ACS Submission to the Office of Science and Technology Policy Request for Information on Public Access to Digital Data Resulting from Federally Funded Research

FR Doc. 2011-28621

Submitted January 12, 2012
by:
John P. Ochs
j_ochs@acs.org
Vice President, Strategic Planning and Analysis
American Chemical Society
Publications Division

The American Chemical Society (ACS) is the world's largest scientific society with more than 164,000 members. ACS advances knowledge and research through scholarly publishing, scientific conferences, information resources for education and business, and professional development efforts. The ACS also plays a leadership role in educating and communicating with public audiences—citizens, students, public leaders, and others—about the important role that chemistry plays in identifying new solutions, improving public health, protecting the environment, and contributing to the economy.

ACS Publications is a division of the American Chemical Society. The Publications Division strives to provide its members and the worldwide scientific community with a comprehensive collection, in any medium, of high-quality information products and services that advance the practice of the chemical and related sciences. Currently, over 40 peer-reviewed journals and magazines are published or co-published by the Publications Division. Over 290,000 pages of research material are published annually, representing over 37,000 research papers. With the introduction of the ACS Journal Archives in 2002 and the C&EN Archives in 2011, we provide searchable access to over one million original chemistry articles dating back to 1879.

ACS Publications offers both sponsored and author-enabled open access to research articles through our ACS Author Choice and ACS Articles on Request programs. In addition, digital data that supports the findings of articles and bibliographic information, including abstracts of research articles, are freely available on our website. Since the beginning of the transition to electronic publishing in the mid- to late-1990s, we have developed, and are continuing to develop, innovative and accessible business models, policies, and practices to support the scholarly communication process and broaden information access.

As a socially responsible organization deeply rooted in the scholarly community, we share the interest of the Federal government in maximizing the dissemination and discoverability of knowledge. ACS believes that success in this area will hinge on these efforts being sustainable for publishers over the long-term. We welcome for the opportunity to respond to the invitation to contribute to the Request for Information (RFI) on Public Access to Digital Data Resulting from Federally Funded Scientific Research published by Office of Science and Technology Policy (OSTP) in the *Federal Register* on November 4, 2011.

Our response is in two parts: first a summary of our overall comments and recommendations, and second, answers to the specific questions posed in the RFI.

I. Summary

ACS supports the view that Federal agencies should work with researchers and other stakeholders to create appropriate policies to make digital data resulting from federally funded scientific research freely available to the public. ACS sees an appropriate role for governmental and other funding agencies to identify standards and best practices for the management of primary scientific data that are generated via taxpayer or other research grant funding that supports independent investigators. This governmental role could also include standards for the interoperability of data repositories with the published research literature. As part of this process, agencies should investigate and establish contacts where appropriate with a number of initiatives already underway or recently concluded that are examining data stewardship issues.

Within the context of standards and best practices that have been identified, the Federal government can develop effective, evidence-based policies to enhance public access to and preservation of digital data. We recommend that these policies be established in collaboration with researchers and other key stakeholders.

Grants should earmark specific funds to support researcher data management and deposit activities. The amount should be determined in collaboration with representative bodies of key stakeholders who are involved in the data preservation and deposit process. It may need to vary with discipline. In parallel with this activity, the government should ask the General Accounting Office to undertake a study of existing federal data archives to determine the full costs required for start-up and ongoing access, preservation, and migration of data depositories.

Federal policies should establish clear rules for citation of data sets and acknowledgement of changes or modifications to source data. They should provide for the establishment of security protocols that protect stored data from unauthorized modification, damage or deletion and liability arrangements if data is lost or affected. Key policy terms should be defined and policies should take into account that there are differences between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course of research or other efforts ('data sets').

Federal intellectual property 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. Databases themselves – i.e. collections of data specifically organized and presented, often at considerable cost, for the ease of viewing, retrieval and analysis – merit intellectual property protection, under copyright or database protection principles.

