March 18, 2015

Re: Notice Number: NOT-ES-15-011

On behalf of the major trade associations of scholarly publishing – the Professional and Scholarly Publishing Division of the Association of American Publishers ("AAP/PSP") and the International Association of Scientific, Technical, and Medical Publishers ("STM"), we are pleased to respond to the National Institutes of Health's ("NIH") February 18, 2015 Request for Information ("RFI") regarding "Input on Sustaining Biomedical Data Repositories."

Our members publish the vast majority of materials used in the U.S. by scholars and professionals in biomedicine and other areas of scholarship, and they include the worldwide disseminators, archivists and shapers of the public record on scientific research via print and electronic means. They include non-profit professional societies, commercial publishers and university presses that create books, journals, computer software, databases and electronic products in virtually all areas of human inquiry and activity.

Collectively, our members represent tens of thousands of publishing employees, professional individuals, editors and authors throughout the country who regularly contribute to the advancement of American science, learning, culture and innovation. They comprise the bulk of a $10 billion commercial and non-profit publishing industry that contributes significantly to the U.S. economy and enhances the U.S. balance of trade.

For our members that publish biomedical journals and other peer-reviewed scholarly publications, the primary goal of their publishing activity is to disseminate information and provide access through a high quality and user-friendly digital environment in which to discover, analyze and link to the latest breakthroughs and developments in scientific and other scholarly research. In particular, publishers of scientific journals have, for more than 100 years, played an integral role in building and documenting the U.S. scientific research enterprise. In addition to their efforts to disseminate publications that report on and analyze the latest research, they also have considerable experience and investment in digital technology, metadata standards and tools to help users understand and work with data. This makes publishers uniquely positioned to help the Federal Government in expanding public access to digital data, ensure the long-term stewardship and discoverability of data and support the innovation and economic development that is derived from scholarly advancements.

It is worth noting at the outset that, in contrast to peer-reviewed publications, which are not the "result" of federally funded research and contain significant publisher added-value, digital data does often directly result from activity funded by the government. Research and publication are both different and unique creative acts. Publishers support better discoverability and reuse of scholarly data and are pleased that in this RFI NIH has recognized both the distinction between data and peer-

reviewed publications and the need to consider sustainability at the outset as the government seeks to develop tools for sharing data. The dissemination of information is an area of publishers' professional expertise, and the development of biomedical data repositories potentially impact our members not only as publishers of peer-reviewed biomedical journals, but also as disseminators of information whose innovative products and services enhance and add value to taxpayer-funded research activities and are expected to do so in the future.

It is with this view that the following comments and recommendations have been submitted on behalf of AAP/PSP and STM. We hope that they will help to facilitate the successful development of sustainable and effective policies on the development of biomedical data repositories that are consistent with the Administration's "Open Government" framework[1] and that NIH will proceed in enhancing the long-term sustainability of biomedical data repositories in a spirit of collaboration with all stakeholders, particularly publishers in recognition of our expertise and investments in these matters.

General Recommendations

Scholarly publishers have long served as integral hubs of America's research enterprise, validating research through the peer review process, producing the scientific record and facilitating scholarly communication through dissemination and preservation of scientific literature and, since the emergence of the digital age, through the presentation and long-term stewardship of digital data that is often submitted to publishers during the publication process. The primary goal of publishing is to facilitate the widest possible dissemination of the information that publishers provide. In the digital age, publishers have invested significantly to enhance the discoverability, public access to and the utility of research data, particularly for the scientific, technological, engineering, social science and medical communities: expanding accessibility, improving interoperability and fuelling innovation.

Publisher investments have created digital platforms with the latest and continually evolving Web capabilities, providing researchers with faster and more robust delivery of scholarly information, new ways to present data and research findings and links that enable information to be found and navigated with ease. Publishers have improved interoperability through new metadata standards and pilot projects, which are driving innovation and providing for better information discovery and expanded use of information.

