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

National Snow and Ice Data Center
Boulder, CO
January 2012

The National Snow and Ice Data Center (NSIDC) is a leader in the data science community with decades of experience supporting science through the archiving of ethically open polar and cryospheric data from around the world, including data produced with NASA, NOAA, NSF and other agency funding. As an organization, we are committed to the principles of open data and data stewardship, and we strongly support the IWGDD and NSTC efforts to further develop and harmonize policies for the ethical sharing, access, and preservation of digital data resulting from federally funded scientific research.

Data and metadata issues - including capture, access, discovery, standardization, security, ethics, preservation, interoperability, and synthesis among others – are all extremely complex and dynamic. Therefore, we believe that policies will play a strong though partial role in shifting the trajectory of research science towards a culture of open and ethical data sharing.

The NSIDC response to RFI questions follows:

(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?

Based on years of experience in data science, NSIDC recommends a clear, government-wide policy supporting open and ethical data sharing through federated repositories with professional expertise.

Data provide greatest value when they are viewed as a common, networked good rather than as individual or intellectual property. As such, data should be as fully open as possible with restrictions based only on ethical rather than proprietary concerns. Our more than thirty years experience, especially our involvement with the very large, interdisciplinary International Polar Year (IPY), has taught us several lessons on policies around open data.

1. Clear, explicit policy mandating that research data be openly available soon after collection helps make data systems simpler, more robust, and more adaptable.
2. Limited ethical restrictions should be clearly identified. For example, the IPY Data Policy notes legitimate ethical restrictions of data about human subjects, local and traditional knowledge, and where data release may cause harm.
3. We see the greatest success in achieving timely and open data access when data are required to be deposited in an open archive. The requirement should be supported by identified, funded archives and professional data scientists working with the data providers.
4. Licenses, contracts, and other legal agreements restrict the usability and interoperability of data and our ability to address interdisciplinary challenges, because they restrict our ability to use machines to discover, manipulate, and

repurpose data. Wherever possible, legal proprietary rights to data should be waived and data should be exposed to the web in a way that machines can readily interpret as open. Ethical considerations such as fair attribution and accurate documentation of quality should be based on scientific norms rather than legal mandate. This is the concept of an information commons. See, for example, polarcommons.org.

5. While various agencies are recognizing the value of shared data services and are funding development of shared repositories, tools, and services, the funding in many cases has followed the traditional term-limited model. Data preservation and access efforts have been hampered by this short-term structure, and data curated by term-funding in some cases have been put at risk when the repository funding ends. Secure and stable funding of infrastructure and expertise would allow the growth of a knowledgebase, a foundation for functionally enabling long-term interdisciplinary data reuse.

How these policies help grow the U.S. economy and improve the productivity of science will be influenced by a number of factors including the rate of cultural change within the domain communities. Policy, in concert with other efforts such as the including basic data management skills in academic curricula and the supporting ongoing data science research, work in concert to promote change and advance scientific understanding.

(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?

In our view, the critical consideration is the advancement of science and not of scientific institutions or individuals per se. As discussed above, publicly funded data (as opposed to creative works) should be viewed as a common, networked good. Proprietary considerations should be minimal to non-existent. In the current domain, data sharing is uneven across individuals and disciplines. This creates an unequal playing field for scientific researchers. Funding agencies can address this. They can demand that researchers all play by the same rules of openness and they can provide greater recognition and support for data work. This de-emphasis of individual knowledge in favor of an information commons approach provides the greatest overall benefit to the U.S. and indeed to humanity.

(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 policy-creation process should remain dynamic and flexible, encouraging participation in governance from multiple stakeholder groups while acknowledging that policy alone is not enough to motivate full compliance. Policy must at some point either align with practices on the ground or remain abstracted; engagement with stakeholders and addressing their differing concerns through an iterative process helps align policy with practices. As practices are in a constant state of change, continual bi-directional communication amongst all stakeholder groups is essential when trying to bridge disparate

domains. "A critical challenge [is] in making policy formation a dynamic, interactive process involving all stakeholders" (Parsons, 2011).

Supporting the work of data scientists is important as well; they break down barriers between projects and domains and bring data into an interdisciplinary context. Domain-based data centers for instance, can translate the data and complex local context of that data from individual PIs into standardized form that others from outside domains can access, understand and reuse. Supporting data scientists at local, domain, and cross-domain levels, as might be implemented through a federated web-of-repositories (Baker and Yarmey, 2009), represents a strategy for preserving data and making them accessible for reuse at multiple levels. Leveraging data science expertise maintains the deep domain knowledge of local work through cross-disciplinary mappings into the broad context of interdisciplinary research.

(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?

First, proposal-stage data management planning should include consideration of resulting data products and costs for maintaining data through their entire lifecycle. "End-to-end data management (which includes acquiring, processing, storing, maintaining, updating, and providing access to data) should be planned and budgeted at the outset of any activity that will generate environmental data. This planning should explicitly address data archiving, data stewardship, and data access responsibilities, and sufficient funds should be provided to archive and provide ready and easy access to the resulting data for extended periods of time" (NRC, 2007).

