scientific programs at federal agencies.

J. Common Format:

Data standards are necessary to ensure adherence to the FAIR principles; without standards, large volumes of data cannot be reused or even reassessed. Several issues that may hinder users from submitting data include limited data formats, heavy reliance on manual entry, and insufficient tools available to export and import data and metadata.

To encourage deployment of user-friendly platforms FASEB recommends coordinating with funding agencies such as NIH and NSF to develop metrics that evaluate and offer guidance about such barriers. Additionally, FASEB encourages OSTP and colleagues to measure the extent to which automation is incorporated in the submission process. Automated features such as auto-fillable fields and saved templates can enhance the submission experience and circumvent several sources of data corruption and loss.

K. Provenance:

Understanding the context by which data is obtained, processed, and analyzed is essential to its appropriate interpretation and application. Because datasets are often reformatted to pursue new research inquiries, data provenance allows researchers to trace newly designed or repurposed data back to their original settings.

Implementation of strong data provenance ensures data creators are held accountable for their work and enables systematic data tracking for a wide range of scenarios that utilize and apply research data. For example, researchers frequently share and adapt data for their individual purposes when collaborating with fellow investigators on research projects. With clear data provenance guidelines, end-users will be able to visualize how a specific dataset was derived and thus more appropriately employ the information that is suitable for their research.

FASEB supports responsible data management and encourages OSTP to engage with the stakeholder community to incorporate data provenance best practices across federal agencies.

L. Other relevant topics

The emergence of “big data” is allowing investigators to pursue more lines of inquiry that could ultimately lead to transformative discoveries. However, as larger quantities and more types of data can be combined in new ways, we must also be cautious of spurious correlations and “over-mining” of datasets. The Federation is concerned that analytical methods and tools do not always keep pace with research opportunities. Rigorous research practices will depend on coordinated efforts among federal agencies, and research stakeholders, ranging from single investigators to large institutions, to generate and support “big data” analytical methods and best practices. FASEB encourages OSTP to take the lead in coordinating these efforts to ensure parity across agencies and scientific disciplines.

II. Additional Considerations for Repositories Storing Human Data (Even if De-identified)

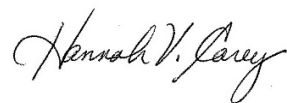
C. Privacy

Current U.S. policy frameworks and privacy proposals are insufficient to ensure the privacy of human research subjects in perpetuity. In comments on the [NIH Genomic Data Sharing Policy](#), FASEB stated that “de-identification cannot be guaranteed for certain types of data, including whole genomic sequences.” FASEB, therefore, recommended the consideration of alternative models to protect human research subjects, such as shifting from a privacy-protection paradigm to “one that provides research subjects with substantive legal protections against the misuse of or inappropriate access to their data.”

OSTP should also consider the risk of harm from inaccurate re-identification or speculation of the identities of participants and their outcomes. There are many other types of data misuse, and OSTP must proactively work with federal agencies and the research community to mitigate these risks.

FASEB appreciates the opportunity to provide input on this important topic. In addition to the comments provided in response to the specific elements of this RFC, links to recent organizational statements on this issue are provided below the signature line. We look forward to working with OSTP, federal research agencies, and other stakeholders on development of a feasible strategy to foster data sharing and reuse across scientific disciplines.

Sincerely,



Hannah V. Carey, PhD
FASEB President

Related FASEB Statements of Interest

1. [FASEB Comments in response to NIH Request for Information \(RFI\) on Draft Data Management and Sharing Policy and Guidance Documents](#) (Issued December 10, 2019)
2. [FASEB Comments on Draft NIH Strategic Plan for Data Science](#) (Issued April 4, 2018)
3. [FASEB Response to NIH RFI, “Registration and Results Reporting Standards for Prospective Basic Science Studies Involving Human Participants”](#) (Issued November 8, 2018)
4. [FASEB Response to NIH RFI, “Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research](#) (Issued December 10, 2018)
5. [FASEB Comments on Next-Generation Data Science Challenges in Health and Biomedicine](#) (Issued November 8, 2017)
6. [FASEB Statement on Data Management and Access](#) (Issued March 1, 2016)
7. [FASEB Response to NIH RFI: Metrics to Assess Value of Biomedical Digital Repositories](#) (Issued September 7, 2016)
8. [Comments on NIH RFI: Strategies for NIH Data Management, Sharing, and Citation](#) (Issued December 7, 2016)