NDSA
NATIONAL DIGITAL STEWARDSHIP ALLIANCE

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

Submitted by the National Digital Stewardship Alliance (NDSA)

January 2, 2012

Introduction to the NDSA

The National Digital Stewardship Alliance (NDSA) was founded in July 2010 to extend work begun in 2001 by the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress. The Alliance has over 100 members from educational institutions, non-profit organizations, businesses and local, state and federal government agencies, as well affiliations with international organizations. Its mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. [1] Members of the Alliance are taking action to preserve access to our national digital heritage by:

- broadening access to our nation's expanding digital resources
- developing and coordinating sustainable infrastructures for the preservation of digital content
- advocating standards for the stewardship of digital objects
- building a community of practice around the management of distributed digital collections
- promoting innovation
- facilitating cooperation between government agencies, educational institutions, non-profit organizations, and commercial entities
- fostering the participation of diverse communities and relationships across boundaries
- raising public awareness of the enduring value of digital resources and the need for active stewardship of these national resources.

Supporting communities of practice for preservation and access

The values of the Alliance are highly relevant to establishing approaches for ensuring long-term stewardship and encouraging broad public access to unclassified digital data that result from federally-funded scientific research. When applied, these values support the practical development of communities of practice capable of gaining consensus to support preservation and access to digital data. The shared expertise and common experience of these communities result in stakeholder buy-in and adoption of policies and

standards. The National Digital Stewardship Alliance member organizations are bound as a community by the following values.

Stewardship. Members of the NDSA are committed to managing digital content for current and long-term use. The members of the NDSA are actively ensuring sustained access to the digital content that constitutes our national legacy and empowers us as leaders in the global knowledge economy. Individually, these organizations support the management of digital resources; the Alliance is committed to protecting our nation's cultural, scientific, scholarly, and business heritage.

Collaboration. Collaborative work is the centering value of the Alliance; it is a value shared by all members and a priority in work with all organizations and associations. Approaching digital stewardship collaboratively allows the NDSA to coordinate effort, avoid duplicate work, build a community of practice, develop new preservation strategies, flexibly respond to a changing economic landscape, and build relationships to increase capacity to manage content beyond institutional boundaries.

Inclusiveness. The NDSA is a collaborative effort to preserve a distributed national digital collection for the benefit of current and future generations. We value the range of experience, the potential for innovation, and the fault-tolerance that heterogeneity brings. We believe the preservation of digital information is a pervasive challenge and that engaging across different communities strengthens the nation’s digital preservation practices and increases the likelihood of preserving content now and into the future.

Exchange. Members of the Alliance encourage the open exchange of ideas, services, and software. This leverages the commitments of each member to increase the capacity of the entire stewardship network. Participation and engagement result in innovations and benefits that can be shared by all. The Alliance is committed to transparency and all products generated or produced by the Alliance will be circulated under open licenses.

Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

Community-based approaches to the challenges of rapid change and high volume within the data domain have proven to be the most successful in the long term. The Blue Ribbon Task Force on Preservation and Access recommended that for research data “Each domain, through professional societies or other consensus making bodies, should set priorities for data selection, level of curation, and length of retention.” [2]

The report validated experience over the last ten years of digital preservation work. A study of the networks developed through the NDIIPP program indicated that participating institutions bring to the network their own resources, interests, and organizational culture. Under the auspices of a neutral convener and honest broker, natural networks emerge over time through participation in shared activities and problem solving. As these networks form, the larger network becomes more complex, but also stronger and better able to withstand stresses and strains. [3]

The Opportunities for Data Exchange (ODE) project supported by the Alliance for Permanent Access and the European Union also takes a cross-cutting community approach to preservation and access to digital data. “The potential answers to grand challenges of our times require…the inclusion of an interoperable data sharing, re-use and preservation layer to the emerging eco-system of e-infrastructures…All stakeholders in the scientific process must be involved in the design of this layer; policy makers, funders, infrastructure operators, data centers, data providers and users, libraries and publishers…” [4]

An exemplar of collaborative community efforts is the Dataverse Network project [5] recently described by the National Research Council of the National Academies as the “State of the Practice in Data Sharing.” [6] The Dataverse Network is “unique in being designed to explicitly support long-term access and permanent preservation. To this end the system supports best practices, such as format migration, human-understandable formats and metadata, persistent identifier assignment and semantic fixity checking. In addition, many threats to long-term access can be fully addressed only by collaborative stewardship of content, and the system supports distributed, policy-based replication of its content across multiple collaborating institutions, to ensure the long-term stewardship of the data against budgetary and other institutional threats.” [7]

Foster public values and support for stewardship of digital data beyond mandating data management plans.