Second, professional data managers should be supported as part of data access and preservation infrastructures (NSB, 2005). This mediating layer of expertise provides a lens through which the costs and benefits of data preservation can be weighed. Data scientists contribute cross-cutting vision and interdisciplinary understanding to multi-stakeholder discussions on maximizing the long-term benefit of valuable data resources.

Third, different levels of service should be created and applied for the preservation and access of each data set depending on importance and community applications (Weaver et al., 2008). These are potential responsibilities for data scientists in collaboration with other stakeholders. Policies should recognize the range of relative contributions and applications of diverse data sets and support strategic delegation of resources by data scientists and curators.

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

Generally - For the successful implementation of data management plans, many stakeholders with a diversity of roles, responsibilities and expertise must come together under a common goal of ethical data sharing. Recognizing the need for a diversity of perspectives at the table – including those of the scientist and research team, data curator, technologists, metadata experts, digital archivist, user services personnel, and others – is an

initial step towards supporting DMP implementation. Within the context of each support organization, roles and responsibilities must be clearly defined and delegated to the stakeholders with the appropriate expertise. All stakeholders should promote policy compliance along with community-based standards-making.

Universities and research organizations - Clarify intellectual property statements regarding data and promote data publication and sharing as part of tenure considerations. Recognize and support coordinated data management and infrastructure systems and services as valuable institutional facilities for researchers. Offer career paths for data managers and data scientists on campus. Include data management training as part of the science curriculum.

Research communities – Recognize researcher efforts to preserve and make data accessible and usable when considering rewards for professional achievement. Proactively contribute to standards-making discussions and the cultural shift towards data sharing and complete metadata capture.

Academic Research Libraries – Consider data as not only part of the scholarly communication cycle, but as a resource to be curated. Apply extensive experience in cataloging and bringing together diverse, interdisciplinary resources to the data preservation and access paradigms.

Publishers – Enable and encourage ethical data sharing and attribution through citations within publications and research documentation. Recommend authors make associated data freely available through open data repositories.

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

With the recognition that data are a common scientific good, it becomes imperative that 1) the substantial investment in generating data for research must be protected through preserving and making openly accessible those data for use in additional research, and 2) that the funding responsibility must be borne across all of the sciences. Quality data stewardship starts in the proposal stages, is enacted through the field- or experiment-based management of generated data, and continues through the useful lifespan of data products. Funding should include all aspects of data management; if you pay to collect the data, pay to preserve them. Funding the support of data management could be viewed as "tax" on scientific research used to maintain valuable data resources for all science.

In many ways the real costs of preserving and making data accessible are just beginning to emerge. Recognizing that data preservation and stewardship are not solely technical problems is one step towards uncovering the true costs involved. The cost of data preservation for reuse includes not only the infrastructure and long-term system maintenance fees, but also the expenditure of defining, capturing and structuring necessary metadata, ongoing standardization work, prioritization of data resource allocation, and user support.

Especially for valuable observational data from the 'small' sciences, the largest cost involved potentially comes when not only making data accessible over the long-term, but making data useful well into the future. The amount and quality of metadata required for reuse potentially dwarf the metadata required for access and preservation. One of the biggest obstacles to data reuse is the capture and standardization of metadata. Metadata is at the core of many data reuse issues, including discovery, trust, cost and data quality. To realize the

goals of growing the economy and improving productivity of the scientific enterprise through the preservation of and access to research outputs, we must have a comprehensive strategy for metadata capture and 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?

Current trends in the academic sciences add to the burden of compliance and verification on PIs. First, changes in research funding, for instance a shift from large, single-domain labs to interdisciplinary collaborations of smaller labs, mean a gap in infrastructure support. The traditional lab-based infrastructure including local expertise and hardware has become more difficult to support. Recognizing this, shared resources are under development by many domain groups, universities and other communities of different sizes and foci. However, many of these efforts are operating in temporary funding environments and data access and preservation services remain piecemeal (ex. NBII).

Second, the continuing changes in technology, while opening up potential avenues of research, mean a constant demand on researchers to keep up with changing practices, communities, and policies. Funding for shared and coordinated resources at different levels can reduce researcher burdens by shifting some of the responsibility for metadata creation and structuring, translating requirements and keeping up with tools and services. For instance, the Advanced Cooperative Arctic Data and Information Service (ACADIS) project supported by NSF offers researchers a central portal for information, infrastructure, tools, and expertise to support their efforts to meet NSF Data Management Plan and data sharing requirements.

Along with community-based data management support, repositories like ACADIS are also a step towards quantifiable verification of policy requirements. The presence of a researcher's data in the ACADIS repository clearly indicates compliance with NSF program requirements to program managers. Generally, policies should require deposit into an openly accessible repository as a condition of continued funding. A federated network of repositories and a clear government-wide policy supporting sharing of ethically open data through use of these repositories enable generalizable measurement, verification, and compliance checking.

(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?

NSIDC recommends a clear, government-wide policy supporting sharing of ethically open data through both human- and machine-accessible systems. Through basic APIs, open data are re-usable at scale and accessible to tools in addition to individuals. This allows people to build new and creative applications and services on top of the data that can provide new markets. Consider, for example, the growth of weather apps, visualizations, etc. that resulted from NOAA making their meteorological data more open and machine accessible.

