AIP American Institute of Physics
One Physics Ellipse, College Park, MD 20740-3843 H. Frederick Dylla, Executive Director and CEO

22 December 2011

Submission for the Record: **Response to November 4, 2011 Federal Register Notice of Request for Information, OFFICE OF SCIENCE AND TECHNOLOGY POLICY, Public Access to Digital Data Resulting From Federally Funded Scientific Research; FR Doc. 2011–28621**

Submitted by: H. Frederick Dylla, Executive Direct and CEO, American Institute of Physics
Tel. +1 301-209-3131; Dylla@aip.org

Electronically submitted to: digitaldata@ostp.gov

The American Institute of Physics (AIP) appreciates this opportunity to submit comments and would be delighted to continue working with OSTP and other federal partners through a process of active engagement.

About AIP

The American Institute of Physics (AIP) is a 501(c)(3) not-for-profit membership corporation created in 1931 for the purpose of "the advancement and diffusion of knowledge of the science of physics and its applications to human welfare." AIP is an organization of 10 physical sciences societies representing more than 135,000 scientists, engineers, and educators. As one of the largest publishers of scientific information in physics, AIP employs innovative publishing technologies and offers publishing services for its Member Societies. AIP's suite of publications includes 15 journals, three of which are published in partnership with other organizations; magazines, including its flagship publication *Physics Today*; and the AIP Conference Proceedings series. AIP delivers valuable resources and expertise in education and student services, science communication, government relations, career services for science and engineering professionals, statistical research, industrial outreach, and the history of physics and other sciences.

Enabled by Internet technologies, AIP disseminates more information, more widely and more affordably, than ever before in its history, reaching more authors, subscribers, and users than ever before. This accomplishment requires heavy investments in technology and infrastructure (such as an online platform) and business-model innovation to deliver the option of free or low-cost access: open access, pay-per-view, or article rental, recognizing that the value of the final published article needs to be paid for to remain sustainable.

Introduction

AIP's highest goal is to achieve the widest possible dissemination of the research results it publishes, including any pertinent associated data and context information. As a scholarly publisher, AIP believes that better discoverability and reuse of original research data are to be encouraged at all levels and among all stakeholders. AIP also believes that data resulting directly from federally funded scientific

research should be made freely available in a sustainable manner and that this is best achieved through appropriate policies that leverage public-private collaboration.

AIP believes that it would be in the best interest of the United States and its government, as well as in the best interest of all other stakeholders, to strike a balance between public access and sustenance of the scholarly publishing industry because of the impact and value it brings to the progress of science and its contributions to American society and economy. Such a balance can be achieved based on shared principles such as the importance of peer review, the recognition of economic realities through adaptable and viable publishing business models, the need to ensure secure archiving and preservation of scholarly information, and the desirability of broad access. Policies should recognize that hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term, has a cost, which, in certain circumstances, the host should be entitled to recover. One way to achieve this balance is for government to adopt a sensible, flexible, and cautious approach to drafting public access policies—an approach that engages all concerned parties, including federal agencies, scientists, university administrators, librarians, publishers, and the public.

Consistent with the recognition of economic realities, it is AIP's position that government agencies should develop their public access policies through voluntary collaborations with nongovernmental stakeholders, including researchers and publishers. Any policies should be guided by the need to foster interoperability of information across multiple databases and platforms. Agencies' efforts then could be directed toward facilitating cyberinfrastructure and collaboration programs with and between agencies and the stakeholders to develop robust standards for the structure of full text and metadata, navigation tools, and other applications to achieve interoperability across the scholarly literature. More detail on this is provided later in the document. AIP believes that any scholarly publication access policy needs to be flexible to accommodate agency-specific needs and have the capacity to evolve in response to the rapidly changing nature of scholarly publishing.

AIP specifically recommends that federal grants set aside funds to support researcher data management and deposit efforts. Federal agencies could also play a role in supporting and encouraging the establishment of discipline-specific data archives where these are currently lacking. The amount and type of support should be determined in collaboration with key stakeholders involved in the deposit, storage, and preservation of data.