As long as the government does not diminish incentives for creative publication, publishers will continue to provide tools that enhance innovative reuse and discovery of scientific information. Publishers look forward to continuing a positive collaboration to enhance science and innovation in the United States, and welcome any partnership with the Administration to harness the power and potential of technology and innovation to spur long-term economic growth and provide cutting-edge solutions to support domestic priorities.

[1] As articulated in Memorandum for the Heads of Executive Departments and Agencies on Transparency and Open Government (January 21, 2009), available at http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment and Memorandum for the Heads of Executive Departments and Agencies on Open Government Directive available at http://www.whitehouse.gov/open/documents/open-government-directive

There are many potential benefits from better sharing and consistent presentation of "Big Data." At the same time, "Big Data" is surrounded by "Big Uncertainty" about the what, the how and the possible payoffs. The experience of publishers informs an understanding that significant challenges remain to wider public access to and long-term preservation of digital data. Unlike publications which can be counted by the number of pages, datasets can be measured in mega-, giga- or even terabytes of data. 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 and validation. It is difficult to control for quality, display and consistency.

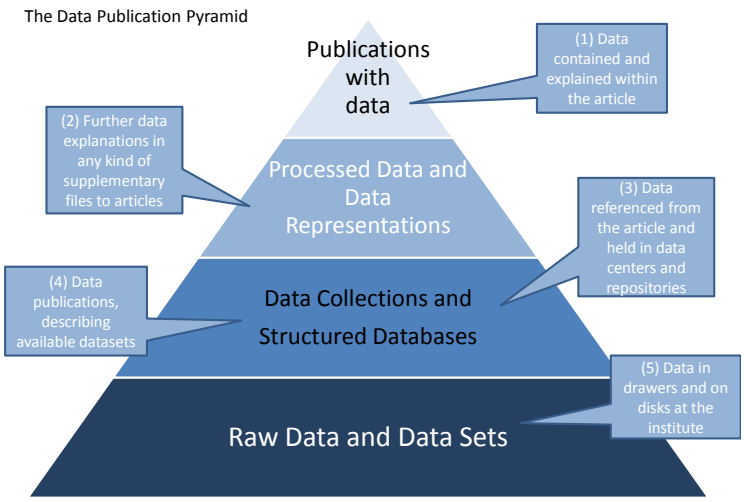
While publication incentives have been embedded in the academic and research process, incentives for complete tagging of individual datasets are limited. There is no consistent approach to presentation, standards for which data should be preserved or overall management and no consistent responsibility for data storage, tagging and dissemination. In addition, there are significant costs associated with storage, distribution bandwidth and overall management and curation. The best way to address lack of incentives, inconsistencies and costs is through collaboration among key stakeholders to promote the development of robust, sustainable and flexible standards that meet the needs of users at all levels. Publishers stand ready to lend their expertise to such a collaborative process to provide value to the research community and to the taxpayer.

Publishers support collaboration among key stakeholders to provide broad access to the digital data that results from federally funded research. At the same time, the government should not invest funding or energy to recreate what is already being achieved by the private sector. The government's best approach is to leverage public-private collaborations to ensure continued innovations in publishing that contribute to the progress of science, allow innovation to flourish and help grow the American economy.

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 plans 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 collaboration within and between federal agencies; 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 by various groups engaged with the issue, and we encourage the continued evolution of programs that are working 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.

It is critical that the federal government continue to distinguish between data and various types of presentation of data and preserve and respect intellectual property protection and copyright ownership as appropriate. The Data Publications Pyramid displayed here,[2] derived from open science pioneer Jim Gray's e-science pyramid, provides a model for understanding how research data can be presented in a variety of ways with increasing levels of curation and analysis.



Federal policies should take into account the differences between information products at different levels of the pyramid and 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. In particular, 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 the precedent it is setting regarding data management and preservation. While the NIH's focus is on biomedical data, other federal agencies are developing policies on data management, and the procedures established by NIH are likely to influence other agencies' approaches to finding solutions just as the NIH policy on public access to research articles reporting on federally funded research has influenced other agencies' newly emerging policies.