Policy should assert the value of research data and provide mechanisms to support the preservation, discoverability and access. To relieve frustration and confusion about actions the policy should provide a clear direction for funders, researchers and stewardship organizations. The Blue Ribbon Task Force recommended “Funders should impose preservation mandates, when appropriate. When mandates are imposed, funders should also specify selection criteria, funds to be used, and responsible organizations to provide archiving. They should explicitly recognize “data under stewardship” as a core indicator of scientific effort and include this information in standard reporting mechanisms.” [8]

Leverage substantial national and international efforts for common practices that support interoperability.

Substantial efforts have been made to pave the way for interoperability, re-use and re-purposing. Emerging practices for data citation, licensing and protocols for data sharing and sustainable re-use are becoming enough to adopt more broadly. Notable in these areas are work on the Data Seal of Approval by the Data Archiving and Networked Services that promotes sustainable access to digital research and provides training and advice about archiving and reuse.[9] LOCKSS is a community initiative that provides libraries with digital preservation tools and support so that they can easily and inexpensively collect and preserve their own copies of authorized e-content. [10] The

Data-PASS organization promotes collaborative, institutional stewardship of research data, permanent data archiving, and citation that permits results to be verified and re-purposed. [11] DataCite collaboratively addresses the challenges of making research data visible and accessible through data citation.[12] The Creative Commons project, Science Commons, has focused on protocols for sharing scientific data that includes licensing and mitigating legal barriers.[13]

Summary of Major Recommendations

- Support sustainable action through policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Establish policy that catalyzes collaborative work on preservation and access within and across scientific disciplines

- Foster public values and support for stewardship of digital data beyond mandating data management plans.

- Leverage substantial national and international efforts for common practices that support interoperability.

Additional Responses on Selected Questions

The principles and recommendations above apply broadly to the set of questions posed by the RFI. The responses below exemplify how the principles can be applied to the individual questions, and highlight relevant NDSA activities in these areas.

(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?

The most effective policies in this regard would mandate data deposit into publicly accessible repositories. In the absence of such a policy, there are already cases of data which have been lost. The Federal policy framework should move public access to data away from the current idiosyncratic environment to a systematic approach that lowers barriers to data access, discovery, sharing and re-use.

Many members of NDSA provide repository services at low cost or through cooperative arrangements. Members of the NDSA also provide repository services that provide legal, technical, procedural and statistical controls necessary to protect data confidentiality while ensuring long. And the NDSA provides a model of institutional collaboration that supports stewardship, discovery and accessibility An example of a free access service is ViewShare.org, a platform for empowering curators, archivists, and librarians to provide access to the digital collections they are preserving through a shared interface. This

service provides the dual benefit of making data more broadly available and accessible while also making it easy for end users to copy and make use of the data in other environments. [14] The NDSA content working group is also working toward developing a clearinghouse for at-risk digital collections to help match data to potential preservation partners.

(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?

Each domain and discipline should be empowered to set priorities for data selection through, level of curation, and length of retention, through professional societies or other consensus making bodies.

Notwithstanding, there are still baseline conditions or requirements that apply to all data regardless of discipline, particularly as they relate to archiving and preservation. For most data, "open access" is needed not only for the short term, but for the long term. And scientific disciplines have focused primarily on short-term access. There are critical standards for metadata exchange, fixity information and verification, and persistent citation that can support long-term access to data, preservation, and the long-term reproducibility of public results. Such baseline standards should be applied all scientific data. Among the range of important new standards for preservation and access there is still little knowledge about which standards are being implemented in which situations. The NDSA Standards working group is working on inventorying these standards and exploring how they are currently being used by NDSA member organizations. More than advocating the need for standards there is a clear need to understand which standards are being used in which situations and use that information to promote the usage of standards that are leading to results.

