A national digital data policy for the United States: to be or not to be?

James L. Mullins
Purdue University

A National Digital Data Policy for the United States: To Be or Not to Be?

James L. Mullins
Purdue University
USA
jmullins@purdue.edu

Abstract

As countries worldwide are coming to terms with establishing a national data policy, the United States is approaching the issue in a piecemeal manner. With numerous federal, state and private agencies in control of funding, it is unlikely that a national policy will emerge for the United States in the near future. Regardless, efforts are moving forward on digital initiatives, including open access to scholarly publications, access to digital data-sets, creation of standards for data-set management, and national repositories for scanned images. Consortiums of research libraries such as the Committee on Institutional Cooperation (CIC), Association of Research Libraries (ARL), Coalition of Networked Information (CNI) and the Digital Library Federation (DLF) are facing these issues and assisting with definition of challenges and options. Several not-for-profit agencies are investigating ways in which they can participate, including OCLC, JSTOR, Portico and LOCKSS. Commercial firms such as Google are establishing partnerships with research libraries. Major federal funding agencies including the National Science Foundation and the National Institutes of Health have issued statements about the need for a digital policy. Federal agencies such as the National Archives and the Library of Congress are participating by actively managing massive amounts of data.

Although there is activity on numerous fronts, there is no forum for a nationally concerted effort. While it is unlikely that a national policy on digital management will emerge in the United States, it is likely that within five to ten years a patch-work quilt of digital policies will emerge.

This paper will explore issues faced by the scientific and technical disciplines and the collaborative approaches developing between the research and library communities to meet these challenges. .

Introduction

In 2008 it is impossible to clearly define the breadth and depth of what a data management policy or plan would be for the United States. For the past ten years focus has been primarily given to the issues associated with maintaining digital files that contained scanned images of items originally published or created in print, or more recently, material that was “born” digital. During the last ten years efforts have culminated in the creation of software that has allowed for the storage, access, and preservation of digital textual data, for instance DSpace [MIT Libraries and Hewlett-Packard Company. 2008] and Fedora [Fedora Wiki. 2008].

It has only been within the last three to four years that the larger issue of discovering, managing, archiving, and preserving data-sets, primarily numeric data, has become a major issue for discussion within the research and library communities. Although there are still important issues associated with maintaining and curating digital text data, the focus of this paper will be primarily on the challenges associated with discovering, accessing and making available, now and into the future, numeric data associated with research and learning in the sciences and engineering. However, there will be a brief discussion of the relationship between the research article and the underlying data-set.

Overview of US Research Funding

The United States is both blessed and cursed by having a myriad of sources for research funding including the federal and state governments, private foundations, corporations, individuals, and not-for-profit organizations. Each of these funding entities has its own application, criteria, assessment,

and guidelines. Little if any mention is made about the availability or sharing of data generated by the research undertaken on its behalf.

The government on the federal and state levels provides funds for research. Federal agencies and departments such as the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Institute for Museum and Library Services (IMLS) are significant players by providing hundreds of millions of dollars each year for basic and applied research. In addition, the Departments of Agriculture, Defense, and Energy, along with several others, provide substantial funding in support of research. The governments of the fifty states also provide funding for projects that relate to economic development including agriculture, and environmental focused on issues involving soil and water. Even though it might seem logical and efficient that agencies that have funding entirely from a governmental taxing authority would have a consistent set of standards for data management, it is not the case. It is not the case for the federal and state, or the federal alone. Each funding agency determines its own specifications and guidelines. To further complicate this picture, the United States Congress, which determines the funding level for the agencies, until very recently had little or no interest in requiring or prescribing data management.

Combined efforts by several of the federal funding agencies, the research and library communities have brought to attention to this major challenge. The major thrust of the argument for better stewardship of the resulting data or findings from federally sponsored research is, 1). Since the taxpayer pays for the research shouldn't they have open access to the results? 2). Data that is generated by a federally funded research project is not being "mined" to its highest potential and thereby insuring the greatest contribution to further research and the advancement of science.

An Illustration:

The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 "to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense…"[National Science Foundation. 2008. *About the National Science Foundation*]. Each year the NSF expends nearly $6.0 billion and is the funding source for approximately twenty percent of all federally supported basic research undertaken by academic research institutions in the United States.

Early in this century, the National Science Foundation became acutely aware that the basis of research in the sciences was changing, from the classical two approaches to scientific research of theoretical/ analytical and experimental/observational, to simulation and modeling to explore new possibilities and to achieve new precision [National Science Foundation. Blue Ribbon Advisory Panel] This report, issued by the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January 2003, often referred to as the Atkins Report, provided in depth analysis of the role computational applications can and would play in the advancement of science and engineering research. One recommendation of the report called for one office within the NSF to coordinate and advance cyberinfrastructure across the disciplines/domains.