Federal policies should also focus on supporting and encouraging the development of community standards for the citation and reuse of data sets, thereby facilitating the creation of a system that gives researchers an incentive to share data resulting from federal grants.

AIP Responses to RFI Questions

Preservation, Discoverability, and Access

(1) What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the U.S. economy and improve the productivity of the American scientific enterprise?

We would make the distinction that it is not "public access" in the broadest sense that is important but rather access by other scientists who can use the digital data for the further advancement of science.

As data are not copyrightable, policies about access become policies about deposit by the data owner or proxy into an accessible system. It should be noted, though, that any policies should recognize and take into account differences between 'databases' (information products created for the specific display and retrieval of data) and 'data sets' (sets or collections of raw relevant data captured in the course of research or other efforts). Policies could require that data generated from federally-funded research be deposited in a certified and openly accessible repository; furthermore, researchers could be encouraged to make these deposits upon submission of their first manuscript showing results that were based on the data set. Although some agencies already have a preservation/access role (for example, DOE Order 241.1B), AIP agrees with the 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." Agency policies should support and encourage such a distributed system for both access and preservation; that is, policies should recognize and build upon the broad set of capabilities that exist for both access and preservation within the library and publishing communities for both documents and data – Portico, LOCKSS.

The integrity of preserved data would also need to be taken into account and supported by any policy.

(2) What specific steps can be taken to protect the intellectual property interests of publishers, scientists, Federal agencies, and other stakeholders, with respect to any existing or proposed policies for encouraging public access to and preservation of digital data resulting from federally funded scientific research?

All policies should comply with current copyright and patent law. Data should be embargoed to the principle researcher until conclusions drawn from the data can be published in the research literature. An additional maximum embargo of one year would also provide for the filing of patents by the grantees (or their institution) as allowed by many, if not all, funding agencies (HR 1249 Sec 102(b)(1)(A)). See also the distinction between databases and datasets as addressed response to question 1.

(3) How could Federal agencies take into account inherent differences between scientific disciplines and different types of digital data when developing policies on the management of data?

Differences between scientific disciplines and different digital data must be taken into account by domain experts at the time of proposal review (note the language used in the Data Management Plan FAQ's of NSF in a variety of instances: "to be determined by the community of interest through the process of peer review and program management.") Only such experts will be able to determine if the data to be generated by the proposed research will be of longer term value to the scientific community of interest and if its type conforms to acceptable community standards.

Metadata—data about the data—which would include information both about what the data is and how it was collected, is addressed further in this response.

(4) How could agency policies consider differences in the relative costs and benefits of long-term stewardship and dissemination of different types of data resulting from federally funded research?

Policies must first recognize that not all data is worth preserving. Every type of data should be assessed regarding long-term stewardship. Policies would have to take into account not just the size of the datasets but also long-term usability, which depends on the rate of technology change, and level of documentation required. Along with the data, enough information needs to be preserved to reproduce the dataset. As noted in the answer to question 3, agencies will need to call upon data experts as well as scientific experts.

(5) How can stakeholders (e.g., research communities, universities, research institutions, libraries, scientific publishers) best contribute to the implementation of data management plans?

There needs to be an interconnected system for access to and sharing and preservation of data based on community-developed standards and best practices. The system needs to encourage innovation and must support multiple solutions—data as an information resource is inherently more complicated than scholarly articles. Each stakeholder will then need to contribute based on their specific skills and expertise. Libraries, through Institutional Repositories, could take on a stronger preservation role. Publishers have been adding value to the research process and providing access to and preservation of the scholarly literature for hundreds of years and could extend this to data, well beyond current support for supplemental material. Universities and research institutions have both scientific domain knowledge and data and information experts. Any system will need to preserve incentives for innovation.

Consider, for example, work being done by the Data Preservation Alliance for Social Sciences through their partnership with the Library of Congress, LOCKSS, and Dataverse to prototype a policy-based replicated data archive.