Rather than imposing an inflexible mandate, federal policies should focus on supporting and encouraging the development of cyber infrastructure, standards for the structure of data and metadata, navigation tools and applications to achieve discoverability and interoperability and ensuring appropriate and sustainable funding for innovation and long-term stewardship. These policies should be developed in collaboration with all key stakeholders involved in the presentation, analysis, deposit, storage and preservation of data.

NIH should promote a comprehensive framework for reliable digital data preservation, access and interoperability through the promotion of standards and clear rules developed by the scholarly community. NIH could also support pilot projects, data curation programs and interpretation initiatives for the relevant scholarly disciplines. Finally, NIH could use its web presence to provide a clearinghouse to the data they hold or which is funded by their grants.

---

[2] As appearing in the October 17, 2011 *Report on Integration of Data and Publications*, a report of Opportunities for Data Exchange which brings together stakeholders including researchers, publishers, libraries and data centers to support a more connected and integrated scholarly record. Full report available at http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

Additional concerns connected to selected prompts in the RFI

In the RFI, the NIH notes that it is seeking information that addresses several specified areas. Our general comments above respond to many of the specified areas and should be seen as a holistic response to the request. However, there are some additional issues to consider in response to the specified categories, which are outlined below.

- Financial Models

Considering sustainability at the outset
AAP/PSP and STM appreciate that NIH has explicitly used the word "sustaining" as the title of this RFI, as financial considerations must be at the center of the any effort to increase access to information. Significant costs are involved in the collection, storage and maintenance of any repository. Data repositories, with the variety of formats, content and quality are particularly challenging.

NIH should work to leverage existing efforts wherever possible, whether based at universities, in non-profit collaboratives, in other government agencies or in the private sector. In addition, NIH must support such efforts with funding as much as possible and should share in the costs of their development and operation. Where the government owns the repository, efforts should be made to use existing standards and ensure unfettered access for all stakeholders. It should also be clearly delineated as to how existing data management and preservation policies will dovetail with the new government guidelines.

Long-term preservation
The challenges for ensuring the availability and integrity of collected content are significant. Even in the short time that digital material has been available, we have seen many changes in formats and storage technology, and approaches that may have seemed cost-effective at the outset can quickly be supplanted by new innovations. NIH must remain flexible as to the approaches it takes and ensure that it does not lock in to one technology that becomes unsustainable over the long term.

- Best Practices

Collaboration as the best way forward
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.

The involvement of each stakeholder will ensure the preservation of incentives for innovation and help improve information sharing and training within each stakeholder community. 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. Finally, stakeholder input is necessary to incentivize the deposit of datasets and minimize the administrative burden on key stakeholders.

Creating systems of attribution and credit

The scholarly community already has a robust attribution and credit system with respect to peer-reviewed publication. This 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 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[3] 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.

Learning from the experience of developing identifiers

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 research.

Digital data standards are newer and still evolving. OSTP should learn from ongoing initiatives and published reports that address real problems through collaborative public-private partnerships with stakeholders, such as:

- DataCite (http://datacite.org/), which is working collaboratively to address the challenges of making research data visible and accessible;
- APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), which is working through a collaborative group of more thirty research institutes, national libraries, IT providers and research funders to create a Network-of-Excellence on digital preservation;
- Opportunities for Data Exchange (ODE, www.ode-project.eu), which is working to promote best practices around the way scientific data are treated;[4]

[3] DataCite is a non-profit organization whose aims are to establish easier access to research data on the Internet; increase acceptance of research data as legitimate, citable contributions to the scholarly record; and support data archiving that will permit results to be verified and re-purposed for future study. DataCite is currently engaged in the process of helping researchers find, identify, and cite research datasets; providing persistent identifiers for datasets, workflows and standards for data publication; and enabling research articles to be linked to the underlying data. To achieve these goals, they are currently working primarily with organizations that host data, such as data centers and libraries.