To reduce legal uncertainty for data users and producers, federal policy should give clear direction as to what data may be shared publicly – e.g. no personal data related to volunteer subjects. Penalties for the misuse or abuse of data should be established, such as grant bans for those who willfully misrepresent or distort the data created by others, and technical measures should be put into place to ensure ongoing data integrity.

Policies should not require researchers to fund the establishment or maintenance of data archives nor should they be required to pay submission fees for deposit. Federal policy should encourage, but not require researchers to supply their data when submitting manuscripts to scientific journals. This is because certain forms of publication, e.g. letters and other short communications, act as early alerts to results of potential interest and the requirement to supply data can add a burden that slows scholarly communication to the detriment of all.

Policies could create an incentives hierarchy for scientists to share their data, with the greatest reward for those who publish data with articles and short communications but also recognition for those who publish data only.

II Response 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?

Before specific policies are adopted, ACS sees an appropriate role for governmental and other funding agencies to identify standards and best practices for the management of primary scientific data that are generated via taxpayer or other research grant funding that supports independent investigators – e.g. recommendations for best practices in the PARSE.Insight report *Insight into Digital Preservation of Research Output in Europe*. This governmental role could also include standards for the interoperability of data repositories with the published research literature.

Once standards and best practices have been identified, the Federal government will be in a stronger position to adopt effective, evidence-based policies to enhance public access to and preservation of digital data. We recommend that these policies be established in collaboration with researchers and other key stakeholders.

Policies should establish clear rules for citation of data sets and acknowledgement of changes or modifications to source data. They should provide for the establishment of security protocols that protect stored data from unauthorized modification, damage or deletion and liability arrangements if data is lost or affected. Key policy terms such as data and data integrity should be defined since, for example, there are differences between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course of research or other efforts ('data sets').

To reduce legal uncertainty for data users and producers, clear direction should be given as to what data may be shared publicly – e.g. no personal data related to volunteer subjects. Penalties for the misuse or abuse of data should be established, such as grant bans for those who willfully misrepresent or distort the data created by others, and technical measures should be put into place to ensure ongoing data integrity.

Policies should not require researchers to fund the establishment or maintenance of data archives nor should they be required to pay submission fees for deposit. Federal policy should encourage, but not require researchers to supply their data when submitting manuscripts to scientific journals. This is because certain forms of publication, e.g. letters and other rapid communication formats, act as early alerts to results of potential interest and the requirement to supply data can add a burden that slows scholarly communication to the detriment of all.

Policies could create an incentives hierarchy for scientists to share their data, with the greatest reward for those who publish data with articles and short communications but also recognition for those who publish data only – i.e. with no discussion, analysis, or interpretation of such material.

(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?

Policies adopted by the federal government should establish clear rules for citation of data sets and acknowledgement of changes or modifications to source data. Penalties should be established for the misuse or abuse of data, e.g. bans on grant eligibility for those who willfully misrepresent or distort the data created by others, and technical measures should be put into place to ensure ongoing data integrity. Key policy terms such as data and data integrity should be clearly defined to differentiate between information products created for the specific display and retrieval of data ('databases') and sets or collections of raw relevant data captured in the course of research or other efforts ('data sets'). To reduce legal uncertainty for data users and producers, clear direction should be given as to what data may be shared publicly – e.g. no personal data related to volunteer subjects.

The ACS endorses the view that researcher-validated primary data should be made freely available but federal intellectual property 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. Databases themselves – i.e. collections of data specifically organized and presented, often at considerable cost, for the ease of viewing, retrieval and analysis – merit intellectual property protection, under copyright or database protection principles. Such databases are often characterized by the sophistication of their data field structuring, searchability tools, and contain valuable and useful information for scholarly research. The value of researcher validated data sets and individual data points is different from specific databases that have been organized and compiled to serve particular research needs.

(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?

The ACS supports the position that researcher-validated primary data should be made freely available and that Federal agencies should work with the scientific community and other stakeholders to create appropriate policies that reflect different standards currently in use or commonly accepted. If no consensus emerges from such efforts, ACS believes that the government has an appropriate role in working with key stakeholders such as researchers and publishers to develop best practices that will advance scholarly communication and the public good.