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

There exists diversity in approaches for data management within various scientific communities, which is healthy for various reasons. In cases where communities have resources for data management, it is worthwhile to build upon existing infrastructure (e.g., the Data-PASS archives). However, it is critical that even in these cases the community service provider demonstrates rather than assert capability. Far too often, terms such as archiving or preservation being used loosely without associated evidence of meeting specific requirements. Memory institutions such as archives, libraries and museums have an extensive track record with these functions and collaborative organizations such as NDSA could serve the essential purpose of developing or implementing frameworks that thoroughly test and certify assertions. In this respect, work from the NDSA innovation working group toward developing a "Neighborhood Watch" system for repository quality assurance could serve as the basis for establishing clear, externally verifiable reporting. [14]. The group has identified a pressing need for

an objective, repeatable, independently verifiable and simple way for an external agent to periodically retrieve content, verify its bit level integrity and publicly announce the results. This is a clear example of how assertions about data management could be tested and certified.

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

The most important step would be to communicate that the real costs of preserving and making digital data accessible are indeed legitimate and necessary costs of the overall research enterprise. Researchers routinely include publication costs within their research proposals -- the costs of ensuring long-term access reuse of data should be treated in the same way.

(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?

One of the key points in this context is that it is easier to verify compliance through systematic approaches. It is easier to verify compliance of library-based or community-based data archives than to check thousands of individual researcher hard drives. Technical infrastructure components such as persistent identifiers and appropriate licenses represent critical mechanisms through which compliance and verification can be automated thereby reducing costs.

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

There is widespread consensus within the research community that it is essential to link publications and underlying or associated data. Many NDSA partners are leaders in this area. The peer-reviewed publication is viewed as the final "snapshot" of the research process and outcome. One of the most important considerations from a policy, practices and standards is a requirement to use persistent, unique identifiers for publications, data, authors, etc. These identifiers not only bolster the linking of publications and data, but also help foster the re-use and development of new services by people and machines. While there are multiple identifier schemes, at this point, perhaps the most important policy decision would be to require using persistent identifiers instead of relying upon existing mechanisms such as website URLs.

References

[1] The National Digital Stewardship Alliance: http://www.digitalpreservation.gov/ndsa

[2] Berman, Francine, and Brian Lavoie, et al. 2010. *Sustainable Economics for a Digital Plant: Ensuring Long-term Access to Digital Information.* Final Report of the Blue

Ribbon Task Force on Sustainable Digital Preservation and Access supported by the National Science Foundation, et al. Washington, DC: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[3] Library of Congress. 2010. *Preserving our Digital Heritage: The National Digital Information Infrastructure and Preservation Program (NDIIPP) 2010 Report*. Washington, DC: http://1.usa.gov/hmw2lj

[4] Alliance for Permanent Access. 2011. "Opportunities for Data Exchange (ODE) Project": http://www.alliancepermanentaccess.org/index.php/current-projects/ode/

[5] King, Gary. 2007. An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research* 36:173-99.

[6] National Research Council. 2011. *Communicating Science and Engineering Data in the Information Age: Panel on Communicating National Science Foundation Science and Engineering Information to Data Users*. Preprint. Washington, D.C.: National Academies Press: http://bit.ly/NCSES

[7] Altman, Micah and Jonathan Crabtree. 2011 "Using the SafeArchive System: TRAC-Based Auditing of LOCKSS," Archiving 2011 Final Program and Proceedings, May 16–19, 2011, Salt Lake City, Utah: 165–170. Society for Imaging Science and Technology: http://bit.ly/tLzUmr

[8] Berman et al. 2010.

[9] Data Seal of Approval: http://www.datasealofapproval.org/

[10] LOCKSS: http://lockss.org

[11] DataPass: http://data-pass.org/

[12] DataCite: http://datacite.org/

[13] Creative Commons project, Science Commons: http://creativecommons.org/science http://wiki.creativecommons.org/Science

[14] ViewShare: http://viewshare.org

[15] Abrams, S, Cruse, P, Kunze, J, Minor, D, Smorul, M. 2011."Neighborhood Watch" for Repository Quality Assurance. Presented at Designing Storage Architectures for Preservation, Washington, DC: http://1.usa.gov/uXj2Mf