[4] ODE's *Report on Integration of Data and Publications* is available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/

- CoData (http://www.codata.org/), an interdisciplinary scientific committee of the International Council for Science Unions (ICSU) currently working on an initiative for a World Data System.
- NISO/NFAIS Supplemental Journal Article Materials Project (http://www.niso.org/workrooms/supplemental), which released a report in 2013 that discusses technical issues and presents recommended practices regarding the definition, publication and linking of journal articles and supplemental materials, including data, as well as archiving, preservation and migration of different file formats;
- PARSE.insight (http://www.parse-insight.eu/), which published a roadmap and recommendations[5] for long-term accessibility and usability of scientific digital information in Europe; and
- ORCID (http://orcid.org/), which provides a persistent, unique digital identifier that distinguishes one researcher from another with the same or similar name.

Considering ethics and confidentiality

When it comes to biomedical information, issues about the best approaches go beyond technology and financial concerns to the characteristics of the data itself. 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 the protections that have been put in place as it works to expand access to data. The communities of practice, particularly scientific societies, can be ideal partners in the development of safeguards to ensure that any data of concern is treated properly. All publishers have experience dealing with such concerns in the review and publication of articles that report on such data and can share their experience.

➢ Technical

Interlinking between data sources, metadata and publications

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.

➢ Human Capital

Support for researchers and their partners

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. Initiatives must be carefully developed to ensure that such tagging and validation can be achieved, and success will depend on a collaborative approach that elicits buy-in from all communities and includes consultation and contributions by key stakeholders. Such an approach must carefully consider how best to create incentives for data management and sharing, and provide support to all involved.

[5] The *Insight into Digital Preservation of Research Output* report is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf and the *Science Data Infrastructure Roadmap* is available at http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf

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.

- Life Cycle

Preventing fraud
From the onset of allowing deposit of data into a repository, 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. NIH should partner with key stakeholders to identify and promote best practices, develop standards and implement policies and practices to detect data fraud.

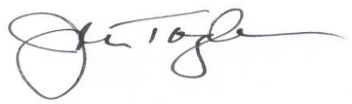
Validating data
Although very different from fraud, the question of quality must also be carefully considered when thinking about long-term support and relevance of data in repositories. Capabilities should be developed for validating data in terms of both quality and utility. For example, data could be tagged if it is replicated by another team, if it is consistent with published results or with other measures of quality. Some of our member publishers have developed the capability to do validation in the peer-review process in certain instances, and would be willing to share their expertise with NIH relative to their rubrics and practices to help develop appropriate criteria for the efforts of various NIH components.

Maintaining repositories
As NIH develops and supports repositories, it must consider not just the size but also the diverse structure of datasets, which can create challenges regarding maintenance. Some collections, such as data from big imaging studies, will be extremely large and could quickly overwhelm existing repositories or servers. Others may require specific formats to communicate the underlying data, formats that may become obsolete and will need translating into new formats. Any consideration of sustainability must take into account the full life cycle of the data sets in the repository, and ensure that entire life cycle can be supported and preserved for the long term.

Conclusion

Ever since the journal was created 350 years ago this month, publishers have been devoted to the goal articulated in this RFI: turning information into knowledge products that can support new innovation. Our members understand better than most how critical the issue of sustainability is to any effort to disseminate information. Throughout our history and into the digital age, publishers have been devoted to the integrity and preservation of the scholarly record, and that requires a commitment to ensuring the sustainability of the products we create. AAP/PSP and STM appreciate that NIH is taking the issue seriously as it considers the many issues involved in supporting new efforts to share data, particularly "Big Data," and the efforts required to populate and sustain new data repositories. We welcome the opportunity to collaborate in the future on integrating your efforts with the broader efforts of all stakeholders in the scholarly communication system.

Sincerely,



John Tagler
Vice President & Executive Director
Professional & Scholarly Publishing
Association of American Publishers, Inc.
71 Fifth Avenue
New York, NY 10003-3004
455 Massachussetts Ave
Washington, DC
jtagler@publishers.org
Phone: 212 255-1407

Michael Mabe
Chief Executive Officer
International Association of STM Publishers

Prins Willem Alexanderhof 5
The Hague, 2595 BE
The Netherlands
Mabe@stm-assoc.org
Phone: +44 1865 339321