AMERICAN ASTRONOMICAL SOCIETY

Kevin B. Marvel
Executive Officer

11 January 2012

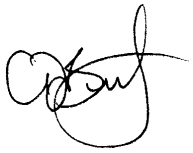
Submitted to digitaldata@ostp.gov

The American Astronomical Society (AAS) appreciates the opportunity to submit comments in response to the Request for Information concerning Public Access to Digital Data Resulting From Federally Funded Research [FR Doc. 2011-28621].

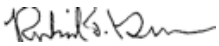
Sincerely yours,

Debra Elmegreen
President, AAS

Kevin B. Marvel
Executive Officer, AAS



Chris Biemesderfer
Director of Publishing, AAS



Richard F. Green
Chair, AAS Publications Board

Request for Information: Public Access to Digital Data Resulting From Federally Funded Scientific Research

[FR Doc. 2011-28621]

Submission from the American Astronomical Society

The mission of the American Astronomical Society is to enhance and share humanity's scientific understanding of the Universe.

The American Astronomical Society (AAS) is the major association for professional astronomers in the United States, with over 7500 members. One of its primary functions is the publication of the key North American scientific journals dedicated to the dissemination of peer-reviewed research in astronomy and astrophysics, the *Astrophysical Journal* and the *Astronomical Journal*. As a society of research and higher education professionals, we have made a concerted effort to conduct our scholarly publishing enterprise with sensitivity to and balance among the need for prompt and inexpensive access to new results, the pressures on the budgets of technical libraries, and the challenges of obtaining grant and institutional funding to support author fees.

The Society's mission has a broad public purpose, but its constituency is primarily professional research astronomers. Consequently, public access to data, while an attractive desideratum, is less of a concern than is ensuring access to data among research professionals engaged in on-going investigations. However, it is reasonable to assume that the mechanisms for sharing research data among professionals will also serve the needs of interested members of the public, much as is the case for access to the scholarly literature.

Questions from the RFI

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?*

   Use of scientific data by the public is a less crucial concern than the accessibility of digital data by other scientists. It is access and re-use by other scientists that will improve the productivity of the American scientific enterprise. Most scientific data themselves are not easy to monetize, so public accessibility follows straightforwardly once data are available to professional researchers. The AAS is in general agreement with the Interagency Working Group on Digital Data (IWGDD) 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." [1] Agency policies should support and encourage a distributed system for both access and preservation. Once community-based repositories are in place and in use by a community, agencies and other entities such as learned societies and journals can insist on deposit of digital data. Deploying mandatory deposit policies in the absence of trustworthy repositories exacerbates challenges in communities already struggling with incompletely coordinated efforts to manage the increasing amount of data being produced. Community-based repositories need to be supported first, and soon.

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 need to be compliant with applicable copyright and patent laws. Most astronomical facilities guarantee a proprietary period for researchers who collect digital data, and this seems a sensible policy broadly. Astronomy, however, is not biomedicine, so there tend to be fairly few secondary IP issues in our discipline.

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?*

   We believe that disciplinary differences are real and important and can't be – and shouldn't be – homogenized away. To that end, the most critical thing for the government to do is to be aware of those differences and to respect them. That will require the maintenance of discipline-specific apparatuses for research prioritization, for reviewing research proposals, and for assessing facility and infrastructure effectiveness. This includes any committees and task forces empowered by the government to oversee data management infrastructure.

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?*

   It is neither practical nor intellectually desirable to keep everything; some data are not worth preserving. So an important element of cost-effectiveness is recognizing this fact, and allowing for the disappearance of insignificant data. Being able to distinguish data that matter from data that are ephemeral is critical: therefore a process of evaluation (by peers, technology experts, etc.) is appropriate. The distinction is not as simple as choosing one type of data over another. The size and complexity of both data sets and any necessary post-processing streams must be taken into account. The availability of high-quality descriptions of data sets is important.

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

   Data management plans (DMPs) need to operate (minimally) at two levels. One level has to arrange for the near-term management of digital data from the current experiment, to ensure its quality and its availability to others investigating the research problem. Another level, a different set of considerations, is needed to address issues of long-term stewardship and preservation. On the question of long-term data management, it would seem that effective DMPs would subject data to a "publishing" process. A publishing perspective would seem quite sensible for considering the curation necessary for the long-term preservation of data sets.

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

   Support for (inevitable) deposit fees must be available for researchers whose data will be deposited in community-based repositories for long-term preservation.