(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?

Agencies should investigate and establish contacts where appropriate with a number of initiatives already underway, or recently concluded, which are examining data stewardship issues. These include:

- Opportunities for Data Exchange (ODE, www.ode-project.eu), whose aim is to gather and promote best practices around the way scientific data are treated. Its *Report on*

Integration of Data and Publications is available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/

- APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), a project of the Alliance for Permanent Access which includes over thirty research institutes, national libraries, IT providers and research funders working together to create a Network-of-Excellence on digital preservation
- PARSE.insight (http://www.parse-insight.eu/), who developed a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. 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
- CoData (http://www.codata.org/), an interdisciplinary scientific committee of the International Council for Science (ICSU) working on an initiative for a World Data System
- DataCite (http://datacite.org/), convening members of the datasets community to collaboratively address the challenges of making research data visible and accessible, and
- NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental), looking at policy and technical issues surrounding the definition, publication and linking of journal articles and supplemental materials, including data, as well as archiving, preservation and migration of different file formats.

Interaction with these and other initiatives should give Federal agencies a good base from which to estimate the relative costs and benefits of long-term stewardship and dissemination of different types of data. Agencies may also find the data sections of the *Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access* (available at http://brtf.sdsc.edu/) to be relevant in evaluating the relative costs and benefits of long-term data preservation and migration.

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

Keys to successfully implementing data management plans from the Federal government, and other funders, include the following:

- Requirements for data management plans should clear, complete and unambiguous. They should specifically address liability issues
- Data management policies established in collaboration with researchers and other stakeholders such as publishers
- They should take into account the practices of different research communities and be developed in collaboration with representative bodies of all stakeholders who will likely be affected – e.g. researchers, funders, publishers, universities, data repositories, etc.
- FAQs, training courses, and e-learning modules should be available for researchers to gain a more complete understanding of data management plan requirements as well as the data deposit process

- Grant funds should be earmarked to support data management and deposit activities
- Incentives to deposit, such as the possibility for receiving research credit for data deposit, should be provided as well as penalties, like grant bans, for noncompliance after a clearly-defined and collaboratively-set time frame
- Data deposit, integrity, provenance, and access at repositories should be fast, efficient and clear.
- Data repositories should be certified and audited to foster trust. Researchers should not be required to maintain the accuracy or integrity of the data once it has been deposited but depositing researchers should have the right to modify or correct data they have deposited. Liability
- The administrative burden on researchers should be kept to the barest minimum possible

Stakeholders should work collaboratively on these issues since more than one stakeholder can contribute to each. There is no one stakeholder that has, or should have, a monopoly on any of these activities.

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

Grants should earmark specific funds to support researcher data management and deposit activities. The amount should be determined in collaboration with representative bodies of key stakeholders who are involved in the data preservation and deposit process and may vary with discipline. In parallel with this activity, the government should ask the General Accounting Office to undertake a study of existing federal data archives to determine the full costs required for start-up and ongoing access, preservation, and migration of data depositories. Agencies could also investigate the Open Archive Information System (OAIS) Reference Model (ISO standard 14721:2003, available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683), used by many as a model for building a sustainable digital archive. Last, agencies note the assessment of funding models from the *Blue Ribbon Task Force on Sustainable Digital Preservation and Access* (available at http://brtf.sdsc.edu/):

"*There is no single "best" funding model for digital preservation. Selection of an appropriate model requires an in-depth knowledge of the circumstances surrounding the effort, preservation goals, the stakeholder community, and so on.*" (p. 44)

(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?

As in other areas related to preservation where significant activity is already underway, the federal government could establish relationships with groups like the ISO *Repository Audit and Certification Working Group* (see http://wiki.digitalrepositoryauditandcertification.org/bin/view) to learn about standards and best practices already in development. Once standards and best practices have been identified, the Federal government will be in a stronger position to adopt effective, evidence-based measures related to the assessment of compliance with its data stewardship policies.

(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?