Other examples include:

- linking between datasets and their resulting scholarly publications based on community-accepted standards, thus ensuring datasets become part of the scientific literature;
- Having clear standards and guidelines for the certification and auditing of data repositories; encouraging a system that incentivizes data repositories to maintain the accuracy or integrity of the data once it has been deposited;
- Incentivizing the deposit of datasets and ensuring that the administrative burden this imposes on researchers minimal.

(6) How could funding mechanisms be improved to better address the real costs of preserving and making digital data accessible?

Require data management plans and coordinate plan requirements across agencies and to community standards (see the Open Archive Information System Reference Model – ISO standard 14721:2003). What constitutes data that needs to be preserved should be clearly identified through the process of peer review and program management. Preserving and disseminating digital data should then be considered “part of the cost” of funding and doing research, not “an additional cost”. Funding agencies could emphasize that proposals must take into account data fit for reuse and preservation. Again, this

should be the approach across agencies. Research labs/institutions/university overhead rates would need to include cost of data preservation.

As pointed out in the final report from the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (*Sustainable Economics for a Digital Planet*): "Policy mechanisms can play an important role in strengthening weak motivations" as there is often "misalignment of incentives between communities that benefit from preservation (and therefore have an incentive to preserve), and those that are in a position to preserve (because they own or control it) but lack incentives to do so."

(7) What approaches could agencies take to measure, verify, and improve compliance with Federal data stewardship and access policies for scientific research? How can the burden of compliance and verification be minimized?

If data is created in the course of federally-funded research, then the funding agency could require that any such data deemed to be "preservation data" be deposited in a recognized archive. Through direct agency involvement in creating a "comprehensive framework for data access and preservation" based on community-accepted standards and best practices for data citation and reuse, agencies would maintain lists of certified repositories. Certified repositories could be similar to the data center members of the DataCite organization (of which DOE's Office of Scientific and Technical Information is a member) or participants in the SafeArchive program of Data-PASS. In addition, grantee data management plans could be required to identify all datasets expected to be produced from funded work.

Certification of compliance would then simply require grantee reporting to include in reports on their funded proposal the data citations and the repository where the data was deposited.

As work is already being carried out to develop standards in this area (i.e. *The ISO 16363 Standard for Trusted Digital Repositories),* it would be more expedient for federal agencies to work within and help support such standards.

(8) What additional steps could agencies take to stimulate innovative use of publicly accessible research data in new and existing markets and industries to create jobs and grow the economy?

AIP agrees with the statement from the Interagency Working Group on Digital Data (IWGDD) in its report, *Harnessing the Power of Digital Data for Science and Society,* that "the current landscape lacks a comprehensive framework for reliable digital [data] preservation, access, and interoperability". We feel that there is a very important role for the federal government and its science funding agencies to play to help create and promulgate such a comprehensive framework.

Federal investment in creating stable, standardized, and accessible data will be an essential base from which innovation can occur. The ease of reuse could then lead to developments akin to IBM Research's "Many Eyes" product for data visualization (www-958.ibm.com), or spur the private sector to offer data services for researchers.

(9) What mechanisms could be developed to assure that those who produced the data are given appropriate attribution and credit when secondary results are reported?

This ecosystem of attribution and credit already exists with respect to scholarly articles. A researcher's standing in their field is largely a result of their list of authored scholarly articles and the number of citations to those published articles. The credit comes in the form of respect from peers, funding for further work, and career advancement, and rests in large part on the underlying quality control provided by peer review. Not providing appropriate attribution is considered unethical scientific behavior and can lead to the retraction of published work.

The mechanisms to be developed would support an extension of this system to cover data. The elements to support are:

- data must be recognized as a primary research output,
- data must have unique and persistent identifiers and be fully citable, thereby allowing its use and reuse to be tracked and recorded in the same way as scholarly publications, and
- data citation information must be used for research evaluation and reward.

Persistent identifiers for data could be handled through use of digital object identifiers already used for scholarly articles or similar (see Datacite.org). There are also examples of recommended practice for citing data. [For example: creator (publication year): Title, Publisher, identifier; see http://datacite.org/whycitedata and DOE's Data ID Service.]

