Public Access to Digital Data Resulting from Federally Funded Scientific Research

http://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific#p-32

RFI #2011-28621
Released 11/04/2011

Responders:

Daniel Crichton
Daniel.J.Crichton@jpl.nasa.gov
Planetary Data System, Engineering Node
NASA Jet Propulsion Laboratory
Pasadena, California 91109

Faith Vilas
fvilas@psi.edu
Planetary Data System, Chief Scientist
Planetary Science Institute
Tucson, AZ

J. Steven Hughes
Steve.Hughes@jpl.nasa.gov
Planetary Data System, Engineering Node
NASA Jet Propulsion Laboratory
Pasadena, California 91109

Susan Slavney
Planetary Data System, Geosciences Node
slavney@wunder.wustl.edu
Washington University
St. Louis, Missouri

Reta Beebe
rbeebe@nmsu.edu
Planetary Data System, Atmospheres Node
New Mexico State University
Las Cruces, NM

Raymond Arvidson
Planetary Data System, Geosciences Node
arvidson@wunder.wustl.edu
Washington University
St. Louis, Missouri

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?

There are a number of policies that would encourage the treatment of data as a national asset in combination with encouraging the use of the data to enhance scientific discovery. These include:

1. Agency requirements for sustainable infrastructures: Agencies should be required to invest in and maintain public archives. These archives should be considered essential "facilities" and provided sustained funding.

2. Requirement for funding: The delivery of data, within a specified amount of time, to national, public archives, should be a requirement for funding public scientific research.

3. Release after a period of time: Scientists and data providers should have a period of time in which they have exclusive access and use of the data prior to delivery to pubic archives.

4. Capture data in reliable formats: Data should be captured in long-term, sustainable data structures that limit the use of proprietary data formats. Ample descriptive information (e.g., metadata) should be provided to support interpretation of the data long-term.

5. Auditing and certification of "official" archives: The U.S. should establish core guidelines for public archives and perform regular auditing for compliance against those guidelines.

6. Ensure active participation of discipline experts: The involvement of discipline scientists in sustaining the data is critical to ensure proper preservation and usability of the data.

7. Policies that encourage the use of data from public archives: Funding should be made available from agencies supporting scientific research that requires analysis of data from public, scientific archives.

(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?

Encouraging the public release of data is a critical step in sharing scientific research results. Often the incentive to share data is the result of a funding requirement. As such, scientists and data providers will do the minimum to prepare and share the data. If sharing the data affects their ability to publish and/or receive proper recognition for the research results, the incentive is diminished further. Incentives need to be in place whereby researchers are rewarded for sharing data. Furthermore, a specific set of practices can be put in place to protect researchers' IP interests better and encourage data sharing if it can be considered on the caliber of a peerreviewed publication. These include:

1. Release after a period of time: Scientists and investigators who have acquired the data,should have exclusive access to the data for a period of time. This should be for the purposes of improving the reliability of the discovery and results as well as corresponding publications.

2. Separate Intellectual Property (IP) from data: Specific algorithms and methodologies used to support the acquisition and generation of the data may be considered intellectual property of the investigator. While capture of the provenance information that produced those data is critical to understanding the heritage, there are opportunities for releasing public data while protecting the specific techniques used by the investigator.

3. Citations: Develop citation of data in public archives as a standard, scientific practice. A citation of high quality data should be considered equivalent to the citation of a scientific publication. Public archives should provide support for data citations.

(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?

Scientific involvement and oversight in establishing and maintaining usable scientific archives is essential.

1. Discipline leaders need to be involved: As prominent experts in their fields, discipline leaders should create a bridge between the scientific community and the public archives ensuring that the archive is scientifically useful and that the local policies are consistent with the scientific needs.
2. Peer review: Scientific review of the data is important for ensuring usability of both the metadata and data itself. Peer review ensures that the metadata can effectively used to annotate and understand the data. For the data itself, it helps in ensuring scientific usability.
3. Local standards: Proper annotation of the data is dependent on having discipline-specific descriptions that can be used to describe the data fully.
4. Sustainable infrastructures (establishment of national archives): Sustainable infrastructures are critical to ensuring long-term stewardship. These infrastructures must address use of the data for specific disciplines and therefore should involve the discipline experts in their implementation and operations.

(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 take a long-term, agency-wide view of preserving data in formats that support long-term analysis and use. Rather than fund each individual research project, agencies should develop sustainable infrastructures that are separately funded. This is essential for treating these infrastructures as facilities.

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

management plans?

The community should participate in the implementation and operations. This can be accomplished in multiple ways including:

1. Open source development: Development of the necessary computing infrastructure and publication as open source software. This is an effective approach for fostering collaboration.
2. Peer review: As mentioned, involvement of the community in review of the data is an effective way to foster collaboration.

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

Cost models need to be developed that take into account the entire data lifecycle including the supporting infrastructure. Funding is essential for ensuring data are captured and preserved for long-term usability through an infrastructure that supports effective access to the data. As a result, recommendations include:

1. Develop cost models: Cost models should be developed for disciplines that address both the cost of preparing and ingesting data as well as the long-term preservation. These costs should identify both the costs for the data supplier as well as the archive itself.
2. Archives as facilities: Archives should be treated as facilities. Their funding should originate from a sustained, operational allocation rather than be funded per experiment or investigation. Individual investigators should have a portion of their funding allocated towards preparing and submitting data to national archives.
3. Audit of archive systems: Archives should be regularly audited to ensure they meet Federal requirements for preserving their data and guaranteeing access.

(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?

Agencies should provide ample documentation and support to investigators to help them plan for archiving their data into public archives. Archive centers should work with the investigators to help them plan for capture of their data early in the lifecycle of a project to ensure the burden is minimized and adequate tool support exists. In addition, both peer review of the data and auditing of the archive itself should be performed. Peer review should assess whether the data meet the necessary requirements for compliance against standards and usability for scientific research. Agencies should consider auditing by an independent ISO-approved auditing body that follows a process by which digital repositories can be formally evaluated in terms of their ability to preserve the digitally-encoded information with which they have been entrusted.

(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?

Funding should be made available from agencies supporting scientific research that requires analysis of data from public, scientific archives.

(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?

Develop citation of data in public archives as a standard, scientific practice. A citation of high quality data should be considered equivalent to the citation of a scientific publication. Public archives should provide support for data citations.

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.

Data standards need to be defined at different levels. Discipline-specific data standards are critical and must be funded as a community effort. These data standards need to be defined and managed as "open standards" that can also allow for international adoption. The NASA Planetary Data System, for example, has developed standards for annotating planetary science data that are used world wide. They are developed as part of a community-wide effort with cross-disciplinary experts from both planetary science and computer science.

Furthermore, standards that support the development and long-term preservation of national archives must also be funded. Examples such as the Open Archive Information System (OAIS) [ISO 14721] are important for defining reusable standards that cross disciplines. These types of discipline independent standards on both the data and computing infrastructure are important to define.

In addition, best practices for the construction of long-term data and archive systems are also important to develop. These should link to agency and national priorities and requirements for preserving data and be used as a basis for auditing the implementation and operations of such systems.

(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 Planetary Data System (PDS) is a good example of a community-driven effort that has been effective for use in capturing and managing the scientific results from NASA solar system missions. The PDS has assembled a data standards design team that includes discipline-specific experts in planetary science who are responsible for the capture and curation of data specific to their area of planetary science. Each of these representatives works with an expert team who designs, implements and operates a set of integrated standards that span the entire discipline. These standards are integrated and published as the basis for capturing planetary science data for long-term archive for assembling, documenting and preserving data in stable formats. NASA Announcements of Opportunity require the use of these standards by investigators who generate data during a mission or investigation.

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

Working internationally can be a challenge due to different priorities and limited budgets. However, working in an "open standards" environment is essential. In addition, creating alignment with the international scientific community is a must. The NASA Planetary Data System worked to establish the International Planetary Data Alliance (IPDA) as a body for distributing its data standards internationally. The IPDA has become a vehicle for distributing the data standards. Many agencies have not had the budgets necessary to develop their own standards and therefore have been open to adopting the standards. Rather than distributing the U.S. standards directly from the U.S., the Planetary Data System is working to distribute these through an international organization for which the agencies are stakeholders. This has helped to pave the way toward international adoption. In addition, the IPDA aligned itself with the international science community through the Committee of Space Research that passed a resolution recognizing the effort. The critical goal was ensuring that there is "one" data standards effort for capturing planetary science data archives, rather than several individual and independent efforts.

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

Data are often not treated as publications. Publications generally have a well-defined structure for data annotation and standard practices for citing data. Data need to be treated as a publication going through a peer review and getting cataloged with well-defined annotations. The registration of data should be done using a standard set of metadata for describing the data. That standard set of metadata should be independent of any one discipline defining a universal common set of data elements/attributes (such as Dublin Core) which can be adopted by agencies implementing data archive systems. In addition to defining a standards set of attributes, a standard scheme should be defined which references the data much like that of a journal paper identifying the authors, title, dates, and location of the data. Researchers that use the data should be required to cite it in their papers as a key reference.