

NSF AND ARL CONDUCT WORKSHOP ON DIGITAL DATA STEWARDSHIP

To explore the challenges of digital data stewardship and preservation, ARL and the National Science Foundation (NSF) conducted a workshop in September 2006 on “New Collaborative Relationships: Academic Libraries in the Digital Data Universe.” The workshop was co-chaired by San Diego Supercomputer Center (SDSC) Director Fran Berman and University Librarian Wendy Lougee from the University of Minnesota, and organized by Prue Adler, ARL Associate Executive Director. The workshop report provides a wealth of information on the issues of digital preservation; the Executive Summary follows.

Executive Summary

The rapid adoption of information technology and ubiquitous networking has transformed the research and education landscape. Central to this transformation are scientific and engineering digital data collections. The life cycle management challenges associated with these intellectual assets are substantial.

This is the Executive Summary of a report of a two-day workshop that examined the role of research and academic libraries with other partners in the stewardship of scientific and engineering digital data. Workshop participants explored issues concerning the need for *new partnerships* and collaborations among domain scientists, librarians, and data scientists to better manage digital data collections; necessary *infrastructure development* to support digital data; and the need for *sustainable economic models* to support long-term stewardship of scientific and engineering digital data for the nation’s cyberinfrastructure.

The workshop builds on prior studies supported by the National Science Foundation (NSF), engaging numerous research communities. It reflects the recognition, voiced in many NSF workshop reports, that digital data stewardship is fundamental to the future of scientific and engineering research and the education enterprise, and hence to innovation and competitiveness. Overall, it is clear that an ecology of institutional arrangements among individuals and organizations, sharing an infrastructure, will be required to address the particularities of heterogeneous digital data and diverse scholarly and professional cultures.

Summary findings and final recommendations are presented below.

Findings

- The ecology of digital data reflects a distributed array of stakeholders, institutional arrangements, and repositories, with a variety of policies and practices.
- The scale of the challenge regarding the stewardship of digital data requires that responsibilities be distributed across multiple

entities and partnerships that engage institutions, disciplines, and interdisciplinary domains.

- Historically, universities have played a leadership role in the advancement of knowledge and shouldered substantial responsibility for the long-term preservation of knowledge through their university libraries. An expanded role for some research and academic libraries and universities, along with other partners, in digital data stewardship is a topic for critical debate and affirmation.
- Responsibility for the stewardship of digital information should be vested in distributed collections and repositories that recognize the heterogeneity of the data while ensuring the potential for federation and interoperability.
- Stakeholder groups have different expertise, outlooks, assumptions, and motivations about the use of data. Forging partnerships will require transcending and reconciling cultural differences. Collaboration models to share expertise and resources will be critical.
- Stewardship of digital resources involves both preservation and curation. Preservation entails standards-based, active management practices that guide data throughout the research life cycle, as well as ensure the long-term usability of these digital resources. Curation involves ways of organizing, displaying, and repurposing preserved data.
- Infrastructure for digital data resources is a shared common good and the digital data produced through federally funded research is a public good.
- The stewardship and sharing of digital data produced by members of the research and education communities requires sustainable models of technical and economic support.
- There is a need for a close linking between digital data archives, scholarly publications, and associated communication. The potential for an expanded role for research libraries in the area of digital data stewardship affords opportunities to address these important linkages.
- A change in both the culture of federal funding agencies and of the research enterprise regarding digital data stewardship is necessary if the programs and initiatives that support the long-term preservation, curation, and stewardship of digital data are to be successful.
- It is critically important that NSF and other funding agencies raise awareness and meet the needs of the research community for the stewardship and sharing of digital data.

Recommendations from the Workshop

Overarching Recommendation

NSF should facilitate the establishment of a sustainable framework for the long-term stewardship of data. This framework should involve multiple stakeholders by:

- supporting the *research and development* required to understand, model, and prototype the technical and organizational capacities needed for data stewardship, including strategies for long-term sustainability, and at multiple scales;
- supporting *training and educational programs* to develop a new workforce in data science both within NSF and in cooperation with other agencies; and
- developing, supporting, and promoting educational efforts to *effect change in the research enterprise* regarding the importance of the stewardship of digital data produced by all scientific and engineering disciplines/ domains.

Three general recommendations emerged around the following themes.

NSF should:

1. *Fund projects that address issues concerning ingest, archiving, and reuse of data by multiple communities.* Promote collaboration and “intersections” between a variety of stakeholders, including research and academic libraries, scholarly societies, commercial partners, science, engineering, and research domains, evolving information technologies, and institutions.
2. *Foster the training and development of a new workforce in data science.* This could include support for new initiatives to train information scientists, library professionals, scientists, and engineers to work knowledgeably on data stewardship projects.
3. *Support the development of usable and useful tools, including*
 - automated services which facilitate understanding and manipulating data;
 - data registration;
 - reference tools to accommodate ongoing documentation of commonly used terms and concepts;
 - automated metadata creation; and
 - rights management and other access control considerations.

These general recommendations and themes are amplified by the following targeted recommendations.

1. NSF should develop a program to fund projects/case studies for digital data stewardship and preservation in science and engineering. Funded awards should involve collaborations

between research and academic libraries, scientific/research domains, extant technologies bases, and other partners. Multiple projects should be funded to experiment with different models.

2. NSF, with other partners such as the Institute of Museum and Library Services and schools of library and information science, should support training initiatives to ensure that information and library professionals and scientists can work more credibly and knowledgeably on data stewardship—data curation, management, and preservation—as members of research teams.
3. NSF should support the development of usable and useful tools and automated services (e.g., metadata creation, capture, and validation) which make it easier to understand and manipulate digital data. Incentives should be developed which encourage community use.
4. Economic and social science experts should be involved in developing economic models for sustainable digital data stewardship. Research in these areas should ultimately generate models which could be tested in practice in a diversity of scientific/research domains over a reasonable period of time in multiple projects.
5. NSF should require the inclusion of data management plans in the proposal submission process and place greater emphasis on the suitability of such plans in the proposal’s review. A data management plan should identify if the data are of broader interest; if there are constraints on potential distribution, and if so, the nature of the constraint; and, if relevant, the mechanisms for distribution, life cycle support, and preservation. Reporting on data management should be included in interim and final reports on NSF awards. Appropriate training vehicles and tools should be provided to ensure that the research community can develop and implement data management plans effectively.
6. NSF should encourage the development of data sharing policies for programs involving community data. Discussion of mechanisms for developing such plans could be included as part of a proposal’s data management plan. In addition, NSF should strive to ensure that all data sharing policies be available and accessible to the public.

The complete report of the workshop, *To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering* (Washington, DC: ARL, 2006), is available on the ARL Web site at <http://www.arl.org/info/events/digdatarpt.pdf>.