NIH Request for Information on Strategies for NIH Data Management, Sharing, and Citation

Please review your comments before submitting

Submitter Name
John Tagler & Michael Mabe *If submitting comments on behalf of another individual, please submit the name and function of that other individual.

Name of Organization
AAP Professional and Scholarly Publishing Division & International Association of STM Publishers

Type of Organization
Professional Org/Association

Role
Other

Other Role
Our members are professional and scholarly publishers

Domain of Research Most Important to You or Your Organization (e.g., cognitive neuroscience, infectious disease epidemiology)
Our members publish journals and other scholarly material in all domains of research

Type of Data That You Primarily Plan to Generate and Share

Type of Data
Other

Human

Other Type
Our members enable the sharing of many types of data, both from human and non-human subjec

Repositories You or Your Organization Primarily Utilize (Maximum: 250 words)

1. The highest-priority types of data to be shared and value in sharing such data (Maximum: 250 words)
The highest-priority types of data to be shared should be those based in or connected to publications, as these have the greatest potential to improve users' understanding of research findings and enable the research community and practitioners to further validate and replicate findings in published articles. These are also data that are already shared in many communities of practice, and therefore are the lowest hanging fruit for encouraging their use and broader dissemination.

In looking beyond such data, it is critical to distinguish between data and various types of presentation of data and appropriately consider a researcher's rights to data generated in his or her research, as well as respect intellectual property protection and copyright laws. We refer you to the Data Publication Pyramid on p. 6 of the "Report on Integration of Data and Publications" (http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf). The report, written by a coalition representing researchers, publishers, libraries and data centers, is a comprehensive look at research data and sharing.

The need to expand incentives for providing broad and timely access to new data must be balanced with the need to preserve incentives for researchers to interpret and analyze their results through curation and peer-reviewed publication. NIH should also be mindful of precedents it may set regarding data management and preservation, including with respect to privacy. While the NIH's focus is on biomedical data, NIH's policies are likely to influence approaches taken in other research and funding communities.

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)
This is a complex issue with significant resource implications, which should be addressed by researchers themselves within their communities of practice. Publishers are already working with the communities we serve to develop standards for linking and sharing data where authors choose to do so, and to provide persistent links to external data collections.

Federal policies should take into account the differences between information products at different levels of the pyramid mentioned in our response to question 1. Information products at the top of the pyramid, those connected to publications, should be persistently preserved in perpetuity for the integrity of the scholarly record, whereas data at lower levels of the pyramid (e.g. some raw data) might not need to be preserved for as long. It will be key for NIH to work with all stakeholders, including primary researchers, secondary researchers, publishers, libraries and data centers, to create clear rules and protocols for the management and sharing of data. A collaborative approach will ensure that the needs of each stakeholder group are addressed and that the progress of science is not impeded.

3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers (Maximum: 250 words)
Unlike publications which can be counted in pages, datasets can be measured in mega-, giga- or even terabytes. Unlike publications, which undergo peer-review and editing, extensive formatting and tagging, data files come in a variety of forms, formats and levels of quality. It is difficult to control for quality, display and consistency, and there are no standards for which data should be preserved and how or where it should be managed.

The compliance costs of sharing data are significant, especially when compared with current practices. In order to maximize its usefulness, data should be tagged, metadata added and the data must be reviewed to determine what can be shared and where. In addition, there are significant costs associated with storage, distribution bandwidth and overall management and curation.

Initiatives must be carefully developed to support storage, dissemination, tagging, and validation. Success will depend on a collaborative approach that elicits buy-in from all communities and includes consultation and

contributions by key stakeholders to develop robust, sustainable and flexible standards. NIH must carefully consider how best to create incentives for data management and sharing, and provide support for such activities. Publishers stand ready to lend their expertise to such a collaborative process to provide value to the research community and to the taxpayer. NIH should not invest resources to recreate what is already being achieved by the private sector, but should leverage public-private collaborations to ensure continued innovations that contribute to the progress of science and innovation and help grow the American economy.

4. Any other relevant issues respondents recognize as important for NIH to consider (Maximum words: 250)
A federal role in expanding access to and the preservation of digital data could include partnering with the scholarly community for the identification of standards and best practices for data management and the interoperability of data repositories; creating clear rules for citation, modification and privacy; improving links between data, research grant reports and peer-reviewed publications; facilitating cyber infrastructure; and advancing policies and funding to ensure the long-term sustainability of data archives. Public access policies should be developed through voluntary collaborations with nongovernmental stakeholders, including researchers and publishers, university administrators, librarians and the public.

NIH could learn from initiatives already underway to standardize metadata and provide links between sources of scientific information. Issues related to expanding access, managing data, minimizing compliance costs and other policy questions are already being worked through, and we encourage the continued evolution of programs to improve data stewardship and public access to data. These include the Research Data Alliance (RDA), CrossRef, DataCite, Opportunities for Data Exchange (ODE), APARSEN and the NISO/NFAIS Supplementary Journal Articles Material Project, among others. Such collaborative approaches provide the best way forward towards broad access to and preservation of digital data.

NIH also needs to put systems in place to monitor and assess the use of shared data. In doing so, both costs and benefits should be considered. Such an assessment would also allow NIH to consider how to incentive or encourage the use of the data made available so that it has the desired impact on research and the public.