(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?

Data citation standards (Ball and Duke, 2011; ESIP, 2011), in concert with administrative metadata standards and publisher acceptance and support of expanding traditional attribution mechanisms are a first step towards proper attribution of data. Research into application of data citation standards is ongoing.

(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 and metadata standards are emerging and require time to be refined and stabilized through implementation and iterative design. For interoperability, reuse and repurposing of digital scientific data, standards are needed at many different levels. Community-driven efforts to refine and formalize language and representation are necessary and important within a domain reuse and repurposing context. Broader, more general standards are necessary to promote interoperability across disciplines as well as nationally and internationally. A package of human- and machine-readable standards along with crosswalks, tools, and open and accessible data and metadata enable reuse and repurposing.

To encourage these combinations, spread support across international standards development and ontology research while recognizing the important role of community-driven standards creation and implementation. In the Earth sciences, the high-level ISO19115 metadata format and the NetCDF data format are emerging as useful standards. At a community level, the QARTOD->OGC effort has been specifically focused on quality specifications, data dictionaries and sensor-based metadata description for the oceanographic sciences. These examples are not only compatible but complimentary. For top-down, high-level standards to effectively intersect with bottom-up, detail-oriented standards and best practices there will need to be a decentralized model of support, communication, and governance driven by the clear goal of formal integration.

(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?

Application tools can often define a standard. For example, KML rapidly emerged as a community standard that was then adopted by OGC because of the popularity of Google Earth and other virtual globes. In another example, UNIDATA worked with the Atmospheric sciences community to develop tools along with netCDF to demonstrate how useful that standard could be.

Standards development needs to be an open, accessible, iterative, and well-supported process with emphasis on communication, community engagement and formal maintenance mechanisms. The Global Change Science Keywords (GCMD) provide an interesting example. They have been very popular and were central to the search interfaces and organization of many Earth science data systems around the world. The keywords were all

defined and openly accessible. The GCMD managed and controlled the list, but they were open to modifications from certain communities, if a need could be demonstrated. The keywords also had a certain amount of internal flexibility with a free form "detailed variable" that could be added anywhere at the end of the hierarchy. Now, however, the GCMD science keywords are losing relevance; they have not evolved well to keep up with modern data systems. The community has repeatedly asked the GCMD to make the keywords and their definitions available as a web service, but the GCMD has yet to provide one. Further, the GCMD made a major revision to the keywords without sufficient community consultation and the revision was not well received. As a result, the GCMD is using the new version internally, but it is not broadly used outside of the organization. External groups have begun to develop and use independent web services based on the older version of the keywords. This is bound to lead to some level of divergence in the "standard". Our point here is not to criticize the GCMD, which remains a valuable resource, but rather to highlight the need for community engagement, formal standard maintenance, and ongoing technical evolution.

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

To promote effective coordination on digital data standards internationally, we recommend a multi-level strategy with the following three elements:

1. Work with international organizations such as ICSU and intergovernmental organizations such as WMO and UNEP to promote harmonization of data policy around ethical openness. The Belmont Forum may be an especially appropriate venue (http://www.igfagcr.org/index.php/belmont-forum).
2. Support national- or discipline-based data coordinators to work with counterparts in other nations around common international initiatives. There is growing international interest in standardizing Arctic observing data for example.
3. Seek collaborative science funding opportunities with individual nations or groups such as the EU around common interests like Arctic observing and then develop a common data policy as part of the collaborative effort.

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

Data citation standards and a clear policy mandating ethically open data accessible through both human- and machine-readable system are the first steps towards linking publications and associated data. With community acceptance of data citation practices and a persistent, machine-readable link to associated data in an open archive, tools to take advantage of the linked environment will emerge in response to specific needs.

References

Baker, KS, and L Yarmey. 2009. Data stewardship: Environmental data curation and a web-of-repositories. *International Journal of Digital Curation*. 4(2). Available online: http://www.ijdc.net/index.php/ijdc/article/view/115

Ball, A, and M Duke. 2011. 'Data Citation and Linking'. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers/

ESIP (Federation of Earth Science Information Partners). 2011. Interagency Data Stewardship/Citations/provider guidelines. Retrieved January 7, 2012 from: http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines

NRC (National Research Council). 2007. *Environmental Data Management at NOAA: Archiving, Stewardship, and Access.* Washington, DC: National Academies Press. 116 pp. Available online: http://books.nap.edu/catalog.php?record_id=12017

NSB (National Science Board). 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* Washington, DC: National Science Foundation. 87 pp.

Parsons, M. 2011. Expert Report on Data Policy and Open Access. GRDI2020. Available online: http://www.grdi2020.eu/Pages/SelectedDocument.aspx?id_documento=e31a1aab-b01e-4e7e-9b10-0fd93d4b710f

Weaver, RLS, Meier, WM, and RM Duerr. 2008. Maintaining Data Records: Practical Decisions Required For Data Set Prioritization, Preservation, and Access. *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International 7-11 July.* 3:III-617 - III-619. Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4779423&isnumber=4779256