In 2004 the National Science Board (the governing body of the NSF) requested that a report be generated on the issues confronting the scientific and engineering community. After a fact-finding tour of university and research labs around the United States, a report was generated that defined the challenges and opportunities the scientific research community were encountering. This report "*Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century*" was released September 2005 [National Science Foundation. 2005]

This report was the first major statement about the necessity of gaining control of the massive amounts of data being generated by researchers in science and engineering. Among several recommendations were ones that called for data management plans for the long-term access and preservation of data and the creation of or a better understanding for the role of professionals in the management and curation of data. The latter called for the creation of data scientists, who could conduct creative inquiry and analysis; enhance through consultation, collaboration, and coordination the ability of other to conduct research and education using digital data collections; be at the forefront in developing innovative concepts in database technology and information sciences, including

methods for data visualization and information discovery, and applying these in the fields of science and education relevant to the collection; implement best practices and technology; serve as a mentor to beginning or transitioning investigators, students, and others interested in pursuing data science; and design and implement education and outreach programs that make the benefits of data collections and digital information science available to the broadest possible range of researchers, educators, students, and the general public [National Science Foundation. 2005. p.2]

On July 29, 2005, Dr. Arden Bement, director of the National Science foundation announced the re-naming of the Division of Shared Cyberinfrastructure to the Office of Cyberinfrastructure with a change in reporting line from the Directorate of Computer and Information Science and Engineering (CISE) to the Office of the Director. Under the leadership of Dr. Sangtae Kim, Donald W. Fedderson Distinguished Professor, Purdue University, the transition from CISE to OCI was completed [HPC Wire. 2005].

As described on the home page of the Office of Cyberinfrastructure (OCI) it … "coordinates and supports the acquisition, development and provision of state-of-the-art cyberinfrastructure resources, tools and services essential to the conduct of 21st century science and engineering research and education."[National Science Foundation. Office of Cyberinfrastructure. 2008] For the sake of understanding the role that OCI now plays in supporting and enabling the advances of e-science (computational enhanced scientific research) the following quotes are taken from the OCI mission:

> *OCI supports cyberinfrastructure resources, tools and related services such as supercomputers, high-capacity mass-storage systems, system software suites and programming environments, scalable interactive visualization tools, productivity software libraries and tools, large-scale data repositories and digitized scientific data management systems, networks of various reach and granularity and an array of software tools and services that hide the complexities and heterogeneity of contemporary cyberinfrastructure while seeking to provide ubiquitous access and enhanced usability.*
>
> *OCI supports the preparation and training of current and future generations of researchers and educators to use cyberinfrastructure to further their research and education goals, while also supporting the scientific and engineering professionals who create and maintain these IT-based resources and systems and who provide essential customer services to the national science and engineering user community* [National Science Foundation. Office of Cyberinfrastructure. 2008 p.2]

Soon after the creation of OCI, Dr. Christopher Greer became the program officer for the OCI. Dr. Greer, a biologist, had knowledge, experience and insights into the issues associated with cyberinfrastructure within the research community. Dr. Greer reached out to the library community through presentations at conferences such as the Coalition for Networked Information (CNI) and other library venues, as well as visiting and speaking at universities around the country. His message was well received by the scientific and engineering research community; the library community listened and saw a role as well.

To start a dialogue among the scientific and engineering researchers, information technologists, computer scientists and librarians, the National Science Foundation and the Association of Research Libraries, brought about forty representatives of these fields together in a workshop in Arlington, Virginia. The goal of the workshop was to develop a blueprint for the stewardship of data sets. The result was the report, "To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering." [Association of Research Libraries. 2006]

This workshop not only identified the challenges that must be met through research, and the development of new tools, a newly recognized collaboration between the scientific/engineering

research community and the library research community were identified. The role that library and archival sciences could play was, possibly for the first time, clearly appreciated. As a participant in the workshop, my observation was that librarians gained a clearer understanding from the scientists and engineers the challenges they were experiencing in managing massive amounts of data. It was apparent, however, that the scientists and engineers knew little about the training and professional skills of librarians, or the principles of library science. It was a learning opportunity for all concerned.