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)
Such increased reporting may create additional burdens on researchers. Publishers would be interested in taking part in discussions with funders and supported researchers to explore if there might be opportunities to reduce these burdens through ongoing projects or new initiatives and standards.

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)
The Digital Object Identifier (DOI), a unique code for each piece of content in a scholarly publication, was developed and adopted through a multi-stakeholder, community-driven approach. It is successful because the standard evolved in response to a real problem in scholarly communication and is providing practical benefits to users of published articles about research. Digital data standards are newer and still evolving.

Publishers have worked throughout the digital era to develop appropriate standards, persistent identifiers and protocols to enable seamless interlinking between publications through the development of the CrossRef organization and the use of standardized digital object identifiers (DOIs). CrossRef and DataCite have already been hard at work to extend these practices to data, but challenges remain. In order to minimize costs and maximize accessibility and usability of data, NIH should work with these existing initiatives and standards organizations like NISO to ensure the widespread adoption of both standardized DOIs and standard metadata protocols for data.

Potential exemplars include DataCite (http://datacite.org/), APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), Opportunities for Data Exchange (ODE, www.ode-project.eu), CoData (http://www.codata.org/), NISO/NFAIS Supplemental Journal Article Materials Project (http://www.niso.org/workrooms/supplemental), PARSE.insight (http://www.parse-insight.eu/), and ORCID (http://orcid.org/).

b. Inclusion of a link to the data/software resource with the citation in the report (Maximum: 250 words)
Any guidance as to the format and location of links, as well as acceptable repositories or locations to which a link might be directed, should be developed in consultation with stakeholders and consistent with the development of broadly-accepted community standards, as discussed above.

c. Identification of the authors of the Data/Software products (Maximum: 250 words)
The scholarly community already has a robust attribution and credit system with respect to peer-reviewed publication, including disambiguation tools like ORCID and systems to identify various types of contributions to a work. Existing systems and tools could be leveraged in a bi-directional manner by linking between datasets and publications on the one hand, and exploring a requirement with key stakeholders that all data which informs the analysis and conclusions of a peer-reviewed publication be cited and attributed according to community standards on the other. The federal government's role could be to help by promoting those standards and provide clear rules for the citation of datasets and acknowledgement of modifications to source data. Such standards should also promote unique and persistent identifiers for data and disambiguate researcher, institution and funder information in metadata. Over the past decade, publishers developed the Digital Object Identifier (DOI), a unique code for each piece of content in a scholarly publication, and similar identifiers are being developed by DataCite for data (www.datacite.org). The work of DataCite, CrossRef, ORCID, and DOE's Data ID Service should be leveraged to ensure data is appropriately archived and recognized as a primary research output.

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)

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)
For the persistence and integrity of the scholarly record, it is important that digital repositories be identified alongside the deposit of the information resource cited. Automated updates to limit the burden on researchers and preserve the integrity of the information would be helpful. Unambiguous identification could support succession planning and location services should the repository cease operation or the resource be moved at a future time.

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)
Many communities of practice are investigating the question of how to support researchers in sharing data and what kinds of incentives work. These studies should inform the development of NIH policies. Empirical evidence should be used to make evidence-based decisions about what efforts should move forward and how best to develop policy.

Involving a broad array of stakeholders in policy development and implementation will ensure the preservation of incentives for innovation and help improve information sharing and training within each stakeholder community, as well as incentivize the sharing itself. Stakeholders can help develop clear standards and guidelines for the availability of research data, certification and auditing of data repositories and metadata standards, which respect each community's standards and practices, working together to create universal policies that work for all communities. Stakeholder input is also important for the integrity of the scholarly record, including the creation of links between datasets and the scholarly publications that analyze and interpret the

data. Supporting such standards will improve researcher buy-in and compliance with requests for sharing. Developing guidance, in consultation with key stakeholders, to minimize the administrative burden on key stakeholders, would also improve compliance.

4. Any other relevant issues respondents recognize as important for NIH to consider (Maximum: 250 words)
Community-based policies: AAP/PSP and STM agree with the OSTP's Interagency Working Group on Digital Data that "data stewardship is best accomplished in a system that includes distributed collections and repositories maintained where the custodian has trusted community-proxy status with the relevant communities of practice." A critical component of any policy needs to be collaboration with researchers, publishers, librarians, universities and research institutions in an interconnected system based on community needs, standards and best practices. Each stakeholder community can contribute its expertise and ensure the creation of data management policies that reflect the different practices of individual research communities.

Confidential data: Research communities have gone to great efforts to develop standards that ensure research subjects are treated ethically and that confidentiality of data is preserved. NIH must make sure that it does not undermine these protections as it works to expand access to data.

Fraud: NIH will need to consider how it can work to detect data fraud. Tools can be deployed to analyze data that is "too perfect" from a statistical standpoint or to analyze images for manipulation, but next generation tools need to be developed to stay ahead of any efforts to mislead the public.

Validating data: Capabilities should be developed for validating data in terms of both quality and utility, especially when considering some of the questions raised in the technical section with respect to long-term support for and relevance of the data.

Attachment

Back to Form Submit

Other Links

- Disclaimer
- Accessibility
- Privacy Notice
- FOIA
- Site Map
- Get Acrobat (PDF) Reader

USA.gov