In addition to the measures already discussed in previous questions, agencies could set aside funds to promote to use of the data depositories or develop special sections of their websites promoting the availability and characteristics of the data they hold.

(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?

Agencies should seek collaborations with DataCite (see http://datacite.org/), 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 active in supporting researchers by helping them to find, identify, and cite research datasets with confidence; supporting data centers by providing persistent identifiers for datasets, workflows and standards for data publication; and support journal publishers by enabling research articles to be linked to the underlying data. They are currently working primarily with organizations that host data, such as data centers and libraries.

Standards for Interoperability, Re-Use 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.

The PARSE.insight *Science Data Infrastructure Roadmap* (available at http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf) notes the following initiatives that may prove to be use examples of digital stewardship: CASPAR (http://www.casparpreserves.eu/), Planets(http://www.planetsproject.eu/), DCC (http://www.dcc.ac.uk/), OAIS (http://public.ccsds.org/publications/archive/650x0b1.pdf), SHAMAN (http://www.shaman-ip.eu/), and nestor (http://www.langzeitarchivierung.de/)

Also the Technical Working Group of the NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental) is preparing an initial draft of its recommendations. The narrative form is expected to contain a table outlining the minimum metadata elements recommended to describe supplemental materials and establish their relationship to the main article, as well as a more detailed discussion of optional elements to more comprehensively characterize the materials for future applications. A non-normative DTD is also expected in draft form. This DTD, once finalized, will not be an official standard. Rather it will be a model to more precisely define a hierarchy for the recommended metadata, and could be used as a starting point for organizations seeking to adhere to the NISO/NFAIS recommendations.

(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?

Most standards development in the field of digital data stewardship is either ongoing or prospective. However, the initiatives cited in the answer to question 4 above are good examples of active projects in this area and are reproduced below for ease of reference:

- Opportunities for Data Exchange (ODE, www.ode-project.eu), whose aim is to gather and promote best practices around the way scientific data are treated. Its *Report on Integration of Data and Publications* is available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/
- APARSEN (http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/), a project of the Alliance for Permanent Access which includes over thirty research institutes, national libraries, IT providers and research funders working together to create a Network-of-Excellence on digital preservation
- PARSE.insight (http://www.parse-insight.eu/), who developed a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. 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
- CoData (http://www.codata.org/), an interdisciplinary scientific committee of the International Council for Science (ICSU) working on an initiative for a World Data System
- DataCite (http://datacite.org/), convening members of the datasets community to collaboratively address the challenges of making research data visible and accessible, and
- NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental), looking at policy and technical issues surrounding the definition, publication and linking of journal articles and supplemental materials, including data, as well as archiving, preservation and migration of different file formats.

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

Federal agencies should join the international community of organizations already actively involved in the development of digital preservation standards, best practices and policies that have been cited in answers to questions 4 and 11.

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

Agencies should become involved with three initiatives already well underway in this area:

- Opportunities for Data Exchange (ODE, www.ode-project.eu) – whose aim is to gather and promote best practices around the way scientific data are treated. See its *Report on Integration of Data and Publications* available at http://www.alliancepermanentaccess.org/index.php/current-projects/ode/outputs/
- The NISO/NFAIS Supplemental Journal Materials Working Group (http://www.niso.org/workrooms/supplemental) which is preparing an initial draft of its recommendations. The narrative form is expected to contain a table outlining the minimum metadata elements recommended to describe supplemental materials and establish their relationship to the main article, as well as a more detailed discussion of optional elements to more comprehensively characterize the materials for future applications. A non-normative DTD is also expected in draft form. This DTD, once finalized, will not be an official standard. Rather it will be a model to more precisely define a hierarchy for the recommended metadata, and could be used as a starting point for organizations seeking to adhere to the NISO/NFAIS recommendations
- DataCite (http://datacite.org/), 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 active in supporting researchers by helping them to find, identify, and cite research datasets with confidence; supporting data centers by providing persistent identifiers for datasets, workflows and standards for data publication; and support journal publishers by enabling research articles to be linked to the underlying data. They are currently working primarily with organizations that host data, such as data centers and libraries.