It should be noted, that one of the models used and referred to during the workshop was from New Zealand. The National Library of New Zealand Preservation Metadata Model helped to define the role of metadata in the preservation of data. [Association of Research Libraries. 2006 p.31]

Response from the Scientific/Engineering and Library Communities:

The stage was now set for a joint effort by the scientific, engineering, computer science and library communities. Proceeding from the recommendations from the report*, To Stand the Test of Time,* both communities launched initiatives.

Beginning Fall 2006 and throughout the early part of 2007, the Office of Cyberinfrastructure of the National Science Foundation worked to define the research project necessary to create the infrastructure to steward data. At the same time the university research library community took it upon itself to continue studying the role of research libraries in this new arena.

The American Research Library Community

In October 2006, the Association of Research Libraries (ARL) Steering Committees for Scholarly Communication and for Research, Teaching, and Learning jointly appointed a task force to address the emerging role of e-science. The Joint Task Force on Library Support for E-Science was charged to focus its attention on the trends and implications of e-science for research libraries, and study the impact upon collections, services, research infrastructure, and professional development. [Association of Research Libraries. Joint Task Force on Library Support for E-Science. 2007]

The Joint Task Force Report summarizes the developments on the national and international level the role of e-science, and the participation and active involvement of libraries in this area. However, it is recognized that the involvement of libraries in the United States has been piecemeal and does not provide a unified approach to solving the challenges. The Joint Task Force report is excerpted below, for the complete report see http://www.arl.org/bm~doc/ARL_EScience_final.pdf [Association of Research Libraries. Joint Task Force on Library Support for E-Science. 2007]

OUTCOME 1: **An ongoing capacity and process within ARL to develop, coordinate, and evaluate an**
e-science program agenda.

STRATEGIES:

- Develop structure and processes for carrying out the ARL e-science program agenda, including a robust education and communication program.
- Coordinate and monitor progress on the outcomes, strategies, and action plans outlined below.

ACTIONS:

1.1 Establish an ARL e-science working group with responsibility for program development (including education and communication programming), and associated resource development, and coordination across the three ARL strategic direction steering committees.
…

1.2 Create a structure to ensure communication channels and coordination between the e-science agenda, the three ARL steering committees, the ARL membership, and external stakeholders.
1.3 Identify key library and technology organizations and scholarly societies/associations with whom to establish communication linkages and a potential structured liaison program.

OUTCOME 2: A widely shared understanding both within research libraries and among other stakeholders in the e-science support community of how libraries can contribute to the development and ongoing evolution of cyberinfrastructure and e-science.

STRATEGIES:

- Build understanding at the level of library leadership of the potential for e-science to transform the process and conduct of research.
- Develop, in collaboration with other stakeholders and experts, a set of principles for research libraries support of e-science. …
- Articulate, both within ARL and with key education and research societies, a vision following from these principles and from existing models and exemplars, for research libraries roles in stewardship of research assets and as a consultant/partner in the full life cycle of scientific data.
- Build understanding at the practitioner level in the library profession of e-science support practices and needs.

ACTIONS:

2.1 The ARL e-science working group … will propose a process for developing, vetting, and sharing a set of principles for research library engagement in e-science.

2.2 Sponsor a program for library directors at an ARL membership meeting to discuss principles, with invited participants outside the library community (e.g., Science Commons, NSF Office of Cyberinfrastructure (OCI)).

2.3 Develop talking points on key issues for use in communicating with campus stakeholders about library roles and engagement for e-science, e.g., dealing with data persistence, methods, etc. Develop separate talking points for different sectors, including:

- University librarians/library directors, to have language to use in talking with campus leadership
- Individual librarians, to use when talking with their disciplinary communities

2.4 Plan programs to explore e-science issues for the ARL membership and consider ways to disseminate these events to a broader audience, e.g., staff at research libraries and interested faculty and administrators. Program types may include:

- Panel of Significant Players in E-Science Projects**.** This might highlight case studies illustrating exemplary library roles (e.g., CERN, Cornell, Illinois, Johns Hopkins, Purdue, Queensland University of Technology, Woods Hole Oceanographic Institute) or discipline-specific projects. Part of the intent of this programming would be to identify and showcase the major federally funded research centers.
- Workshop for Self-Selected Teams with a particular motivation or an emergent e-science project. This could take the form of institutional teams or representatives from multi-institutional projects. Focus on exploring the roles and challenges of team science. Using knowledge gained through team workshops, design and deliver programs for library service practitioners with a focus on building understanding and strategies to engage with faculty to support their research agenda.

2.5 Initiate conversations within the Association of American Universities, EDUCAUSE, the National Association of State Universities and Land-Grant Colleges (NASULGC) and the

ISchools Project to promote the role of research libraries as significant players in e-science. Seek to establish a formal liaison network with these organizations.