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?*

   There are two aspects to issues of compliance: the quality of the data repositories themselves as "trustworthy" parties for the management of data sets, and whether or not individual researchers are complying with mandates for data deposit in trustworthy repositories. On the first concern, it seems to be appropriate for the government to allow the academy to manage and maintain these certifications. Organizations exist that either already perform these tasks, or could perform them with nominal broadening of scope and governance (the ICSU World Data System [2], e.g.). We anticipate that the trustworthy repositories will be oriented along disciplinary lines, and will necessarily be international, and for those reasons it makes sense for the US government to use a light touch at most in exercising control over these resources. We presume that individual researchers will be subjected to some level of mandatory deposition of data gathered in the course of federally-funded research. (See our comments to question 4.) The policies employed by the National Institutes of Health (NIH) for ensuring that researchers comply with mandated deposit of articles in the PubMedCentral repository could serve as a model for the policies and procedures that might be used for mandatory data deposit, with the proviso that they may need refinement on a disciplinary, and possibly a repository, basis.

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?*

   As the IWGDD has noted, there is a lack of a comprehensive framework for long-term data management in most disciplines. In astronomy, there have been efforts to resolve that shortcoming over the years, most recently in the form of "virtual observatories", and these have resulted in fairly effective channels of communication as well as a collection of standards and procedures for managing digital data across wide scales. National governments (not just the US) have a role to play in ensuring the development of the comprehensive framework envisioned by the IWGDD. This should take the form of continued support for efforts in broad disciplinary organization (like the virtual observatories and international alliances), and also support for the creation of trustworthy repositories in appropriate niches in the academy.

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?*

   In general, citation and attribution of data resources should follow the examples set in the academy for citing articles in the scholarly literature. There is a good deal of discussion in academic circles about the need for data to be regarded as "first-class objects", and it is important for that to happen. The feeling among many astronomers is that the astronomy community is already well down that road, with data set creation being considered (albeit on an ad hoc basis) by tenure and promotion committees. The broader interests of attribution are served in the community by an efficient and well-understood mechanism for data citation, akin to citing the literature. DataCite [3] is an international coalition whose purpose to build a framework for persistent identification of data sets and for the evolution of policies and practices for citing data so that appropriate credit can be assigned to data set "authors".

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.*

    There are good examples of data formats in many disciplines: in astronomy, the data format of choice is called "FITS", which stands for Flexible Image Transport System [4]. We agree that the purposes stated in the question are important, but rather than trying to name specific standards that are good for those purposes, it might be better to consider the *properties* of the formats, such as FITS, that work well. In our experience, those properties (certainly as they relate to FITS) are that: the standard is community-sourced (defined by the community and governed by on-going community efforts); the standard should be well-documented, and the definitive documentation should be openly and permanently available (the FITS standards are published in the astronomical literature); and the format needs to be widely adopted by both the researchers in the community and the groups in those communities that build the tools for managing and analyzing data. In addition to pure format considerations, for broad data interoperability it is also necessary to have agreed-upon metadata elements and semantics. Metadata semantics should be defined in a way that is nominally independent of specific data formats to permit multiple data formats to co-exist in research niches, and so that data formats can evolve over time.

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?*

    In addition to FITS, we would also cite the development of the Digital Object Identifier (DOI) [5] (mostly by the publishing industry) for persistent digital object identification, and the creation of the Dublin Core [6] (spearheaded by the library community) for core metadata semantics. Those standards by and large have the properties we described in our response to question 10.

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

    The committees and task forces that are assigned to oversee data management infrastructure (question 3) should be charged with maintaining awareness of standards efforts, and for participating in appropriate forums for standards development. The government's committees have to be well-enough informed so that the government can (credibly) endorse effective international programs.

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

    Conventions and mechanisms for these purposes are being investigated in the academy today. The most prominent coalition is DataCite (question 9); the AAS supports and participates in DataCite through an alliance with the California Digital Library. The technological standard being proposed by DataCite to support linking is the persistent

identifier, of which the DOI is an important example because of its use in the scholarly literature for these same purposes.

References

1. IWGDD report, *Harnessing the Power of Digital Data for Science and Society*, January 2009, http://www.nitrd.gov/About/Harnessing_Power_Web.pdf
2. http://www.icsu-wds.org/
3. http://www.datacite.org/
4. Wells, D. C., Greisen, E. W., and Harten, R. H. 1981, *Astronomy and Astrophysics Supplement*, vol.44, p.363.
5. http://www.doi.org/
6. http://dublincore.org/