Publishers could support the development of such a system by requiring that all data needed to reproduce the results and conclusions of a published scholarly article must be cited according to community standards.

Funding agencies could support the development of such a system by recognizing data that has been archived and made available to the research community as "first class research objects" at the same level as articles. Agencies should also recognize any reuse of these data which could then be counted via citations.

See the Australian National Data Center's "Building a Culture of Data Citation" poster available at http://ands.org.au/cite-data/index.html.

For a hybrid example spanning the world of digital data and scholarly publication, see the *Journal of Physical and Chemical Reference Data*, a long and successful collaboration between AIP and the National Institute of Standards and Technology.

Standards for Interoperability, Reuse and Re-Purposing

(10) What digital data standards would enable interoperability, reuse, and repurposing of digital scientific data? For example, MIAME (minimum information about a microarray experiment; see Brazma et al., 2001, Nature Genetics 29, 371) is an example of a community-driven data standards effort.

First, it is important to separate metadata standards from data format standards. Metadata standards could be developed that are lightweight enough to be widely interoperable and extensible so as to accommodate discipline-specific needs (within the XML publishing standard). These standards would need to cover both bibliographic information (data creator, date of creation, what the data describes, where it can be accessed, etc.), and how it was collected (experimental apparatus, experimental conditions, location, etc.).

Data format standards that would enable reuse and repurposing would need to be developed at the discipline-specific level. There need not be one solution per discipline: it may be that the communities in question need a handful of solutions that correspond to the various types of data and/or modes of scientific research that produces the data. So while it is true that actual data solutions need to be discipline appropriate, there may be logical clusters of solutions for the connections between publishing and data depending on the nature of the data.

There is a role for federal agencies in coordinating across discipline boundaries (covering all funded areas) and internationally. In its October 2011 report, *Federal Engagement in Standards Activities to Address National Priorities: Background and Proposed Policy Recommendations,* the Subcommittee on Standards of the National Science and Technology Council noted that "There was agreement among respondents that the US government should continue to play the role of participant in private sector standards setting processes. There was also general agreement that the effectiveness of government participation depends on the level and consistency of involvement and commitment of resources, both staff and budgetary, to the process. Lack of coordination among agencies…was cited by many respondents as having a negative impact on government effectiveness. "

(11) What are other examples of standards development processes that were successful in producing effective standards and what characteristics of the process made these efforts successful?

The Digital Object Identifier, or DOI, is an example of a successful standard. Its development and adoption involved a multi-stakeholder, community-driven approach that solved a practical problem and provided benefit to the end-user.

(12) How could Federal agencies promote effective coordination on digital data standards with other nations and international communities?

AIP supports the recommendation of the Interagency Working Group on Digital Data (IWGDD) that an NSTC Subcommittee for digital data preservation, access, and interoperability be created. This subcommittee would then be able to provide coordination among the US funding agencies and collaborate with its international counterparts. Coordination at the national level should extend beyond

science funding agencies as relevant work is being done elsewhere within the US government (for example, the work of the Library of Congress through its National Digital Information and Infrastructure Program [NDIIP], particularly its "partnership with the National Science Foundation in 2005 to undertake a program of pioneering research to support advanced research into the long-term management of digital information").

In addition, this subcommittee could ensure that each Federal agency is itself required to adopt and implement digital data standards developed within the global community.

Federal agencies can support conferences and other initiatives on a discipline level by funding standards and preservation work as well as pure research.

(13) What policies, practices, and standards are needed to support linking between publications and associated data?

See answer to question 9. The mechanism for linking between publications and associated data essentially exists with the digital object identifier, which is already used widely for linking between publications. The federal government could provide additional logistics and financial support for making this mechanism standard practice with respect to data and coordinating/aligning policies across federal agencies to encourage use of those standards by grantees.

Agency involvement and/or support of current initiatives such as the NISO/NFAIS Working Group on Supplementary Journal Information (www.niso.org), which is working on recommended practices for publishers who are increasingly attaching data sets as supplementary information appended to publications, would also help address some of the issues at a practical level.