2.6 Develop a communications agenda related to data preservation directed to scientists and researchers similar to the successful Council on Library Resources' "Slow Fires" campaign for issues of preservation. This should involve other partners with preservation interests.

OUTCOME 3: Knowledgeable and skilled research library professionals with capacity to contribute to e-science and to shape new roles and models of service.

STRATEGIES:

- Highlight exemplary programs and lessons learned.
- Identify gaps in library services and recommend steps that research libraries can take to address support needs of team science, including inter-institutional team science.
- Build a library workforce with relevant new skills and knowledge about emergent forms of documentation and research dissemination.

ACTIONS:

3.1 Establish a process within ARL to develop and sustain e-science education and communication resources to include:

- **Glossary**—a glossary defining key e-science terms to foster a common understanding.
- **Resource bibliography**—an annotated bibliography to identify and link to the major reports on e-science and cyberinfrastructure in various disciplines.
- **Inventory**—an inventory of important discipline-based e-science centers and large-scale projects. Identify the major groups dealing with data issues (e.g., Committee on Data for Science and Technology/ CODATA and CENDI) at national and international levels.
- **Wiki**—an ARL-hosted wiki to gather and build a knowledgebase of resources surrounding e-science topics. Use to gather together information about relevant projects and to document emerging, innovative types of library staff positions. This tool will be used initially among working group and interested members to communicate findings. Eventually, interest and momentum may build to the extent that it may become a useful tool for member communication.

3.2 ♦Pursue programs to develop science librarian skills to meet the needs of e-science. Collaboration with IMLS may be one strategy.

OUTCOME 4: Research libraries as active participants in the conceptualization and development of research infrastructure, including systems and services to support the processes of research and the full life cycle of research assets.

STRATEGIES:

- Actively monitor, understand, and engage in activity around emergent models in publishing, particularly publication with associated primary research data.
- Monitor developments in research tools and systems, e.g., electronic laboratory notebook systems.
- Monitor and document development of collaboration environments (e.g., through requests for proposals) to identify logical points in which research librarians and research libraries might play a role.
- Document development of discipline-based repositories.
- Support new forms of scientific data publication.
- Support long-term access to scientific data as part of the scientific record.

ACTIONS:

4.1 Work with relevant professional organizations and disciplinary societies to explore issues associated with new forms of publication "packages" and genres that include data. The National Academy may be relevant here.

4.2 Partner with CNI to pursue potential "Friday Forum" or executive roundtable opportunities to explore research infrastructure associated with data and related applications.

4.3 Partner with CNI to conduct an analysis of the unmet infrastructure needs related to research collaboration that might be met by libraries.

OUTCOME 5: **Influence on policy, standards, and resource allocation decisions that support ARL principles.**

STRATEGIES:

- Promote research library involvement in shaping policy and protocols with respect to emerging scholarly communication models that integrate data and publications.
- Develop mechanisms to be an active participant in the open data movement.

ACTIONS:

5.1 Inventory and document policies of government agencies, foundations, and other organizations funding e-science projects.

5.2 Develop an education and communication program for ARL libraries to assist university officials with new research council regulations about data deposit and access and to support compliance officials on local campuses regarding these policies.

5.3 Identify and share models of library support to assist scientists in complying with data management policies of relevant funding agencies (e.g., the "concierge" model practiced at University of California, San Francisco).

5.4 Align the policy apparatus and resources of both ARL and SPARC to work in concert in support of
open data principles and policies [Association of Research Libraries. Joint Task Force on Library Support for E-Science. 2007]

The *Final Report and Recommendations of the Joint Task Force of ARL* speaks specifically to what the ARL should and can do to move this agenda forward, it does speak more broadly to the individual American research libraries as well as those in other countries. The Report identifies the challenges associated with the inadequacies confronting research libraries in launching a concerted and collaborative response to the issues associated with data management. For instance, the lack of a clear understanding of the role of librarians and archivists in data curation, the lack of professionals within the libraries who have an interest in or a commitment to working to solve these challenges.

The National Science Foundation, Office of Cyberinfrastructure Response

On September 28, 2007, the Office of Cyberinfrastructure (OCI) issued a request for proposals (RFP) titled: Sustainable Digital Data Preservation and Access Network Partners (DataNet). The goal of this project is clearly stated and is best described by the document itself, inserted below:

Science and engineering research and education are increasingly digital and increasingly data-intensive. Digital data are not only the output of research but provide input to new hypotheses, enabling new scientific insights and driving innovation. Therein lies one of the major challenges of this scientific generation: how to develop the new methods, management structures and technologies to manage the diversity, size, and complexity of current and future data sets and data streams. This solicitation addresses that challenge by creating a set of exemplar national and global data research

infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning.

