To Stand the Test of Time

Long-term Stewardship of Digital Data Sets in Science and Engineering

A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe

DigCCurr 2007
April 20, 2007

Presented by:
Tyler Walters
Associate Director, Technology and Resource Services
Georgia Institute of Technology Library

Prue Adler
Associate Executive Director
Association of Research Libraries
prue@arl.org

Workshop Goals

- To examine issues associated with **sustainable economic models for long-term preservation and curation of digital data** and to articulate recommendations for further work, including identification of sources of funding;
- To examine the **structure of new partnerships** to facilitate seamless capture, processing, storage, and management of heterogeneous scientific and engineering data; and
- To examine the **infrastructure requirements necessary to support long-term management of digital data** in distributed yet federated collections, recognizing the rapid pace of technological change and the need for unfettered access.

Focus

- Infrastructure
- Partnerships
- Sustainable Economic Models

Workshop Participants

- Thirty-two scientists, librarians, information scientists, researchers
- Social sciences
- Geosciences
- Oceanography
- Computer science
- Astronomy
- Ecology

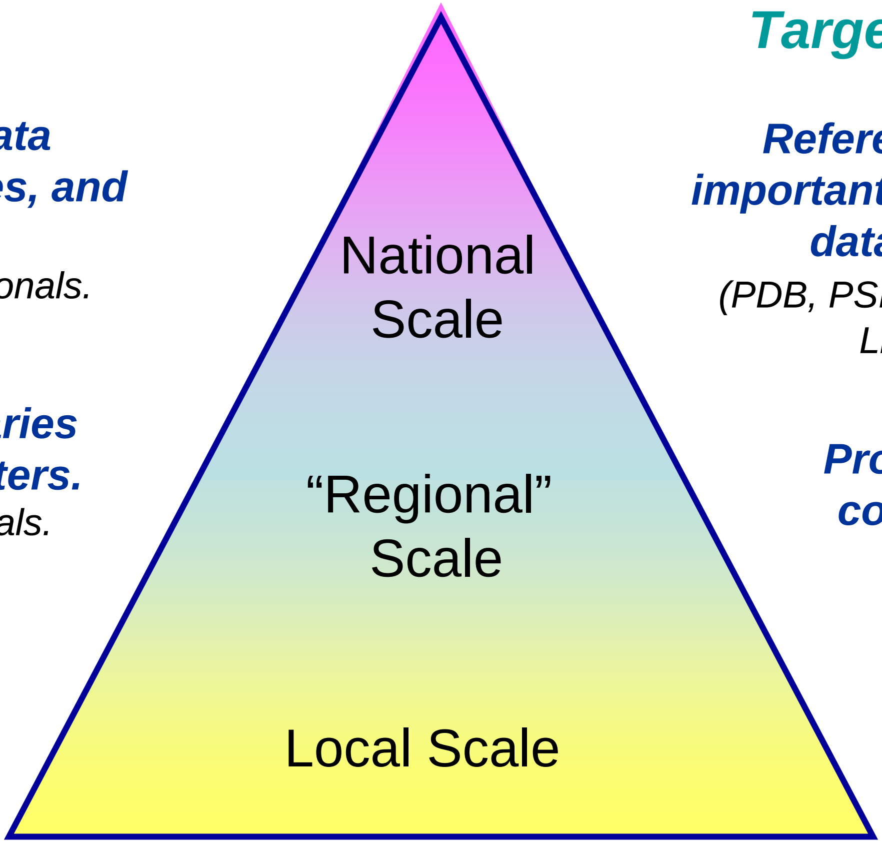
The “Data Pyramid”: An Organizational Structure for Talking about Research Data

Facilities

National-scale data repositories, archives, and libraries.
Maintained by professionals.

“Regional”-scale libraries and targeted data centers.
Maintained by professionals.

Private repositories.
Supported by users or their proxies.



National Scale

“Regional” Scale

Local Scale

The Data Pyramid

Target Collections

Reference, nationally important, and irreplaceable data collections.
(PDB, PSID, Shoah, Presidential Libraries, etc.)

Project and some community data collections

Personal data collections

Workshop 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, interdisciplinary domains.

- Historically, universities have played a leadership role in 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.

http://www.arl.org/pp/access/nsfworkshop.shtml

Workshop Findings

- Responsibility for the stewardship of digital information should be vested in **distributed collections** and repositories that recognize: **heterogeneity** of **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.

http://www.arl.org/pp/access/nsfworkshop.shtml

Workshop Findings

- **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.
- **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.

http://www.arl.org/pp/access/nsfworkshop.shtml

Workshop Findings

- 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.

Workshop Recommendations

NSF should facilitate the establishment of a sustainable institutional framework for 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 data produced by all science and engineering disciplines/domains.

http://www.arl.org/pp/access/nsfworkshop.shtml

Workshop Recommendations

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.

Workshop Recommendations

3. Support the development of usable and useful tools including:

- automated services and standards which facilitate understanding and manipulating data;
- data registration;
- reference tools to accommodate ongoing documentation of commonly used terms and concept;
- automated metadata creation; and
- rights management and other access control considerations.

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.

Targeted Recommendations

3. NSF should support the **development of usable and useful tools**, 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.

Targeted Recommendations

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.

Targeted Recommendations

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.

For a complete copy of the report:

To Stand the Test of Time
Long-term Stewardship of Digital Data Sets in Science and Engineering

See: **http://www.arl.org/bm~doc/digdatarpt.pdf**