The new types of organizations envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to:

- *provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;*
- *continuously anticipate and adapt to changes in technologies and in user needs and expectations;*
- *engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and*
- *serve as component elements of an interoperable data preservation and access network.*

By demonstrating feasibility, identifying best practices, establishing viable models for long term technical and economic sustainability, and incorporating frontier research, these exemplar organizations can serve as the basis for rational investment in digital preservation and access by diverse sectors of society at the local, regional, national, and international levels, paving the way for a robust and resilient national and global digital data framework.

These organizations will provide:

- *a vision and rationale that meet critical data needs, create important new opportunities and capabilities for discovery, innovation, and learning, improve the way science and engineering research and education are conducted, and guide the organization in achieving long-term sustainability;*
- *an organizational structure that provides for a comprehensive range of expertise and cyberinfrastructure capabilities, ensures active participation and effective use by a wide diversity of individuals, organizations, and sectors, serves as a capable partner in an interoperable network of digital preservation and access organizations, and ensures effective management and leadership; and*
- *activities to provide for the full data management life cycle, facilitate research as resource and object, engage in computer science and information science research critical to DataNet functions, develop new tools and capabilities for learning that integrate research and education at all levels, provide for active community input and participation in all phases and all aspects of Partner activities, and include a vigorous and comprehensive assessment and evaluation program* [National Science Foundation. Office of Cyberinfrastructure. 2007. p.2]

The NSF OCI intends to award five $20,000,000 grants up to five years depending upon the quality of the proposals. There will be two review cycles, one commenced January 7th, 2008, with the submission of the preliminary proposal, with the full proposal submitted by March 21st, 2008. A second round has a submission date of October 6th, 2008, with the full proposal submitted no later than February 16, 2009. There will be two or three funded in the first cycle and two or three funded in the second cycle, depending upon the number and quality of proposals submitted.

A critical statement evidencing the new relationship between library and archival sciences is in the second paragraph on page 2 of the DataNet RFP, where it states, “The new types of organization envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise … [National Science Foundation. Office of Cyberinfrastructure. 2007. p.2]

It is probably the first time in the history of the National Science Foundation that the principles of library and archival science were highlighted specifically as being a necessary part of a scientific/technical infrastructure project. This is a result of a collaborative exploration of the

problems and challenges facing the scientific and engineering research community, and the opportunities to work collaboratively by librarians and archivists with computer scientists, information technologists, and computational scientists to solve the challenges.

Will this lead to a data policy for the United States, possibly, at least it is a major first step to gain collaboration among many partners to strive for a way to manage data, and thereby, de facto create a national data policy.

References:

Association of Research Libraries. Joint Task Force on Library Support for E-Science. 2007. *Agenda for Developing E-Science in Research Libraries. Final Report and Recommendations to the Scholarly Communication Steering Committee, the Public Policies Affecting Research Libraries Steering Committee, and the Research, Teaching, and Learning Steering Committee. November 2007.* http://www.arl.org/bm~doc/ARL_EScience_final.pdf

Association of Research Libraries. 2006. To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering. Arlington, VA, September 26-27,2006. http://www.arl.org/bm~doc/digdatarpt.pdf

Fedora Wiki. 2008 http://fedoraproject.org/wiki/Overview

HPC Wire. 2005 *NSF Creates Office of Cyberinfrastructure*. July 29, 2005. http://www.hpcwire.com/hpc/439330.html

MIT Libraries and Hewlett-Packard Company. 2008. http://libraries.mit.edu/dspace-mit/about/definition.html.

National Science Foundation. *About the National Science Foundation. 2008. NSF at a Glance*. http://www.nsf.gov/about/

National Science Foundation. Blue Ribbon Advisory Panel on Cyberinfrastructure. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure. http://www.nsf.gov/od/oci/reports/atkins.pdf*

National Science Foundation. 2005. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.* http://www.nsf.gov/pubs/2005/nsb0540/

National Science Foundation. Office of Cyberinfrastructure. 2008. *About the Office of Cyberinfrastructure (OCI)*. http://www.nsf.gov/od/oci/about.jsp

National Science Foundation. Office of Cyberinfrastructure. (2007). *Sustainable Digital Data Preservation and Access Network Partners (DataNet)* (07-601). http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm